SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 1, pp. 1–28
© 1999 Society for Industrial and Applied Mathematics

SMOOTHNESS OF MULTIPLE REFINABLE FUNCTIONS AND MULTIPLE WAVELETS*

RONG-QING JIA†, SHERMAN D. RIEMENSCHNEIDER†, AND DING-XUAN ZHOU‡

Abstract. We consider the smoothness of solutions of a system of refinement equations written in the form

$$\phi = \sum\nolimits_{\alpha\in\mathbb{Z}} a(\alpha)\phi(2\cdot-\alpha),$$

where the vector of functions $\phi = (\phi_1, \dots, \phi_r)^T$ is in $(L_p(\mathbb{R}))^r$ and a is a finitely supported sequence of $r \times r$ matrices called the refinement mask. We use the generalized Lipschitz space $\mathrm{Lip}^*(\nu, L_p(\mathbb{R}))$, $\nu > 0$, to measure smoothness of a given function.

Our method is to relate the optimal smoothness, $\nu_p(\phi)$, to the p-norm joint spectral radius of the block matrices A_ε, $\varepsilon = 0, 1$, given by $A_\varepsilon = (a(\varepsilon + 2\alpha - \beta))_{\alpha,\beta}$, when restricted to a certain finite dimensional common invariant subspace V. Denoting the p-norm joint spectral radius by $\rho_p(A_0|_V, A_1|_V)$, we show that $\nu_p(\phi) \ge 1/p - \log_2 \rho_p(A_0|_V, A_1|_V)$ with equality when the shifts of $\phi_1, \dots, \phi_r$ are stable and the invariant subspace is generated by certain vectors induced by difference operators of sufficiently high order. This allows an effective use of matrix theory. Also the computational implementation of our method is simple.

When $p = 2$, the optimal smoothness is also given in terms of the spectral radius of the transition matrix associated with the refinement mask.

To illustrate the theory, we give a detailed analysis of two examples where the optimal smoothness can be given explicitly. We also apply our methods to the smoothness analysis of multiple wavelets. These examples clearly demonstrate the applicability and practical power of our approach.

Key words. refinement equations, multiple refinable functions, multiple wavelets, vector subdivision schemes, joint spectral radii, transition operators

AMS subject classifications. Primary, 39B12, 41A25, 42C15, 65F15

PII. S089547989732383X

1. Introduction.

The purpose of this paper is to investigate the smoothness properties of multiple refinable functions and multiple wavelets.

Suppose $\phi_1, \dots, \phi_r$ are compactly supported distributions on $\mathbb{R}$. Denote by ϕ the vector $(\phi_1, \dots, \phi_r)^T$, the transpose of $(\phi_1, \dots, \phi_r)$. We say that ϕ is *refinable* if it satisfies the following *refinement equation*:

$$\phi = \sum_{\alpha\in\mathbb{Z}} a(\alpha)\phi(2\cdot-\alpha), \tag{1.1}$$

where each $a(\alpha)$ is an $r \times r$ matrix of complex numbers and $a(\alpha) = 0$ except for finitely many α. We view a as a sequence from $\mathbb{Z}$ to $\mathbb{C}^{r\times r}$ and call it the *refinement mask*.

In our previous paper [20], we gave a characterization for the accuracy of a vector of multiple refinable functions in terms of the corresponding mask. In another paper [21], we characterized the L_p-convergence ($1 \le p \le \infty$) of a subdivision scheme in

*Received by the editors July 2, 1997; accepted for publication (in revised form) by A. Edelman March 9, 1998; published electronically August 3, 1999. This research was supported in part by NSERC Canada under grants OGP 121336 and A7687.

http://www.siam.org/journals/simax/21-1/32383.html

†Department of Mathematical Sciences, University of Alberta, Edmonton, AB, Canada T6G 2G1 (jia@xihu.math.ualberta.ca, sherm@approx.math.ualberta.ca).

‡Department of Mathematics, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong (mazhou@math.cityu.edu.hk).

terms of the p-norm joint spectral radius of two matrices derived from the mask. (See [17] for the definition of the p-norm joint spectral radius of a finite collection of matrices.) In this paper, we will take the same approach as we did in [21] to a study of the smoothness properties of the solutions of the refinement equation (1.1).

Taking the Fourier transform of both sides of (1.1), we obtain

$$\hat{\phi}(\xi) = H(\xi/2)\hat{\phi}(\xi/2), \qquad \xi \in \mathbb{R}, \tag{1.2}$$

where

$$H(\xi) := \sum_{\alpha\in\mathbb{Z}} a(\alpha)e^{-i\alpha\xi}/2, \qquad \xi \in \mathbb{R}.$$

Evidently, H is 2π-periodic. Let

$$M := H(0) = \sum_{\alpha\in\mathbb{Z}} a(\alpha)/2.$$

If ϕ is a solution of (1.1), then it follows from (1.2) that $\hat{\phi}(0) = M\hat{\phi}(0)$. In other words, either $\hat{\phi}(0) = 0$, or $\hat{\phi}(0)$ is an eigenvector of M corresponding to the eigenvalue 1.

Before proceeding, we introduce some notation. For $1 \le p \le \infty$, let $(L_p(\mathbb{R}))^r$ denote the linear space of all vectors $f = (f_1, \ldots, f_r)^T$ such that $f_1, \ldots, f_r \in L_p(\mathbb{R})$. The norm on $(L_p(\mathbb{R}))^r$ is defined by

$$\|f\|_p := \left(\sum_{j=1}^r \|f_j\|_p^p\right)^{1/p}, \qquad f = (f_1, \ldots, f_r)^T \in (L_p(\mathbb{R}))^r.$$

By $(C(\mathbb{R}))^r$ we denote the linear space of all $r \times 1$ vectors of continuous functions.

The shifts of functions $\phi_1, \ldots, \phi_r \in L_p(\mathbb{R})$ are said to be *stable* if there exist two positive constants C_1 and C_2 such that, for arbitrary $b_1, \ldots, b_r \in \ell_p(\mathbb{Z})$,

$$C_1 \sum_{j=1}^r \|b_j\|_p \le \left\| \sum_{j=1}^r \sum_{\alpha\in\mathbb{Z}} b_j(\alpha)\phi_j(\cdot - \alpha) \right\|_p \le C_2 \sum_{j=1}^r \|b_j\|_p.$$

It was proved by Jia and Micchelli [19] that the shifts of the functions $\phi_1, \ldots, \phi_r$ are stable if and only if, for any $\xi \in \mathbb{R}$, the sequences $(\hat{\phi}_j(\xi + 2\pi\beta))_{\beta\in\mathbb{Z}}$ $(j = 1, \ldots, r)$ are linearly independent.

Let $\ell_0(\mathbb{Z})$ denote the linear space of all finitely supported sequences on $\mathbb{Z}$. Similarly, we denote by $\ell_0(\mathbb{Z} \to \mathbb{C}^r)$ (resp., $\ell_0(\mathbb{Z} \to \mathbb{C}^{r\times r})$) the linear space of all finitely supported sequences of $r \times 1$ vectors (resp., $r \times r$ matrices). We identify $\ell_0(\mathbb{Z} \to \mathbb{C}^r)$ with $(\ell_0(\mathbb{Z}))^r$ and identify $\ell_0(\mathbb{Z} \to \mathbb{C}^{r\times r})$ with $(\ell_0(\mathbb{Z}))^{r\times r}$.

For $\beta \in \mathbb{Z}$, we use δ_β to denote the sequence given by

$$\delta_\beta(\alpha) = \begin{cases} 1 & \text{for } \alpha = \beta, \\ 0 & \text{for } \alpha \in \mathbb{Z} \setminus \{\beta\}. \end{cases}$$

In particular, we write δ for δ_0.

We denote by ∇ the difference operator on $\ell_0(\mathbb{Z})$:

$$\nabla v := v - v(\cdot - 1), \qquad v \in \ell_0(\mathbb{Z}).$$

The domain of the difference operator ∇ can be naturally extended to include $(\ell_0(\mathbb{Z}))^r$ and $(\ell_0(\mathbb{Z}))^{r\times r}$.

Let a be an element of $(\ell_0(\mathbb{Z}))^{r\times r}$. For $\varepsilon = 0, 1$, let A_ε be the linear operator on $(\ell_0(\mathbb{Z}))^r$ given by

$$A_\varepsilon v(\alpha) = \sum_{\beta\in\mathbb{Z}} a(\varepsilon + 2\alpha - \beta)v(\beta), \quad \alpha \in \mathbb{Z}, \quad v \in (\ell_0(\mathbb{Z}))^r. \tag{1.3}$$

Our first concern is the existence and uniqueness of the refinement equation (1.1) with the mask a. Under the condition that $\lim_{n\to\infty} M^n$ exists, Heil and Colella [13] established existence and uniqueness of distributional solutions of (1.1). In this case, convergence of the subdivision scheme was studied by Cohen, Dyn, and Levin in [4]. Without assuming the condition that $\lim_{n\to\infty} M^n$ exists, the existence and uniqueness of distributional solutions of (1.1) were studied by several authors, including Cohen, Daubechies, and Plonka [3], Jiang and Shen [23], and Zhou [34].

In section 2, we will investigate the existence of L_p-solutions of the refinement equation (1.1) with the mask a. For $j = 1, \dots, r$, we use e_j to denote the jth column of the $r \times r$ identity matrix. Let A_0 and A_1 be the linear operators given by (1.3) and V the minimal common invariant subspace of A_0 and A_1 generated by $e_j(\nabla\delta)$, $j = 1, \dots, r$, in $(\ell_0(\mathbb{Z}))^r$. We will prove that, for an eigenvector y of the matrix $M := \sum_{\alpha\in\mathbb{Z}} a(\alpha)/2$ corresponding to the eigenvalue 1, there exists a compactly supported solution $\phi \in (L_p(\mathbb{R}))^r$ ($\phi \in (C(\mathbb{R}))^r$ in the case $p = \infty$) of the refinement equation (1.1) subject to $\hat{\phi}(0) = y$, provided

$$\rho_p(A_0|_V, A_1|_V) < 2^{1/p},$$

where $\rho_p(A_0|_V, A_1|_V)$ denotes the p-norm joint spectral radius of $A_0|_V$ and $A_1|_V$. This condition is necessary if, in addition, the shifts of $\phi_1, \dots, \phi_r$ are stable.

We use the generalized Lipschitz space to measure smoothness of a given function. Let us recall from [8] the definition of the generalized Lipschitz space. For $y \in \mathbb{R}$, the difference operator ∇_y is defined by

$$\nabla_y f = f - f(\cdot - y),$$

where f is a function from $\mathbb{R}$ to $\mathbb{C}$. The *modulus of continuity* of a function f in $L_p(\mathbb{R})$ ($1 \le p \le \infty$) is defined by

$$\omega(f, h)_p := \sup_{|y|\le h} \left\|\nabla_y f\right\|_p, \qquad h \ge 0.$$

Let k be a positive integer. The kth *modulus of smoothness* of $f \in L_p(\mathbb{R})$ is defined by

$$\omega_k(f, h)_p := \sup_{|y|\le h} \left\|\nabla_y^k f\right\|_p, \qquad h \ge 0.$$

For $\nu > 0$, let k be an integer greater than ν. The *generalized Lipschitz space* $\mathrm{Lip}^*(\nu, L_p(\mathbb{R}))$ consists of those functions $f \in L_p(\mathbb{R})$ for which

$$\omega_k(f, h)_p \le Ch^\nu \qquad \forall\, h > 0,$$

where C is a positive constant independent of h.

By $(\mathrm{Lip}^*(\nu, L_p(\mathbb{R})))^r$ we denote the linear space of all vectors $f = (f_1, \ldots, f_r)^T$ such that $f_1, \ldots, f_r \in \mathrm{Lip}^*(\nu, L_p(\mathbb{R}))$. The optimal smoothness of a vector $f \in (L_p(\mathbb{R}))^r$ in the L_p-norm is described by its *critical exponent* $\nu_p(f)$ defined by

$$\nu_p(f) := \sup\Big\{ \nu : f \in \big(\mathrm{Lip}^*(\nu, L_p(\mathbb{R}))\big)^r \Big\}.$$

In section 3, we will establish our main result on characterization of the smoothness of multiple refinable functions. Suppose $\phi = (\phi_1, \ldots, \phi_r)^T$ is a compactly supported solution of the refinement equation (1.1) with the mask a. Let k be a positive integer and V the minimal common invariant subspace of A_0 and A_1 generated by $e_j(\nabla^k \delta)$, $j = 1, \ldots, r$. If ϕ lies in $(L_p(\mathbb{R}))^r$ for $1 \le p < \infty$ (ϕ lies in $(C(\mathbb{R}))^r$ for $p = \infty$), then

$$\nu_p(\phi) \ge 1/p - \log_2 \rho_p(A_0|_V, A_1|_V). \tag{1.4}$$

If, in addition, the shifts of $\phi_1, \ldots, \phi_r$ are stable, and if $k > 1/p - \log_2 \rho_p(A_0|_V, A_1|_V)$, then equality holds in (1.4). When $p = 2$, the critical exponent is also given in terms of the spectral radius of the transition operator associated with the refinement mask a.

Regularity of multiple refinable functions was studied by Cohen, Daubechies, and Plonka in [3] and by Micchelli and Sauer in [25]. Both approaches are based on the factorization technique introduced by Plonka [26]. Our approach is different from theirs and does not rely on factorization. Thus, our methods can be applied to multiple refinable functions and multiple wavelets of several variables. For smoothness analysis of a single multivariate refinable function, the reader is referred to [18] and [27]. Even in the univariate case our methods have advantages over the factorization technique. Indeed, our methods use the joint spectral radius of finite matrices. This allows a more effective use of matrix theory to reduce the size of the matrices by a restriction to a certain common invariant subspace. Thus, the computational implementation of our method becomes much simpler. In fact, in the multiple case, the factorization would usually enlarge the support of the mask making the order of the matrices larger, hence computationally more complex. For a discussion of the size of the support of vector scaling functions, see So and Wang [30].

To illustrate the general theory, we shall give detailed analysis of smoothness for two examples in section 4. One example is taken from [10], the other from [21]. In particular, for the example of Donovan et al. [10], our method gives explicitly the exact smoothness in all p-norms. In comparison, Cohen, Daubechies, and Plonka [3] partially recovered the result of [10] for the regularity in the L_∞-norm, while Micchelli and Sauer [25] gave a crude estimate for the regularity in the L_1-norm for a special case.

In section 5, applying our study to multiple wavelets, we construct a family of orthogonal double wavelets which includes the one of Chui and Lian [2]. We give a complete smoothness analysis in L_2 for this family, and in all L_p for the example of Chui and Lian (who did not discuss smoothness).

The examples in sections 4 and 5 clearly demonstrate the applicability and practical power of our approach.

2. Existence of L_p-solutions. In order to solve the refinement equation (1.1), we introduce the linear operator Q_a on $(L_p(\mathbb{R}))^r$ $(1 \le p \le \infty)$ as follows:

$$Q_a f := \sum_{\alpha \in \mathbb{Z}} a(\alpha) f(2 \cdot - \alpha), \qquad f \in (L_p(\mathbb{R}))^r. \tag{2.1}$$

If ϕ is a fixed point of Q_a, i.e., $Q_a\phi = \phi$, then ϕ is a solution of the refinement equation (1.1).

Let Q_a be the linear operator given in (2.1). For an initial vector $f \in (L_p(\mathbb{R}))^r$, we have

$$Q_a^n f = \sum_{\alpha\in\mathbb{Z}} a_n(\alpha) f(2^n \cdot -\alpha), \qquad n = 1, 2, \ldots,$$

where each a_n is independent of the choice of f. In particular, $a_1 = a$. Consequently, for $n > 1$ we have

$$\begin{aligned} Q_a^n f = Q_a^{n-1}(Q_a f) &= \sum_{\beta\in\mathbb{Z}} a_{n-1}(\beta)(Q_a f)(2^{n-1}\cdot -\beta) \\ &= \sum_{\beta\in\mathbb{Z}}\sum_{\alpha\in\mathbb{Z}} a_{n-1}(\beta)a(\alpha) f(2^n\cdot -2\beta-\alpha) \\ &= \sum_{\alpha\in\mathbb{Z}} \left[\sum_{\beta\in\mathbb{Z}} a_{n-1}(\beta)a(\alpha-2\beta)\right] f(2^n\cdot-\alpha). \end{aligned}$$

This establishes the following iteration relation for a_n $(n = 1, 2, \ldots)$:

$$a_1 = a \quad \text{and} \quad a_n(\alpha) = \sum_{\beta\in\mathbb{Z}} a_{n-1}(\beta)a(\alpha-2\beta), \quad \alpha\in\mathbb{Z}. \tag{2.2}$$

For $\varepsilon \in \mathbb{Z}$, we denote by $A_\varepsilon = (A_\varepsilon(\alpha,\beta))_{\alpha,\beta\in\mathbb{Z}}$ the bi-infinite block matrix given by

$$A_\varepsilon(\alpha,\beta) := a(\varepsilon + 2\alpha - \beta), \qquad \alpha,\beta\in\mathbb{Z}. \tag{2.3}$$

For $a \in (\ell_0(\mathbb{Z}))^{r\times r}$ and $n = 1, 2, \ldots$, let $a_n \in (\ell_0(\mathbb{Z}))^{r\times r}$ be given by the iteration relation (2.2). If $\alpha = \varepsilon_1 + 2\varepsilon_2 + \cdots + 2^{n-1}\varepsilon_n + 2^n\gamma$, where $\varepsilon_1, \ldots, \varepsilon_n, \gamma \in \mathbb{Z}$, then

$$a_n(\alpha-\beta) = A_{\varepsilon_n}\cdots A_{\varepsilon_1}(\gamma,\beta) \qquad \forall\,\beta\in\mathbb{Z}. \tag{2.4}$$

This can be proved easily by induction on n. For $n = 1$ and $\alpha = \varepsilon_1 + 2\gamma$, where $\varepsilon_1, \gamma \in \mathbb{Z}$, we have

$$a_1(\alpha-\beta) = a(\varepsilon_1 + 2\gamma - \beta) = A_{\varepsilon_1}(\gamma,\beta).$$

Suppose $n > 1$ and (2.4) has been verified for $n-1$. For $\alpha = \varepsilon_1 + 2\alpha_1$, where $\alpha_1, \varepsilon_1 \in \mathbb{Z}$, by the iteration relation (2.2) we have

$$a_n(\alpha-\beta) = \sum_{\eta\in\mathbb{Z}} a_{n-1}(\eta)a(\alpha-\beta-2\eta) = \sum_{\eta\in\mathbb{Z}} a_{n-1}(\alpha_1-\eta)a(\varepsilon_1+2\eta-\beta). \tag{2.5}$$

Suppose $\alpha_1 = \varepsilon_2 + \cdots + 2^{n-2}\varepsilon_n + 2^{n-1}\gamma$, where $\varepsilon_2, \ldots, \varepsilon_n, \gamma \in \mathbb{Z}$. Then by the induction hypothesis we have

$$a_{n-1}(\alpha_1-\eta) = A_{\varepsilon_n}\cdots A_{\varepsilon_2}(\gamma,\eta).$$

This in connection with (2.5) gives

$$a_n(\alpha-\beta) = \sum_{\eta\in\mathbb{Z}} A_{\varepsilon_n}\cdots A_{\varepsilon_2}(\gamma,\eta)A_{\varepsilon_1}(\eta,\beta) = A_{\varepsilon_n}\cdots A_{\varepsilon_2}A_{\varepsilon_1}(\gamma,\beta),$$

thereby completing the induction procedure.

The relation (2.4) motivates us to consider the joint spectral radius of a finite multiset of linear operators. The uniform joint spectral radius was introduced by Rota and Strang [28]. The use of the joint spectral radius to obtain regularity results was initiated by Daubechies and Lagarias [6, 7] for the scalar case. Colella and Heil [5] used the joint spectral radius to characterize continuous solutions of scalar refinement equations.

The p-norm joint spectral radius was introduced by Jia in [17]. Let us recall from [17] the definition of the p-norm joint spectral radius. Let V be a *finite-dimensional* vector space equipped with a vector norm $\|\cdot\|$. For a linear operator A on V, define

$$\|A\| := \max_{\|v\|=1} \big\{\|Av\|\big\}.$$

Let $\mathcal{A}$ be a finite multiset of linear operators on V. For a positive integer n we denote by $\mathcal{A}^n$ the nth Cartesian power of $\mathcal{A}$:

$$\mathcal{A}^n = \big\{(A_1, \ldots, A_n) : A_1, \ldots, A_n \in \mathcal{A}\big\}.$$

For $1 \le p < \infty$, let

$$\|\mathcal{A}^n\|_p := \left(\sum_{(A_1,\ldots,A_n)\in\mathcal{A}^n} \|A_1 \cdots A_n\|^p\right)^{1/p},$$

and, for $p = \infty$, define

$$\|\mathcal{A}^n\|_\infty := \max\big\{\|A_1 \cdots A_n\| : (A_1, \ldots, A_n) \in \mathcal{A}^n\big\}.$$

For $1 \le p \le \infty$, the *p-norm joint spectral radius* of $\mathcal{A}$ is defined to be

$$\rho_p(\mathcal{A}) := \lim_{n\to\infty} \|\mathcal{A}^n\|_p^{1/n}.$$

It is easily seen that this limit indeed exists, and

$$\lim_{n\to\infty} \|\mathcal{A}^n\|_p^{1/n} = \inf_{n\ge 1} \|\mathcal{A}^n\|_p^{1/n}.$$

Clearly, $\rho_p(\mathcal{A})$ is independent of the choice of the vector norm on V.

If $\mathcal{A}$ consists of a single linear operator A, then $\rho_p(\mathcal{A}) = \rho(A)$, where $\rho(A)$ denotes the spectral radius of A, which is independent of p. It is easily seen that $\rho(A) \le \rho_\infty(\mathcal{A})$ for any element A in $\mathcal{A}$.

Now let $\mathcal{A}$ be a finite multiset of linear operators on a normed vector space V, which is not necessarily finite dimensional. A subspace W of V is said to be invariant under $\mathcal{A}$, or *$\mathcal{A}$-invariant*, if it is invariant under every operator A in $\mathcal{A}$. For a vector $w \in V$, we define

$$\|\mathcal{A}^n w\|_p := \begin{cases} \left(\sum_{(A_1,\ldots,A_n)\in\mathcal{A}^n} \|A_1 \cdots A_n w\|^p\right)^{1/p} & \text{for } 1 \le p < \infty, \\ \max\big\{\|A_1 \cdots A_n w\| : (A_1, \ldots, A_n) \in \mathcal{A}^n\big\} & \text{for } p = \infty. \end{cases} \tag{2.6}$$

If the minimal $\mathcal{A}$-invariant subspace W generated by w is finite dimensional, then we have

$$\lim_{n\to\infty} \big\|\mathcal{A}^n w\big\|_p^{1/n} = \rho_p(\mathcal{A}|_W), \qquad 1 \le p \le \infty.$$

See [12, Lemma 2.4] for a proof of this result.

If $V = (\ell_0(\mathbb{Z}))^r$, it is often convenient to choose the ℓ_p-norm as the underlying vector norm in (2.6). We denote by $\ell_p(\mathbb{Z} \to \mathbb{C}^r)$ the linear space of all sequences $u : \mathbb{Z} \to \mathbb{C}^r$ such that $u(\alpha) = (u_1(\alpha), \dots, u_r(\alpha))^T$ for some $u_1, \dots, u_r \in \ell_p(\mathbb{Z})$ and all $\alpha \in \mathbb{Z}$. Obviously, $u \mapsto (u_1, \dots, u_r)^T$ is a canonical isomorphism between $\ell_p(\mathbb{Z} \to \mathbb{C}^r)$ and $(\ell_p(\mathbb{Z}))^r$. Thus, we may identify $\ell_p(\mathbb{Z} \to \mathbb{C}^r)$ with $(\ell_p(\mathbb{Z}))^r$. The norm of $u = (u_1, \dots, u_r)^T$ is given by

$$\|u\|_p := \Big(\sum\nolimits_{j=1}^r \|u_j\|_p^p\Big)^{1/p}.$$

Equipped with this norm, $(\ell_p(\mathbb{Z}))^r$ becomes a Banach space.

We denote by $\ell_p(\mathbb{Z} \to \mathbb{C}^{r\times r})$ the linear space of all sequences $b : \mathbb{Z} \to \mathbb{C}^{r\times r}$ such that $b(\alpha) = (b_{jk}(\alpha))_{1\le j,k\le r}$ for some $b_{jk} \in \ell_p(\mathbb{Z})$ $(j, k = 1, \dots, r)$ and all $\alpha \in \mathbb{Z}$. We also identify $\ell_p(\mathbb{Z} \to \mathbb{C}^{r\times r})$ with $(\ell_p(\mathbb{Z}))^{r\times r}$. The norm of $b = (b_{jk})_{1\le j,k\le r}$ is defined by

$$\|b\|_p := \left(\sum_{j=1}^r \sum_{k=1}^r \|b_{jk}\|_p^p\right)^{1/p}.$$

Let a be an element of $(\ell_0(\mathbb{Z}))^{r\times r}$. The bi-infinite block matrices A_ε $(\varepsilon \in \mathbb{Z})$ defined in (2.3) may be viewed as the linear operators on $(\ell_0(\mathbb{Z}))^r$ given by

$$A_\varepsilon v(\alpha) = \sum_{\beta\in\mathbb{Z}} a(\varepsilon + 2\alpha - \beta)v(\beta), \quad \alpha \in \mathbb{Z}, \quad v \in (\ell_0(\mathbb{Z}))^r. \tag{2.7}$$

Suppose $y \in \mathbb{C}^r$ and $\alpha = \varepsilon_1 + 2\varepsilon_2 + \cdots + 2^{n-1}\varepsilon_n + 2^n\gamma$, where $\varepsilon_1, \dots, \varepsilon_n, \gamma \in \mathbb{Z}$. Then it follows from (2.4) that

$$a_n(\alpha - \beta)y = A_{\varepsilon_n} \cdots A_{\varepsilon_1}(y\delta_\beta)(\gamma) \qquad \forall\, \beta \in \mathbb{Z}. \tag{2.8}$$

For a bounded subset K of $\mathbb{R}$, we use $\ell(K)$ to denote the linear space of all sequences on $\mathbb{Z}$ supported in $K \cap \mathbb{Z}$. Suppose a is supported on $[0, N]$, where N is a positive integer. Then, for $j \le 0$ and $k \ge N - 1$, $(\ell([j, k]))^r$ is invariant under both A_0 and A_1. Consequently, the minimal common invariant subspace of A_0 and A_1 generated by a finite subset of $(\ell_0(\mathbb{Z}))^r$ is finite dimensional.

THEOREM 2.1. *For an element $a \in (\ell_0(\mathbb{Z}))^{r\times r}$, let y be an eigenvector of the matrix $M := \sum_{\alpha\in\mathbb{Z}} a(\alpha)/2$ corresponding to the eigenvalue 1, and let V be the minimal common invariant subspace of A_0 and A_1 generated by $e_j(\nabla\delta)$, $j = 1, \dots, r$, in $(\ell_0(\mathbb{Z}))^r$. If*

$$\rho_p(A_0|_V, A_1|_V) < 2^{1/p}, \tag{2.9}$$

then there exists a compactly supported solution $\phi \in (L_p(\mathbb{R}))^r$ ($\phi \in (C(\mathbb{R}))^r$ in the case $p = \infty$) of the refinement equation (1.1) with the mask a subject to $\hat\phi(0) = y$. Conversely, if $\phi = (\phi_1, \dots, \phi_r)^T \in (L_p(\mathbb{R}))^r$ ($\phi \in (C(\mathbb{R}))^r$ in the case $p = \infty$) is a compactly supported solution of (1.1) such that the shifts of $\phi_1, \dots, \phi_r$ are stable, then (2.9) holds true.

Proof. The proof follows the lines of [21, Theorem 5.3]. Let $f := yg$, where g is the hat function supported on $[0, 2]$ satisfying $g(x) = x$ for $0 \le x \le 1$ and $g(x) = 2 - x$ for $1 < x \le 2$. Since $\hat g(0) = 1$, we have $\hat f(0) = y$.

Consider $f_n := Q_a^n f$, $n = 1, 2, \ldots$. We have

$$f_n = \sum_{\alpha\in\mathbb{Z}} a_n(\alpha) f(2^n \cdot -\alpha),$$

where the sequences a_n $(n = 1, 2, \ldots)$ are given by the iteration relation (2.2). Since $g = g(2\cdot)/2 + g(2\cdot -1) + g(2\cdot -2)/2$, we have

$$f_n = \sum_{\alpha\in\mathbb{Z}} \frac{1}{2}\big[a_n(\alpha)y + a_n(\alpha-1)y\big] g(2^{n+1}\cdot -2\alpha) + \sum_{\alpha\in\mathbb{Z}} \big[a_n(\alpha)y\big] g(2^{n+1}\cdot -2\alpha -1).$$

Moreover,

$$f_{n+1} = \sum_{\alpha\in\mathbb{Z}} \big[a_{n+1}(2\alpha)y\big] g(2^{n+1}\cdot -2\alpha) + \sum_{\alpha\in\mathbb{Z}} \big[a_{n+1}(2\alpha+1)y\big] g(2^{n+1}\cdot -2\alpha -1).$$

Subtracting the first equation from the second, we obtain

$$f_{n+1} - f_n = \sum_{\alpha\in\mathbb{Z}} \big[b_n(\alpha)y\big] g(2^{n+1}\cdot -2\alpha) + \sum_{\alpha\in\mathbb{Z}} \big[c_n(\alpha)y\big] g(2^{n+1}\cdot -2\alpha -1),$$

where

$$b_n(\alpha) := a_{n+1}(2\alpha) - \frac{1}{2}a_n(\alpha) - \frac{1}{2}a_n(\alpha-1) \quad \text{and} \quad c_n(\alpha) := a_{n+1}(2\alpha+1) - a_n(\alpha), \quad \alpha\in\mathbb{Z}.$$

It follows that

$$\|f_{n+1} - f_n\|_p \le 2^{1-(n+1)/p}\big(\|b_n y\|_p + \|c_n y\|_p\big). \tag{2.10}$$

Let us estimate $\|b_n y\|_p$ and $\|c_n y\|_p$. Suppose $\alpha = \varepsilon_1 + 2\varepsilon_2 + \cdots + 2^{n-1}\varepsilon_n + 2^n\gamma$, where $\gamma\in\mathbb{Z}$ and $\varepsilon_1, \ldots, \varepsilon_n \in \{0, 1\}$. Then $2\alpha = 0 + 2\varepsilon_1 + 2^2\varepsilon_2 + \cdots + 2^n\varepsilon_n + 2^{n+1}\gamma$, and an application of (2.8) gives

$$\begin{aligned} b_n(\alpha)y &= a_{n+1}(2\alpha)y - \frac{1}{2}a_n(\alpha)y - \frac{1}{2}a_n(\alpha-1)y \\ &= A_{\varepsilon_n}\cdots A_{\varepsilon_1}A_0(y\delta)(\gamma) - \frac{1}{2}A_{\varepsilon_n}\cdots A_{\varepsilon_1}(y\delta + y\delta_1)(\gamma) \\ &= A_{\varepsilon_n}\cdots A_{\varepsilon_1}u(\gamma), \end{aligned}$$

where $u := A_0(y\delta) - (y\delta + y\delta_1)/2$. Similarly, we have

$$c_n(\alpha)y = A_{\varepsilon_n}\cdots A_{\varepsilon_1}v(\gamma),$$

where $v := A_1(y\delta) - y\delta$.

Let $\mathcal{A} := \{A_0, A_1\}$. The norm in (2.6) is chosen to be the ℓ_p-norm. The discussion above tells us that

$$\|b_n y\|_p = \|\mathcal{A}^n u\|_p \quad \text{and} \quad \|c_n y\|_p = \|\mathcal{A}^n v\|_p. \tag{2.11}$$

Write ρ for $\rho_p(A_0|_V, A_1|_V)$. In order to prove that the sequence $(f_n)_{n=1,2,\ldots}$ converges in the L_p-norm, it suffices to show that

$$\lim_{n\to\infty} \|\mathcal{A}^n u\|_p^{1/n} \le \rho \quad \text{and} \quad \lim_{n\to\infty} \|\mathcal{A}^n v\|_p^{1/n} \le \rho. \tag{2.12}$$

Indeed, since $\rho < 2^{1/p}$, we may pick a number σ such that $2^{-1/p}\rho < \sigma < 1$. Hence, there exists a constant C independent of n such that

$$2^{-n/p}\|\mathcal{A}^n u\|_p \le C\sigma^n \quad \text{and} \quad 2^{-n/p}\|\mathcal{A}^n v\|_p \le C\sigma^n.$$

This together with (2.10) and (2.11) yields

$$\|f_{n+1} - f_n\|_p \le 2^{1-(n+1)/p}\big(\|b_n y\|_p + \|c_n y\|_p\big) \le 4C\sigma^n \qquad \forall\, n \in \mathbb{N}.$$

Since $\sigma < 1$, this shows that $(f_n)_{n=1,2,\dots}$ converges to some $\phi \in (L_p(\mathbb{R}))^r$ in the L_p-norm. In the case $p = \infty$, since each f_n is continuous, the limit ϕ is also continuous. Furthermore, since y is an eigenvector of the matrix M corresponding to the eigenvalue 1, we have

$$\hat{f}_n(0) = M^n \hat{f}(0) = M^n y = y.$$

Taking the limit as $n \to \infty$ in the above equation, we obtain $\hat{\phi}(0) = y$.

Let us verify (2.12). For this purpose, we set $v_j := A_j(y\delta) - y\delta$ for $j = 0, 1$. Then $v = v_1$ and $u = v_0 + (y\nabla\delta)/2$. But

$$\lim_{n\to\infty} \|\mathcal{A}^n(y\nabla\delta)\|_p^{1/n} \le \rho.$$

Thus, it suffices to show that

$$\lim_{n\to\infty} \|\mathcal{A}^n(v_0 + v_1)\|_p^{1/n} \le \rho \quad \text{and} \quad \lim_{n\to\infty} \|\mathcal{A}^n(v_0 - v_1)\|_p^{1/n} \le \rho. \tag{2.13}$$

To verify the first inequality in (2.13), we observe that

$$v_0 + v_1 = A_0(y\delta) - y\delta + A_1(y\delta) - y\delta = \sum_{\alpha\in\mathbb{Z}} \big[a(2\alpha) + a(2\alpha+1)\big] y\delta_\alpha - 2y\delta.$$

But $\sum_{\alpha\in\mathbb{Z}}[a(2\alpha) + a(2\alpha+1)]\, y = 2My = 2y$. Hence it follows that

$$v_0 + v_1 = \sum\nolimits_{\alpha\in\mathbb{Z}} \big[a(2\alpha) + a(2\alpha+1)\big] y(\delta_\alpha - \delta).$$

Note that only finitely many terms in the above sum do not vanish, while $\delta_\alpha - \delta$ can be written as $-\sum_{\beta=0}^{\alpha-1} \nabla\delta_\beta$. Therefore, $v_0 + v_1$ can be written as a finite linear combination of $e_j(\nabla\delta_\beta)$, $j = 1, \dots, r$, $\beta \in \mathbb{Z}$. We claim that

$$\|\mathcal{A}^n(e_j\nabla\delta_\beta)\|_p = \|\mathcal{A}^n(e_j\nabla\delta)\|_p \qquad \forall\, \beta \in \mathbb{Z}.$$

Indeed, for $w \in \mathbb{C}^r$ and $\alpha = \varepsilon_1 + 2\varepsilon_2 + \cdots + 2^{n-1}\varepsilon_n + 2^n\gamma$, where $\varepsilon_1, \dots, \varepsilon_n \in \{0, 1\}$ and $\gamma \in \mathbb{Z}$, by (2.8) we have

$$A_{\varepsilon_n} \cdots A_{\varepsilon_1}\big[w(\delta_\beta - \delta_{\beta+1})\big](\gamma) = a_n(\alpha - \beta)w - a_n(\alpha - \beta - 1)w.$$

Note that $\|a_n(\cdot - \beta)w - a_n(\cdot - \beta - 1)w\|_p = \|a_n w - a_n(\cdot - 1)w\|_p$. Consequently,

$$\|\mathcal{A}^n(w\nabla\delta_\beta)\|_p = \|\nabla a_n w\|_p = \|\mathcal{A}^n(w\nabla\delta)\|_p \qquad \forall\, \beta \in \mathbb{Z}. \tag{2.14}$$

This verifies our claim, and thereby establishes the first inequality in (2.13).

As to the second inequality in (2.13), we observe that

$$A_0(y\delta) = \sum_{\alpha\in\mathbb{Z}} a(2\alpha) y\delta_\alpha = A_1(y\delta_1).$$

It follows that

$$v_0 - v_1 = A_0(y\delta) - A_1(y\delta) = A_1(y\delta_1) - A_1(y\delta) = -A_1(y\nabla\delta).$$

Hence, for $n = 1, 2, \ldots$, we have

$$\|\mathcal{A}^n(v_0 - v_1)\|_p \le \|\mathcal{A}^{n+1}(y\nabla\delta)\|_p.$$

This verifies the second inequality in (2.13). The proof for the sufficiency part of the theorem is complete.

It remains to prove the necessity part of the theorem. Suppose $\phi = (\phi_1, \ldots, \phi_r)^T$ is a solution of (1.1), where $\phi_1, \ldots, \phi_r$ are compactly supported functions in $L_p(\mathbb{R})$ ($\phi_1, \ldots, \phi_r$ are continuous in the case $p = \infty$). Iterating the refinement equation (1.1) n times, we obtain

$$\phi = \sum_{\alpha\in\mathbb{Z}} a_n(\alpha)\phi(2^n \cdot -\alpha),$$

where the sequences a_n ($n = 1, 2, \ldots$) are given by (2.2). It follows that

$$\phi - \phi(\cdot - 1/2^n) = \sum_{\alpha\in\mathbb{Z}} \nabla a_n(\alpha)\phi(2^n \cdot -\alpha).$$

If the shifts of $\phi_1, \ldots, \phi_r$ are stable, then there exists a constant $C_1 > 0$ such that

$$2^{-n/p}\|\nabla a_n\|_p \le C_1\|\phi - \phi(\cdot - 1/2^n)\|_p \qquad \forall\, n \in \mathbb{N}.$$

Consequently, there exists a constant $C > 0$ such that

$$2^{-n/p}\|\nabla a_n e_j\|_p \le C\|\phi - \phi(\cdot - 1/2^n)\|_p \qquad \forall\, j = 1, \ldots, r;\ n \in \mathbb{N}. \tag{2.15}$$

But (2.14) tells us that

$$\lim_{n\to\infty} \|\nabla a_n e_j\|_p^{1/n} = \lim_{n\to\infty} \|\mathcal{A}^n(e_j\nabla\delta)\|_p^{1/n}.$$

Note that $\rho = \max_{1\le j\le r}\{\lim_{n\to\infty} \|\mathcal{A}^n(e_j\nabla\delta)\|_p^{1/n}\}$. Since $\lim_{n\to\infty} \|\phi - \phi(\cdot - 1/2^n)\|_p = 0$, (2.15) holds true only if $2^{-1/p}\rho < 1$. This shows $\rho < 2^{1/p}$, as desired. □

3. Characterization of smoothness. In this section, we give a characterization for the smoothness of solutions of the refinement equation (1.1) in terms of the corresponding refinement mask. Our work is based on the following results from approximation theory: For a function f in $L_p(\mathbb{R})$ (f is continuous in the case $p = \infty$), f lies in $\mathrm{Lip}^*(\nu, L_p(\mathbb{R}))$ ($\nu > 0$) if and only if, for some integer $k > \nu$, there exists a constant $C > 0$ such that

$$\|\nabla_{2^{-n}}^k f\|_p \le C2^{-n\nu} \qquad \forall\, n \in \mathbb{N}.$$

For these results we refer the reader to the work of Boman [1] and Ditzian [9].

Suppose $\phi = (\phi_1, \dots, \phi_r)^T$ is a solution of the refinement equation (1.1). Iterating (1.1) n times, we obtain

$$\phi = \sum_{\alpha\in\mathbb{Z}} a_n(\alpha)\phi(2^n\cdot - \alpha), \tag{3.1}$$

where a_n $(n = 1, 2, \dots)$ are given by (2.2). Applying the difference operator $\nabla_{2^{-n}}$ to both sides of (3.1), we obtain

$$\nabla_{2^{-n}}\phi = \sum_{\alpha\in\mathbb{Z}} a_n(\alpha)\big[\phi(2^n\cdot - \alpha) - \phi(2^n\cdot - \alpha - 1)\big] = \sum_{\alpha\in\mathbb{Z}} \nabla a_n(\alpha)\phi(2^n\cdot - \alpha).$$

For $k = 1, 2, \dots$, an induction argument tells us that

$$\nabla^k_{2^{-n}}\phi = \sum_{\alpha\in\mathbb{Z}} \nabla^k a_n(\alpha)\phi(2^n\cdot - \alpha). \tag{3.2}$$

Suppose $\phi_1, \dots, \phi_r$ are compactly supported functions in $L_p(\mathbb{R})$. It follows from (3.2) that

$$2^{n/p}\big\|\nabla^k_{2^{-n}}\phi\big\|_p \le C_1\big\|\nabla^k a_n\big\|_p \qquad \forall\, n \in \mathbb{N}, \tag{3.3}$$

where C_1 is a constant independent of n. If, in addition, the shifts of $\phi_1, \dots, \phi_r$ are stable, then there exists a constant $C_2 > 0$ such that

$$2^{n/p}\big\|\nabla^k_{2^{-n}}\phi\big\|_p \ge C_2\big\|\nabla^k a_n\big\|_p \qquad \forall\, n \in \mathbb{N}, \tag{3.4}$$

but, for $0 < \nu < k$, ϕ lies in $(\mathrm{Lip}^*(\nu, L_p(\mathbb{R})))^r$ if and only if there exists a constant $C > 0$ such that

$$\|\nabla^k_{2^{-n}}\phi\|_p \le C2^{-n\nu} \qquad \forall\, n \in \mathbb{N}.$$

Thus, we have established the following result.

LEMMA 3.1. *Suppose $\phi = (\phi_1, \dots, \phi_r)^T \in (L_p(\mathbb{R}))^r$ ($\phi \in (C(\mathbb{R}))^r$ in the case $p = \infty$) is a compactly supported solution of the refinement equation* (1.1) *with mask a. For $n = 1, 2, \dots$, let a_n be given by the iteration relation* (2.2). *Let $k > \nu > 0$, where k is an integer. If there exists a constant $C > 0$ such that*

$$\|\nabla^k a_n\|_p \le C2^{-n(\nu-1/p)} \qquad \forall\, n \in \mathbb{N}, \tag{3.5}$$

then ϕ belongs to $(Lip^*(\nu, L_p(\mathbb{R})))^r$. Conversely, if ϕ lies in $(Lip^*(\nu, L_p(\mathbb{R})))^r$, then (3.5) *holds true, provided the shifts of $\phi_1, \dots, \phi_r$ are stable.*

For two elements b and c in $\ell_0(\mathbb{Z})$, the *discrete convolution* of b and c, denoted by $b*c$, is defined by

$$b*c(\alpha) = \sum_{\beta\in\mathbb{Z}} b(\alpha - \beta)c(\beta), \qquad \alpha \in \mathbb{Z}.$$

Evidently, $b*\delta = b$ for any $b \in \ell_0(\mathbb{Z})$. If $b \in (\ell_0(\mathbb{Z}))^{r\times r}$ and $c \in (\ell_0(\mathbb{Z}))^r$, then $b*c$ is defined in a similar way.

LEMMA 3.2. *Let a be an element of $(\ell_0(\mathbb{Z}))^{r\times r}$, and let a_n $(n = 1, 2, \dots)$ be given by the iteration relation* (2.2). *For $\varepsilon = 0, 1$, let A_ε be the linear operator on $(\ell_0(\mathbb{Z}))^r$ given by* (2.7). *Then, for each integer $k \ge 0$,*

$$\lim_{n\to\infty}\big\|\nabla^k a_n\big\|_p^{1/n} = \rho_p(A_0|_V, A_1|_V),$$

where V is the minimal common invariant subspace of A_0 and A_1 generated by $e_j(\nabla^k\delta)$, $j = 1, \ldots, r$.

Proof. Write $\mathcal{A}$ for $\{A_0, A_1\}$. For an element $w \in (\ell_0(\mathbb{Z}))^r$, the quantity $\|\mathcal{A}^n w\|_p$ is defined as in (2.6) with the ℓ_p-norm being the underlying vector norm on $(\ell_0(\mathbb{Z}))^r$.

Let v be an element in $(\ell_0(\mathbb{Z}))^r$. We observe that

$$(\nabla^k a_n) * v = a_n * (\nabla^k v).$$

Suppose $\alpha = \varepsilon_1 + 2\varepsilon_2 + \cdots + 2^{n-1}\varepsilon_n + 2^n\gamma$, where $\varepsilon_1, \ldots, \varepsilon_n \in \{0, 1\}$ and $\gamma \in \mathbb{Z}$. Then by (2.4) we have

$$\begin{aligned}\big(\nabla^k a_n * v\big)(\alpha) &= \sum_{\beta\in\mathbb{Z}} a_n(\alpha - \beta)\nabla^k v(\beta) \\ &= \sum_{\beta\in\mathbb{Z}} A_{\varepsilon_n}\cdots A_{\varepsilon_1}(\gamma, \beta)\nabla^k v(\beta) = A_{\varepsilon_n}\cdots A_{\varepsilon_1}(\nabla^k v)(\gamma).\end{aligned}$$

It follows that

$$\|\nabla^k a_n * v\|_p = \|\mathcal{A}^n(\nabla^k v)\|_p. \tag{3.6}$$

Choosing $v = e_j\delta$ in (3.6), we obtain

$$\|(\nabla^k a_n)e_j\|_p = \|(\nabla^k a_n) * (e_j\delta)\|_p = \|\mathcal{A}^n(e_j\nabla^k\delta)\|_p.$$

This shows that

$$\lim_{n\to\infty} \|(\nabla^k a_n)e_j\|_p^{1/n} = \lim_{n\to\infty} \|\mathcal{A}^n(e_j\nabla^k\delta)\|_p^{1/n} = \rho_p(A_0|_{V_j}, A_1|_{V_j}),$$

where V_j is the minimal common invariant subspace of A_0 and A_1 generated by $e_j(\nabla^k\delta)$. But

$$\|\nabla^k a_n\|_p = \left(\sum_{j=1}^r \|(\nabla^k a_n)e_j\|_p^p\right)^{1/p}.$$

Therefore, we arrive at the conclusion that

$$\lim_{n\to\infty} \big\|\nabla^k a_n\big\|_p^{1/n} = \max\big\{\rho_p(A_0|_{V_j}, A_1|_{V_j}) : j = 1, \ldots, r\big\} = \rho_p(A_0|_V, A_1|_V),$$

where V is the sum of $V_1, \ldots, V_r$. ∎

We are in a position to prove the main result of this paper.

THEOREM 3.3. *Let a be an element of $(\ell_0(\mathbb{Z}))^{r\times r}$. For $\varepsilon = 0, 1$, let A_ε be the linear operator on $(\ell_0(\mathbb{Z}))^r$ given by (2.7). Let k be a positive integer and V the minimal common invariant subspace of A_0 and A_1 generated by $e_j(\nabla^k\delta)$, $j = 1, \ldots, r$. Suppose $\phi = (\phi_1, \ldots, \phi_r)^T$ is compactly supported and lies in $(L_p(\mathbb{R}))^r$ for $1 \le p < \infty$ (f lies in $(C(\mathbb{R}))^r$ for $p = \infty$). If ϕ is a solution of the refinement equation (1.1) with the mask a, then*

$$\nu_p(\phi) \ge 1/p - \log_2 \rho_p(A_0|_V, A_1|_V). \tag{3.7}$$

In addition, if the shifts of $\phi_1, \ldots, \phi_r$ are stable and if $k > 1/p - \log_2 \rho_p(A_0|_V, A_1|_V)$, then equality holds in (3.7).

Proof. Write ρ for $\rho_p(A_0|_V, A_1|_V)$. By Lemma 3.2, we have $\lim_{n\to\infty} \|\nabla^k a_n\|_p^{1/n} = \rho$. Thus, for $\varepsilon > 0$, there exists a constant $C > 0$ such that, $\forall n \in \mathbb{N}$,

$$\|\nabla^k a_n\|_p \le C(\rho+\varepsilon)^n = C2^{n\log_2(\rho+\varepsilon)} = C2^{-n(\nu-1/p)},$$

where $\nu := 1/p - \log_2(\rho+\varepsilon)$. By Lemma 3.1, ϕ belongs to $(\mathrm{Lip}^*(\nu, L_p(\mathbb{R})))^r$. This shows that

$$\nu_p(\phi) \ge 1/p - \log_2(\rho+\varepsilon),$$

but $\varepsilon > 0$ can be arbitrarily small; hence, we obtain

$$\nu_p(\phi) \ge 1/p - \log_2 \rho.$$

Now suppose $k > 1/p - \log_2\rho$ and the shifts of $\phi_1, \dots, \phi_r$ are stable. We wish to show $\nu_p(\phi) \le 1/p - \log_2\rho$. If this is not true, then there exists μ such that $1/p - \log_2\rho < \mu < k$ and $\phi \in (\mathrm{Lip}^*(\mu, L_p(\mathbb{R})))^r$. By Lemma 3.1, there exists a constant $C > 0$ such that

$$\|\nabla^k a_n\|_p \le C2^{-n(\mu-1/p)} \qquad \forall\, n \in \mathbb{N}.$$

By Lemma 3.2 we get

$$\rho = \lim_{n\to\infty} \|\nabla^k a_n\|_p^{1/n} \le 2^{-\mu+1/p}.$$

It follows that

$$\mu \le 1/p - \log_2\rho,$$

which contradicts the assumption $\mu > 1/p - \log_2\rho$. Therefore, we obtain the desired result $\nu_p(\phi) \le 1/p - \log_2\rho$. □

The case $p = 2$ is of particular interest. In this case, the smoothness is usually measured by using Sobolev spaces. For $\nu \ge 0$ we denote by $W_2^\nu(\mathbb{R})$ the Sobolev space of all functions $f \in L_2(\mathbb{R})$ such that

$$\int_{\mathbb{R}} \big|\hat f(\xi)\big|^2 \big(1+|\xi|^\nu\big)^2\, d\xi < \infty.$$

It is well known that, for $\nu > \varepsilon > 0$, the inclusion relations

$$W_2^\nu(\mathbb{R}) \subseteq \mathrm{Lip}^*(\nu, L_2(\mathbb{R})) \subseteq W_2^{\nu-\varepsilon}(\mathbb{R})$$

hold true. Therefore, for a vector $f = (f_1, \dots, f_r)^T \in (L_2(\mathbb{R}))^r$, we have

$$\nu_2(f) = \sup\big\{\nu : f \in (W_2^\nu(\mathbb{R}))^r\big\}.$$

In [29] Shen obtained lower bounds for the L_2-smoothness of refinable vectors.

When $p = 2$, the joint spectral radius in (3.7) can be computed by finding the spectral radius of a certain finite matrix associated to the mask a (see [11] and [21]).

Let us review some related results from [21]. For an element $a \in (\ell_0(\mathbb{Z}))^{r\times r}$, define the *transition operator* F_a to be the linear mapping from $(\ell_0(\mathbb{Z}))^{r\times r}$ to $(\ell_0(\mathbb{Z}))^{r\times r}$ given by

$$F_a w(\alpha) := \sum_{\beta,\gamma\in\mathbb{Z}} a(2\alpha-\beta)w(\beta+\gamma)a(\gamma)^*/2, \qquad \alpha \in \mathbb{Z},\ w \in (\ell_0(\mathbb{Z}))^{r\times r}, \tag{3.8}$$

where $a(\gamma)^*$ denotes the complex conjugate transpose of $a(\gamma)$. For $n = 1, 2, \ldots$, let a_n be the sequences given by the iteration relation (2.2). It was proved [21, Lemma 7.3] that, for any $v \in (\ell_0(\mathbb{Z}))^r$,

$$\lim_{n\to\infty} \|a_n * v\|_2^{1/n} = \sqrt{2\rho(F_a|_W)}, \tag{3.9}$$

where W is the minimal invariant subspace of F_a generated by the element $w \in (\ell_0(\mathbb{Z}))^{r\times r}$ given by

$$w(\beta) := \sum_{\gamma\in\mathbb{Z}} v(\beta+\gamma)v(\gamma)^*, \qquad \beta \in \mathbb{Z}.$$

Let Δ denote the difference operator on $\ell_0(\mathbb{Z})$ given by

$$\Delta v := -v(\cdot - 1) + 2v - v(\cdot + 1), \qquad v \in \ell_0(\mathbb{Z}).$$

In particular, $\Delta\delta := -\delta_{-1} + 2\delta - \delta_1$. Suppose V is the minimal common invariant subspace of A_0 and A_1 generated by $e_j(\nabla^k\delta)$, $j = 1, \ldots, r$. Then Lemma 3.2 and (3.9) tell us that

$$\rho_2(A_0|_V, A_1|_V) = \sqrt{2\rho(F_a|_W)},$$

where W is the minimal invariant subspace of F_a generated by $e_j e_j^T(\Delta^k\delta)$, $j = 1, \ldots, r$.

Thus, for the case $p = 2$, Theorem 3.3 can be strengthened as follows.

THEOREM 3.4. *Suppose $\phi = (\phi_1, \ldots, \phi_r)^T \in (L_2(\mathbb{R}))^r$ is a compactly supported solution of the refinement equation* (1.1) *with mask a. Let F_a be the transition operator given in* (3.8)*. Then, for any positive integer k,*

$$\nu_2(\phi) \ge -\log_2 \sqrt{\rho(F_a|_W)},$$

where W is the minimal invariant subspace of F_a generated by $e_j e_j^T(\Delta^k\delta)$, $j = 1, \ldots, r$. Moreover, $\nu_2(\phi) = -\log_2 \sqrt{\rho(F_a|_W)}$, provided $k > -\log_2 \sqrt{\rho(F_a|_W)}$ and the shifts of $\phi_1, \ldots, \phi_r$ are stable.

In order to apply Theorems 3.3 and 3.4 to smoothness analysis, one must check the stability of refinable functions in terms of the refinement mask. For the scalar case ($r = 1$), Jia and Wang [22] gave a characterization for the stability and linear independence of the shifts of a refinable function in terms of the refinement mask. Their results were extended by Zhou [33] to the case where the scaling factor is an arbitrary integer greater than 1. For the vector case ($r > 1$), stability of the shifts of multiple refinable functions was discussed by Hervé [15], Hogan [16], and Wang [32]. Assuming the vector of refinable functions lies in $(L_2(\mathbb{R}))^r$, Shen [29] gave a characterization for L_2-stability.

4. Examples. In this section, we give two examples to illustrate the general theory.

Let A be a linear operator on a linear space V with $\{v_1, \ldots, v_s\}$ as its basis. Suppose $Av_j = \sum_{k=1}^s a_{jk}v_k$ for $1 \le j \le s$. Then the matrix $(a_{jk})_{1\le j,k\le s}$ is said to be the matrix representation of A.

The definition of joint spectral radius given in section 2 also applies to a finite multiset of square matrices of the same size. Indeed, an $s \times s$ matrix can be viewed as a linear operator on $\mathbb{C}^s$. Obviously, the p-norm joint spectral radius of a finite multiset

of linear operators is the same as that of the multiset of the matrices representing those linear operators.

Now suppose $\mathcal{A} = \{A_0, A_1\}$, where $A_0 = (\lambda)$ and $A_1 = (\mu)$ are two 1×1 matrices. For $\varepsilon_1, \ldots, \varepsilon_n \in \{0, 1\}$, we have

$$A_{\varepsilon_1} \cdots A_{\varepsilon_n} = \prod_{j=1}^{n} (\lambda^{1-\varepsilon_j} \mu^{\varepsilon_j}).$$

Hence for $1 \le p < \infty$,

$$\|\mathcal{A}^n\|_p = \big(|\lambda|^p + |\mu|^p\big)^{n/p},$$

while

$$\|\mathcal{A}^n\|_\infty = \big(\max\{|\lambda|, |\mu|\}\big)^n.$$

Therefore we obtain

$$\rho_p(A_0, A_1) = \big(|\lambda|^p + |\mu|^p\big)^{1/p}, \qquad 1 \le p \le \infty, \tag{4.1}$$

where the right-hand side of (4.1) is interpreted as $\max\{|\lambda|, |\mu|\}$ for the case $p = \infty$.

Suppose $\mathcal{A} = \{A_1, \ldots, A_m\}$ and each A_j is a block triangular matrix:

$$A_j = \begin{pmatrix} E_j & 0 \\ G_j & F_j \end{pmatrix}, \qquad j = 1, \ldots, m,$$

where $E_1, \ldots, E_m$ are square matrices of the same size, and so are $F_1, \ldots, F_m$. It was proved [21, Lemma 4.2] that

$$\rho_p(A_1, \ldots, A_m) = \max\{\rho_p(E_1, \ldots, E_m), \rho_p(F_1, \ldots, F_m)\}, \qquad 1 \le p \le \infty. \tag{4.2}$$

Let A_0 and A_1 be two triangular matrices of the same type:

$$A_0 = \begin{bmatrix} \lambda_{11} & & & \\ \lambda_{12} & \lambda_{22} & & \\ \vdots & \vdots & \ddots & \\ \lambda_{s1} & \lambda_{s2} & \cdots & \lambda_{ss} \end{bmatrix} \quad \text{and} \quad A_1 = \begin{bmatrix} \mu_{11} & & & \\ \mu_{12} & \mu_{22} & & \\ \vdots & \vdots & \ddots & \\ \mu_{s1} & \mu_{s2} & \cdots & \mu_{ss} \end{bmatrix}.$$

Then (4.1) and (4.2) tell us that

$$\rho_p(A_0, A_1) = \max_{1 \le j \le s} \big(|\lambda_{jj}|^p + |\mu_{jj}|^p\big)^{1/p}, \qquad 1 \le p \le \infty. \tag{4.3}$$

Let us analyze the following example considered by Donovann et al. [10]. Suppose a is a sequence on $\mathbb{Z}$ supported on $[0, 3]$ and

$$a(0) = \begin{bmatrix} h_1 & 1 \\ h_2 & h_3 \end{bmatrix}, \qquad a(1) = \begin{bmatrix} h_1 & 0 \\ h_4 & 1 \end{bmatrix},$$

$$a(2) = \begin{bmatrix} 0 & 0 \\ h_4 & h_3 \end{bmatrix}, \qquad a(3) = \begin{bmatrix} 0 & 0 \\ h_2 & 0 \end{bmatrix},$$

where

$$h_1 = -\frac{s^2-4s-3}{2(s+2)}, \qquad h_2 = -\frac{3(s^2-1)(s^2-3s-1)}{4(s+2)^2},$$

$$h_3 = \frac{3s^2+s-1}{2(s+2)}, \qquad h_4 = -\frac{3(s^2-1)(s^2-s+3)}{4(s+2)^2}.$$

The matrix $M := \sum_{\alpha=0}^{3} a(\alpha)/2$ has two eigenvalues, 1 and s. We assume that $|s| < 1$. The eigenvectors of M corresponding to the eigenvalue 1 are cy, where $c \neq 0$ and

$$y = \begin{bmatrix} 1 \\ (s-1)^2/(s+2) \end{bmatrix}.$$

EXAMPLE 4.1. *Let $\phi = (\phi_1, \phi_2)^T$ be the solution of the refinement equation with the mask a such that $\hat{\phi}(0) = y$. Then*

$$\nu_p(\phi) = \begin{cases} 1 + 1/p & \text{if } |s| < 2^{-1-1/p}, \\ -\log_2 |s| & \text{if } 2^{-1-1/p} \le |s| < 1. \end{cases}$$

Proof. First, we prove that the solution ϕ is continuous, provided $|s| < 1$. For this purpose, let A_ε ($\varepsilon = 0, 1$) be the linear operators on $(\ell_0(\mathbb{Z}))^2$ given by

$$A_\varepsilon v(\alpha) = \sum_{\beta \in \mathbb{Z}} a(\varepsilon + 2\alpha - \beta) v(\beta), \qquad \alpha \in \mathbb{Z},\ v \in (\ell_0(\mathbb{Z}))^2.$$

Since a is supported on $[0, 3]$, the linear space $(\ell([0,3]))^2$ is invariant under both A_0 and A_1. Choose

$$\{e_1\delta_0, e_2\delta_0, e_1\delta_1, e_2\delta_1, e_1\delta_2, e_2\delta_2, e_1\delta_3, e_2\delta_3\}$$

as a basis for $(\ell([0,3]))^2$. With respect to this basis, the matrix representations of A_0 and A_1 are $(a(2\beta-\alpha)^T)_{0\le\alpha,\beta\le3}$ and $(a(1+2\beta-\alpha)^T)_{0\le\alpha,\beta\le3}$, respectively. We have

$$(a(2\beta-\alpha)^T)_{0\le\alpha,\beta\le3} = \begin{bmatrix} h_1 & h_2 & 0 & h_4 & & & & \\ 1 & h_3 & 0 & h_3 & & & & \\ & & h_1 & h_4 & 0 & h_2 & & \\ & & 0 & 1 & 0 & 0 & & \\ & & h_1 & h_2 & 0 & h_4 & & \\ & & 1 & h_3 & 0 & h_3 & & \\ & & & & h_1 & h_4 & 0 & h_2 \\ & & & & 0 & 1 & 0 & 0 \end{bmatrix}$$

and

$$(a(1+2\beta-\alpha))^T_{0\le\alpha,\beta\le3} = \begin{bmatrix} h_1 & h_4 & 0 & h_2 & & & & \\ 0 & 1 & 0 & 0 & & & & \\ h_1 & h_2 & 0 & h_4 & & & & \\ 1 & h_3 & 0 & h_3 & & & & \\ & & h_1 & h_4 & 0 & h_2 & & \\ & & 0 & 1 & 0 & 0 & & \\ & & h_1 & h_2 & 0 & h_4 & 0 & 0 \\ & & 1 & h_3 & 0 & h_3 & 0 & 0 \end{bmatrix}.$$

We observe that s is a common eigenvalue of both A_0 and A_1. Corresponding to this eigenvalue, A_0 and A_1 have a common eigenvector v_1 given by

$$v_1 := \begin{bmatrix} 1 \\ -\mu \end{bmatrix} \delta + \begin{bmatrix} 0 \\ -\mu \end{bmatrix} \delta_1,$$

where

$$\mu := \frac{3(1-s^2)}{2(s+2)}.$$

This motivates us to choose

$$v_2 := \begin{bmatrix} 1 \\ -\mu \end{bmatrix} \delta_1 + \begin{bmatrix} 0 \\ -\mu \end{bmatrix} \delta_2.$$

It is easily verified that $A_0 v_2 = s v_2$ and $A_1 v_2 = s v_1$. From Theorem 2.1, we need to add the generators $e_j \nabla\delta$, $j = 1, 2$, which motivates us to set $v_3 := e_2 \nabla\delta$ and $v_4 := e_2 \nabla\delta_1$, i.e.,

$$v_3 := \begin{bmatrix} 0 \\ 1 \end{bmatrix} \delta - \begin{bmatrix} 0 \\ 1 \end{bmatrix} \delta_1 \qquad \text{and} \qquad v_4 := \begin{bmatrix} 0 \\ 1 \end{bmatrix} \delta_1 - \begin{bmatrix} 0 \\ 1 \end{bmatrix} \delta_2.$$

Denote by V the linear span of v_1, v_2, v_3, and v_4. Then $e_1 \nabla\delta = v_1 - v_2 + \mu(v_3 + v_4) \in V$ and $e_2 \nabla\delta = v_3 \in V$. Using the matrix representations of A_0 and A_1 we find that

$$A_0 \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} s & & & \\ 0 & s & & \\ 1 & 0 & 0.5 & \\ 0 & -1 & 0 & 0.5 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}$$

and

$$A_1 \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} s & & & \\ s & 0 & & \\ -1 & 0 & 0.5 & \\ 1 & 0 & 0.5 & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}.$$

Thus, V is invariant under both A_0 and A_1. Applying (4.3) to the two 4×4 matrices above, we obtain

$$\rho_\infty(A_0|_V, A_1|_V) = \max\{1/2, |s|\}.$$

Thus, $\rho_\infty(A_0|_V, A_1|_V) < 1$ for $|s| < 1$. Therefore, by Theorem 2.1, the solution ϕ is continuous, provided $|s| < 1$.

We claim that the shifts of ϕ_1 and ϕ_2 are stable. For this purpose it suffices to show that the shifts of ϕ_1 and ϕ_2 are linearly independent (see [19]), that is,

$$\sum_{\alpha\in\mathbb{Z}} b(\alpha)\phi_1(\cdot - \alpha) + \sum_{\alpha\in\mathbb{Z}} c(\alpha)\phi_2(\cdot - \alpha) = 0 \quad \Longrightarrow \quad b(\alpha) = c(\alpha) = 0 \quad \forall\, \alpha \in \mathbb{Z}. \tag{4.4}$$

In order to verify (4.4), we first compute $\phi(\alpha)$ for $\alpha \in \mathbb{Z}$. Since ϕ is supported on $[0, 3]$, we have $\phi(\alpha) = 0$ for $\alpha \in \mathbb{Z} \setminus \{1, 2\}$. The vector $\phi = (\phi_1, \phi_2)^T$ satisfies the refinement equation

$$\phi(x) = \sum_{\alpha\in\mathbb{Z}} a(\alpha)\phi(2x - \alpha) \qquad \forall\, x \in \mathbb{R}. \tag{4.5}$$

In particular,

$$\phi(\beta) = \sum_{\alpha=1}^{2} a(2\beta - \alpha)\phi(\alpha) \qquad \text{for } \beta = 1, 2.$$

Solving the above system of linear equations, we get

$$\phi(1) = (0, t)^T \quad \text{and} \quad \phi(2) = 0,$$

where t is a nonzero constant. Moreover, it follows from (4.5) that

$$\phi(\beta + 1/2) = \sum_{\alpha=1}^{2} a(2\beta + 1 - \alpha)\phi(\alpha) \qquad \forall\, \beta \in \mathbb{Z}.$$

Consequently, $\phi(1/2) = t(1, h_3)^T$, $\phi(3/2) = t(0, h_3)^T$, and $\phi(\beta + 1/2) = 0$ for $\beta \in \mathbb{Z} \setminus \{0, 1\}$. Suppose

$$\sum_{\alpha\in\mathbb{Z}} b(\alpha)\phi_1(x - \alpha) + \sum_{\alpha\in\mathbb{Z}} c(\alpha)\phi_2(x - \alpha) = 0 \qquad \forall x \in \mathbb{R}. \tag{4.6}$$

Choosing $x = \beta$ for $\beta \in \mathbb{Z}$ in (4.6), we obtain $c(\beta - 1) = 0$. This is true for all $\beta \in \mathbb{Z}$. Hence (4.6) implies

$$\sum_{\alpha\in\mathbb{Z}} b(\alpha)\phi_1(x - \alpha) = 0 \qquad \forall x \in \mathbb{R}.$$

For $\beta \in \mathbb{Z}$, setting $x = \beta + 1/2$ in the above equation gives $b(\beta) = 0$. Thus, (4.4) has been verified.

We are in a position to determine the smoothness of ϕ. For the subspace W, we retain the first two generators, v_1, v_2, but replace the others by $e_j(\nabla^2\delta)$, $j = 1, 2$, as required by Theorem 3.3:

$$w_1 := \begin{bmatrix} 1 \\ -\mu \end{bmatrix} \delta + \begin{bmatrix} 0 \\ -\mu \end{bmatrix} \delta_1, \qquad w_2 := \begin{bmatrix} 1 \\ -\mu \end{bmatrix} \delta_1 + \begin{bmatrix} 0 \\ -\mu \end{bmatrix} \delta_2,$$
$$w_3 := e_2(\nabla^2\delta), \qquad \text{and} \qquad w_4 := e_1(\nabla^2\delta).$$

Using the matrix representations of A_0 and A_1, we find

$$A_0 \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} = \begin{bmatrix} s & & & \\ 0 & s & & \\ 1 & 1 & 0.5 & \\ h_1 & -h_1 & \mu/2 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix}$$

and

$$A_1 \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} = \begin{bmatrix} s & & & \\ s & 0 & & \\ -2 & 0 & 0 & \\ -h_1 & h_1 & \mu/2 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix}.$$

Let W be the linear span of w_1, w_2, w_3, and w_4. Then W is the minimal common invariant subspace of A_0 and A_1 generated by $e_1(\nabla^2\delta)$ and $e_2(\nabla^2\delta)$. Applying (4.3) to the two 4×4 matrices above, we obtain

$$\rho_p(A_0|_W, A_1|_W) = \max\{2^{1/p}|s|, 1/2\}.$$

If $|s| < 2^{-1-1/p}$, then $\rho_p(A_0|_W, A_1|_W) = 1/2$; hence $\nu_p(\phi) = 1 + 1/p$ for $1 < p \le \infty$, by Theorem 3.3. When $p = 1$, we have $\nu_1(\phi) \ge 2$. But $\nu_1(\phi) > 2$ is impossible. Indeed, $\nu_1(\phi) > 2$ would imply $\nu_p(\phi) > 2$ for some $p > 1$, by the embedding theorem. But we know that $\nu_p(\phi) < 2$ for $1 < p \le \infty$. Therefore, $\nu_1(\phi) = 2$ for $|s| < 1/4$. If $2^{-1-1/p} \le |s| < 1$, then $\rho_p(A_0|_W, A_1|_W) = 2^{1/p}|s|$. Theorem 3.3 tells us that $\nu_p(\phi) = -\log_2 |s|$ for $1 \le p \le \infty$. Thus, for $1 \le p \le \infty$, we have found the optimal L_p-smoothness of ϕ explicitly. □

Using fractal interpolation, Donovan et al. [10] showed that $\phi \in \operatorname{Lip} 1$ for $|s| < 1/2$ and $\phi \in \operatorname{Lip} \nu$ for $1/2 < |s| < 1$, where $\nu = -\log_2 |s|$. In [3], Cohen, Daubechies, and Plonka established the continuity of ϕ for $|s| < 1/2$. The case $p = 1$ was considered by Micchelli and Sauer [25], who obtained $\nu_1(\phi) > 1.1087$ for $s = -0.2$. In comparison with their result, our method gives $\nu_1(\phi) = 2$ for $|s| < 1/4$.

Our second example is taken from [20, 21]. Let a be the element in $(\ell_0(\mathbb{Z}))^{2\times 2}$ supported in $[0, 2]$ given by

$$a(0) = \begin{bmatrix} \frac{1}{2} & \frac{s}{2} \\ t & \lambda \end{bmatrix}, \qquad a(1) = \begin{bmatrix} 1 & 0 \\ 0 & \mu \end{bmatrix}, \quad \text{and} \quad a(2) = \begin{bmatrix} \frac{1}{2} & -\frac{s}{2} \\ -t & \lambda \end{bmatrix}. \tag{4.7}$$

The matrix $M := \sum_{\alpha=0}^{2} a(\alpha)/2$ has two eigenvalues: 1 and $\lambda + \mu/2$. We assume that $|2\lambda + \mu| < 2$. Then there exists a unique distributional solution $\phi = (\phi_1, \phi_2)^T$ of the refinement equation with the mask a subject to $\hat{\phi}(0) = (1, 0)^T$. The distribution ϕ_1 is symmetric about 1, and ϕ_2 is antisymmetric about 1. It was proved [20, Example 4.3] that the shifts of ϕ_1 and ϕ_2 reproduce all quadratic polynomials if and only if

$$t \ne 0, \quad \mu = 1/2, \quad \text{and} \quad \lambda = 1/4 + 2st. \tag{4.8}$$

In this case, the condition $|2\lambda + \mu| < 2$ reduces to $-3/4 < st < 1/4$.

EXAMPLE 4.2. *Let a be the mask given in* (4.7) *with $-3/4 < st < 1/4$. Let $\phi = (\phi_1, \phi_2)^T$ be the solution of the refinement equation with the mask a such that $\hat{\phi}(0) = (1, 0)^T$. Suppose the conditions in* (4.8) *are satisfied. Then, for $s \ne 0$ we have*

$$\nu_p(\phi) = \begin{cases} 2 + \frac{1}{p} & \text{if } |st + 1/4| \le 2^{-3-1/p}, \\ -\log_2 |\frac{1}{2} + 2st| & \text{if } 2^{-3-1/p} < |st + 1/4| < 1/2. \end{cases}$$

In the case $s = 0$, $\mu = 1/2$, and $\lambda = 1/4$, we have $\nu_p(\phi) = 1 + 1/p$.

Proof. First, we investigate the case $s = 0$. Under the conditions in (4.8), the refinement equation

$$\phi(x) = \sum_{\alpha \in \mathbb{Z}} a(\alpha)\phi(2x - \alpha) \qquad \forall\, x \in \mathbb{R} \tag{4.9}$$

can be solved explicitly (see [20, Example 4.3]). The solution $\phi = (\phi_1, \phi_2)^T$ of (4.9) subject to $\hat{\phi}(0) = (1, 0)^T$ is given by

$$\phi_1(x) = \begin{cases} x & \text{for } 0 \le x < 1, \\ 2 - x & \text{for } 1 \le x \le 2, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\phi_2(x) = \begin{cases} 4tx(1 - x) & \text{for } 0 \le x < 1, \\ -4t(2 - x)(x - 1) & \text{for } 1 \le x \le 2, \\ 0 & \text{otherwise.} \end{cases}$$

Consequently, $\nu_p(\phi) = 1+1/p$ for $1 \le p \le \infty$. In this case, the shifts of ϕ_2 are linearly *dependent*.

Second, we consider the case $s \neq 0$. Under the conditions in (4.8), the solution $\phi = (\phi_1, \phi_2)^T$ is continuous, provided $-3/4 < st < 1/4$ (see [21, Example 6.3]). In this case, we claim that the shifts of ϕ_1 and ϕ_2 are linearly independent. To justify our claim, we find $\phi(\alpha)$ for $\alpha \in \mathbb{Z}$. Since ϕ is supported on $[0, 2]$, we have $\phi(\alpha) = 0$ for $\alpha \in \mathbb{Z} \setminus \{1\}$. From [20, Example 3.2] we see that $\sum_{\alpha\in\mathbb{Z}} \phi_1(\alpha) = 1$. Hence $\phi_1(1) = 1$. Moreover, it follows from (4.9) that $\phi(1) = a(1)\phi(1)$, which implies $\phi_2(1) = 0$. Next, we find $\phi(\beta + 1/2)$ for $\beta \in \mathbb{Z}$. Using the refinement equation (4.9), we obtain

$$\phi(\beta + 1/2) = \sum_{\alpha\in\mathbb{Z}} a(\alpha)\phi(2\beta + 1 - \alpha) = a(2\beta)[1, 0]^T.$$

Therefore, $\phi(1/2) = [1/2, t]^T$, $\phi(3/2) = [1/2, -t]^T$, and $\phi(\beta+1/2) = 0\ \forall\beta \in \mathbb{Z}\setminus\{0, 1\}$. Furthermore, we can use (4.9) to find $\phi(\gamma + 1/4)\ \forall\gamma \in \mathbb{Z}$. As a result, we obtain $\phi(\gamma + 1/4) = 0\ \forall\gamma \in \mathbb{Z} \setminus \{0, 1\}$,

$$\phi(1/4) = \sum_{\alpha\in\mathbb{Z}} a(\alpha)\phi(1/2 - \alpha) = a(0)\phi(1/2),$$

and

$$\phi(5/4) = \sum_{\alpha\in\mathbb{Z}} a(\alpha)\phi(5/2 - \alpha) = a(1)\phi(3/2) + a(2)\phi(1/2).$$

Suppose

$$\sum_{\alpha\in\mathbb{Z}} b(\alpha)\phi_1(x - \alpha) + \sum_{\alpha\in\mathbb{Z}} c(\alpha)\phi_2(x - \alpha) = 0 \qquad \forall x \in \mathbb{R}. \tag{4.10}$$

Choosing $x = \beta$ for $\beta \in \mathbb{Z}$ in the above equation, we obtain $b(\beta - 1) = 0$. This is true for all $\beta \in \mathbb{Z}$. Hence (4.10) implies

$$\sum_{\alpha\in\mathbb{Z}} c(\alpha)\phi_2(x - \alpha) = 0 \qquad \forall x \in \mathbb{R}. \tag{4.11}$$

For $\beta \in \mathbb{Z}$, setting $x = \beta + 1/2$ in the above equation gives

$$t\,[c(\beta) - c(\beta - 1)] = 0.$$

Since $t \neq 0$, we have $c(\beta) = c(\beta - 1)\ \forall\beta \in \mathbb{Z}$. Setting $x = 5/4$ in (4.11), we get

$$c(0)[\phi_2(1/4) + \phi_2(5/4)] = 0, \quad \text{i.e.}, \quad c(0)(2\lambda - \mu)t = 0.$$

But $(2\lambda - \mu)t = 4st^2 \neq 0$; hence $c(0) = 0$. So $c(\beta) = c(0) = 0$ for all $\beta \in \mathbb{Z}$. This justifies our claim that the shifts of ϕ_1 and ϕ_2 are linearly independent.

We are in a position to determine the smoothness of ϕ. Let A_ε ($\varepsilon = 0, 1$) be the linear operators on $(\ell_0(\mathbb{Z}))^2$ given by

$$A_\varepsilon v(\alpha) = \sum_{\beta\in\mathbb{Z}} a(\varepsilon + 2\alpha - \beta)v(\beta), \qquad \alpha \in \mathbb{Z},\ v \in (\ell_0(\mathbb{Z}))^2.$$

We observe that $1/2 + 2st$ is a common eigenvalue of both A_0 and A_1. Corresponding to this eigenvalue, A_0 and A_1 have a common eigenvector v_1 given by

$$v_1 := \begin{bmatrix} 1 \\ 4t \end{bmatrix} \delta + \begin{bmatrix} -1 \\ 4t \end{bmatrix} \delta_1.$$

This motivates us to choose

$$v_2 := \begin{bmatrix} 1 \\ 4t \end{bmatrix} \delta_1 + \begin{bmatrix} -1 \\ 4t \end{bmatrix} \delta_2.$$

Then we have $A_0 v_2 = (1/2 + 2st)v_2$ and $A_1 v_2 = (1/2 + 2st)v_1$. We also must add $e_j(\nabla^3\delta)$, $j = 1, 2$, by Theorem 3.3 and from the action of A_0, A_1 on those vectors, we are led to our choice of v_3, v_4, v_5:

$$v_3 := e_2(\nabla^2\delta), \qquad v_4 := e_1(\nabla^3\delta), \qquad \text{and} \qquad v_5 := e_2(\nabla^3\delta).$$

By computation we find that

$$A_0 \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{bmatrix} = \begin{bmatrix} 1/2 + 2st & & & & \\ 0 & 1/2 + 2st & & & \\ s/2 & s/2 & 1/4 & & \\ 1/2 & -1/2 & -t & 0 & \\ s/2 & 3s/2 & 1/4 & 0 & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{bmatrix}$$

and

$$A_1 \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{bmatrix} = \begin{bmatrix} 1/2 + 2st & & & & \\ 1/2 + 2st & 0 & & & \\ -s & 0 & 0 & & \\ -1/2 & 1/2 & -t & 0 & \\ -3s/2 & -s/2 & -1/4 & 0 & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{bmatrix}.$$

Let V be the linear span of v_j, $j = 1, \ldots, 5$. Then V is the minimal common invariant subspace of A_0 and A_1 generated by $e_1(\nabla^3\delta)$ and $e_2(\nabla^3\delta)$. Applying (4.3) to the two 5×5 matrices above, we obtain

$$\rho_p(A_0|_V, A_1|_V) = \max\{2^{1/p}|1/2 + 2st|, 1/4\}.$$

If $|1/2 + 2st| < 2^{-2-1/p}$, then $\rho_p(A_0|_V, A_1|_V) = 1/4$; hence $\nu_p(\phi) = 2 + 1/p$ for $1 < p \le \infty$, by Theorem 3.3. When $p = 1$, we have $\nu_1(\phi) \ge 3$. But $\nu_1(\phi) > 3$ is impossible. Indeed, $\nu_1(\phi) > 3$ would imply $\nu_p(\phi) > 3$ for some $p > 1$, by the embedding theorem. But we know that $\nu_p(\phi) < 3$ for $1 < p \le \infty$. Therefore, $\nu_1(\phi) = 3$ for $|1/2 + 2st| < 2^{-2-1/p}$. If $2^{-2-1/p} \le |1/2 + 2st| < 1$, then $\rho_p(A_0|_V, A_1|_V) = 2^{1/p}|1/2 + 2st|$. Theorem 3.3 tells us that $\nu_p(\phi) = -\log_2 |1/2 + 2st|$ for $1 \le p \le \infty$. Thus, for $1 \le p \le \infty$, we have found the optimal L_p-smoothness of ϕ explicitly. □

The special case $s = 3/2$, $t = -1/8$, $\lambda = -1/8$, and $\mu = 1/2$ was discussed by Heil, Strang, and Strela [14]. In this case, ϕ can be solved explicitly as follows:

$$\phi_1(x) = \begin{cases} x^2(-2x + 3) & \text{for } 0 \le x \le 1, \\ (2 - x)^2(2x - 1) & \text{for } 1 < x \le 2, \\ 0 & \text{for } x \in \mathbb{R} \setminus [0, 2] \end{cases}$$

and

$$\phi_2(x) = \begin{cases} x^2(x - 1) & \text{for } 0 \le x \le 1, \\ (2 - x)^2(x - 1) & \text{for } 1 < x \le 2, \\ 0 & \text{for } x \in \mathbb{R} \setminus [0, 2]. \end{cases}$$

It is evident that $\nu_p(\phi) = 2 + 1/p$.

5. Multiple wavelets. In this section we apply the general theory to smoothness analysis of orthogonal multiple wavelets.

In Example 4.1, if $\phi = (\phi_1, \phi_2)^T$ is the solution of the refinement equation corresponding to the parameter $s = -0.2$, then the shifts of ϕ_1 and ϕ_2 are orthogonal. It was shown in the last section that the optimal smoothness of ϕ is $\nu_p(\phi) = 1 + 1/p$. This example of continuous symmetric orthogonal double refinable functions was first constructed by Donovan et al. [10] by means of fractal interpolation. On the basis of their work, Strang and Strela constructed symmetric orthogonal double wavelets in [31].

In this section we shall use refinement equations to study multiple wavelets. Let a be an element in $(\ell_0(\mathbb{Z}))^{r\times r}$ such that the matrix $M = \sum_{\alpha\in\mathbb{Z}} a(\alpha)/2$ has the following form:

$$M = \begin{bmatrix} 1 & 0 \\ 0 & \Lambda \end{bmatrix} \qquad \text{and} \qquad \lim_{n\to\infty} \Lambda^n = 0.$$

There exists a unique solution ϕ of the refinement equation

$$\phi = \sum_{\alpha\in\mathbb{Z}} a(\alpha)\phi(2\cdot - \alpha)$$

such that $\hat{\phi}(0) = (1, 0, \ldots, 0)^T$. This solution is called the *normalized solution*.

The following theorem summarizes the general theory on orthogonal multiple wavelets (see, e.g., [21]). Some different forms of this result were obtained by Long, Chen, and Yuan [24] and Shen [29].

THEOREM 5.1. *Let $\phi = (\phi_1, \ldots, \phi_r)^T$ be the normalized solution of the refinement equation with mask a. Then $\{\phi_j(\cdot - \alpha) : j = 1, \ldots, r;\ \alpha \in \mathbb{Z}\}$ forms an orthonormal system in $L_2(\mathbb{R})$ if and only if*

a. *$\sum_{\alpha\in\mathbb{Z}} a(\alpha)a(\alpha + 2\gamma)^* = 2\delta_{\gamma,0} I_r$ $\forall\gamma \in \mathbb{Z}$, where I_r denotes the $r \times r$ identity matrix, and*

b. *$\rho(F_a|_W) < 1$ where F_a is the linear operator on $(\ell_0(\mathbb{Z}))^{r\times r}$ given by (3.8) and W is the minimal invariant subspace of F_a generated by $e_1 e_1^T(\Delta\delta)$, $e_2 e_2^T\delta, \ldots, e_r e_r^T\delta$.*

Furthermore, if ψ is given by $\psi = (\psi_1, \ldots, \psi_r)^T = \sum_{\alpha\in\mathbb{Z}} b(\alpha)\phi(2\cdot - \alpha)$, where b is a sequence in $(\ell_0(\mathbb{Z}))^{s\times s}$ satisfying

$$\sum_{\alpha\in\mathbb{Z}} a(\alpha)b(\alpha + 2\gamma)^* = 0 \qquad \forall\gamma \in \mathbb{Z}$$

and

$$\sum_{\alpha\in\mathbb{Z}} b(\alpha)b(\alpha + 2\gamma)^* = 2\delta_{\gamma,0} I_r \qquad \forall\gamma \in \mathbb{Z},$$

then $\{\sqrt{2}^k \psi_j(2^k\cdot - \alpha) : j = 1, \ldots, r;\ k, \alpha \in \mathbb{Z}\}$ forms an orthonormal basis for $L_2(\mathbb{R})$. In other words, $\psi_1, \ldots, \psi_r$ are orthogonal multiple wavelets.

In [21] we constructed a class of continuous orthogonal double wavelets with symmetry. In our construction the mask a is supported on $[0, 2]$ and

$$a(0) = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ t & t \end{bmatrix}, \quad a(1) = \begin{bmatrix} 1 & 0 \\ 0 & \sqrt{2 - 4t^2} \end{bmatrix}, \quad \text{and} \quad a(2) = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -t & t \end{bmatrix},$$

where the parameter t is in the range $-1/\sqrt{2} \le t < -1/2$. Let $\phi = (\phi_1, \phi_2)^T$ be the normalized solution of the refinement equation with mask a. Then ϕ_1 is symmetric

about 1, ϕ_2 is antisymmetric about 1, and $\{\phi_j(\cdot-\alpha) : j = 1, 2,\ \alpha \in \mathbb{Z}\}$ forms an orthonormal system in $L_2(\mathbb{R})$. Moreover, for the coefficients

$$b(0) := \begin{bmatrix} -\frac{1}{2} & -\frac{1}{2} \\ \frac{\mu}{2} & \frac{\mu}{2} \end{bmatrix}, \quad b(1) := \begin{bmatrix} 1 & 0 \\ 0 & -2t \end{bmatrix}, \quad \text{and} \quad b(2) := \begin{bmatrix} -\frac{1}{2} & \frac{1}{2} \\ -\frac{\mu}{2} & \frac{\mu}{2} \end{bmatrix},$$

where $\mu = \sqrt{2-4t^2}$, the vector

$$\psi = (\psi_1, \psi_2)^T := \sum_{\alpha\in\mathbb{Z}} b(\alpha)\phi(2\cdot-\alpha)$$

gives orthogonal double wavelets ψ_1 and ψ_2 which are continuous. Furthermore, ψ_1 is symmetric about 1, and ψ_2 is antisymmetric about 1.

The special case $t = -\sqrt{7}/4$ was studied by Chui and Lian [2]. The following example gives a detailed analysis for the smoothness of the corresponding double-refinable functions.

EXAMPLE 5.2. *Let $\phi = (\phi_1, \phi_2)^T$ be the normalized solution of the refinement equation with mask a, where a is supported on $[0, 2]$ and*

$$a(0) = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ t & t \end{bmatrix}, \quad a(1) = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix}, \quad a(2) = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -t & t \end{bmatrix}$$

with $t = -\sqrt{7}/4$. Then

$$\nu_p(\phi) = \frac{1}{p} - \frac{1}{p}\log_2\left[\Big(\frac{\sqrt{7}}{4}\Big)^p + \Big(\frac{\sqrt{7}-2}{4}\Big)^p\right], \qquad 1 \le p \le \infty. \tag{5.1}$$

Proof. Denote by A_ε $(\varepsilon = 0, 1)$ the linear operators on $(\ell_0(\mathbb{Z}))^2$ given by

$$A_\varepsilon v(\alpha) = \sum_{\beta\in\mathbb{Z}} a(\varepsilon + 2\alpha - \beta)v(\beta), \qquad \alpha \in \mathbb{Z},\ v \in (\ell_0(\mathbb{Z}))^2.$$

Set

$$v_1 := \begin{bmatrix} 1 \\ 2t \end{bmatrix}\delta + \begin{bmatrix} -1 \\ 2t-1 \end{bmatrix}\delta_1, \qquad v_2 := \begin{bmatrix} 1 \\ 2t-1 \end{bmatrix}\delta + \begin{bmatrix} -1 \\ 2t \end{bmatrix}\delta_1,$$
$$v_3 := \begin{bmatrix} 1 \\ 2t \end{bmatrix}\delta_1 + \begin{bmatrix} -1 \\ 2t-1 \end{bmatrix}\delta_2, \qquad \text{and} \qquad v_4 := \begin{bmatrix} 1 \\ 2t-1 \end{bmatrix}\delta_1 + \begin{bmatrix} -1 \\ 2t \end{bmatrix}\delta_2.$$

By computation we obtain

$$A_0\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} t+\frac{1}{2} & 0 & & \\ t & 0 & & \\ & & 0 & t \\ & & 0 & t+\frac{1}{2} \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}$$

and

$$A_1\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} 0 & t & & \\ 0 & t+\frac{1}{2} & & \\ t+\frac{1}{2} & 0 & 0 & 0 \\ t & 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}.$$

Let V be the linear span of v_1 and v_2, and let W be the linear span of v_j, $j = 1, \ldots, 4$. Then V and W are invariant under both A_0 and A_1. It is easily seen that W is the minimal common invariant subspace of A_0 and A_1 generated by $e_1(\nabla^2\delta)$ and $e_2(\nabla^2\delta)$. By the remark made at the beginning of section 4 we see that

$$\rho_p(A_0|_W, A_1|_W) = \max\big\{\rho_p(A_0|_V, A_1|_V), |t + 1/2|\big\}.$$

It remains to compute $\rho_p(A_0|_V, A_1|_V)$. For this purpose, we choose the norm on V as follows:

$$\|\xi_1 v_1 + \xi_2 v_2\| := \max\{|\xi_1|, |\xi_2|\} \qquad \text{for } \xi_1, \xi_2 \in \mathbb{C}.$$

In particular, $\|v_1\| = \|v_2\| = 1$. Suppose $1 \le p < \infty$. Then

$$\|\mathcal{A}^n v_1\|_p^p = \sum_{\varepsilon_1, \ldots, \varepsilon_n \in \{0,1\}} \|A_{\varepsilon_n} \cdots A_{\varepsilon_1} v_1\|^p.$$

Write s for $t + 1/2$. We claim that

$$\sum_{\varepsilon_1, \ldots, \varepsilon_n \in \{0,1\}} \|A_{\varepsilon_n} \cdots A_{\varepsilon_1} v_1\|^p = (|s|^p + |t|^p)^n. \tag{5.2}$$

This will be proved by induction on n. For $n = 1$, we have $A_0 v_1 = s v_1$ and $A_1 v_1 = t v_2$. Hence

$$\|A_0 v_1\|^p + \|A_1 v_1\|^p = |s|^p + |t|^p.$$

This verifies (5.2) for $n = 1$. Suppose (5.2) is valid for n. We wish to establish it for $n + 1$. Recall that $A_0 v_2 = t v_1$ and $A_1 v_2 = s v_2$. Thus, either $A_{\varepsilon_n} \cdots A_{\varepsilon_1} v_1 = \xi v_1$ for some $\xi \in \mathbb{C}$, or $A_{\varepsilon_n} \cdots A_{\varepsilon_1} v_1 = \eta v_2$ for some $\eta \in \mathbb{C}$. In the former case, we have

$$A_0 A_{\varepsilon_n} \cdots A_{\varepsilon_1} v_1 = s\xi v_1 \qquad \text{and} \qquad A_1 A_{\varepsilon_n} \cdots A_{\varepsilon_1} v_1 = t\xi v_2.$$

It follows that

$$\|A_0 A_{\varepsilon_n} \cdots A_{\varepsilon_1} v_1\|^p + \|A_1 A_{\varepsilon_n} \cdots A_{\varepsilon_1} v_1\|^p = (|s|^p + |t|^p)|\xi|^p.$$

But $|\xi| = \|A_{\varepsilon_n} \cdots A_{\varepsilon_1} v_1\|$. Therefore, by the induction hypothesis, we obtain

$$\sum_{\varepsilon_1, \ldots, \varepsilon_n \in \{0,1\}} \big[\|A_0 A_{\varepsilon_n} \cdots A_{\varepsilon_1} v_1\|^p + \|A_1 A_{\varepsilon_n} \cdots A_{\varepsilon_1} v_1\|^p\big] = (|s|^p + |t|^p)^{n+1}. \tag{5.3}$$

In the latter case, we have

$$A_0 A_{\varepsilon_n} \cdots A_{\varepsilon_1} v_1 = t\eta v_1 \qquad \text{and} \qquad A_1 A_{\varepsilon_n} \cdots A_{\varepsilon_1} v_1 = s\eta v_2.$$

Hence (5.3) is also valid. This completes the induction procedure.

Finally, we derive from (5.2) that

$$\rho_p(A_0|_V, A_1|_V) = \lim_{n\to\infty} \|\mathcal{A}^n v_1\|_p^{1/n} = \big(|s|^p + |t|^p\big)^{1/p}, \qquad 1 \le p < \infty.$$

For the case $p = \infty$, a similar argument gives

$$\rho_\infty(A_0|_V, A_1|_V) = \max\{|s|, |t|\}.$$

Note that $t = -\sqrt{7}/4$ and $s = t + 1/2 = (2 - \sqrt{7})/4$. By Theorem 3.3, we obtain the desired result (5.1). In particular,

$$\nu_\infty(\phi) = -\log_2 \frac{\sqrt{7}}{4} \approx 0.59632,$$

$$\nu_2(\phi) = \frac{1}{2} - \frac{1}{2}\log_2 \frac{9 - 2\sqrt{7}}{8} \approx 1.05458,$$

and

$$\nu_1(\phi) = 1 - \log_2 \frac{\sqrt{7} - 1}{2} \approx 1.28125. \qquad \square$$

We now return to the general mask and discuss the L_2-smoothness of the normalized solution ϕ.

EXAMPLE 5.3. *Let a be the mask*

$$a(0) = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ t & t \end{bmatrix}, \quad a(1) = \begin{bmatrix} 1 & 0 \\ 0 & \mu \end{bmatrix}, \quad \textit{and} \quad a(2) = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -t & t \end{bmatrix},$$

where $\mu := \sqrt{2 - 4t^2}$ and the parameter t is in the range $-1/\sqrt{2} \le t < -1/2$. Then

$$\nu_2(\phi) = -\log_2 \sqrt{\rho(B(t))}, \qquad t \ne -\sqrt{7}/4,$$

where $B(t)$ is the matrix

$$B(t) := \begin{bmatrix} 1/4 & t^2 & 0 & 0 \\ 1/4 & t\mu - t^2 & 1 & 0 \\ \frac{2t+2-\mu-4t\mu}{8} & -\frac{2t^2\mu - t\mu^2 + 4t^2 - 8t^3\mu}{4} & 1 & \frac{4t^2-1-8t^2\mu+2\mu}{8} \\ 1/2 & -\mu t & 0 & \frac{\mu+2t}{4} \end{bmatrix}.$$

Proof. To apply Theorem 3.4 with $k = 1$, we need to determine the minimal invariant subspace W of F_a generated by $e_1 e_1^T(\Delta\delta) =: w_1$ and $e_2 e_2^T(\Delta\delta) =: w_2$. From a computation, we find that W can be described as $W := \operatorname{span}\{w_1, w_2, w_3, w_4\}$ using the additional matrices

$$w_3 := \begin{bmatrix} 0 & (t/2 - \mu/4)(\delta_{-1} - \delta_1) \\ -(t/2 - \mu/4)(\delta_{-1} - \delta_1) & (2 - 4t\mu)\Delta\delta \end{bmatrix}$$

and

$$w_4 := \begin{bmatrix} 0 & \delta_{-1} - \delta_1 \\ \delta_1 - \delta_{-1} & 0 \end{bmatrix}.$$

Furthermore, $F_a|_W$ has the matrix representation given by

$$F_a \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} = B(t) \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix}.$$

One of the eigenvalues of $B(t)$ is zero. The absolute value of the dominant eigenvalue has minimum $1/4$ precisely at $t = -\sqrt{7}/4$. Hence when $t \ne -\sqrt{7}/4$, $-\log_2 \sqrt{\rho(F_a|_W)} < 1$ and $\nu_2(\phi) = -\log_2 \sqrt{\rho(B(t))}$ by Theorem 3.4. $\square$

When $t = -\sqrt{7}/4$, the matrix B takes the form

$$B\left(-\frac{\sqrt{7}}{4}\right) = \begin{bmatrix} 1/4 & 7/16 & 0 & 0 \\ 1/4 & -\frac{2\sqrt{7}+7}{16} & 1 & 0 \\ 3/16 & -\frac{35+8\sqrt{7}}{64} & 1 & 0 \\ 1/2 & \frac{\sqrt{7}}{8} & 0 & \frac{1-\sqrt{7}}{8} \end{bmatrix}.$$

It is evident that in this case, the minimal invariant subspace W of F_a generated by $e_1 e_1^T(\Delta\delta) = w_1$ and $e_2 e_2^T(\Delta\delta) = w_2$ is span $\{w_1, w_2, w_3\}$. Moreover, the eigenvalues of the matrix representation of F_a on W are 0, $(9-2\sqrt{7})/16$, and $1/4$. From Example 5.2, we know that $\nu_2(\phi) = -\log_2 \sqrt{(9-2\sqrt{7})/16}$ for $t = -\sqrt{7}/4$; hence, this case provides an example to show that the condition $k > -\log_2 \sqrt{\rho(F_a|_W)}$ is necessary for the equality in Theorem 3.4.

We have seen above that the choice $t = -\sqrt{7}/4$ of the parameter t yields the best smoothness in $L_2(\mathbb{R})$ among all the orthonormal wavelets generated by the masks in Example 5.3. That same choice of the parameter also is the only one for which the resulting vector of functions $\phi = (\phi_1, \phi_2)^T$ achieves accuracy 2. However, if we measure the smoothness in L_∞, then the choice $t = -\sqrt{7}/4$ no longer provides optimal smoothness. Recall from [21] that the operators A_0 and A_1, when restricted to the subspace V generated by

$$\big\{2e_2\delta, \quad 2e_2\delta_1, \quad -e_1\nabla\delta + e_2\delta + e_2\delta_1\big\},$$

have the matrix representation

$$A_0|_V := \begin{bmatrix} \frac{1}{2}+t & \frac{1}{2}+t & -1 \\ 0 & \sqrt{2-4t^2} & 0 \\ 0 & \frac{2t+\sqrt{2-4t^2}}{2} & 0 \end{bmatrix} \quad \text{and} \quad A_1|_V := \begin{bmatrix} \sqrt{2-4t^2} & 0 & 0 \\ \frac{1}{2}+t & \frac{1}{2}+t & -1 \\ \frac{2t+\sqrt{2-4t^2}}{2} & 0 & 0 \end{bmatrix}.$$

Clearly, V contains the subspace generated by $e_1\nabla\delta$, $e_2\nabla\delta$.

A lower bound for the joint spectral radius $\rho_\infty(A_0|_V, A_1|_V)$ is given by the maximal spectral radius of the square root of the $\rho_\infty(A_{\varepsilon_1}|_V A_{\varepsilon_2}|_V)$, $\varepsilon_1, \varepsilon_2 \in \{0, 1\}$. Now

$$f_1(t)^2 := \rho_\infty(A_0|_V^2) = \rho_\infty(A_1|_V^2) = \max\big(|t+1/2|, \mu\big)^2,$$

and

$$\begin{aligned} f_2(t)^2 &:= \rho_\infty(A_0|_V A_1|_V) = \rho_\infty(A_1|_V A_0|_V) \\ &= \max\big(0, |\eta \pm \sqrt{\eta^2 - 64t^2(\mu-1)^2}|/8\big), \end{aligned}$$

where $\eta = 8\mu t + 1 + 4t^2 - 4t$. In the interval $-1/\sqrt{2} \le t < -1/2$, the function f_1 is increasing, the function f_2 is decreasing, and they are equal in the point $t \approx -0.64268764$ at which the minimum value, $.5897545\ldots$, of $g := \max(f_1, f_2)$ is achieved. Perhaps surprisingly, the lower bound g for the joint spectral radius is exact for the point $t = -\sqrt{7}/4$. This suggests that the value $t \approx -0.64268764$ should give rise to ϕ which is smoother when measured in the L_∞ norm. This is indeed the case as a numerical computation shows that

$$\|\mathcal{A}^n\|^{\frac{1}{n}} < .6064 < \sqrt{7}/4 \quad \text{for} \quad t = -0.64268764 \quad \text{and} \quad n = 28.$$

Thus,

$$\nu_\infty(\phi) \ge -\log_2(.6064) = 0.721658\ldots \quad \text{for} \quad t = -0.64268764.$$

Acknowledgment. We thank Bin Han for the numerical computation for the above joint spectral radius.

REFERENCES

[1] J. Boman, *On a problem concerning moduli of smoothness*, in Fourier Analysis and Approximation, Vol. I, (Proc. Colloq., Budapest, 1976), North-Holland, Amsterdam, New York, 1978, pp. 175–179.

[2] C. K. Chui and J. A. Lian, *A study of orthonormal multi-wavelets*, J. Appl. Numer. Math., 20 (1996), pp. 273–298.

[3] A. Cohen, I. Daubechies, and G. Plonka, *Regularity of refinable function vectors*, J. Fourier Anal. Appl., 3 (1997), pp. 295–324.

[4] A. Cohen, N. Dyn, and D. Levin, *Stability and inter-dependence of matrix subdivision schemes*, in Advanced Topics in Multivariate Approximation, F. Fontanella, K. Jetter, and P.-J. Laurent, eds., World Scientific, Singapore, 1996, pp. 33–45.

[5] D. Colella and C. Heil, *Characterizations of scaling functions: Continuous solutions*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 496–518.

[6] I. Daubechies and J. Lagarias, *Two-scale difference equations* I. *Existence and global regularity of solutions*, SIAM J. Math. Anal., 22 (1991), pp. 1388–1410.

[7] I. Daubechies and J. Lagarias, *Two-scale difference equations* II. *Local regularity, infinite products of matrices and fractals*, SIAM J. Math. Anal., 23 (1992), pp. 1031–1079.

[8] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*, Springer-Verlag, Berlin, 1993.

[9] Z. Ditzian, *Moduli of smoothness using discrete data*, J. Approx. Theory, 49 (1987), pp. 115–129.

[10] G. Donovan, J. S. Geronimo, D. P. Hardin, and P. R. Massopust, *Construction of orthogonal wavelets using fractal interpolation functions*, SIAM J. Math. Anal., 27 (1996), pp. 1158–1192.

[11] T. N. T. Goodman, R.-Q. Jia, and C. A. Micchelli, *On the spectral radius of a bi-infinite periodic and slanted matrix*, Southeast Asian Bull. Math., 22 (1998), pp. 115–134.

[12] B. Han and R.-Q. Jia, *Multivariate refinement equations and convergence of subdivision schemes*, SIAM J. Math. Anal., 29 (1998), pp. 1177–1199.

[13] C. Heil and D. Colella, *Matrix refinement equations: Existence and uniqueness*, J. Fourier Anal. Appl., 2 (1996), pp. 363–377.

[14] C. Heil, G. Strang, and V. Strela, *Approximation by translates of refinable functions*, Numer. Math., 73 (1996), pp. 75–94.

[15] L. Hervé, *Multi-resolution analysis of multiplicity d: Applications to dyadic interpolation*, Appl. Comput. Harmon. Anal., 1 (1994), pp. 299–315.

[16] T. A. Hogan, *Stability and linear independence of the shifts of finitely many refinable functions*, J. Fourier Anal. Appl., 3 (1997), pp. 757–774.

[17] R.-Q. Jia, *Subdivision schemes in L_p spaces*, Adv. Comput. Math., 3 (1995), pp. 309–341.

[18] R.-Q. Jia, *Characterization of smoothness of multivariate refinable functions in Sobolev spaces*, Trans. Amer. Math. Soc., to appear.

[19] R.-Q. Jia and C. A. Micchelli, *On linear independence of integer translates of a finite number of functions*, Proc. Edinburgh Math. Soc., 36 (1992), pp. 69–85.

[20] R.-Q. Jia, S. Riemenschneider, and D. X. Zhou, *Approximation by multiple refinable functions*, Canad. J. Math., 49 (1997), pp. 944–962.

[21] R.-Q. Jia, S. Riemenschneider, and D. X. Zhou, *Vector subdivision schemes and multiple wavelets*, Math. Comp., 67 (1998), pp. 1533–1563.

[22] R.-Q. Jia and J. Z. Wang, *Stability and linear independence associated with wavelet decompositions*, Proc. Amer. Math. Soc., 117 (1993), pp. 1115–1124.

[23] Q. T. Jiang and Z. W. Shen, *On existence and weak stability of matrix refinable functions*, Constr. Approx., 15 (1999), pp. 337–353.

[24] R. L. Long, W. Chen, and S. L. Yuan, *Wavelets generated by vector multiresolution analysis*, Appl. Comput. Harmon. Anal., 4 (1997), pp. 317–350.

[25] C. A. Micchelli and T. Sauer, *Regularity of multiwavelets*, Adv. Comput. Math., 7 (1997), pp. 455–545.

[26] G. Plonka, *Approximation order provided by refinable function vectors*, Constr. Approx., 13 (1997), pp. 221–244.

[27] S. D. Riemenschneider and Z. W. Shen, *Multidimensional interpolatory subdivision schemes*, SIAM J. Numer. Anal., 34 (1997), pp. 2357–2381.

[28] G.-C. Rota and G. Strang, *A note on the joint spectral radius*, Indag. Math. (N.S.), 22 (1960), pp. 379–381.
[29] Z. Shen, *Refinable function vectors*, SIAM J. Math. Anal., 29 (1998), pp. 235–250.
[30] W. So and J. Wang, *Estimating the support of a scaling vector*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 66–73.
[31] G. Strang and V. Strela, *Short wavelets and matrix dilation equations*, IEEE Trans. Signal Process., 43 (1995), pp. 108–115.
[32] J. Wang, *Stability and linear independence associated with scaling vectors*, SIAM J. Math. Anal., 29 (1998), pp. 1140–1156.
[33] D.-X. Zhou, *Stability of refinable functions, multiresolution analysis, and Haar bases*, SIAM J. Math. Anal., 27 (1996), pp. 891–904.
[34] D. X. Zhou, *Existence of multiple refinable distributions*, Michigan Math. J., 44 (1997), pp. 317–329.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 1, pp. 29–44
© 1999 Society for Industrial and Applied Mathematics

REDUCED SYSTEMS FOR THREE-DIMENSIONAL ELLIPTIC EQUATIONS WITH VARIABLE COEFFICIENTS*

CHEN GREIF†

Abstract. We consider large sparse nonsymmetric linear systems arising from finite difference discretization of three-dimensional (3D) convection-diffusion equations with variable coefficients. We show that performing one step of cyclic reduction yields a system of equations which is well conditioned and for which fast convergence can be obtained. A certain block ordering strategy is applied, and analytical results concerning symmetrizability conditions and bounds on convergence rates are given. The analysis is accompanied by numerical examples.

Key words. cyclic reduction, 3D elliptic problems, variable coefficients

AMS subject classifications. 65F10, 65N22

PII. S0895479897329849

1. Introduction.

Consider the following three-dimensional (3D) convection-diffusion equation

$$-[(pu_x)_x + (qu_y)_y + (ru_z)_z] + su_x + tu_y + vu_z = w \ , \tag{1.1}$$

on a domain $\Omega \subset \mathbb{R}^3$, subject to Dirichlet, Neumann, or mixed boundary conditions, where all the functions in (1.1) are trivariate, and $p,\ q,\ r > 0$ on Ω. Several discretization schemes are possible. See Morton [11] for a comprehensive survey on numerical solution of the convection-diffusion problem. In this work we use a seven-point discretization technique as a starting point and extend the analysis of Elman and Golub [4], done for the two-dimensional (2D) variable coefficient case, and the analysis of Greif and Varah [8], [9] for the 3D problem with constant coefficients to the 3D problem with variable coefficients.

Let h denote the width of a uniform mesh. In the description that follows we use the notation $G_{i,j,k} \equiv G(ih, jh, kh)$, where G is a trivariate function. The seven-point discretization is done as follows (see, e.g., [4] for the analogous 2D case). For the first term in (1.1) we have

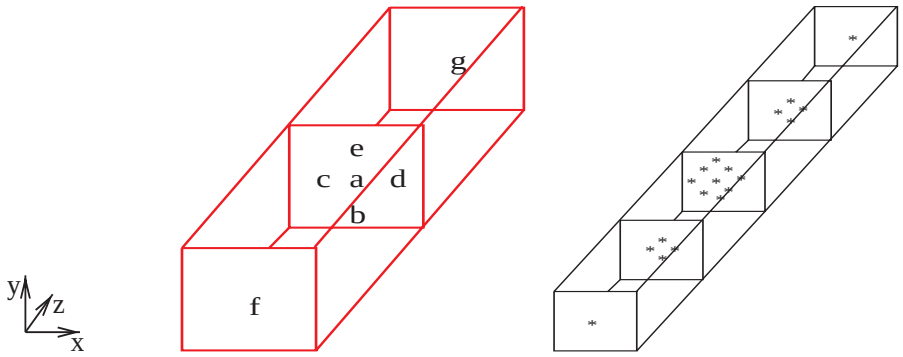
$$(pu_x)_x \approx \frac{p_{i+\frac{1}{2},j,k}u_{i+1,j,k} - (p_{i+\frac{1}{2},j,k} + p_{i-\frac{1}{2},j,k})u_{i,j,k} + p_{i-\frac{1}{2},j,k}u_{i-1,j,k}}{h^2} \ ,$$

and an analogous discretization is performed for $(qu_y)_y$ and $(ru_z)_z$. For the convective terms $su_x,\ tu_y$, and vu_z we use either upwind or centered difference schemes.

Let F denote the corresponding difference operator, scaled by h^2, and denote the values of the associated computational molecule by $a_{i,j,k},\ b_{i,j,k},\ c_{i,j,k},\ d_{i,j,k},\ e_{i,j,k},\ f_{i,j,k}$, and $g_{i,j,k}$, in the following manner: if (i,j,k) is a gridpoint not next to the boundary, then

$$\begin{aligned} F\,u_{i,j,k} = {} & a_{i,j,k}\,u_{i,j,k} + b_{i,j,k}\,u_{i,j-1,k} + c_{i,j,k}\,u_{i-1,j,k} + d_{i,j,k}\,u_{i+1,j,k} \\ & + e_{i,j,k}\,u_{i,j+1,k} + f_{i,j,k}\,u_{i,j,k-1} + g_{i,j,k}\,u_{i,j,k+1} \ . \end{aligned} \tag{1.2}$$

*Received by the editors November 10, 1997; accepted for publication (in revised form) by Z. Strakoš August 20, 1998; published electronically August 3, 1999.
http://www.siam.org/journals/simax /21-1/32984.html
†Department of Computer Science, The University of British Columbia, Vancouver, BC, V6T 1Z4 Canada (greif@cs.ubc.ca).



(**a**) seven-point operator (**b**) reduced operator

FIG. 1.1. *Computational molecules of the unreduced and the reduced operators.*

The computational molecule is graphically illustrated in Figure 1.1(a) (in the figure the subscripts are dropped).

If centered differences are used to discretize the convective terms, the values of the computational molecule are given by

$$
\begin{aligned}
(1.3)\qquad a_{i,j,k} &= p_{i+\frac{1}{2},j,k} + p_{i-\frac{1}{2},j,k} + q_{i,j+\frac{1}{2},k} + q_{i,j-\frac{1}{2},k} + r_{i,j,k+\frac{1}{2}} + r_{i,j,k-\frac{1}{2}};\\
b_{i,j,k} &= -q_{i,j-\frac{1}{2},k} - \frac{t_{i,j,k}h}{2}; \quad e_{i,j,k} = -q_{i,j+\frac{1}{2},k} + \frac{t_{i,j,k}h}{2};\\
c_{i,j,k} &= -p_{i-\frac{1}{2},j,k} - \frac{s_{i,j,k}h}{2}; \quad d_{i,j,k} = -p_{i+\frac{1}{2},j,k} + \frac{s_{i,j,k}h}{2};\\
f_{i,j,k} &= -r_{i,j,k-\frac{1}{2}} - \frac{v_{i,j,k}h}{2}; \quad g_{i,j,k} = -r_{i,j,k+\frac{1}{2}} + \frac{v_{i,j,k}h}{2}.
\end{aligned}
$$

If one uses upwind schemes, then the type of scheme depends on the sign of the convective terms. Assuming that s, t, and v do not change sign in the domain, if they are positive one can use the backward scheme, and if they are negative one can use the forward scheme. Discretizing using backward differences yields

$$
\begin{aligned}
(1.4)\qquad a_{i,j,k} &= p_{i+\frac{1}{2},j,k} + p_{i-\frac{1}{2},j,k} + q_{i,j+\frac{1}{2},k} + q_{i,j-\frac{1}{2},k} + r_{i,j,k+\frac{1}{2}} + r_{i,j,k-\frac{1}{2}} + s_{i,j,k}h\\
&\quad + t_{i,j,k}h + v_{i,j,k}h;\\
b_{i,j,k} &= -q_{i,j-\frac{1}{2},k} - t_{i,j,k}h; \quad e_{i,j,k} = -q_{i,j+\frac{1}{2},k};\\
c_{i,j,k} &= -p_{i-\frac{1}{2},j,k} - s_{i,j,k}h; \quad d_{i,j,k} = -p_{i+\frac{1}{2},j,k};\\
f_{i,j,k} &= -r_{i,j,k-\frac{1}{2}} - w_{i,j,k}h; \quad g_{i,j,k} = -r_{i,j,k+\frac{1}{2}},
\end{aligned}
$$

and for forward differences (1.4) needs to be modified in an obvious manner.

The sparsity structure of the matrix representing the system of equations depends on the ordering of the unknowns. A common strategy is the red/black ordering, which is depicted in Figure 1.2: the gridpoints are colored using two colors in a checkerboard fashion, and the points that correspond to one of the colors (say, red) are numbered first. In this case the corresponding matrix can be written as

$$
(1.5)\qquad \begin{pmatrix} B & C \\ D & E \end{pmatrix} \begin{pmatrix} u^{(r)} \\ u^{(b)} \end{pmatrix} = \begin{pmatrix} w^{(r)} \\ w^{(b)} \end{pmatrix},
$$

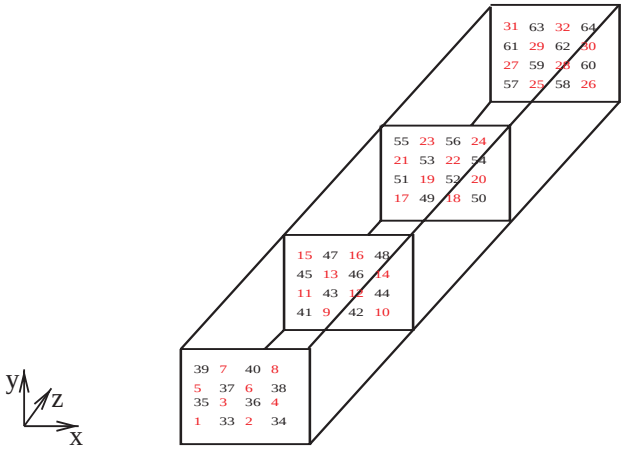


FIG. 1.2. *Red/black ordering of the 3D grid.*

where both B and E are diagonal. In (1.5) superscripts (r) and (b) are attached to denote the associated colors. A simple process of block Gaussian elimination leads to a smaller system, for the black points only, which is called a *reduced system* [2]:

$$[E - DB^{-1}C]u^{(b)} = w^{(b)} - DB^{-1}w^{(r)} \ . \tag{1.6}$$

Since B is diagonal, the matrix of (1.6) is sparse. In the 3D case the corresponding difference operator has a computational molecule which consists of 19 points, as illustrated in Figure 1.1(b). Once the solution for the black points is computed, the solution for the red points corresponds to solving a diagonal system and thus is readily obtained. Moving from system (1.5) to system (1.6) amounts to performing *one step of cyclic reduction* [2]. This procedure can be repeated until a small system of equations is obtained, which can be solved directly. An overview of the idea of cyclic reduction and several references are given in [5, pp. 177–180].

The elimination of half of the unknowns is accompanied by permutation of the matrix (equivalently, reordering of the unknowns). Once the permuted reduced system is formed, an iterative method can be used to find the solution. The procedure of performing one step of cyclic reduction for a non-self-adjoint problem and solving the resulting system using an iterative solver was extensively investigated by Elman and Golub for 2D problems [2], [3], [4]. They showed that one step of cyclic reduction leads to systems with valuable properties, such as symmetrizability by a real diagonal nonsingular matrix for a large set of the underlying PDE coefficients (which is effectively used to derive bounds on the convergence rates of iterative solvers), and fast convergence. In [2] the univariate case, which is naturally more transparent, is used to illustrate the advantages of this technique. Many of the highly effective techniques presented and used by Elman and Golub can be generalized to the 3D case and will be mentioned throughout this paper.

An outline of the rest of this paper follows. In section 2 we introduce the cyclically reduced operator. In section 3 we discuss block orderings and present a family of orderings for the reduced grid. In section 4 we present symmetrization results. In section 5 we use the results of section 4 to derive bounds on convergence rates for block stationary methods. In section 6 we present numerical examples, which include solving the systems using Krylov subspace solvers. Finally, in section 7 we draw some conclusions.

2. The reduced operator. One step of cyclic reduction for a 3D model problem with constant coefficients has been described in [9], where full details on the construction of the matrix are given. Convergence analysis and techniques for solving the resulting system of equations using block stationary methods are described in [8], [9]. In the case of constant coefficients, the values of the computational molecule associated with a given gridpoint do not depend on the point's coordinates, as opposed to the variable coefficient case. Nevertheless, the sparsity structure of the reduced matrix is identical in both cases.

The explicit difference equation associated with the reduced operator for the 3D variable coefficient case can be found in [7] and should be used for constructing the reduced matrix. The alternative of performing the matrix products in (1.6) might be significantly more costly, especially in the 3D case, and in particular, in programming environments where vectorization is crucial (e.g., Matlab).

Consider the following constant coefficient model problem: $p(x,y,z) = q(x,y,z) = r(x,y,z) \equiv 1,\ s(x,y,z) = \sigma,\ t(x,y,z) = \tau,\ v(x,y,z) = \mu$. After scaling by ah^2, the difference equation has the form

$$\begin{aligned}(2.1)\qquad \tilde{R}\,u_{i,j,k} = {} & (a^2 - 2be - 2cd - 2fg)\,u_{i,j,k} - f^2\,u_{i,j,k-2} - 2ef\,u_{i,j+1,k-1} \\ & -2cf\,u_{i-1,j,k-1} - 2df\,u_{i+1,j,k-1} - 2bf\,u_{i,j-1,k-1} - e^2\,u_{i,j+2,k} \\ & -2de\,u_{i+1,j+1,k} - c^2\,u_{i-2,j,k} - d^2\,u_{i+2,j,k} - 2bc\,u_{i-1,j-1,k} \\ & -b^2\,u_{i,j-2,k} - 2eg\,u_{i,j+1,k+1} - 2cg\,u_{i-1,j,k+1} - 2ce\,u_{i-1,j+1,k} \\ & -2bd\,u_{i+1,j-1,k} - 2dg\,u_{i+1,j,k+1} - 2bg\,u_{i,j-1,k+1} - g^2\,u_{i,j,k+2}\ .\end{aligned}$$

Denote the continuous operator corresponding to this model problem

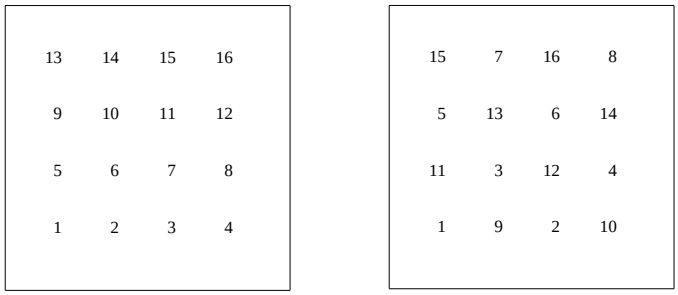
$$(2.2)\qquad L = -\Delta + (\sigma,\tau,\mu)^T\nabla\ .$$

The reduced operator can be derived directly as a discretization scheme of the original PDE, with $O(h^2)$ correction terms in the case of centered difference discretization and $O(h)$ correction terms if upwind discretization is used. This can be done by means analogous to the techniques used for the 2D case (see Elman and Golub [3] and Golub and Tuminaro [6]). Consider the centered difference discretization. Expanding (2.2) in a multivariate Taylor expansion about the gridpoint (ih, jh, kh) yields, after scaling by $2ah^2$,

$$\begin{aligned}(2.3)\qquad 2\tilde{R}\,u = {} & L\,u - \frac{1}{6}\,h^2\,u_{yyyy} - \frac{1}{6}\,h^2\,u_{yyzz} - \frac{1}{6}\,h^2\,\tau\,\mu\,u_{yz} + \frac{1}{6}\,h^2\,\mu\,u_{yyz} + \frac{1}{6}\,h^2\,\sigma u_{xzz} \\ & + \frac{1}{6}\,h^2\,\tau\,u_{xxy} + \frac{1}{6}\,h^2\,\sigma\,u_{xyy} + \frac{1}{3}\,h^2\,\sigma\,u_{xxx} - \frac{1}{6}\,h^2\,\sigma\,\mu\,u_{xz} - \frac{1}{6}\,h^2\,\sigma\,\tau\,u_{xy} + \frac{1}{3}\,h^2\,\mu\,u_{zzz} \\ & - \frac{1}{12}\,h^2\,\mu^2\,u_{zz} + \frac{1}{3}\,h^2\,\tau\,u_{yyy} - \frac{1}{12}\,h^2\,\tau^2\,u_{yy} + \frac{1}{6}\,h^2\,\tau\,u_{yzz} - \frac{1}{6}\,h^2\,u_{xxyy} - \frac{1}{6}\,h^2\,u_{xxzz} \\ & + \frac{1}{6}\,h^2\,\mu\,u_{xxz} - \frac{1}{12}\,h^2\,\sigma^2\,u_{xx} - \frac{1}{6}\,h^2\,u_{xxxx} - \frac{1}{6}\,h^2\,u_{zzzz} + O(h^3)\ .\end{aligned}$$

The above computation was carried out using Maple V. The $O(h^2)$ terms contain, among other terms, the expression $-\frac{h^2}{12}(\sigma^2 u_{xx} + \tau^2 u_{yy} + \mu^2 u_{zz})$, which can be thought of as addition of artificial viscosity to the original equation. The reduced right-hand side is equal to $w_{i,j,k}$ with an $O(h^2)$ error. Gaussian elimination yields the following right-hand side,

$$w_{i,j,k} - \frac{b}{a}\,w_{i,j-1,k} - \frac{c}{a}\,w_{i-1,j,k} - \frac{d}{a}\,w_{i+1,j,k} - \frac{e}{a}\,w_{i,j+1,k} - \frac{f}{a}\,w_{i,j,k-1} - \frac{g}{a}\,w_{i,j,k+1}\ ,$$



(**a**) Natural (**b**) Red/black

FIG. 3.1. *Two possible orderings of the block grid. Each point in these grids corresponds to a one-dimensional (1D) block of gridpoints in the underlying 3D grid.*

whose Taylor expansion about the gridpoint (ih, jh, kh), after scaling by $2ah^2$, is $w - \frac{h^2}{12}[\,-\Delta w + (\sigma w_x + \tau w_y + \mu w_z)\,]$, evaluated at the point (ih, jh, kh). This is another similarity with the 2D case [3].

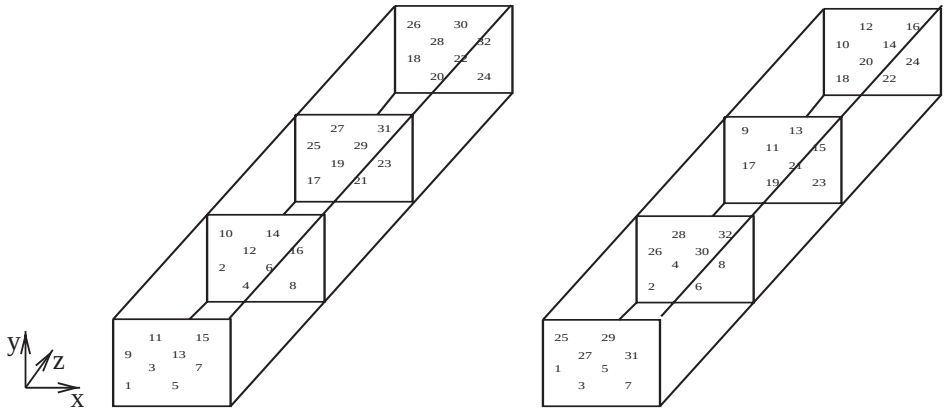
3. Block ordering strategies for the reduced grid. The question of ordering is of major importance, as a good ordering strategy can lead to fast convergence. An excellent overview of the literature that deals with ordering strategies is found in a recent report by Benzi, Szyld, and van Duin [1]. For 3D problems it seems useful to consider the ordering of *blocks* of unknowns, rather than "pointwise" ordering. Such a strategy could be particularly useful for the cyclically reduced problem, as the reduced grid is somewhat irregular. Instead of ordering the unknowns directly in the 3D reduced grid, the problem of ordering is divided into two parts. First, define blocks of gridpoints and order them in a tensor-product 2D "block grid." Once the ordering of the blocks is determined, the task of the "inner" ordering in each of the blocks is relatively simple.

We can define an x-oriented "1D block of gridpoints" by referring to a set of gridpoints whose collection of all x-coordinate values include all the possible values $\{jh\}$ on the grid. A simple example is a single horizontal line of gridpoints in a tensor-product 3D grid. Similarly, y-oriented and z-oriented 1D blocks can be defined. Once the 1D blocks of gridpoints are defined, a *block computational molecule* can be defined as follows.

DEFINITION 3.1. *For a certain given 3D grid and a 1D block of gridpoints in it, the associated* block computational molecule *is defined as the computational molecule in the corresponding block grid. That is, its components are the 1D blocks in the block grid, each of which contains at least one gridpoint which belongs to the (point) computational molecule associated with the 3D problem.*

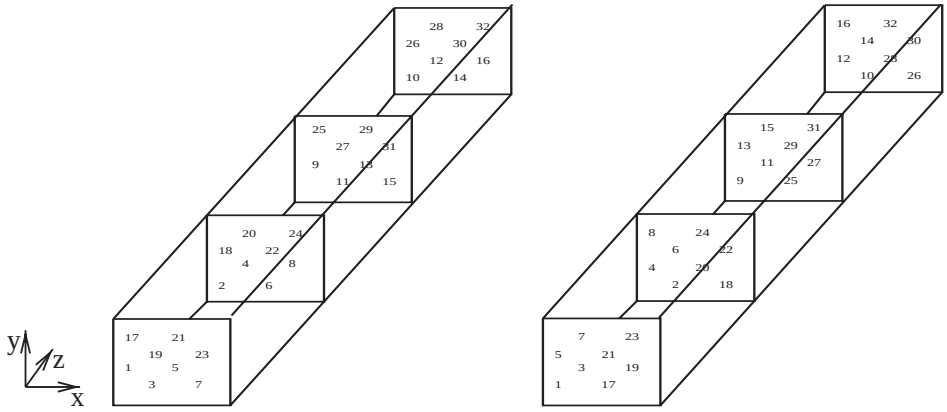
Using the above, we can now easily define different families and types of orderings. For example, a certain ordering strategy is a *natural* block ordering strategy relative to the 1D blocks of gridpoints if these blocks are ordered in the block grid using natural lexicographic ordering. Similarly, one can define a red/black block ordering strategy, and so on (see Figure 3.1).

Below we focus on a particular family of orderings for the reduced grid, which we call the *two-plane* ordering. This ordering corresponds to defining each of the 1D blocks of gridpoints as a collection of $2n$ gridpoints from two horizontal lines and two adjacent planes (here n is the number of gridpoints in a single line in the original



(**a**) 2PNxy (**b**) 2PRBxy

FIG. 3.2. *Two types of two-plane ordering.*

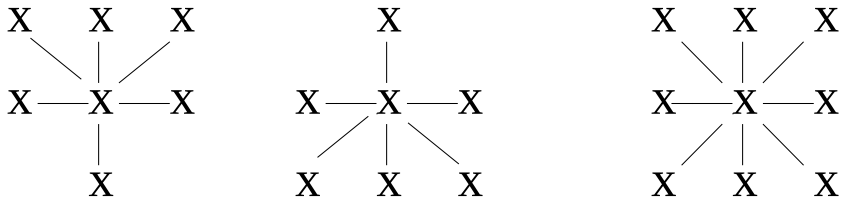


(**a**) 2PNxz (**b**) 2PNyz

FIG. 3.3. *Possible orientations of the* 1*D blocks of gridpoints in the set of natural two-plane orderings.*

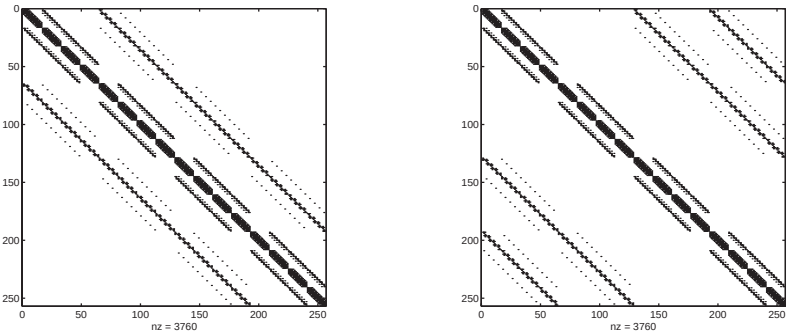
unreduced grid). A single member of this family was introduced in [9]. In Figure 3.2 two members of the family are depicted: natural two-plane ordering and red/black two-plane ordering. For notational convenience, we label them "2PN" and "2PRB," respectively. Two additional letters are added in order to distinguish between different orientations of the $2n$-item 1D blocks of gridpoints. Let us illustrate this for the specific case depicted in Figure 3.2(a). Here $n = 4$. Indices 1–8, 9–16, 17–24, and 25–32 are each an x-oriented 1D block. The block grid is of size 2×2 and its components are ordered in natural lexicographic fashion. Each of the sets of indices 1–16 and 17–32 forms an x-y–oriented pair of planes. Hence the name 2PNxy. In Figure 3.3 other orientations of the blocks in the natural two-plane ordering are depicted.

Figures 3.4(a) and (b) illustrate what blocks are associated with a single gridpoint. In these figures, each "X" corresponds to a 1D block which contains at least one gridpoint in the "point" computational molecule associated with the selected gridpoint (the one that is at the center of the computational molecule). As is evident, the



(**a**) even $k = \frac{z}{h}$ (**b**) odd $k = \frac{z}{h}$ (**c**) computational molecule

FIG. 3.4. *Block computational molecule associated with the two-plane ordering.*



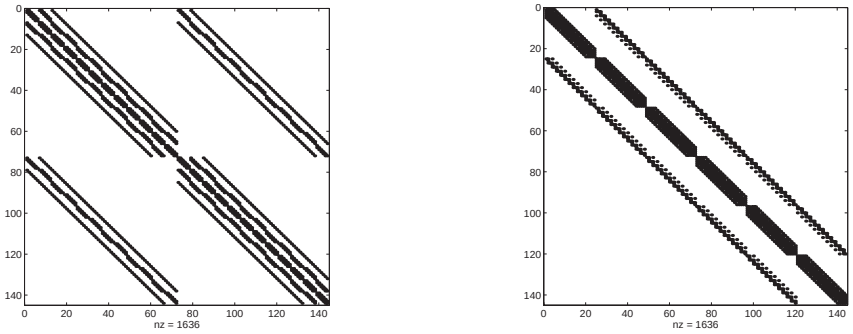
(**a**) natural (**b**) red/black based on 2D blocks

FIG. 3.5. *Sparsity structures of two matrices which belong to the family of two-plane orderings.*

structure depends on the parity of this gridpoint's index. The block computational molecule (Figure 3.4(c)) is obtained by taking the union of all the 1D blocks associated with each of the gridpoints in the block, and thus it is identical to the computational molecule of the classical nine-point operator. This allows one to conclude, e.g., that the reduced matrix does not have block property A [14] relative to partitioning into $2n \times 2n$ blocks. On the other hand, applying a 4-color scheme to the blocks of gridpoints can be effective for parallelization.

The same ideas as above can be applied to $2D$ *blocks* of gridpoints. The block grid in this case is univariate. The reduced matrix is consistently ordered relative to partitioning associated with 2D blocks. In Figure 3.5 the sparsity structures of 256×256 matrices corresponding to natural two-plane ordering and red/black block ordering relative to 2D blocks are depicted.

In order to illustrate the effectiveness of the two-plane ordering, we present in Figure 3.6 a single 2D block of a natural two-plane matrix vs. one that corresponds to ("point") natural lexicographic ordering. The matrices in the figure are associated with a $12 \times 12 \times 12$ grid. As is evident, the main diagonal block of the two-plane matrix is more dense. Compared to the natural lexicographic ordering, there are significantly more nonzero entries in the block diagonal submatrix whose bandwidth does not depend on n. As a result, direct preconditioner solves will be more efficient due to less fill-in. If a stationary scheme such as block Jacobi is used, by [13, Thm. 3.15] it is guaranteed that if the reduced matrix is an M-matrix, then the convergence of a scheme associated with the two-plane matrix is faster than that of a scheme with



(**a**) lexicographic (**b**) natural two-plane

FIG. 3.6. *A zoom on 2D blocks of the matrices corresponding to two ordering strategies of the reduced grid.*

the lexicographic ordering. (The circumstances in which the reduced matrix is an M-matrix in the constant coefficient case are discussed in [9].)

4. Symmetrization of the reduced matrix. In order to obtain bounds on convergence rates for block stationary methods, we consider the following technique, suggested and used in [2], [3], [4] and also effectively applied in [8], [9]: if there exists a real diagonal matrix Q, such that $\hat{S} = Q^{-1}SQ$ is symmetric, then for a splitting $S = D - C$ and an analogous splitting of $\hat{S}$, namely $\hat{S} = Q^{-1}DQ - Q^{-1}CQ \equiv \hat{D} - \hat{C}$, we have

$$\rho(D^{-1}C) = \rho(\hat{D}^{-1}\hat{C}) \leq ||\hat{D}^{-1}||_2||\hat{C}||_2 = \frac{\rho(\hat{C})}{\lambda_{min}(\hat{D})} \ . \tag{4.1}$$

A bound for the spectral radius of the iteration matrix can thus be obtained by evaluating the expression on the right-hand side of (4.1). Denote the entries of S by $\{s_{i,j}\}_{i,j=1}^{n^3/2}$. Since S is sparse and has a block structure, a small amount of work is needed in order to find Q—by requiring for the entries of $\hat{S}$, which we denote by $\{\hat{s}_{i,j}\}_{i,j=1}^{n^3/2}$, that $\hat{s}_{i,j} = \hat{s}_{j,i}$. Since matrices that correspond to different orderings are merely symmetric permutations of one another, we can pick a matrix that corresponds to a specific ordering strategy and do all the work for it. This will result in obtaining general symmetrization conditions for the reduced matrices (regardless of the ordering used). Thus we pick the specific ordering strategy 2PNxz.

Let q_ℓ denote the ℓth diagonal entry in Q. Then

$$\hat{s}_{j,\ell} = \frac{s_{j,\ell}q_\ell}{q_j} \ , \qquad 1 \leq j, \ell \leq \frac{n^3}{2} \ , \tag{4.2}$$

and the symmetry conditions can be expressed as

$$\frac{q_\ell}{q_j} = \sqrt{\frac{s_{\ell,j}}{s_{j,\ell}}} \ , \qquad 1 \leq \ell \leq \frac{n^3}{2}, \quad 1 \leq j \leq \ell \ . \tag{4.3}$$

It is sufficient to look at $2n \times 2n$ blocks. We start with the main diagonal blocks. We have to examine all the entries of the main block that appear in the ℓth row of

the matrix, namely $s_{\ell,\ell-4}, s_{\ell,\ell-3}, \ldots, s_{\ell,\ell}, \ldots, s_{\ell,\ell+4}$. For $s_{\ell,\ell-4}$, if $\ell \bmod 2n \geq 5$ or is equal to 0, then $\ell - 4$ corresponds to the $(i-2, j, k)$ mesh point. Thus

$$s_{\ell-4,\ell} = - \left. \frac{d_{\tilde{\alpha}+1,\tilde{\beta},\tilde{\gamma}} d_{\tilde{\alpha},\tilde{\beta},\tilde{\gamma}}}{a_{\tilde{\alpha}+1,\tilde{\beta},\tilde{\gamma}}} \right|_{\tilde{\alpha}=i-2,\tilde{\beta}=j,\tilde{\gamma}=k}, \tag{4.4}$$

and from this it follows that

$$\left(\frac{q_\ell}{q_{\ell-4}}\right)^2 = \frac{-\left(\frac{c_{i-1,j,k}c_{i,j,k}}{a_{i-1,j,k}}\right)}{-\left(\frac{d_{i-2,j,k}d_{i-1,j,k}}{a_{i-1,j,k}}\right)} = \frac{c_{i-1,j,k}c_{i,j,k}}{d_{i-2,j,k}d_{i-1,j,k}} \,. \tag{4.5}$$

In this case the values associated with the center of the computational molecule (namely, $a_{i,j,k}$) are canceled, but this happens only for rows that involve the $(i \pm 2, j, k)$, $(i, j \pm 2, k)$, and $(i, j, k \pm 2)$ gridpoints. Applying the same procedure to the rest of the entries of the main diagonal block, we obtain the following:

$$\left(\frac{q_\ell}{q_{\ell-3}}\right)^2 = \frac{\frac{c_{i,j,k-1}f_{i,j,k}}{a_{i,j,k-1}} + \frac{c_{i,j,k}f_{i-1,j,k}}{a_{i-1,j,k}}}{\frac{g_{i,j,k-1}d_{i-1,j,k-1}}{a_{i,j,k-1}} + \frac{g_{i-1,j,k-1}d_{i-1,j,k}}{a_{i-1,j,k}}} \quad \ell \bmod 2n = 0, 4, 6, \ldots, 2n-2; \tag{4.6}$$

$$\left(\frac{q_\ell}{q_{\ell-2}}\right)^2 = \frac{\frac{c_{i,j+1,k}e_{i,j,k}}{a_{i,j+1,k}} + \frac{c_{i,j,k}e_{i-1,j,k}}{a_{i-1,j,k}}}{\frac{b_{i,j+1,k}d_{i-1,j+1,k}}{a_{i,j+1,k}} + \frac{d_{i-1,j,k}b_{i-1,j+1,k}}{a_{i-1,j,k}}} \qquad \ell \bmod 4 = 2 \text{ or } 3; \tag{4.7}$$

$$\left(\frac{q_\ell}{q_{\ell-2}}\right)^2 = \frac{\frac{b_{i-1,j,k}c_{i,j,k}}{a_{i-1,j,k}} + \frac{c_{i,j-1,k}b_{i,j,k}}{a_{i,j-1,k}}}{\frac{d_{i-1,j,k}e_{i-1,j-1,k}}{a_{i-1,j,k}} + \frac{e_{i,j-1,k}d_{i-1,j-1,k}}{a_{i,j-1,k}}} \qquad \ell \bmod 4 = 0 \text{ or } 1; \tag{4.8}$$

$$\left(\frac{q_\ell}{q_{\ell-1}}\right)^2 = \frac{\frac{g_{i-1,j,k}c_{i,j,k}}{a_{i-1,j,k}} + \frac{c_{i,j,k+1}g_{i,j,k}}{a_{i,j,k+1}}}{\frac{d_{i-1,j,k}f_{i-1,j,k+1}}{a_{i-1,j,k}} + \frac{d_{i-1,j,k+1}f_{i,j,k+1}}{a_{i,j,k+1}}} \qquad \ell \bmod 2n = 3, \ldots, 2n-1; \tag{4.9}$$

$$\left(\frac{q_\ell}{q_{\ell-1}}\right)^2 = \frac{\frac{e_{i,j,k-1}f_{i,j,k}}{a_{i,j,k-1}} + \frac{f_{i,j+1,k}e_{i,j,k}}{a_{i,j+1,k}}}{\frac{g_{i,j,k-1}b_{i,j+1,k-1}}{a_{i,j,k-1}} + \frac{b_{i,j+1,k}f_{i,j+1,k-1}}{a_{i,j+1,k}}} \qquad \ell \bmod 4 = 2; \tag{4.10}$$

$$\left(\frac{q_\ell}{q_{\ell-1}}\right)^2 = \frac{\frac{b_{i,j,k-1}f_{i,j,k}}{a_{i,j,k-1}} + \frac{f_{i,j-1,k}b_{i,j,k}}{a_{i,j-1,k}}}{\frac{g_{i,j,k-1}e_{i,j-1,k-1}}{a_{i,j,k-1}} + \frac{e_{i,j-1,k}g_{i,j-1,k-1}}{a_{i,j-1,k}}} \qquad \ell \bmod 4 = 0. \tag{4.11}$$

As is evident, (4.5)–(4.11) overdetermine the nonzero values of the matrix Q. Indeed, (4.9)–(4.11) are sufficient to determine all the diagonal entries, except the first entry in each $2n \times 2n$ block, which at this stage can be arbitrarily chosen. We have to make sure, therefore, that (4.5)–(4.8) are consistent with these three equations, and this requirement imposes some additional conditions. In the constant coefficient case there is unconditional consistency. The problematic nature of the variable coefficient case can be demonstrated simply by looking at one of the consistency conditions.

Consider a gridpoint (i,j,k) whose associated index, ℓ, satisfies $\ell \bmod 4 = 0$. Applying (4.9) to $\ell - 1$ means looking at the row corresponding to the $(i, j-1, k-1)$ gridpoint, and multiplying (4.9), applied to $\ell - 1$, by (4.11) results in an equation for $\frac{q_\ell}{q_{\ell-2}}$, which should be consistent with (4.8). There are three additional consistency conditions for the main block and then eight additional conditions for the rest of the blocks of the reduced matrix. In the consistency condition, if we equate variables that belong to the same location in the computational molecule, we find that sufficient conditions for the above-mentioned consistency conditions to hold are $b_{i-1,j,k} = b_{i,j,k-1}$, $c_{i,j-1,k-1} = c_{i,j,k}$, $d_{i-1,j-1,k-1} = d_{i-1,j,k}$, $e_{i,j-1,k-1} = e_{i-1,j-1,k}$, $f_{i,j,k} = f_{i-1,j-1,k}$, $g_{i,j,k-1} = g_{i-1,j-1,k-1}$. If $b_{i,j,k} = b_j$ for all i, k, and similarly $c_{i,j,k} = c_i$, $d_{i,j,k} = d_i$, $e_{i,j,k} = e_j$, $f_{i,j,k} = f_k$, $g_{i,j,k} = g_k$, the consistency condition becomes

$$\frac{c_i g_{k-1}}{f_k d_{i-1}} \cdot \frac{b_j f_k}{e_{j-1} g_{k-1}} = \frac{b_j c_i}{d_{i-1} e_{j-1}} \ , \tag{4.12}$$

which is obviously satisfied. The actual meaning of these conditions is that the continuous problem is separable.

The analysis for off-diagonal blocks is identical, and the following additional conditions are obtained:

$$\left(\frac{q_\ell}{q_{\ell-2n}}\right)^2 = \frac{f_{i,j,k} f_{i,j,k-1}}{g_{i,j,k-1} g_{i,j,k-2}} \ , \qquad 2n \le \ell \le \frac{n^3}{2} \ ; \tag{4.13}$$

$$\left(\frac{q_\ell}{q_{\ell-n^2}}\right)^2 = \frac{b_{i,j-1,k} b_{i,j,k}}{e_{i,j-1,k} e_{i,j-2,k}} \ , \qquad n^2 \le \ell \le \frac{n^3}{2} \ . \tag{4.14}$$

The two equations above determine the rest of the entries of the matrix, and only the first entry in the symmetrizer can be determined arbitrarily.

Last, in order for the symmetrizer to be real, we must require that the products $c_i d_{i-1}$, $b_j e_{j-1}$, and $f_k g_{k-1}$ have the same sign.

All that has been said can be summarized in the following theorem, which demonstrates another point of similarity between the 2D [4] and the 3D problems with variable coefficients.

THEOREM 4.1. *Suppose the operator of* (1.1) *is separable. If* $c_i d_{i-1}$, $b_j e_{j-1}$, *and* $f_k g_{k-1}$ *are all nonzero and have the same sign for all* i, j, *and* k, *then there exists a real nonsingular diagonal matrix* Q *such that* $Q^{-1} S Q$ *is symmetric.*

The symmetrized computational molecule can be derived *without* actually performing the similarity transformation. For example, the symmetrized value corresponding to $-\frac{c_{i,j,k} c_{i-1,j,k}}{a_{i-1,j,k}}$ is $-\sqrt{\frac{c_{i-1,j,k} c_{i,j,k} d_{i-2,j,k} d_{i-1,j,k}}{a_{i-1,j,k}}}$, and so on. The symmetrization operation should not actually be performed in order to solve the linear system, as the symmetrizing matrix has entries that are unbounded as h goes to zero. The symmetrization should be done for the mere purpose of convergence analysis.

5. Bounds on convergence rates for block stationary methods. The reduced matrix which corresponds to the family of natural two-plane orderings is of size $(n^3/2) \times (n^3/2)$ and can be thought of as a block tridiagonal matrix, relative to $n^2 \times n^2$ blocks:

$$S = tri[S_{j,j-1}, S_{j,j}, S_{j,j+1}] \ . \tag{5.1}$$

$S_{i,j}$ are block tridiagonal matrices with respect to $2n \times 2n$ blocks.

In [8] two partitionings are considered: a partitioning into 1D blocks ($2n \times 2n$ blocks) and a partitioning into 2D blocks (of size $n^2 \times n^2$). In order to find a bound, if symmetrization is possible, then the strategy in [4] can be applied. In Theorem 5.1, we refer to the ordering strategy 2PNxz. Since other orderings are symmetric permutations of 2PNxz, finding the bounds for other ordering strategies discussed in this paper is straightforward.

THEOREM 5.1. *Suppose the continuous problem is separable and $c_{i+1}d_i, b_{j+1}e_j$, and $f_{k+1}g_k$ are all positive and bounded by β_x, β_y, and β_z, respectively. Suppose also that $a_{i,j,k} \geq \alpha$ for all i, j, and k. Denote $\tilde{h} = \frac{1}{\frac{n}{2}+1}$. Then the spectral radii of the iteration matrices associated with the block Jacobi scheme which correspond to 1D splitting and 2D splitting are bounded by $\frac{\phi+\xi}{\eta}$ and $\frac{\phi}{\eta-\xi}$, respectively, where η, ξ, and ϕ are defined as follows:*

$$\eta = \alpha^2 - 2\beta_y - 2\beta_z - 2\sqrt{\beta_y\beta_z} - 4(\sqrt{\beta_x\beta_y} + \sqrt{\beta_x\beta_z})\cos(\pi h) - 4\beta_x \cos^2(\pi h); \tag{5.2a}$$

$$\xi = 2\beta_z \cos(\pi\tilde{h}) + \sqrt{4\beta_y\beta_z + 16\beta_x\beta_z\cos^2(\pi h) + 16\beta_z\sqrt{\beta_x\beta_y}\cos(\pi h)}\,; \tag{5.2b}$$

$$\phi = 4\sqrt{\beta_y\beta_z} + 4\sqrt{\beta_x\beta_y}\cdot\cos(\pi h) + 2\beta_y\cos(\pi\tilde{h})\,. \tag{5.2c}$$

Proof. The proof follows by using the technique of [4, pp. 346–347]. The conditions stated in the theorem guarantee that the matrix is symmetrizable. Denote the reduced matrix by S and the symmetrized matrix by $\hat{S}$. Suppose S^* is obtained by modifying $\hat{S}$ in the following manner: replace each occurrence of c_i and d_i by $\sqrt{\beta_x}$, replace each occurrence of b_j and e_j by $\sqrt{\beta_y}$, replace each occurrence of f_k and g_k by $\sqrt{\beta_z}$, and replace each occurrence of $a_{i,j,k}$ by α. Denote by $S^* = D^* - C^*$ the splitting, which is analogous, as far as sparsity structure is concerned, to the splitting $S = D - C$. For the 1D splitting, the matrix D^* is block diagonal with semibandwidth 4, its sparsity structure is identical to that of $\hat{D}$, and moreover, it is componentwise smaller than or equal to the entries of $\hat{D}$. By [9, Lem. 3.3], D^* is an irreducible diagonally dominant M-matrix.

The matrix C^* is nonnegative and satisfies $C^* \geq \hat{C}$. Thus the Perron–Frobenius theorem [13, p. 30] can be used to obtain an upper bound on the convergence rate for this splitting. Since the matrix S^* can now be referred to as a symmetrized version of a matrix that is associated with a constant coefficient case, the bound on the convergence rate is readily obtained from [9, Thm. 3.15]. For the 2D splitting the procedure is completely analogous and the bound is obtained from [8, Thm. 3.6]. ∎

We remark that estimates for the convergence rates of block Gauss–Seidel and block SOR schemes can be obtained by using the "near property A" analysis presented in [7], [8].

6. Numerical experiments. In the examples that follow, we begin with some results which validate the convergence analysis of section 5 for stationary methods. We then compare the performance of Krylov subspace solvers for the reduced and the unreduced systems. The experiments were performed on an SGI Origin 2000 machine. The program is written in Matlab 5.

6.1. Test problem 1. Consider the separable problem

$$-\Delta u + p_1\, x\, u_x + p_2\, y\, u_y + p_3\, z\, u_z = w(x, y, z) \tag{6.1}$$

TABLE 6.1
Comparison between the computed spectral radii of the block Jacobi iteration matrices and the bounds, using centered differences for the two splittings, with $p_1 = p_2 = p_3 = 1$.

Splitting		1D			2D		
n	n^3	ρ	Bound	Ratio	ρ	Bound	Ratio
8	512	0.793	0.894	1.13	0.682	0.826	1.21
12	1728	0.895	0.946	1.06	0.825	0.908	1.10
16	4096	0.937	0.968	1.03	0.892	0.944	1.06
20	8000	0.958	0.979	1.02	0.927	0.962	1.04
24	13824	0.970	0.985	1.02	0.948	0.973	1.03

on $\Omega = (0,1) \times (0,1) \times (0,1)$, with Dirichlet boundary conditions, where $w(x,y,z)$ is constructed so that the solution is

$$u(x,y,z) = x\ y\ z\ (1-x)\ (1-y)\ (1-z)\ \exp(x+y+z)\ . \tag{6.2}$$

For notational convenience, let $\gamma = \frac{p_1 h}{2}$, $\delta = \frac{p_2 h}{2}$, and $\mu = \frac{p_3 h}{2}$. Suppose h is sufficiently small and centered difference discretization is performed. Then

$$c_{i+1} d_i = (1+\gamma x_{i+1})(1-\gamma x_i) = 1 + \gamma h - \gamma^2 h^2 (i^2 + i)\ , \tag{6.3}$$

hence

$$1 + \gamma h - \gamma^2 (1-h) \le c_{i+1} d_i \le 1 + \gamma h - 2\gamma^2 h^2\ . \tag{6.4}$$

If $-1 < \gamma < \frac{1}{1-h}$, then $c_{i+1} d_i > 0$. For $b_{j+1} e_j$ and $f_{k+1} g_k$ the bounds are obtained in an identical manner. The center of the computational molecule is $a = 6$. In terms of the PDE coefficients, the condition on γ means that the convergence analysis of section 5 is applicable if the PDE coefficients are $O(n)$.

If the above conditions hold, the matrix is symmetrizable. Using the notation of the previous sections, let $\hat{S}$ be the symmetrized matrix, let $\beta_x = 1 + \gamma h - 2\gamma^2 h^2, \beta_y = 1 + \delta h - 2\delta^2 h^2,\ \beta_z = 1 + \mu h - 2\mu^2 h^2$, and let S^* be a modified version of $\hat{S}$, such that each occurrence of $c_{i+1} d_i,\ b_{j+1} e_j$, and $f_{k+1} g_k$ in $\hat{S}$ is replaced by the upper bounds, namely $\beta_x,\ \beta_y$, and β_z, respectively. Since $S^* \ge \hat{S}$, S^* is a symmetrized version of a matrix corresponding to the constant coefficient case, and by [9, Lem. 3.3], it is a diagonally dominant M-matrix. Using Theorem 5.1, the bounds on the convergence rate of the block Jacobi method are given in Table 6.1. As is evident, the bounds are tight even for small n. It should be noted, however, that as the PDE coefficients grow larger, the bounds are not expected to be as tight, as the inequalities in (6.4) become less effective.

Next, the spectral radii of the block Jacobi and block Gauss–Seidel iteration matrices and an approximation to the optimal SOR parameter have been computed for both the upwind scheme and the centered scheme. The relaxation parameter was computed according to the formula $\frac{2}{1+\sqrt{1-\rho_J^2}}$. It is only an approximation to the optimal SOR parameter since the matrix is only "close" to being consistently ordered [8]. The experiments were done on an $8 \times 8 \times 8$ grid (512 gridpoints). In Tables 6.2 and 6.3 the superscripts c and u stand for centered or upwind, respectively, R and U stand for reduced system and unreduced system, respectively, and the subscripts J and GS stand for Jacobi and Gauss–Seidel, respectively. We present the results for

TABLE 6.2
Spectral radius of the block Jacobi, block Gauss–Seidel, and approximate optimal relaxation parameter for the reduced system, using both upwind and centered differences and $1D$ *splitting.*

$p_1 = p_2 = p_3$	$\rho_J^u(R)$	$\rho_{GS}^u(R)$	$\tilde{\omega}^u(R)$	$\rho_J^c(R)$	$\rho_{GS}^c(R)$	$\tilde{\omega}^c(R)$
10	0.77	0.60	1.23	0.77	0.59	1.22
100	0.36	0.14	1.04	> 1	0.35	-

TABLE 6.3
Spectral radii of the block Jacobi, block Gauss–Seidel, and optimal relaxation parameter for the unreduced system, using both upwind and centered differences with $1D$ *splitting.*

$p_1 = p_2 = p_3$	$\rho_J^u(U)$	$\rho_{GS}^u(U)$	$\tilde{\omega}^u(U)$	$\rho_J^c(U)$	$\rho_{GS}^c(U)$	$\tilde{\omega}^c(U)$
10	0.90	0.81	1.39	0.91	0.82	1.40
100	0.66	0.44	1.14	> 1	> 1	-

two different cases, one of which has convection of moderate size, and the other with large convection, for which the upwind scheme is more effective than the centered scheme.

Note that for the unreduced system, in most cases the matrix satisfies all the conditions required for Young's SOR analysis [14]; thus, the spectral radius of the Gauss–Seidel matrix and the optimal relaxation parameter can be computed from the spectral radius of the block Jacobi matrix. By comparing Tables 6.2 and 6.3 it is evident that stationary solvers for the reduced system converge faster than for the unreduced system. In one case there is convergence for the reduced system and divergence for the unreduced system.

Moving to consider Krylov subspace solvers, in Table 6.4 we make a comparison between the performance of solvers for the two systems. The stopping criterion was relative residual smaller than 10^{-10}. The method that is used is nonpreconditioned Bi-CGSTAB. The table presents information on the complete process, namely, construction of the systems and the iterative solves. The increase in iteration counts as the grid is refined agrees with theory, at least if one assumes that for this well conditioned and mildly nonsymmetric system, the condition number is of magnitude $O(h^{-2})$ and the convergence rate is similar to that of the conjugate gradient method for symmetric positive definite systems. When one step of cyclic reduction is applied, the savings become more dramatic as the systems grow larger. An explanation for this is that the construction of the reduced system, which requires significantly more floating point operations compared with the construction of the unreduced system, becomes a less significant factor in the overall computation as the grid becomes finer. In general, since the iterative solve is the costly component of the computation, it is significant that the number of iterations until convergence of the unreduced solver is larger by a factor of approximately 2 compared with the reduced solver. Figure 6.1 illustrates the saving and the convergence behavior for this problem.

In Table 6.5 we provide some numerical evidence which suggests that the good performance of reduced solvers is due to effective preconditioning of the original matrix. In the table, estimates of the condition numbers for $p_1 = 500$, $p_2 = 200$, $p_3 = 100$ with upwind difference discretization are presented. The estimates were obtained using Matlab's command "condest." The factor of approximately 2 has been obtained for several additional cases that have been tested. More observations on the condition

TABLE 6.4
Comparison between the performance of the unreduced and the reduced solvers for increasing mesh size, using nonpreconditioned Bi-CGSTAB, for $p_1 = 50,\ p_2 = 20,\ p_3 = 10$.

n	n^3	Iterations		Mflops		Time (sec.)	
—	—	Unreduced	Reduced	Unreduced	Reduced	Unreduced	Reduced
64	262,144	153	79	3,556.8	1,956.1	958.1	502.7
80	512,000	191	90	8,637.4	4,346.7	2,325.9	1,125.2
96	884,736	224	113	17,497.2	9,376.7	4,689.8	2,454.9

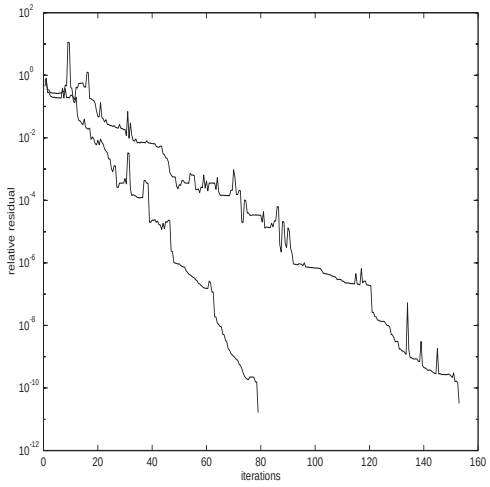


FIG. 6.1. *Relative residuals for nonpreconditioned Bi-CGSTAB applied to linear systems arising from discretization of the test problem, with* $p_1 = 50,\ p_2 = 20,\ p_3 = 10$, *on a* $64 \times 64 \times 64$ *grid. The residual associated with the reduced solver is the lower curve.*

number of the reduced matrix can be found in [7].

6.2. Test problem 2. Consider the nonseparable problem

$$-\Delta u + p_1 x\ \exp(x+y+z)u_x + p_2 y\ \exp(x+y+z)u_y + p_3\ z\ \exp(x+y+z)u_z = w(x,y,z)$$

on $\Omega = (0,1) \times (0,1) \times (0,1)$, with Dirichlet boundary conditions, where $w(x,y,z)$ is constructed so that the solution is (6.2). For this problem the convergence analysis of section 5 does not apply. Results are given in Table 6.6. The experiments were done for $n = 24$, so that the tensor-product grid has $13,824$ gridpoints. GMRES(5) [12] was used, preconditioned by ILU with drop tolerance of 10^{-3}. The stopping criterion was $||r_i||/||r_0|| < 10^{-7}$. In all cases that have been tested, setting up and solving the reduced system is faster compared with setting up and solving the unreduced system. It should be noted that the CPU times are affected by a long preconditioner setup time. When convection dominates, the centered scheme performs poorly. (In general, this scheme suffers numerical instability when the Reynolds numbers are large [10].) Additional numerical experiments indicate that both solvers have the property that for centered difference discretization the solver of mildly nonsymmetric systems converges faster than the solver of a close-to-symmetric system. This phenomenon was proved analytically for stationary methods for the constant coefficient case in [2] (2D) and in [9] (3D).

TABLE 6.5
Comparison between estimates of condition numbers of the unreduced matrix (denoted by U) vs. the reduced matrix (R), for $p_1 = 500,\ p_2 = 200,\ p_3 = 100$.

n	$\kappa_2(U)$	$\kappa_2(R)$
8	420.9	195.0
12	1,049.0	489.8
16	1,859.4	866.5
20	2,693.7	1,258.6
24	3,578.4	1,657.1

TABLE 6.6
Comparison of CPU times (seconds) for setting up and solving the unreduced system and the reduced system using ILU+GMRES. R and U stand for reduced and unreduced, respectively, and the subscripts c and u stand for centered and upwind, respectively.

$p_1 = p_2 = p_3$	U_c	U_u	R_c	R_u
10	58.0	66.2	39.9	47.5
50	91.2	43.5	48.7	32.7
100	148.2	37.3	62.1	23.8

6.3. Test problem 3. Consider the nonseparable problem

$$-0.1\Delta u + yzu_x + xzu_y + xyu_z = w,$$

with Neumann boundary conditions $u_z = 0$ on $z = 0$ and zero Dirichlet conditions for $x = 0$ and $y = 0$ on the unit cube. w was constructed so that the exact solution is $u(x, y, z) = \sin(\pi x)\sin(\pi y)\cos(\pi z)$. Here we compare the performance of a few Krylov subspace solvers; thus, the focus is on the actual iterative solve time, once the systems and the preconditioners were set up.

The results in Table 6.7 are for a $20 \times 20 \times 20$ grid. ILU(0) was used as a preconditioner. The stopping criterion was $||r_i||/||r_0|| < 10^{-7}$. In all cases the reduced solver converges faster than the unreduced solver. The factor of approximately 2 can be a good indication for the gain in performing one step of cyclic reduction in much finer grids. Bi-CGSTAB is slightly faster than CGS. These two schemes are faster than BiCG. The differences in performance between the solvers are qualitatively similar for the reduced and the unreduced systems.

TABLE 6.7
Iteration counts and solving times for three Krylov solvers.

Method	Iterations		Time (sec.)	
—-	Unreduced	Reduced	Unreduced	Reduced
BiCG	32	19	13.9	6.9
CGS	23	14	8.7	4.1
Bi-CGSTAB	19	11	8.2	3.8

7. Concluding remarks. A cyclically reduced operator for a 3D convection-diffusion equation with variable coefficients has been derived. Block orderings have been discussed, some solving techniques for the reduced system have been examined,

and numerical experiments illustrate the fact that the reduced system is easier to solve than the unreduced system.

The results presented in this work show that one step of cyclic reduction can be effectively used as a preconditioning technique for solving the convection-diffusion equation with variable coefficients. The questions of parallelism and applications are topics for further investigation.

Acknowledgments. I am very grateful to Jim Varah for carefully reading this manuscript and for his valuable advice. I would also like to thank Gene Golub, who has held helpful discussions with me and has pointed out some useful references. Many thanks to the two referees for their helpful comments and suggestions.

REFERENCES

[1] M. Benzi, D. B. Szyld, and A. van Duin, *Orderings for incomplete factorization preconditioning of nonsymmetric problems.* SIAM J. Sci. Comput., 20 (1999), pp. 1652–1670.
[2] H. C. Elman and G. H. Golub, *Iterative methods for cyclically reduced non-self-adjoint linear systems,* Math. Comp., 54 (1990), pp. 671–700.
[3] H. C. Elman and G. H. Golub, *Iterative methods for cyclically reduced non-self-adjoint linear systems* II, Math. Comp., 56 (1991), pp. 215–242.
[4] H. C. Elman and G. H. Golub, *Line iterative methods for cyclically reduced discrete convection-diffusion problems,* SIAM J. Sci. Comput., 13 (1992), pp. 339–363.
[5] G. H. Golub and C. F. Van Loan, *Matrix Computations,* 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[6] G. H. Golub and R. S. Tuminaro, *Cyclic Reduction/Multigrid,* Tech. report, Stanford University, Palo Alto, CA, 1993.
[7] C. Greif, *Analysis of Cyclic Reduction for the Numerical Solution of Three-Dimensional Convection-Diffusion Equations*, Ph.D. thesis, The University of British Columbia, Vancouver, BC, Canada, 1998.
[8] C. Greif and J. M. Varah, *Block stationary methods for nonsymmetric cyclically reduced systems arising from three-dimensional elliptic equations* SIAM J. Matrix Anal. Appl., 20 (1999), pp. 1038–1059.
[9] C. Greif and J. M. Varah, *Iterative solution of cyclically reduced systems arising from discretization of the three-dimensional convection diffusion equation,* SIAM J. Sci. Comput., 19 (1998), pp. 1918–1940.
[10] P. M. Gresho and R. L. Lee, *Don't suppress the wiggles—They're telling you something!,* Comput. & Fluids, 9 (1981), pp. 223–253.
[11] K. W. Morton, *Numerical Solution of Convection-Diffusion Problems*, Chapman and Hall, London, 1996.
[12] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing Company, Boston, MA, 1996.
[13] R. S. Varga, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.
[14] D. M. Young, *Iterative Solution of Large Linear Systems*, Academic Press, New York, 1971.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 1, pp. 45–66
© 1999 Society for Industrial and Applied Mathematics

A SMOOTHING NEWTON METHOD FOR EXTENDED VERTICAL LINEAR COMPLEMENTARITY PROBLEMS*

HOU-DUO QI† AND LI-ZHI LIAO‡

Abstract. In this paper, we reformulate the extended vertical linear complementarity problem (EVLCP(m,q)) as a nonsmooth equation $H(t,x)=0$, where $H:\mathbb{R}^{n+1}\to\mathbb{R}^{n+1}$, $t\in\mathbb{R}$ is a parameter variable, and $x\in\mathbb{R}$ is the original variable. H is continuously differentiable except at such points (t,x) with $t=0$. Furthermore H is strongly semismooth. The reformulation of EVLCP(m,q) as a nonsmooth equation is based on the so-called aggregation (smoothing) function. As a result, a Newton-type method is proposed which generates a sequence $\{w^k=(t^k,x^k)\}$ with all $t^k>0$. We prove that every accumulation point of this sequence is a solution of EVLCP(M,q) under the assumption of row $\mathcal{W}_0$-property. If row $\mathcal{W}$-property holds at the solution point, then the convergence rate is quadratic. Promising numerical results are also presented.

Key words. aggregation function, smoothing Newton method, semismoothness, global convergence

AMS subject classifications. 90C33, 47H11, 65H10

PII. S0895479897329837

1. Introduction.

Let $M_0,M_1,\dots,M_k$ be $k+1$ linear transformations from $\mathbb{R}^n$ to itself, and let $q_0,q_1,\dots,q_k$ be $k+1$ vectors in $\mathbb{R}^n$. Denote

$$M:=(M_0,M_1,\dots,M_k),\qquad q:=(q_0,q_1,\dots,q_k). \tag{1.1}$$

The extended vertical linear complementarity problem (EVLCP) associated with (M,q) is to find $x\in\mathbb{R}^n$ such that

$$G(x):=\min\{M_0x+q_0,M_1x+q_1,\dots,M_kx+q_k\}=0, \tag{1.2}$$

where min is the component minimum operator. We denote the above problem as EVLCP(M,q), which is considered by Gowda and Sznajder [15]. When M_0 equals I and q_0 is the zero vector, the resulting EVLCP(M,q), as observed by Gowda and Sznajder [15], is equivalent to the generalized complementarity problem considered by Cottle and Dantzig [7], which is now known as the vertical linear complementarity problem (VLCP). It is easy to see that when $k=1$, $M_0=I$, and $q_0=0$ EVLCP reduces to the standard linear complementarity problem (LCP) (see the book by Cottle, Pang, and Stone [8]). The VLCP has many applications such as control theory [28], the generalized bimatrix game problem [16], the generalized Leontief input-output model [11], and the nonlinear network [12]. For applications of EVLCP(M,q), see Gowda [14] and the references therein.

Several papers address both VLCP and EVLCP; see [15, 14, 29, 21, 16, 19] and the references therein. Most of these papers pay attention primarily to the theoretical

*Received by the editors November 10, 1997; accepted for publication (in revised form) by R. Cottle September 13, 1998; published electronically August 3, 1999.

http://www.siam.org/journals/simax/21-1/32983.html

†Institute of Computational Mathematics and Scientific/Engineering Computing, Chinese Academy of Sciences, P.O. Box 2719, Beijing, China (qhd@lsec.cc.ac.cn). The work of this author was supported in part by Hong Kong Baptist University grant FRG/96-97/II-105.

‡Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong (liliao@hkbu.edu.hk). The work of this author was supported in part by Hong Kong Baptist University grant FRG/96-97/II-105.

aspects including the existence and uniqueness of a solution, the boundedness of the solution set, the stability analysis of the problem involved, and extending the basic properties of LCP to their counterparts in VLCP and EVLCP. For example Mohan, Neogy, and Sridhar [21] reformulate VLCP as an LCP and extend many well-known results in LCP to VLCP. The analysis presented in [14] depends, in a somewhat crucial way, on the properties of bilinear programs and on the fact that there are only two variables in the complementarity condition (i.e., $k = 1$). In [28], Sun applied Mangasarian's iterative algorithm to EVLCP(M, q) with $k = 1$. De Schutter and De Moor [9] studied a more general version of EVLCP and developed an algorithm to find all solutions of their problem. The feasible/infeasible interior point method for EVLCP(M, q) by Sznajder and Gowda [29] considers $(k+1)n$ equations with $(k+1)n$ variables, where they extend P_0- and P-properties for square matrix to the block matrix M defined in (1.1). The new extended properties in [29] play important roles in our smoothing Newton method for EVLCP. These properties include $\mathcal{W}_0$-property and $\mathcal{W}$-property of M. Our method is based on the reformulation of EVLCP(M, q) as $(n + 1)$ equations with $(n + 1)$ variables. This reformulation, in turn, is based on the so-called aggregation function for a max-type function.

The definition of the aggregation function for a max-type function is given as follows. Let $g_i : \mathbb{R}^n \to \mathbb{R}, i = 1, \ldots, m$, be continuously differentiable functions. Define

$$g(x) := \max\{g_1(x), g_2(x), \ldots, g_m(x)\}. \tag{1.3}$$

Then $g(x)$ is continuous in $\mathbb{R}^n$ but not differentiable everywhere. Furthermore, if each g_i is an affine function, then g is a piecewise linear function. For any $t > 0$, the aggregation function $g(t, x) : \mathbb{R}^{n+1} \to \mathbb{R}^n$ is defined by

$$g(t, x) := t \ln \sum_{i=1}^{m} \exp(g_i(x)/t). \tag{1.4}$$

The function, viewed as an exponential penalty function, is intensively studied in an excellent paper [13] by Goldstein, who attributed the basic approximation formula (3.4) to his former student Chang [2]. When a multiplier method with an exponential penalty function was applied to $g(x)$, Bertsekas obtained a slight variant of $g(t, x)$ [1]. Independently, the function was studied and named as the aggregation function by Li [18]. The exponential penalty has been employed by others. For example, see Tseng and Bertsekas [30] and Kachiyan [17]. An interesting feature of $g(t, x)$ (see (3.4) below) is

$$\lim_{t \downarrow 0} g(t, x) = g(x). \tag{1.5}$$

Therefore, it is natural to define $g(0, x) = g(x)$. It can be seen that $g(t, x)$ is continuously differentiable on $\mathbb{R}_{++} \times \mathbb{R}^n$, where $\mathbb{R}_{++}$ denotes the set of all positive real numbers. Notice that

$$G(x) = -\max\{-(M_0x + q_0), -(M_1x + q_1), \ldots, -(M_kx + q_k)\}.$$

By using the aggregation technique, we can approximate nonsmooth (1.2) by the following equation:

$$G(t, x) := 0,$$

where $G(t,x):\mathbb{R}\times\mathbb{R}^n\to\mathbb{R}^n$, its ith component is defined by

$$(G(t,x))_i := \begin{cases} -|t|\ln(\sum_{j=1}^k \exp(-(M_jx+q_j)_i/|t|)) & \text{if } t\neq 0,\\ G(x) & \text{if } t=0, \end{cases} \tag{1.6}$$

and $(M_jx+q_j)_i$ denotes the ith component of vector M_jx+q_j. Furthermore, it follows from (1.5) that

$$\lim_{t\downarrow 0} G(t,x) = G(x),$$

and $G(t,x)$ is continuously differentiable whenever $t\neq 0$. Similarly, it is natural to define $G(0,x)=G(x)$. The above smoothing procedure suggests the possibility to extend the smoothing Newton methods, which were proposed recently in [26, 27, 23] for nonlinear complementarity problems and box constrained variational inequality problems, to EVLCP. The first globally and superlinearly (quadratically) convergent smoothing Newton method was proposed by Chen, Qi, and Sun in [5], where the authors exploited a Jacobian consistence property and applied this property to an infinite sequence of smoothing approximation functions to develop high-order convergent methods. We note that Chen, Qi, and Sun's smoothing Newton method is applicable to the general order complementarity problem, but their smoothing approach, as indicated in the last part of their paper, is very complicated. In this paper, we use only one smoothing approximation function $G(t,x)$ instead of an infinite sequence of functions as in [5]. In this way, our smoothing Newton method becomes much simpler.

Based on the approximation function $G(t,x)$ on $\mathbb{R}^{n+1}$, the sequence generated from our algorithm, $w^k=(t^k,x^k)$, satisfies $t^k>0$ for all k. This attractive feature stems from the fact that the smoothing parameter t is defined as the first component of the following function:

$$H(t,x) := \begin{pmatrix} t \\ G(t,x) \end{pmatrix}. \tag{1.7}$$

The equation $H(t,x)=0$ can be solved by invoking the Newton method and enforcing the parameter t positively converging to zero. Such consideration is not new (see [26, 27, 23] for details). In a recent report, Peng and Lin [22] used the aggregation function to VLCP. They proposed a continuation method via neighborhood strategy. Hence their method is quite different from ours.

The paper is organized as follows. In section 2 we introduce some generalized properties with respect to M defined in (1.1) and some preliminary properties on a semismooth function. In section 3 we study the continuity properties of the aggregation function and show that function $G(t,x)$ is a strongly semismooth function. In section 4 we present our smoothing Newton method for EVLCP(M,q). Section 5 is devoted to the global and quadratic convergence analysis of our method. In section 6 we report some preliminary numerical results on several EVLCP problems. Some concluding remarks are made in section 7.

To help the later discussion, we introduce some notation here. Let $\mathbb{R}_+$ be the set of nonnegative real numbers, $\mathbb{R}_{++}$ be the set of all positive real numbers, and e denote the vector of all ones in $\mathbb{R}^n$. For $u\in\mathbb{R}^n$, diag(u) is the diagonal matrix whose ith diagonal element is u_i. For $X\in\mathbb{R}^{n\times n}$, diag$(X)$ is the vector in $\mathbb{R}^n$ consisting of the diagonal elements of X. For a continuously differentiable function $\Phi:\mathbb{R}^n\to\mathbb{R}^m$,

we denote the Jacobian of Φ at $x \in \mathbb{R}^n$ by $\Phi'(x)$, whereas the transposed Jacobian as $\nabla\Phi(x)$. $\|\cdot\|$ denotes the Euclidean norm. If $f : \mathbb{R}^n \to \mathbb{R}$ is directionally differentiable in the classic sense, we denote $f'(x;h)$ as the derivative of f at x along direction $h \in \mathbb{R}^n$. If $L(t,x) : \mathbb{R}^{n+1} \to \mathbb{R}$ is continuously differentiable, we denote the partial gradient with respect to t and x by $\nabla_t L(t,x)$ and $\nabla_x L(t,x)$, respectively. For $R \in \mathbb{R}^{n\times n}$, R_j denotes the jth row of R. For some index set $I(x)$, $|I(x)|$ denotes the cardinality of $I(x)$. Finally let e_i denote the ith column of the identity matrix I.

2. Preliminaries. For M defined in (1.1), a matrix $R \in \mathbb{R}^{n\times n}$ is called a row representative of M if for each $j = 1, 2, \ldots, n$ the jth row of R is the jth row of some M_i, i.e.,

$$R_{j\cdot} \in \{(M_0)_{j\cdot}, (M_1)_{j\cdot}, \ldots, (M_k)_{j\cdot}\} \quad \text{for } j = 1, 2, \ldots, n.$$

Now we state a generalization of P_0- and P-properties for matrix M in (1.1).

DEFINITION 2.1 (see [29]). *Consider M given by* (1.1). *We say that M has the*

(a) *row $\mathcal{W}_0$-property if one of the following holds:*
   (i) *Determinants of all row representative matrices of M are nonnegative and there is at least one such determinant which is positive.*
   (ii) *Determinants of all row representative matrices of M are nonpositive and there is at least one such determinant which is negative.*

(b) *row $\mathcal{W}$-property if one of the following holds:*
   (i) *Determinants of all row representative matrices of M are positive.*
   (ii) *Determinants of all row representative matrices of M are negative.*

(c) *R_0-property if*

$$\min\{M_0x, M_1x, \ldots, M_kx\} = 0 \quad \Longrightarrow \quad x = 0.$$

In fact R_0-property is introduced by Gowda and Sznajder in [15]. A direct consequence of [15, Theorem 17] is that any M satisfying $\mathcal{W}$-property also satisfies R_0-property. The following two propositions reveal some results for row $\mathcal{W}$- and $\mathcal{W}_0$-properties which will be used later.

PROPOSITION 2.2 (see [29, Theorem 3]). *For M given in* (1.1), *the following are equivalent:*

(a) *M has row $\mathcal{W}$-property.*
(b) *For arbitrary nonnegative diagonal matrices $X_0, X_1, \ldots, X_k \in \mathbb{R}^{n\times n}$ with $\operatorname{diag}(X_0 + X_1 + \cdots + X_k) > 0$,*

$$\det(X_0M_0 + X_1M_1 + \cdots + X_kM_k) \neq 0.$$

PROPOSITION 2.3 (see [29, Theorem 3]). *For M given in* (1.1), *the following are equivalent:*

(1) *M has row $\mathcal{W}_0$-property.*
(2) *For arbitrary positive diagonal matrices $X_0, \ldots, X_k \in \mathbb{R}^{n\times n}$,*

$$\det(X_0M_0 + X_1M_1 + \cdots + X_kM_k) \neq 0.$$

The concept of semismoothness was originally introduced by Mifflin [20] for functionals and later was extended by Qi and Sun [25] to vector valued functions. Suppose $\Phi : \mathbb{R}^{m_1} \to \mathbb{R}^{m_2}$ is a locally Lipschitzian function. According to Rademacher's theorem, Φ is differentiable almost everywhere. Let us denote the set of points at which

Φ is differentiable by D_F. Then we define

$$\partial_B\Phi(x) := \left\{ V \in \mathbb{R}^{m_2\times m_1} |\ V = \lim_{\substack{x'\to x\\ x'\in D_F}} \Phi'(x') \right\}.$$

Let $\partial\Phi(x)$ be the generalized Jacobian defined by Clarke [6]. Then $\partial\Phi(x) = \mathrm{co}\partial_B\Phi(x)$. Suppose that Φ is directionally differentiable at x; then Φ is said to be *semismooth* at x if

$$Vh - \Phi'(x;h) = o(\|h\|), \qquad d \to 0,$$

and Φ is said to be *strongly semismooth* if

$$Vh - \Phi'(x;h) = o(\|h\|^2), \qquad d \to 0,$$

where $V \in \partial\Phi(x+h)$. If Φ is strongly semismooth at x, then for any $h \to 0$ [25]

$$\Phi(x+h) - \Phi(x) - \Phi'(x;h) = O(\|h\|^2). \tag{2.1}$$

A function Φ is said to be a (strongly) semismooth function if it is (strongly) semismooth everywhere.

3. Aggregation function and its properties. Throughout this section, we assume that each $g_i : \mathbb{R}^n \to \mathbb{R},\ i = 1, \cdots, m$, is continuously differentiable. Let $g(t,x)$ be defined in (1.4). The first result in the following proposition can be calculated directly and the other is by Chang [2].

PROPOSITION 3.1. *Let $g(t,x)$ be defined in* (1.4). *Then*

(i) *$g(t,x)$ is continuously differentiable at any point $(t,x) \in \mathbb{R}_{++}\times\mathbb{R}^n$ and its partial gradients with respect to t and x are given by*

$$\nabla_t g(t,x) = \ln\sum_{i=1}^m \exp(g_i(x)/t) - \frac{1}{t}\sum_{i=1}^m \lambda_i(t,x) g_i(x) \tag{3.1}$$

and

$$\nabla_x g(t,x) = \sum_{i=1}^m \lambda_i(t,x)\nabla g_i(x), \tag{3.2}$$

where

$$\lambda_i(t,x) = \frac{\exp(g_i(x)/t)}{\sum_{i=1}^m \exp(g_i(x)/t)} \in (0,1),\ i = 1,\ldots,m \quad \textit{and} \quad \sum_{i=1}^m \lambda_i(t,x) = 1. \tag{3.3}$$

(ii) *For any $(t,x) \in \mathbb{R}_+\times\mathbb{R}^n$,*

$$0 \le g(t,x) - g(x) \le t\ln m. \tag{3.4}$$

Let $I(x)$ denote the set of the active indices of g at x, i.e.,

$$I(x) := \{i \mid g_i(x) = g(x),\ i = 1,\ldots,m\}.$$

We note that

$$\lambda_i(t,x) = \frac{\exp((g_i(x) - g(x))/t)}{\sum_{i=1}^{m} \exp((g_i(x) - g(x))/t)}$$

and

$$\exp((g_i(x) - g(x))/t) = 1 \quad \text{if } i \in I(x), \tag{3.5}$$

$$\lim_{t\downarrow 0} \exp((g_i(x) - g(x))/t) = 0 \quad \text{if } i \notin I(x), \tag{3.6}$$

$$\lambda_i(t,x) = O(t^2), \qquad t \downarrow 0 \text{ and } i \notin I(x). \tag{3.7}$$

Therefore from (3.2) we have

$$\lim_{t\downarrow 0} \nabla_x g(t,x) = \frac{1}{|I(x)|} \sum_{i\in I(x)} \nabla g_i(x). \tag{3.8}$$

In order to show the semismoothness property of $G(t,x)$, we need to define the following. Let $\Gamma : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ be defined by

$$\Gamma(w) := \begin{cases} g(|t|,x) & \text{if } t \neq 0, \\ g(0,x) & \text{if } t = 0, \end{cases} \tag{3.9}$$

where $w = (t,x) \in \mathbb{R} \times \mathbb{R}^n$. Then it is easy to see that Γ is continuously differentiable at any point w with $t \neq 0$ and for $t < 0$

$$\nabla\Gamma(t,x) = \begin{pmatrix} -1 & 0_{1\times n} \\ 0_{n\times 1} & I_n \end{pmatrix} \nabla\Gamma(-t,x). \tag{3.10}$$

Note that $g(x)$ is a locally Lipschitz continuous function since $g(x)$ is a piecewise smooth function.

PROPOSITION 3.2. *The function $\Gamma(t,x)$ is a locally Lipschitz continuous function.*

Proof. Let $w_1 = (t_1, x_1) \in \mathbb{R} \times \mathbb{R}^n$ be given. If $t_1 \neq 0$, then Γ is continuously differentiable near w_1. Hence it is Lipschitz continuous near w_1. Now we assume $t_1 = 0$. Consider another point $w_2 = (t_2, x_2)$ near w_1. If $t_2 = 0$, then

$$\Gamma(w_2) - \Gamma(w_1) = g(x_2) - g(x_1) \leq L\|x_2 - x_1\| \leq L\|w_2 - w_1\|,$$

where L is the Lipschitz constant of g near x_1. If $t_2 \neq 0$, without loss of generality, we assume $t_2 > 0$. Then from (3.4)

$$\begin{aligned} &\Gamma(w_2) - \Gamma(w_1) = \Gamma(t_2, x_2) - \Gamma(t_2, x_1) + \Gamma(t_2, x_1) - \Gamma(0, x_1) \\ &\leq L\|x_2 - x_1\| + t_2 \ln m \leq \sqrt{2}(L + \ln m)\|w_2 - w_1\|, \end{aligned}$$

where L is the uniform Lipschitz constant of $\Gamma(t_2, \cdot)$ near w_1. Hence Γ is locally Lipschitz continuous. □

Since we will apply the above results to EVLCP(M,q), we are interested in the case where functions $g_i,\ i=1,\ldots,m$, are affine functions. Let $a_i\in\mathbb{R}^n,\ b_i\in\mathbb{R},\ i=1,\ldots,m$. We assume that

$$g_i(x)=a_i^Tx+b_i,\quad i=1,\ldots,m.$$

Then $g(x)$ defined by (1.3) is a continuously piecewise linear function and hence it is strongly semismooth. It is easy to verify that the directional derivative of g at x along any direction $d\in\mathbb{R}^n$ is given by

$$g'(x;d)=\max_{i\in I(x)}a_i^Td. \tag{3.11}$$

In the rest of the paper, we assume that each g_i is an affine function. In order to establish the strong semismoothness of functions $G(t,x)$ and $H(t,x)$, we need the following two results.

LEMMA 3.3. *Let $x\in\mathbb{R}^n, d\in\mathbb{R}^n$ be given. Then*

(i) *$I(x+h)\subseteq I(x)$ holds for all $\|h\|$ sufficiently small.*

(ii) *$g_i'(x+h;d)=g'(x;d)$ for all $i\in I(x+h)$ holds for all $\|h\|$ sufficiently small.*

Proof. (i) If the statement is not true, there exists a sequence $\{h^k\}\subset\mathbb{R}^n$ with $\|h^k\|\to 0$ and an index $i(k)\in I(x+h^k)$ but $i(k)\notin I(x)$. By subsequencing, we assume, without loss of generality, that $i(k)=i$ for all k. From the definition of $I(x+h^k)$, we have

$$g_i(x+h^k)=g(x+h^k).$$

From the continuity of g and g_i, we obtain $g_i(x)=g(x)$, which indicates that $i\in I(x)$. This contradiction establishes (i).

(ii) Let h be sufficiently small such that $I(x+h)\subseteq I(x)$. Let $i\in I(x+h)$ and $j\in I(x)$. We have

$$a_i^T(x+h)+b_i\geq a_j^T(x+h)+b_j. \tag{3.12}$$

Since $i\in I(x)$, we have

$$a_i^Tx+b_i=a_j^Tx+b_j. \tag{3.13}$$

Equations (3.12) and (3.13) imply that

$$a_i^Th\geq a_j^Th.$$

Hence

$$a_i^Th=\max_{j\in I(x)}\{a_j^Th\}.$$

This equality establishes (ii). ∎

PROPOSITION 3.4. *Assume that $g_i,\ i=1,\ldots,m$, are affine functions. Then $\Gamma:\mathbb{R}\times\mathbb{R}^n$ defined by (3.9) is a strongly semismooth function.*

Proof. Let $w=(t,x),\ h=(h_1,h_2)\in\mathbb{R}\times\mathbb{R}^n$. To show Γ is strongly semismooth at w, it suffices to verify

$$Vh-\Gamma'(w;h)=O(\|h\|^2)\quad\text{for any } V\in\partial\Gamma(w+h), h\to 0. \tag{3.14}$$

From (3.9) and (3.10), it is easy to see that Γ is at least twice continuously differentiable at any $w = (t, x) \in \mathbb{R} \times \mathbb{R}^n$ with $t \neq 0$. Thus we need only to verify (3.14) at any $w = (0, x)$. There are two cases.

Case 1. $h_1 = 0$. Then

$$\Gamma'(w;h) = \lim_{\tau \downarrow 0} \frac{\Gamma(w+\tau h) - \Gamma(w)}{\tau} = \lim_{\tau \downarrow 0} \frac{g(x+\tau h_2) - g(x)}{\tau} = g'(x;h_2). \tag{3.15}$$

Let $V = \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} \in \partial\Gamma(w+h)$, it follows from $h_1 = 0$ that

$$Vh = V_2 h_2. \tag{3.16}$$

From (3.8) and (3.10) it can be seen that

$$V_2 \in \operatorname{co}\left\{ \frac{1}{|I(x+h_2)|} \sum_{i \in I(x+h_2)} \nabla g_i(x+h_2), \partial_B g(x+h_2) \right\}.$$

Since $\partial_B g(x+h_2)$ is finite, (3.15) and (3.16) indicate that (3.14) is equivalent to

$$V_2 h_2 - g'(x;h_2) = O(\|h\|^2) \tag{3.17}$$

for $h \to 0$ and

$$V_2 = \frac{1}{|I(x+h_2)|} \sum_{i \in I(x+h_2)} \nabla g_i(x+h_2) \quad \text{or} \quad V_2 \in \partial_B g(x+h_2). \tag{3.18}$$

Since $g(x)$ is strongly semismooth (continuously piecewise linear), (3.17) holds for $V_2 \in \partial_B g(x+h_2)$. If V_2 takes the first value in (3.18), we have

$$V_2 h_2 = \frac{1}{|I(x+h_2)|} \sum_{i \in I(x+h_2)} g_i'(x+h_2;h_2).$$

By Lemma 3.3, we have

$$V_2 h_2 - g'(x;h) = 0 = O(\|h\|^2) \quad \text{for all} \quad h \to 0.$$

Hence (3.17) is true for Case 1.

Case 2. $h_1 \neq 0$. We assume $h_1 > 0$; the case $h_1 < 0$ can be proved similarly. We note that

$$\begin{aligned}
\Gamma'(w;h) &= \lim_{\tau \downarrow 0} (\Gamma(w+\tau h) - \Gamma(w))/\tau \\
&= \lim_{\tau \downarrow 0} (\Gamma(\tau h_1, x+\tau h_2) - \Gamma(0,x))/\tau \\
&= \lim_{\tau \downarrow 0} \frac{\tau h_1 \ln\left(\sum_{i=1}^m \exp(g_i(x+\tau h_2)/(\tau h_1))\right) - g(x)}{\tau} \\
&= \lim_{\tau \downarrow 0} h_1 \ln\left(\sum_{i=1}^m \exp[(g_i(x+\tau h_2) - g(x))/(\tau h_1)]\right) \\
&= h_1 \ln \sum_{i \in I(x)} \exp(g_i'(x;h_2)/h_1) \\
&= h_1 \ln \sum_{i \in I(x)} \exp(a_i^T h_2/h_1).
\end{aligned}$$

The fifth equality involves (3.6). In addition, we have

$$\begin{aligned}\Delta_1 :&= h_1 \ln \sum_{i=1}^{m} \exp(g_i(x+h_2)/h_1) - g(x)\\ &= h_1 \ln \sum_{i=1}^{m} \exp((g_i(x+h_2) - g(x))/h_1)\\ &= h_1 \ln \left[\sum_{i\in I(x)} \exp((g_i(x+h_2) - g(x))/h_1) + \sum_{i\notin I(x)} \exp((g_i(x+h_2) - g(x))/h_1) \right]\\ &= h_1 \ln \left[\sum_{i\in I(x)} \exp(a_i^T h_2/h_1) + \sum_{i\notin I(x)} \exp((g_i(x+h_2) - g(x))/h_1) \right]\end{aligned}$$

and

$$\begin{aligned}\Delta_2 :&= \sum_{i=1}^{m} \lambda_i(h_1, x+h_2)(\nabla g_i(x+h_2))^T h_2 + g(x) - \sum_{i=1}^{m} \lambda_i(h_1, x+h_2) g_i(x+h_2)\\ &= \sum_{i=1}^{m} \lambda_i(h_1, x+h_2)(\nabla g_i(x+h_2))^T h_2 - \sum_{i=1}^{m} \lambda_i(h_1, x+h_2)[g_i(x+h_2) - g(x)]\\ &= \sum_{i\in I(x)} \lambda_i(h_1, x+h_2)(\nabla g_i(x+h_2))^T h_2 - \sum_{i\in I(x)} \lambda_i(h_1, x+h_2)[g_i(x+h_2) - g(x)]\\ &\quad + \sum_{i\notin I(x)} \lambda_i(h_1, x+h_2)(\nabla g_i(x+h_2))^T h_2 - \sum_{i\notin I(x)} \lambda_i(h_1, x+h_2)[g_i(x+h_2) - g(x)]\\ &= \sum_{i\notin I(x)} \lambda_i(h_1, x+h_2)(a_i^T h_2) - \sum_{i\notin I(x)} \lambda_i(h_1, x+h_2)[g_i(x+h_2) - g(x)].\end{aligned}$$

The second equality in Δ_2 observes the fact $\sum_{i=1}^{m} \lambda_i(h_1, x+h_2) = 1$. From (3.7), we have for each $i \notin I(x)$

$$\lambda_i(h_1, x+h_2) = O(h_1^2) \quad \text{for } \ h \to 0.$$

Hence

$$\Delta_2 = O(\|h\|^2). \tag{3.19}$$

Let $j \notin I(x),\ i \in I(x),\ \epsilon > 0$; then there exists a $\delta > 0$ such that

$$g_j(x+h_2) - g_i(x+h_2) \le \epsilon \tag{3.20}$$

for all $\|h\| \le \delta$. Then from (3.20)

$$\begin{aligned}&\exp[(g_j(x+h_2) - g(x))/h_1] / \exp(a_i^T h_2/h_1)\\ &= \exp[(g_j(x+h_2) - g(x) - a_i^T h_2)/h_1]\\ &= \exp[(g_j(x+h_2) - g_i(x+h_2))/h_1]\\ &\le \exp(-\epsilon/h_1)\end{aligned}$$

for all $\|h\| \le \delta$. Thus we have for all h with $\|h\|$ sufficiently small

$$\begin{aligned}
&\Delta_1 - \Gamma'(w;h)\\
&= h_1 \ln\left[\sum_{i\in I(x)} \exp(a_i^T h_2/h_1) + \sum_{i\notin I(x)} \exp((g_i(x+h_2)-g(x))/h_1)\right]\\
&\quad -h_1 \ln \sum_{i\in I(x)} \exp(a_i^T h_2/h_1)\\
&= h_1 \ln\left(1 + \frac{\displaystyle\sum_{i\notin I(x)} \exp((g_i(x+h_2)-g(x))/h_1)}{\displaystyle\sum_{i\in I(x)} \exp(a_i^T h_2/h_1)}\right)\\
&\le h_1 \ln\left(1 + \sum_{i\notin I(x)} \exp(-\epsilon/h_1)\right)\\
&= O(\|h_1\|^2) = O(\|h\|^2). \qquad (3.21)
\end{aligned}$$

It follows from (3.19) and (3.21) that

$$\begin{aligned}
&Vh - \Gamma'(w;h)\\
&= h_1 \ln \sum_{i=1}^m \exp(g_i(x+h_2)/h_1) - \sum_{i=1}^m \lambda_i(h_1, x+h_2) g_i(x+h_2)\\
&\quad + \sum_{i=1}^m \lambda_i(h_1, x+h_2)(\nabla g_i(x+h_2))^T h_2 - \Gamma'(w;h)\\
&= h_1 \ln \sum_{i=1}^m \exp(g_i(x+h_2)/h_1) - g(x) - \Gamma'(w;h)\\
&\quad + \sum_{i=1}^m \lambda_i(h_1, x+h_2)(\nabla g_i(x+h_2))^T h_2 - \sum_{i=1}^m \lambda_i(h_1, x+h_2) g_i(x+h_2) + g(x)\\
&= \Delta_1 - \Gamma'(w;h) + \Delta_2 = O(\|h\|^2).
\end{aligned}$$

Hence (3.17) is true for Case 2. □

Now we can state the main result of this section.

THEOREM 3.5. *Function $G(t,x)$ defined in* (1.6) *is a strongly semismooth function. Consequently function $H(t,x)$ is also a strongly semismooth function.*

Proof. We note that the ith component of $G(t,x)$ is defined by $-\Gamma(t,x)$ with $m = k$ and $g_j(x) = -[(M_j)_{i\cdot}x + (q_j)_i]$, $j = 1,\dots,k$. Hence each g_j is an affine function. By Proposition 3.4, $(G(t,x))_i$ is strongly semismooth. Then $G(t,x)$ is strongly semismooth since a function is strongly semismooth if and only if its component functions are [25, Corollary 2.4]. Linear function t is strongly semismooth; hence the strong semismoothness of $G(t,x)$ implies the strong semismoothness of $H(t,x)$. □

4. A smoothing Newton method for EVLCP. Theorem 3.5 indicates that we can approximate the nonsmooth function $G(x)$ by a strongly semismooth function $H(t,x)$ given in (1.7) through the aggregation technique. It is easy to see that a solution of EVLCP(M,q) must be a zero point of $H(t,x)$. This observation plays a key role in the development of our smoothing Newton method. The idea behind

our smoothing Newton method is to solve equation $H(t,x) = 0$ approximately for every smoothing parameter $t > 0$ and gradually reduce t to zero. Hopefully, as t approaches zero, the iterate approaches to a solution of EVLCP(M,q). Note that our reformulation function (1.7) includes the parameter t as a function component; our method will adjust t to zero in the limit. Since the parameter t always remains positive in our method, we assume $t > 0$ in the rest of the paper. To develop our new method, we define $g^j : \mathbb{R}^n \to \mathbb{R}^n$ as

$$g^j(x) = M_j x + q_j, \quad j = 0, 1, \ldots, k,$$

where M_j and q_j are given in (1.1). Hence each component of $G(t,x)$ has the form

$$(G(t,x))_i = -t \ln \sum_{j=0}^{k} \exp(-(g^j(x))_i/t).$$

Since the Jacobian matrix $H'(t,x)$ will play an important role in the later convergence analysis, it would be beneficial to exploit its structure. By Proposition 3.1(i), we have for each $i = 1, \ldots, n$

$$\frac{\partial (G(t,x))_i}{\partial t} = -\ln \left(\sum_{j=0}^{k} \exp(-(g^j(x))_i/t) \right) - \frac{1}{t} \sum_{j=0}^{k} \lambda_j^i(t,x)(g^j(x))_i$$

and

$$\frac{\partial (G(t,x))_i}{\partial x} = \sum_{j=0}^{k} \lambda_j^i(t,x)(g^j(x))_i',$$

where

$$\lambda_j^i(t,x) = \frac{\exp(-(g^j(x))_i/t)}{\sum_{j=0}^{k} \exp(-(g^j(x))_i/t)} \in (0,1) \quad \text{and} \quad \sum_{j=0}^{k} \lambda_j^i(t,x) = 1.$$

Note that for each $j = 0, 1, \ldots, k$

$$(\nabla g^j(x))^T = M_j,$$

then

$$(\nabla_x G(t,x))^T = \sum_{j=0}^{k} \Lambda_j M_j,$$

where

$$\Lambda_j = \operatorname{diag}(\lambda_j^1(t,x), \lambda_j^2(t,x), \ldots, \lambda_j^n(t,x)), \ j = 0, 1, \ldots, k. \tag{4.1}$$

Let

$$r = \left(\frac{\partial (G(t,x))_1}{\partial t}, \ldots, \frac{\partial (G(t,x))_n}{\partial t} \right)^T \in R^n.$$

Then the Jacobian matrix of H has the form

$$H'(t,x) = \begin{pmatrix} 1 & 0_{1\times n} \\ r & \sum_{j=0}^{k} \Lambda_j M_j \end{pmatrix}. \tag{4.2}$$

PROPOSITION 4.1. *Let $H(t,x)$ be defined in* (1.7). *Then $H(t,x)$ is continuously differentiable at any point $(t,x) \in \mathbb{R}_{++} \times \mathbb{R}^n$. If M satisfies row $\mathcal{W}_0$-property, then the Jacobian matrix $H'(t,x)$ is nonsingular at any point $(t,x) \in \mathbb{R}_{++} \times \mathbb{R}^n$.*

Proof. For any point $(t,x) \in \mathbb{R}_{++} \times \mathbb{R}^n$, it has been observed that $H(t,x)$ is continuously differentiable at such point. If M satisfies row $\mathcal{W}_0$-property, it follows from Proposition 2.3 and (4.2) that

$$\det H'(t,x) = \det(\Lambda_0 M_0 + \Lambda_1 M_1 + \cdots + \Lambda_k M_k) \neq 0$$

since $\sum_{j=0}^{k} \Lambda_j = I$. Therefore, $H'(t,x)$ is nonsingular at any point $(t,x) \in \mathbb{R}_{++} \times \mathbb{R}^n$. □

Now select $\bar{t} \in \mathbb{R}_{++}$, $\gamma \in (0,1)$ such that $\gamma\bar{t} < 1$, and let $\bar{w} := (\bar{t}, 0) \in \mathbb{R}_{++} \times \mathbb{R}^n$. Then we define the merit function $\psi : \mathbb{R}^{n+1} \to \mathbb{R}_+$ by

$$\psi(w) := \|H(w)\|^2 = t^2 + \|G(t,x)\|^2,$$

where $w = (t,x) \in \mathbb{R}^{n+1}$, and we define $\beta : \mathbb{R}^{n+1} \to \mathbb{R}_+$ by

$$\beta(w) := \gamma \min\{1, \psi(w)\}.$$

It is easy to see that

$$\beta(w) \leq \gamma\psi(w)^{1/2}. \tag{4.3}$$

Let

$$\Omega := \{w = (t,x) \in \mathbb{R} \times \mathbb{R}^n \mid t \geq \beta(w)\bar{t}\}.$$

Then, because for any $w \in \mathbb{R}^{n+1}, \beta(w) \leq \gamma < 1$, it follows that for any $x \in \mathbb{R}^n$,

$$(\bar{t}, x) \in \Omega.$$

The above discussion leads to our new algorithm, which is a modification of the algorithm used in [26, 23, 27] for nonlinear complementarity problems and box constrained variational inequality problems. The modification allows the original algorithm to be applicable for EVLCP. Our new algorithm will generate a sequence in Ω. In section 5 we will show that, even if Ω is unbounded, the generated sequence remains bounded under mild conditions.

ALGORITHM 4.2 (smoothing Newton method for EVLCP).

(S.0) (Initialization)

Choose constants δ, $\gamma \in (0,1)$, $\bar{t} > 0$ with $\gamma\bar{t} < 1$ and $\sigma \in (0, 1/2)$. Let $t^0 = \bar{t}$, $x^0 \in \mathbb{R}^n$ be arbitrary point, $\bar{w} = (\bar{t}, 0)$, $w^0 = (t^0, x^0)$, and $k = 0$.

(S.1) (Termination criterion)

If $H(w^k) = 0$ then stop. Otherwise, let $\beta_k := \beta(w^k)$.

(S.2) (Modified Newton direction)

Compute $\Delta w^k := (\Delta t^k, \Delta x^k) \in \mathbb{R} \times \mathbb{R}^n$ by

$$H(w^k) + H'(w^k)\Delta w^k = \beta_k \bar{w}. \tag{4.4}$$

(S.3) (Line search strategy)
Let ℓ_k be the smallest nonnegative integer ℓ satisfying

$$\psi(w^k + \delta^\ell \Delta w^k) \le [1 - 2\sigma(1 - \gamma\bar{t})\delta^\ell]\psi(w^k). \tag{4.5}$$

Define $w^{k+1} := w^k + \delta^{\ell_k}\Delta w^k$, $k := k + 1$, go to (S.1).

Remarks. The direction Δw^k computed in (4.4) is a little bit different from the conventional Newton direction because $\beta_k \bar{w}$ is introduced in the right-hand side of (4.4). However, as $H(w^k)$ approaches zero, β_k goes to zero as well. Therefore, direction Δw^k would converge to Newton direction in the limit. The special structure of the Jacobian matrix as shown in (4.4) and the extra term $\beta_k \bar{w}$ in (4.4) benefit Algorithm 4.2 in at least the following two aspects: (a) they guarantee that all t^k are positive, and (b) they ensure that the whole iteration sequence $\{w^k\}$ always stays in Ω (see Propositions 4.3–4.5 for proofs). Before any proofs, the following two questions must be answered:

(i) Is Algorithm 4.2 well defined?

(ii) What are the convergence (global and local) results of Algorithm 4.2?

Qi, Sun, and Zhou [26] have established a standard procedure in addressing (i) and (ii) for their methods. The following discussion on Algorithm 4.2 follows the similar argument as in [26]. To avoid the interruption of our presentation, the proofs of Propositions 4.3 and 4.4 are moved to the appendix.

PROPOSITION 4.3. *Suppose that M satisfies row $\mathcal{W}_0$-property. For any $k \ge 0$, if $w^k = (t^k, x^k) \in \mathbb{R}_{++} \times \mathbb{R}^n$, then Algorithm 4.2 is well defined at the kth iteration and $w^{k+1} = (t^k, x^k) \in \mathbb{R}_{++} \times \mathbb{R}^n$.*

PROPOSITION 4.4. *Suppose that M satisfies row $\mathcal{W}_0$-property. For each fixed $k \ge 0$, if $t^k > 0$, $w^k \in \Omega$, then for any $\alpha \in [0, 1]$ such that*

$$\psi(w^k + \alpha\Delta w^k) \le [1 - 2\sigma(1 - \gamma\bar{t})\alpha]\psi(w^k), \tag{4.6}$$

we have $w^k + \alpha\Delta w^k \in \Omega$.

PROPOSITION 4.5. *Suppose that M satisfies row $\mathcal{W}_0$-property. Then any sequence $\{w^k = (t^k, x^k)\}$ (properly initialized) generated by Algorithm 4.2 satisfies $t^k \in \mathbb{R}_{++}$ and $w^k \in \Omega$ for all $k \ge 0$.*

Proof. First, because $w^0 = (\bar{t}, x^0) \in \Omega$, we have from Propositions 4.3 and 4.4 that w^1 is well defined, $t^1 > 0$, and $w^1 \in \Omega$. Then, by repeatedly resorting to Propositions 4.3 and 4.4, we can prove that an infinite sequence $\{w^k\}$ can be generated with $t^k > 0$ and $w^k \in \Omega$ for the whole sequence. ∎

5. Global and quadratic convergence. Thus far, we have answered the first question raised in section 4 for Algorithm 4.2. In this section, we will address the second one. First, let us focus on the global convergence. The proof of the following global convergence result is relatively long, so we have put it in the appendix.

THEOREM 5.1. *Suppose that M satisfies row $\mathcal{W}_0$-property and an infinite sequence $\{w^k = (t^k, x^k)\}$ is generated by Algorithm 4.2. Then each accumulation point $\tilde{w}$ of $\{w^k\}$ is a solution of $H(w) = 0$.*

Theorem 5.1 states that any accumulation point of $\{w^k\}$ generated by Algorithm 4.2, if it exists, is a solution of $H(w) = 0$; hence the x-part of any accumulation point is a solution of EVLCP(M, q). It is necessary and essential to know when such an accumulation point exists. In the following, we will show that R_0-property of M provides a sufficient condition. First, we define the level set of $\psi(w)$ as

$$\mathcal{L}(w^0) := \{w \in \mathbb{R}^{n+1} \,|\, \psi(w) \le \psi(w^0)\}.$$

THEOREM 5.2. *Suppose that M satisfies R_0-property; then $\mathcal{L}(w^0)$ is bounded.*

Proof. Since $\psi(w) = t^2 + \|G(t,x)\|^2$, for all $w = (t,x) \in \mathcal{L}(w^0)$, we have

$$|t| \le \psi(w^0)^{1/2} \quad \text{and} \quad \|G(t,x)\| \le \psi(w^0)^{1/2}. \tag{5.1}$$

From Proposition 3.1(ii), for each component of $G(t,x)$ we have

$$-(G(0,x))_i \le -(G(t,x))_i \le -(G(0,x))_i + |t| \ln k, \ i = 1, \ldots, n.$$

For all $w = (t,x) \in \mathcal{L}(w^0)$, from (5.1), we have

$$-\psi(w)^{1/2} \le (G(0,x))_i \le (1 + \ln k)\psi(w)^{1/2}, \ i = 1, \ldots, n. \tag{5.2}$$

Then, for all $w = (t,x) \in \mathcal{L}(w^0)$, (5.2) is equivalent to the system

$$-\psi(w)^{1/2} e \le \min\{M_0 x + q_0, M_1 x + q_1, \ldots, M_k x + q_k\} \le (1 + \ln k)\psi(w)^{1/2} e. \tag{5.3}$$

If this theorem is not true, there must exist a sequence $w^k = (t^k, x^k) \in \mathcal{L}(w^0)$ such that $\|w^k\| \to \infty$. Since t^k is bounded by $\psi(w^0)^{1/2}$, it must hold that $\|x^k\| \to \infty$. Without loss of generality, we assume that $x^k/\|x^k\| \to \bar{x}$. Then $\|\bar{x}\| = 1$. Replacing x by x^k and dividing inequality (5.3) by $\|x^k\|$ on both sides and taking the limit, we obtain

$$\min\{M_0 \bar{x}, M_1 \bar{x}, \ldots, M_k \bar{x}\} = 0.$$

However, R_0-property of M implies $\bar{x} = 0$, and the contradiction establishes the result. □

An M satisfying row $\mathcal{W}_0$-property does not necessarily satisfy R_0-property, as the following example shows.

EXAMPLE 5.3. Let $M = (M_1, M_2)$, where

$$M_1 = \begin{bmatrix} 2 & 1 \\ 0 & 0 \end{bmatrix}, \qquad M_2 = \begin{bmatrix} 0 & 0 \\ 1 & 2 \end{bmatrix}.$$

By simple computation, the determinants of all row representative matrices of M are nonnegative and the determinant of the row representative matrix R is positive; here

$$R = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

Hence M satisfies $\mathcal{W}_0$-property. The solution set of the system $\min\{M_1 x, M_2 x\} = 0$ is the nonnegative orthant R^2_+; hence M does not satisfy R_0-property.

Now we state the quadratic convergence result of Algorithm 4.2. The proof is provided in the appendix.

THEOREM 5.4. *Suppose that M satisfies row $\mathcal{W}_0$-property and w^* is an accumulation point of the infinite sequence $\{w^k\}$ generated by Algorithm 4.2. If all elements $V \in \partial H(w^*)$ are nonsingular, then the sequence $\{w^k\}$ converges to w^* quadratically, i.e.,*

$$\|w^{k+1} - w^*\| = O(\|w^k - w^*\|^2), \quad t^{k+1} = O((t^k)^2). \tag{5.4}$$

In addition, $l_k = 0$, i.e., the full step will be accepted, for sufficiently large k.

Now we show that row $\mathcal{W}$-property ensures the nonsingularity of all elements in $\partial H(w^*)$.

PROPOSITION 5.5. *Suppose that $w^* = (t^*, x^*) \in \mathbb{R}\times\mathbb{R}^n$ is a solution of $H(w) = 0$. If M satisfies row $\mathcal{W}$-property, then all elements $V \in \partial H(w^*)$ are nonsingular.*

Proof. Let $\mathcal{S}$ be the set of all $n \times n$ matrices V with the form

$$V = \sum_{j=0}^{k} X_j R_j,$$

where $X_j,\ j = 0, 1, \ldots, k$, are nonnegative diagonal matrices; $\sum_{j=0}^{k} \operatorname{diag}(X_j) > 0$; and $R_j,\ j = 0, 1, \ldots, k$, are row representative matrices of M. Since M satisfies $\mathcal{W}$-property, from Proposition 2.2, such matrix V is nonsingular. Then it is easy to see from (3.10) that

$$\partial_x H(w^*) \subset \mathcal{S}.$$

Hence all elements in $\partial H(w^*)$ are nonsingular. □

Besides row $\mathcal{W}$-property, other sufficient conditions for the nonsingularity of the elements in $\partial H(w^*)$ do exist—for example, the strict complementarity condition. Roughly speaking, sufficient conditions studied in nonlinear complementarity problems can be extended to their EVLCP counterparts. These extensions should be straightforward. To conclude the section, we stress that the results obtained for EVLCP can be obtained similarly for VLCP since the latter is a special case of the former.

6. Preliminary numerical results. In this section, some numerical experiments will be reported for Algorithm 4.2. First we would like to replace the classical monotone Armijo condition (4.5) in Algorithm 4.2 by a nonmonotone line search. The new (S.3) for Algorithm 4.2 is given as follows.

(S.3′) Let ℓ_k be the smallest nonnegative integer ℓ satisfying

$$w^k + \delta^\ell \Delta w^k \in \Omega \tag{6.1}$$

and

$$\psi(w^k + \delta^\ell \Delta w^k) \le \mathcal{W} - 2\sigma(1 - \gamma\bar{t})\delta^\ell \psi(w^k), \tag{6.2}$$

where $\mathcal{W}$ is any value satisfying

$$\psi(w^k) \le \mathcal{W} \le \max_{j=0,1,\ldots,M^k} \psi(w^{k-j})$$

and M^k are nonnegative integers bounded above for all k such that the occurrence of nonnegative indices does not happen. Define $w^{k+1} := w^k + \delta^{\ell_k}\Delta w^k$, $k := k + 1$, go to (S.1).

The reason we choose a nonmonotone line search here is that in most cases it increases the stability of algorithms. The requirement (6.1), which automatically holds for Algorithm 4.2, is to guarantee the global convergence of the algorithm. In our implementation, $\mathcal{W}$ is chosen as follows:

(1) Set $\mathcal{W} = \psi(w^0)$ at the beginning of the algorithm.

(2) Keep the value of $\mathcal{W}$ fixed as long as

$$\psi(w^k) \le \min_{j=0,1,\ldots,5} \psi(w^{k-j}). \tag{6.3}$$

(3) If (6.3) is not satisfied at the kth iteration, set $\mathcal{W} = \psi(w^k)$.

Algorithm 4.2 with above adjustments was implemented in MATLAB on an SGI INDIGO workstation. In all computation, the termination criterion is $\psi(w^k) \le 10^{-12}$ and the setting for all parameters is $\sigma = 0.5 \times 10^{-4}$, $\delta = 0.5$, $\bar{t} = 1$, $\gamma = 0.2$. Our numerical results are summarized in Table 6.1 where the following data are reported: Example—number of test example, Dim—dimension of the test problem, SP—starting point, Iter—number of iterations needed until termination, NF—number of function evaluation needed until termination, FF—value of $\psi(\cdot)$ at the final iterate.

The following is a brief description of the test problems.

EXAMPLE 6.1. This is a problem from [10] with $n = 2,\ k = 2,\ M_0 = I,\ q_0 = 0$, and

$$M_1 = \begin{pmatrix} 3 & -2 \\ -4 & 5 \end{pmatrix},\ M_2 = \begin{pmatrix} 4 & -4 \\ -1 & 2 \end{pmatrix},\ q_1 = q_2 = (-1,-1)^T.$$

M satisfies $\mathcal{W}$-property; hence EVLCP(M,q) has a unique solution $x = (2.25, 2)^T$.

EXAMPLE 6.2. This is a VLCP with $k = 4,\ M_0 = I,\ q_0 = 0,\ q_2 = q_3 = (-1,\ldots,-1)^T$,

$$M_1 = \begin{bmatrix} 4 & -2 & 0 & \cdots & 0 \\ -2 & 4 & -2 & \cdots & 0 \\ 0 & -2 & 4 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -2 \\ 0 & 0 & 0 & \cdots & 4 \end{bmatrix}, \quad M_2 = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 2 & 1 & \cdots & 0 \\ 0 & 1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 2 \end{bmatrix},$$

and M_3 is same as M_1 except $M_3(1,1) = 2,\ M_3(2,1) = 2,\ M_3(2,3) = 2$, and $M_3(n,n) = 2$. M satisfies $\mathcal{W}$-property.

EXAMPLE 6.3. This is an example satisfying $\mathcal{W}_0$-property. $M_0 = I, q_0 = 0$, and

$$M_1 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 1 \end{bmatrix}, \quad M_2 = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad q_1 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad q_2 = \begin{bmatrix} 3/4 \\ -1/2 \\ 0 \end{bmatrix}.$$

The solution set is $\{(0, x_2, 0) |\, 1/2 \le x_2 \le 3/4\} \cup \{(x_1, 3/4, 0) |\, x_1 \ge 0\}$.

We note that the matrices M_2, M_3 in Example 6.3 are singular; however, our method can still satisfactorily solve this problem from various starting points. Based on the limited numerical results presented above, it can be seen that the proposed algorithm works very well for all test problems with various dimensions.

7. Concluding remarks. In this paper, we reformulate EVLCP(M,q) as a nonsmooth equation via the aggregation function. Strong semismoothness of the aggregation function is proved for the linear case. As a result, we are able to propose a smoothing Newton method for the nonsmooth equation. Both the global and the local convergence of the proposed method under mild conditions are proved. The numerical results confirm our theoretical analysis.

The aggregation function plays an important role in this paper. A very interesting feature of the function can be seen as follows. Notice that

$$\min\{x, f(x)\} = -\max\{-x, -f(x)\} = x - (x - f(x))_+;$$

TABLE 6.1
Numerical results of Examples 6.1–6.3.

Example	Dim	SP	Iter	NF	FF
6.1	2	(1, 1)	6	7	2.3e-21
		(0, 0)	8	9	2.5e-18
6.2	6	(3,$\cdots$,3)	8	17	3.1e-20
		(1,$\cdots$,1)	6	8	1.5e-13
	12	(1,$\cdots$,1)	7	9	1.4e-16
		(3,$\cdots$,3)	8	17	4.5e-15
	24	(1,$\cdots$,1)	8	10	4.8e-19
	48	(1,$\cdots$,1)	12	19	4.4e-13
	100	(1,$\cdots$,1)	14	23	3.2e-13
6.3	3	(0, 0, 0)	7	8	2.2e-15
		(1, 1, 1)	5	6	3.0e-16
		(1, 2, 3)	6	9	4.1e-14
		(-1,-1,-1)	7	8	8.6e-24

if the aggregation function is applied directly to the above function, we obtain

$$\begin{aligned} g(t,x) &= -t\ln(\exp(-x/t)+\exp(-f(x)/t)) \\ &= x - t\ln(1+\exp((x-f(x))/t)). \end{aligned}$$

This function is exactly the smoothing function obtained by using the neural network smoothing function to the function $x-(x-f(x))_+$ in [4]. In fact, the aggregation function and the smoothing function in [4] share many similar properties [22]. It would be very useful and practical to explain why they have so many similarities.

Appendix.

Proof of Proposition 4.3. If $t^k \in \mathbb{R}_{++}$, it is easy to see $\beta_k = \beta(w^k) > 0$. Since M satisfies $\mathcal{W}_0$-property, it follows from Proposition 4.1 that $H'(w^k)$ is nonsingular. Hence Δw^k can be computed by (4.4). From the structure of $H'(w^k)$, we have

$$\Delta t^k = -t^k + \beta_k \bar{t}. \tag{A.1}$$

Hence for any $\alpha \in [0,1]$, we have

$$t^k + \alpha \Delta t^k = (1-\alpha)t^k + \alpha\beta_k\bar{t}. \tag{A.2}$$

Then, from (4.3) and (A.2), for any $\alpha \in [0,1]$, we have

$$\begin{aligned} &(t^k+\alpha\Delta t^k)^2 \\ &= (1-\alpha)^2(t^k)^2 + 2\alpha(1-\alpha)\beta_k t^k\bar{t} + \alpha^2\beta_k^2\bar{t}^2 \\ &\le (1-\alpha)^2(t^k)^2 + 2\alpha\beta_k t^k\bar{t} + O(\alpha^2) \\ &\le (1-\alpha)^2(t^k)^2 + 2\alpha\gamma\bar{t}\psi(w^k)^{1/2}\|H(w^k)\| + O(\alpha^2) \\ &= (1-\alpha)^2(t^k)^2 + 2\alpha\gamma\bar{t}\psi(w^k) + O(\alpha^2). \end{aligned} \tag{A.3}$$

Define

$$f(\alpha) := G(w^k+\alpha\Delta w^k) - G(w^k) - \alpha G'(w^k)\Delta w^k.$$

Since $G(\cdot)$ is continuously differentiable at $w^k \in \mathbb{R}_{++}\times\mathbb{R}^n$, then

$$f(\alpha) = o(\alpha).$$

From the second part of (4.4), we have

$$G'(w^k)\Delta w^k = -G(w^k).$$

Hence

$$\begin{aligned}
&\|G(w^k + \alpha\Delta w^k)\|^2 \\
&= \|G(w^k) + \alpha G'(w^k)\Delta w^k + f(\alpha)\|^2 \\
&= \|(1-\alpha)G(w^k) + f(\alpha)\|^2 \\
&= (1-\alpha)^2\|G(w^k)\|^2 + 2(1-\alpha)G(w^k)^T f(\alpha) + \|f(\alpha)\|^2 \\
&= (1-\alpha)^2\|G(w^k)\|^2 + o(\alpha). \qquad \text{(A.4)}
\end{aligned}$$

It then follows from (A.4) and (A.3) that for all $\alpha \in [0,1]$, we have

$$\begin{aligned}
&\psi(w^k + \alpha w^k) \\
&= (t^k + \alpha\Delta t^k)^2 + \|G(w^k + \alpha\Delta w^k)\|^2 \\
&\le (1-\alpha)^2(t^k)^2 + 2\alpha\gamma\bar{t}\psi(w^k) + (1-\alpha)^2\|G(w^k)\|^2 + o(\alpha) + O(\alpha^2) \\
&= (1-\alpha)^2\psi(w^k) + 2\alpha\gamma\bar{t}\psi(w^k) + o(\alpha) \\
&= (1-2\alpha)\psi(w^k) + 2\alpha\gamma\bar{t}\psi(w^k) + o(\alpha) \\
&= [1 - 2(1-\gamma\bar{t})\alpha]\psi(w^k) + o(\alpha). \qquad \text{(A.5)}
\end{aligned}$$

Inequality (A.5) indicates that there exists a positive number $\bar{\alpha} \in (0, \alpha]$ such that for all $\alpha \in [0, \bar{\alpha}]$,

$$\psi(w^k + \alpha\Delta w^k) \le [1 - 2\sigma(1-\gamma\bar{t})\alpha]\psi(w^k),$$

which means that ℓ_k in (S.3) of Algorithm 4.2 can be always found after a finite number of tries. Thus w^{k+1} can be computed if $t^k \in \mathbb{R}_{++}$. □

Proof of Proposition 4.4. We prove this proposition by considering the following two cases:

(i) $\psi(w^k) > 1$. Then $\beta_k = \gamma$. It follows from $w^k \in \Omega$ and $\beta(w) \le \gamma$ for any $w \in \mathbb{R}^{n+1}$ that for all $\alpha \in [0,1]$, we have

$$\begin{aligned}
&t^k + \alpha\Delta t^k - \beta(w^k + \alpha\Delta w^k)\bar{t} \ge (1-\alpha)t^k + \alpha\beta_k\bar{t} - \gamma\bar{t} \\
&\ge (1-\alpha)\beta_k\bar{t} + \alpha\beta_k\bar{t} - \beta_k\bar{t} = 0. \qquad \text{(A.6)}
\end{aligned}$$

(ii) $\psi(w^k) \le 1$. Then for any $\alpha \in [0,1]$ satisfying (4.6), we have

$$\psi(w^k + \alpha\Delta w^k) \le 1. \qquad \text{(A.7)}$$

Thus for any $\alpha \in [0,1]$ satisfying (4.6),

$$\beta(w^k + \alpha\Delta w^k) = \gamma\psi(w^k + \alpha\Delta w^k).$$

Hence, again because $w^k \in \Omega$, for any $\alpha \in [0,1]$ satisfying (4.6), we have

$$\begin{aligned}
&t^k + \alpha\Delta t^k - \beta(w^k + \alpha w^k)\bar{t} \\
&= (1-\alpha)t^k + \alpha\beta_k\bar{t} - \gamma\psi(w^k + \alpha\Delta w^k)\bar{t} \\
&\ge (1-\alpha)\beta_k\bar{t} + \alpha\beta_k\bar{t} - \gamma[1 - 2\sigma(1-\gamma\bar{t})\alpha]\psi(w^k)\bar{t} \\
&= \beta_k\bar{t} - \gamma[1 - 2\sigma(1-\gamma\bar{t})\alpha]\psi(w^k)\bar{t} \\
&= \gamma\psi(w^k)\bar{t} - \gamma[1 - 2\sigma(1-\gamma\bar{t})\alpha]\psi(w^k)\bar{t} \\
&= 2\gamma\sigma(1-\gamma\bar{t})\alpha\psi(w^k)\bar{t} \ge 0. \qquad \text{(A.8)}
\end{aligned}$$

Thus, by combining (A.6) and (A.8), we finish the proof. □

Proof of Theorem 5.1. Since M satisfies $\mathcal{W}_0$-property, it follows from Proposition 4.5 that an infinite sequence $\{w^k\}$ can be generated by Algorithm 4.2. From (4.5), $\psi(w^{k+1}) < \psi(w^k)$ for all $k \geq 0$. Hence the two sequences $\{\psi(w^k)\}$ and $\{\beta(w^k)\}$ are monotonically decreasing. Since $\psi(w^k),\ \beta(w^k) \geq 0$ $(k \geq 0)$, there exist $\tilde{\psi},\ \tilde{\beta} \geq 0$ such that $\psi(w^k) \to \tilde{\psi}$ and $\beta(w^k) \to \tilde{\beta}$ as $k \to \infty$. Let $\tilde{w}$ be any accumulation point of $\{w^k\}$. If $\tilde{\psi} = 0$, then from the continuity of ψ we must have $\psi(\tilde{w}) = 0$; therefore $\tilde{w}$ is a solution of $H(w) = 0$. Now we assume $\tilde{\psi} > 0$. It is easy to see

$$\tilde{\beta} = \gamma \min\{1, \tilde{\psi}\} > 0.$$

By taking a subsequence if necessary, we may assume that $\{w^k\}$ converges to $\tilde{w}$. It is easy to see that

$$\tilde{\psi} = \psi(\tilde{w}),\ \beta(\tilde{w}) = \tilde{\beta} \quad \text{and} \quad \tilde{w} \in \Omega,\ \tilde{t} \geq \tilde{\beta}\bar{t} > 0.$$

Then from Proposition 4.1, $H'(\tilde{w})$ exists and is nonsingular. Therefore there exists a small closed neighborhood $\mathcal{N}(\tilde{w})$ of $\tilde{w}$ such that for any $w = (t, x) \in \mathcal{N}(\tilde{w})$ we have $t \in \mathbb{R}_{++}$ and that $H'(w)$ is nonsingular. For any $w \in \mathcal{N}(\tilde{w})$, let $\Delta w = (\Delta t, \Delta x) \in \mathbb{R} \times \mathbb{R}^n$ be the unique solution of the equation

$$H(w) + H'(w)\Delta w = \beta(w)\bar{w}, \tag{A.9}$$

and for any $\alpha \in [0, 1]$ define

$$g_w(\alpha) = G(w + \alpha\Delta w) - G(w) - \alpha G'(w)\Delta w.$$

It is easy to see that $\Delta w \to \Delta\tilde{w}$ as $w \to \tilde{w}$ for all $w \in \mathcal{N}(\tilde{w})$. From (A.9), for any $w \in \mathcal{N}(\tilde{w})$,

$$\Delta t = -t + \beta(w)\bar{t}.$$

Then for all $\alpha \in [0, 1]$ and all $w \in \mathcal{N}(\tilde{w})$,

$$t + \alpha\Delta t = (1 - \alpha)t + \alpha\beta(w)\bar{t} \in \mathbb{R}_{++}. \tag{A.10}$$

It follows from the mean value theorem that

$$g_w(\alpha) = \alpha \int_0^1 [G'(w + \theta\alpha\Delta w) - G'(w)]\Delta w d\theta.$$

Since $G'(\cdot)$ is uniformly continuous on $\mathcal{N}(\tilde{w})$ for all $w \in \mathcal{N}(\tilde{w})$ we have

$$\lim_{\alpha \downarrow 0} \|g_w(\alpha)\|/\alpha = 0.$$

Then, from (A.9), (A.10), and the fact that $\beta(w) \leq \gamma\psi(w)^{1/2}$ for all $\alpha \in [0, 1]$ and all $w \in \mathcal{N}(\tilde{w})$, we have by repeating the arguments in (A.3)–(A.5) that

$$(t + \alpha\Delta t)^2 \leq (1 - \alpha)^2 t^2 + 2\alpha\gamma\psi(w)\bar{t} + O(\alpha^2), \tag{A.11}$$

$$\|G(w + \alpha\Delta w)\|^2 = (1 - \alpha)^2\|G(w)\|^2 + o(\alpha), \tag{A.12}$$

$$\psi(w + \alpha\Delta w) \leq [1 - 2(1 - \gamma\bar{t})\alpha]\psi(w) + o(\alpha). \tag{A.13}$$

Then from inequality (A.13) we can find a positive number $\bar{\alpha} \in [0,1]$ such that for all $\alpha \in [0, \bar{\alpha}]$ and all $w \in \mathcal{N}(\tilde{w})$,

$$\psi(w + \alpha \Delta w) \leq [1 - 2\sigma(1 - \gamma \bar{t})\alpha]\psi(w).$$

Therefore, for a nonnegative integer ℓ such that $\delta^\ell \in (0, \bar{\alpha}]$, we have

$$\psi(w^k + \delta^\ell \Delta w^k) \leq [1 - 2\sigma(1 - \gamma \bar{t})\delta^\ell]\psi(w^k)$$

for all sufficiently large k. Then for every sufficiently large k, we see that $\ell_k \leq \ell$ and hence $\delta^{\ell_k} \geq \delta^\ell$. Therefore

$$\psi(w^{k+1}) \leq [1 - 2\sigma(1 - \gamma \bar{t})\delta^{\ell_k}]\psi(w^k) \leq [1 - 2\sigma(1 - \gamma \bar{t})\delta^\ell]\psi(w^k)$$

for all sufficiently large k. This contradicts the fact that sequence $\{\psi(w^k)\}$ converges to $\tilde{\psi} > 0$. Thus we have proved that $\tilde{\psi} = 0$, which means that $\tilde{w}$ is a solution of $H(w) = 0$. □

Proof of Theorem 5.4. First, from Theorem 5.1, w^* is a solution of $H(w) = 0$. Then, from Proposition 3.1 in [25], for all w^k sufficiently close to w^*,

$$\|(H'(w^k))^{-1}\| = O(1).$$

Theorem 3.5 implies H is strongly semismooth at w^*. From (2.1), for w^k sufficiently close to w^*, we have

$$\begin{aligned} &\|w^k + \Delta w^k - w^*\| \\ &= \|w^k + H'(w^k)^{-1}[-H(w^k) + \beta_k \bar{w}] - w^*\| \\ &= O(\|H(w^k) - H(w^*) - H'(w^k)(w^k - w^*)\| + \beta_k \bar{t} \\ &= O(\|w^k - w^*\|^2) + O(\psi(w^k))). \end{aligned} \tag{A.14}$$

Since H is strongly semismooth at w^*, H is locally Lipschitz continuous near w^*. Then for all w^k sufficiently close to w^*, we have

$$\psi(w^k) = \|H(w^k)\|^2 = O(\|w^k - w^*\|^2).$$

Thus, from (A.14), we have for all w^k sufficiently close to w^*,

$$\|w^k + \Delta w^k - w^*\| = O(\|w^k - w^*\|^2).$$

Since each $V \in \partial H(w^*)$ is nonsingular, following the proof of Theorem 3.1 of [24], we have for all w^k sufficiently close to w^*,

$$\|w^k - w^*\| = O(\|H(w^k) - H(w^*)\|).$$

Hence, for all w^k sufficiently close to w^*, we have

$$\begin{aligned} &\psi(w^k + \Delta w^k) = \|H(w^k + \Delta w^k)\|^2 \\ &= O(\|w^k + \Delta w^k - w^*\|^2) = O(\|w^k - w^*\|^4) \\ &= O(\|H(w^k) - H(w^*)\|^4) = O(\psi(w^k)^2). \end{aligned} \tag{A.15}$$

Therefore, for all w^k sufficiently close to w^*, we have

$$w^{k+1} = w^k + \Delta w^k, \tag{A.16}$$

which means that $\ell_k = 0$ for all sufficiently large k. Equations (A.16) and (A.14) prove the first result in (5.4).

Next, from the definition of β_k and the fact $H(w^k) \to 0$ as $k \to \infty$ for all k sufficiently large, $\beta_k = \gamma\|H(w^k)\|^2$. In addition for k sufficiently large, $w^{k+1} = w^k + \Delta w^k$, then

$$t^{k+1} = t^k + \Delta t^k = \beta_k \bar{t} = \gamma\|H(w^k)\|^2\bar{t},$$

which, together with (A.15), leads to the second result in (5.4). □

Acknowledgments. We are much indebted to Paul Tseng, who drew our attention to [13, 17]. We are grateful to Al Goldstein for providing us with his excellent paper [13] and for his up-to-date comments on exponential penalty method. We are also very thankful to M. Seetharama Gowda for his valuable comments on the solution structure of the EVLCP. We would like to thank Defeng Sun for his detailed comments on an earlier version of this paper. Finally we would like to thank both referees; their opinions and comments improved the presentation of the paper greatly.

REFERENCES

[1] D. P. Bertsekas, *Approximation procedures based on the method of multipliers*, J. Optim. Theory Appl., 23 (1977), pp. 487–510.
[2] P.-L. Chang, *A Minimax Approach to Nonlinear Programming*, Ph.D. dissertation, Department of Mathematics, University of Washington, Seattle, WA, 1980.
[3] B. Chen and P. T. Harker, *Smooth approximations to nonlinear complementarity problems*, SIAM J. Optim., 7 (1997), pp. 403–420.
[4] C. Chen and O. L. Mangasarian, *Smoothing methods for convex inequalities and linear complementarity problems*, Math. Programming, 71 (1995), pp. 51–69.
[5] X. Chen, L. Qi, and D. Sun, *Global and superlinear convergence of the smoothing Newton method and its application to general box constrained variational inequalities*, Math. Comp., 67 (1998), pp. 519–540.
[6] F. H. Clarke, *Optimization and Nonsmooth Analysis*, 2nd ed., Classics Appl. Math. 5, SIAM, Philadelphia, PA, 1990.
[7] R. W. Cottle and G. B. Dantzig, *A generalization of the linear complementarity problem*, J. Combin. Theory, 8 (1970), pp. 79–90.
[8] R. W. Cottle, J.-S. Pang, and R. E. Stone, *The Linear Complementarity Problem*, Academic Press, Boston, MA, 1992.
[9] B. De Schutter and B. De Moor, *The extended linear complementarity problem*, Math. Programming, 71 (1995), pp. 289–325.
[10] A. A. Ebiefung, *Nonlinear mappings associated with the generalized linear complementarity problem*, Math. Programming, 69 (1995), pp. 255–268.
[11] A. A. Ebiefung and M. M. Kostreva, *The generalized Leontief input-output model and its application to the choice of the new technology*, Ann. Oper. Res., 44 (1993), pp. 161–172.
[12] T. Fujisawa and E. S. Kuh, *Piecewise-linear theory of nonlinear networks*, SIAM J. Appl. Math., 22 (1972), pp. 307–328.
[13] A. A. Goldstein, *Chebyshev Approximation and Linear Inequalities via Exponentials*, Report, Department of Mathematics, University of Washington, Seattle, WA, 1997.
[14] M. S. Gowda, *On the extended linear complementarity problem*, Math. Programming, 72 (1996), pp. 33–50.
[15] M. S. Gowda and R. Sznajder, *The generalized order linear complementarity problem*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 779–795.
[16] M. S. Gowda and R. Sznajder, *A generalization of the Nash equilibrium theorem on bimatrix games*, Internat. J. Game Theory, 25 (1996), pp. 1–12.
[17] L. Kachiyan, *Coordination complexity of parallel price-directive decomposition*, Math. Oper. Res., 21 (1996), pp. 321–340.
[18] X.-S. Li, *An aggregate function method for nonlinear programming*, Sci. China Ser. A, 34 (1991), pp. 1467–1473.
[19] O. L. Mangasarian and J.-S. Pang, *The extended linear complementarity problem*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 359–368.

[20] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim., 15 (1977), pp. 957–972.
[21] S. R. Mohan, S. K. Neogy, and R. Sridhar, *The generalized linear complementarity problem revisited*, Math. Programming, 74 (1996), pp. 197–218.
[22] J.-M. Peng and Z.-H. Lin, *A Non-Interior Continuation Method for Generalized Complementarity Problems*, Tech. Report, Institute of Computational Mathematics and Scientific/Engineering Computing, Beijing, China, 1997.
[23] H.-D. Qi, *A Regularized Smoothing Newton Method for Box Constrained Variational Inequality Problem with P_0 Functions*, Tech. Report, Institute of Computational Mathematics and Scientific/Engineering Computing, Beijing, China, 1997.
[24] L. Qi, *Convergence analysis of some algorithms for solving nonsmooth equation*, Math. Oper. Res., 18 (1993), pp. 227–244.
[25] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Programming, 58 (1993), pp. 691–714.
[26] L. Qi, D. Sun, and G. Zhou, *A New Look at Smoothing Newton Methods for Nonlinear Complementarity Problems and Box Constrained Variational Inequalities*, Report AMR 97/13, School of Mathematics, University of New South Wales, Sydney, Australia, 1997.
[27] D. Sun, *A regularization Newton method for solving nonlinear complementarity problems*, Appl. Math. Optim., to appear.
[28] M. Sun, *Monotonicity of Mangasarian's iterative algorithm for generalized linear complementarity problems*, J. Math. Anal. Appl., 144 (1989), pp. 474–485.
[29] R. Sznajder and M. S. Gowda, *Generalizations of P_0- and P-properties, extended vertical and horizontal LCP's*, Linear Algebra Appl., 223/224 (1995), pp. 695–716.
[30] P. Tseng and D. P. Bertsekas, *On the convergence of the exponential multiplier method for convex programming*, Math. Programming, 60 (1993), pp. 1–19.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 1, pp. 67–78
© 1999 Society for Industrial and Applied Mathematics

ON THE CONVERGENCE OF THE MULTISPLITTING METHODS FOR THE LINEAR COMPLEMENTARITY PROBLEM*

ZHONG-ZHI BAI†

Abstract. The convergence properties of a variant of the multisplitting methods for solving the large sparse linear complementarity problems presented by Machida, Fukushima, and Ibaraki [*J. Comput. Appl. Math.*, 62 (1995), pp. 217–227] are further discussed when the system matrices are nonsymmetric and the weighting matrices are nonnegative and diagonal. This directly results in several novel sufficient conditions for guaranteeing the convergence of these multisplitting methods. Moreover, some applicable parallel multisplitting relaxation methods and their corresponding convergence properties are discussed in detail.

Key words. linear complementarity problem, matrix multisplitting, relaxed method, convergence

AMS subject classifications. 65H10, 65W05

PII. S0895479897324032

1. Introduction.

The linear complementarity problem LCP(M,q) consists of finding a vector $z \in \mathbb{R}^n$ such that

$$Mz + q \geq 0, \qquad z \geq 0, \qquad z^T(Mz + q) = 0, \tag{1.1}$$

where $M \in \mathbb{R}^{n\times n}$ and $q \in \mathbb{R}^n$ are a given real matrix and a vector, respectively. This problem usually arises in (linear and) convex quadratic programming, in the problem of finding a Nash equilibrium point of a bimatrix game (e.g., Cottle and Dantzig [3] and Lemke [12]), and also in a number of free boundary problems of fluid mechanics (e.g., Cryer [7]). Therefore, it has various practical applications. To get a numerical solution to this linear complementarity problem, we often utilize the splitting methods, in particular when the system matrix $M \in \mathbb{R}^{n\times n}$ is large and sparse. This class of splitting methods originates from the matrix splitting methods such as the Jacobi, Gauss–Seidel, and SOR iterations for solving systems of linear equations. It was first introduced into the linear complementarity problem by Cottle and Sacher [6] and Cottle, Golub, and Sacher [4] and was further developed by many authors, e.g., Mangasarian [15], Pang [19], Luo and Tseng [13], and Iusem [10, 11]. Now, it has become one of the most practical and efficient iterative solvers for the large sparse linear complementarity problems, possessing systematic and deepened mathematical theories. For details, one can refer to Cottle, Pang, and Stone [5], Berman and Plemmons [2], and references therein.

To suit the requirements of high-speed parallel multiprocessor systems, Machida, Fukushima, and Ibaraki [14] recently presented a class of parallel multisplitting iterative methods for solving the symmetric linear complementarity problems in the synchronous parallel computing environments. These methods are based on several splittings of the system matrix $M \in \mathbb{R}^{n\times n}$ and are constructed with a suitable weighting combination of the solutions of the linear complementarity problems resulting

*Received by the editors July 7, 1997; accepted for publication (in revised form) by R. Cottle September 17, 1998; published electronically August 3, 1999. This work was previously known as project 19601036 and was supported by the National Natural Science Foundation of China.
http://www.siam.org/journals/simax/21-1/32403.html

†State Key Laboratory of Scientific/Engineering Computing, Institute of Computational Mathematics and Scientific/Engineering Computing, Chinese Academy of Sciences, P.O. Box 2719, Beijing 100080, People's Republic of China (bzz@lsec.cc.ac.cn).

from the splittings. Therefore, they are technical extensions of the matrix multisplitting methods for the system of linear equations in O'Leary and White [17] to large sparse symmetric linear complementarity problems, and they can attain high parallel efficiency in practical computations. Under suitable restrictions on the weighting matrices and the multiple splittings, Machida, Fukushima, and Ibaraki [14] proved the convergence of these methods for the case where the system matrices are symmetric.

In fact, the convergence of the multisplitting methods in [14] is guaranteed for many symmetric linear complementarity problems with multiple solutions, including cases in which the system matrix $M \in \mathbb{R}^{n\times n}$ is just positive semidefinite, and for many matrix multisplittings with standard restrictions on the splitting matrices and certain necessary requirements on the weighting matrices. However, the convergence of these multisplitting methods for nonsymmetric linear complementarity problems and for general nonnegative diagonal weighting matrices still requires study. In this paper, we will discuss further the convergence of a variant of these multisplitting methods, which includes the multisplitting methods in [14] as special cases, and develop the convergence theory of Machida, Fukushima, and Ibaraki [14] to the nonsymmetric linear complementarity problems. For some nonsymmetric matrix class and arbitrary nonnegative diagonal weights, we establish the convergence theory of this new method under suitable constraints of the multiple splittings of the system matrix $M \in \mathbb{R}^{n\times n}$. This directly results in several novel sufficient conditions for guaranteeing the convergence of the new multisplitting method. Since the nonsymmetric problem is considerably harder than the symmetric one, a price is also paid for this development, which imposes more demanding conditions on the multiple splittings in order to ensure the convergence of the new multisplitting method as well as the existence of such splitting matrices. Moreover, as applications, we present some applicable parallel multisplitting relaxation methods which are naturally induced by the above multisplitting method, and we discuss the convergence of these relaxed methods in detail.

2. Preliminaries. In this section, we briefly review the multisplitting method discussed by Machida, Fukushima, and Ibaraki in [14] for solving the LCP(M,q).

First, we introduce some necessary notations and concepts. A matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ is called a monotone matrix if it is nonsingular and satisfies $A^{-1} \geq 0$; an M-matrix if it is a monotone matrix and satisfies $a_{ij} \leq 0$ for $i \neq j$, $i, j = 1, 2, \ldots, n$; an H-matrix if its comparison matrix $\langle A\rangle$ is an M-matrix, where $\langle A\rangle = (\langle a_{ij}\rangle) \in \mathbb{R}^{n\times n}$ is defined by $\langle a_{ii}\rangle = |a_{ii}|$ for $i = 1, 2, \ldots, n$, and $\langle a_{ij}\rangle = -|a_{ij}|$ for $i \neq j$, $i, j = 1, 2, \ldots, n$; an H_+-matrix if it is an H-matrix having positive diagonal elements; and a Q-matrix if the LCP(A,b) has a solution for any $b \in \mathbb{R}^n$. A sufficient condition for $A \in \mathbb{R}^{n\times n}$ to be a Q-matrix is that either A is an H_+-matrix [1] or A is a strictly copositive matrix [5].

For a given matrix $A \in \mathbb{R}^{n\times n}$, let $F, G \in \mathbb{R}^{n\times n}$ be such that $A = F + G$. Then (F, G) is called a splitting of the matrix A. The splitting (F, G) is called a convergent splitting if the spectral radius of the matrix $(F^{-1}G)$ is less than one, i.e., $\rho(F^{-1}G) < 1$. It is called a weak regular splitting if $F^{-1} \geq 0$ and $F^{-1}G \leq 0$; a regular splitting if $F^{-1} \geq 0$ and $G \leq 0$; an M-splitting if F is an M-matrix and $G \leq 0$; an H-splitting if $\langle F\rangle - |G|$ is an M-matrix; an H-compatible splitting if $\langle A\rangle = \langle F\rangle - |G|$; and a Q-splitting if F is a Q-matrix. In particular, the splitting (F, G) is called an H_+-splitting and H_+-compatible splitting if it is an H-splitting and H-compatible splitting, respectively, with F an H_+-matrix. Let (B, C) be a splitting of the matrix A such that $\langle B\rangle$ is nonsingular, and let (F, G) be a matrix pair. Then we call the

matrix pair (F, G) a majorizing pair of the splitting (B, C) if F is an H-matrix and it holds that $\langle B\rangle^{-1} \leq \langle F\rangle^{-1}$ and $|C| \leq |G|$. In such a case, we say that (B, C) is majorized by (F, G). Note that here (F, G) is not necessarily a splitting of the matrix A. Evidently, if $A = B + C$ is an H-splitting, then A and B are H-matrices and $\rho(B^{-1}C) \leq \rho(\langle B\rangle^{-1}|C|) < 1$; if it is an H-compatible splitting and A is an H-matrix, then it is an H-splitting and thus convergent; and if (F, G) is a majorizing pair of the splitting (B, C) such that $\langle F\rangle - |G|$ is an M-matrix, then (B, C) is a convergent splitting.

A matrix $A \in \mathbb{R}^{n\times n}$ is symmetric if $A^T = A$, where A^T denotes the transposed matrix of A. It is positive semidefinite if $z^T Az \geq 0$ for all $z \in \mathbb{R}^n$ and positive definite if $z^T Az > 0$ for all $z \in \mathbb{R}^n \setminus \{0\}$. Let $N_0 = \{0, 1, 2, \dots\}$ and $\{A_p\}_{p\in N_0}$ be a sequence of matrices in $\mathbb{R}^{n\times n}$. Then we call $A_p (p \in N_0)$ positive definite uniformly in p if there exists a positive constant c, independent of p, such that $z^T A_p z \geq cz^T z$ holds for all $z \in \mathbb{R}^n$.

We now cite some lemmas which were proved in [8, 9, 18], respectively, and which will frequently be used in what follows.

LEMMA 2.1 (see [8, 18]). *Let $A \in \mathbb{R}^{n\times n}$ have nonpositive off-diagonal entries. Then A is an M-matrix if and only if there exists a positive vector $u \in \mathbb{R}^n$ such that $Au \geq 0$.*

LEMMA 2.2 (see [9, 18]). *Let $A \in \mathbb{R}^{n\times n}$ be an H-matrix, $D = \mathrm{diag}(A)$, and $A = D - B$. Then*

(a) *A is nonsingular,*

(b) *$|A^{-1}| \leq \langle A\rangle^{-1}$, and*

(c) *$|D|$ is nonsingular and $\rho(|D|^{-1}|B|) < 1$.*

The following result deals with the existence and uniqueness of the solution of the LCP(M,q).

LEMMA 2.3 (see [1]). *Let $M \in \mathbb{R}^{n\times n}$ be an H_+-matrix. Then the LCP(M,q) has a unique solution for any $q \in \mathbb{R}^n$.*

A multisplitting of a matrix $M \in \mathbb{R}^{n\times n}$ is a collection of triples $(B_k, C_k, E_k)(k = 1, 2, \dots, K)$, which satisfies

(1) $M = B_k + C_k (k = 1, 2, \dots, K)$ are Q-splittings; and

(2) $E_k (k = 1, 2, \dots, K)$ are nonnegative diagonal matrices with $\sum_{k=1}^{K} E_k = I$ (the $n \times n$ identity matrix).

Let $(B_{k,p}, C_{k,p}, E_{k,p})(k = 1, 2, \dots, K)$, $p \in N_0$, be a sequence of multisplittings of the matrix M. Then we consider the following multisplitting method for solving the LCP(M,q).

METHOD 2.1 (multisplitting method for the LCP(M,q)).

Step 1. *Choose an arbitrary starting vector $z^0 \in \mathbb{R}^n$. Set $p := 0$.*

Step 2. *For each $k \in \{1, 2, \dots, K\}$, let $z^{k,p}$ be an arbitrary solution of LCP$(B_{k,p}, q_{k,p})$:*

$$z \geq 0, \qquad B_{k,p}z + q_{k,p} \geq 0, \qquad z^T(B_{k,p}z + q_{k,p}) = 0,$$

where $q_{k,p} = C_{k,p}z^p + q$.

Step 3. $z^{p+1} = \sum_{k=1}^{K} E_{k,p} z^{k,p}$.

Step 4. *If $z^{p+1} = z^p$, then stop. Otherwise, set $p := p + 1$ and return to Step* 2.

This multisplitting method has quite good parallel computational properties. At every iterate step p, each subproblem LCP$(B_{k,p}, C_{k,p})(k = 1, 2, \dots, K)$ can be solved independently on one processor of the multiprocessor system, and hence Method 2.1 can be implemented in parallel. The K splitting matrices $B_{k,p}(k = 1, 2, \dots, K)$ and

the K weighting matrices $E_{k,p}(k = 1, 2, \ldots, K)$ can be chosen in such a way that the tasks distributed on the K processors of the multiprocessor system are well balanced so that Method 2.1 achieves high parallel efficiency. Moreover, considerable saving on the computational workload is available because the entries of $z^{k,p}$ corresponding to the zero elements of the weighting matrix $E_{k,p}$ need not be computed.

We remark that Method 2.1 is somewhat more general than Algorithm 3.2 in [14]. Here, the multiple splittings $M = B_{k,p} + C_{k,p}(k = 1, 2, \ldots, K, p \in N_0)$ of the system matrix $M \in \mathbb{R}^{n\times n}$ are permitted to vary with p, the iteration index, and the weighting matrices $E_{k,p}(k = 1, 2, \ldots, K, p \in N_0)$ are allowed to be arbitrary nonnegative diagonal matrices. Moreover, the stopping instruction of Method 2.1 is issued by the host processor, and the other processors can get this instruction directly to terminate their computations. This greatly decreases the unnecessary data exchanges among the processors and hence improves the parallel efficiency of Method 2.1.

When the system matrix $M \in \mathbb{R}^{n\times n}$ is symmetric, an analogous demonstration to Machida, Fukushima, and Ibaraki [14] immediately gives the following convergence theorem of Method 2.1.

THEOREM 2.1. *Let $M \in \mathbb{R}^{n\times n}$ be a symmetric matrix, and for every $p \in N_0$, let $(B_{k,p}, C_{k,p}, E_{k,p})(k = 1, 2, \ldots, K)$ be a multisplitting of the matrix M such that*

(a) *$B_{k,p} - C_{k,p}(k = 1, 2, \ldots, K)$ are positive definite uniformly in p; and*

(b) *$f(\sum_{k=1}^{K} E_{k,p} z^{k,p}) \le \max_{1\le k\le K} f(z^{k,p})$, where $f(z) = \frac{1}{2} z^T M z + z^T q$.*

Assume that $f(z)$ is bounded below on $z \ge 0$ and that $0 \ne z \ge 0$, $Mz \ge 0$, and $z^T M z = 0$ imply $q^T z > 0$. Then the sequence $\{z^p\}_{p\in N_0}$ generated by Method 2.1 is bounded and any accumulation point of it solves LCP(M,q).

We remark again that when $B_{k,p} = B_k(k = 1, 2, \ldots, K)$ and $C_{k,p} = C_k(k = 1, 2, \ldots, K)$ for all $p \in N_0$, Method 2.1 reduces to Algorithm 3.2 in [14] and Theorem 2.1 recovers Theorem 3.6 in [14], respectively. Various choices of the weighting matrices $E_{k,p}(k = 1, 2, \ldots, K, p \in N_0)$ satisfy condition (b) in Theorem 2.1. As shown in [14], one of the possibilities is given by the choices of $E_{k,p} = \alpha_{k,p} I$ $(k = 1, 2, \ldots, K, p \in N_0)$, where $\alpha_{k,p}(k = 1, 2, \ldots, K, p \in N_0)$ are nonnegative real numbers satisfying $\sum_{k=1}^{K} \alpha_{k,p} = 1$ $(p \in N_0)$. In this case, condition (b) is automatically satisfied if any of the following three classes of restrictions is further imposed:

(1) For a positive integer sequence $\{k_p\} \in \{1, 2, \ldots, K\}$,

$$\alpha_{k,p} = \begin{cases} 1 & \text{for } k = k_p, \\ 0 & \text{for } k \ne k_p, \end{cases} \qquad k = 1, 2, \ldots, K, \quad p \in N_0,$$

where the indices $k_p(p \in N_0)$ are chosen either randomly at every iteration or in a certain predetermined order such as the cyclic rule, or they are based on the function values $f(z^{k,p})(k = 1, 2, \ldots, K)$ such that, for $p \in N_0$, $f(z^{k_p,p}) = \min_{1\le k\le K} f(z^{k,p})$.

(2) $\alpha_{k,p}(k = 1, 2, \ldots, K, p \in N_0)$ are the minimizers of the functions $g(\alpha_{1,p}, \alpha_{2,p}, \ldots, \alpha_{K,p}) = f(\sum_{k=1}^{K} \alpha_{k,p} z^{k,p})$, $p \in N_0$.

(3) $M \in \mathbb{R}^{n\times n}$ is a positive semidefinite matrix.

For detailed discussions about the choices of the splitting and the weighting matrices, one can refer to [1, 9, 14, 16, 17, 21] and references therein.

3. New convergence theories. In this section, we will discuss the convergence of Method 2.1 for the case where the system matrix $M \in \mathbb{R}^{n\times n}$ belongs to the H_+-matrix class. We first establish a general convergence result.

THEOREM 3.1. *Let $M \in \mathbb{R}^{n\times n}$ be an H_+-matrix, and for every $p \in N_0$ let $(B_{k,p}, C_{k,p}, E_{k,p})(k = 1, 2, \dots, K)$ be a multisplitting of the matrix M such that*

(a) *$B_{k,p}(k = 1, 2, \dots, K)$ are H_+-matrices;*

(b) *for each k, there exists a majorizing pair $(\widehat{B}_{k,p}, \widehat{C}_{k,p})$ of the splitting $(B_{k,p}, C_{k,p})$ satisfying*

$$\widehat{H}_p u \le \theta u, \qquad \widehat{H}_p = \sum_{k=1}^{K} E_{k,p}\langle \widehat{B}_{k,p}\rangle^{-1}|\widehat{C}_{k,p}| \qquad \forall p \in N_0 \tag{3.1}$$

for some nonnegative constant $\theta \in [0, 1)$ and some positive vector $u \in \mathbb{R}^n$.

Then the sequence $\{z^p\}_{p\in N_0}$ generated by Method 2.1 converges to the unique solution of the LCP(M,q).

Proof. Because $M \in \mathbb{R}^{n\times n}$ is an H_+-matrix, the LCP(M,q) has a unique solution $z^* \in \mathbb{R}^n$. That is to say, it holds that

$$z^* \ge 0, \qquad Mz^* + q \ge 0, \qquad (z^*)^T(Mz^* + q) = 0. \tag{3.2}$$

On the other hand, since for each $k \in \{1, 2, \dots, K\}$ and each $p \in N_0$, $B_{k,p}$ is an H_+-matrix, by Lemma 2.3 the LCP$(B_{k,p}, q_{k,p})$ has a unique solution $z^{k,p}$, where $q_{k,p} = C_{k,p}z^p + q$. That is to say, it holds that

$$z^{k,p} \ge 0, \qquad B_{k,p}z^{k,p} + q_{k,p} \ge 0, \qquad (z^{k,p})^T(B_{k,p}z^{k,p} + q_{k,p}) = 0. \tag{3.3}$$

Now, we claim that the following estimates hold:

$$|z^{p+1} - z^*| \le \sum_{k=1}^{K} E_{k,p}\langle B_{k,p}\rangle^{-1}|C_{k,p}||z^p - z^*| \qquad \forall p \in N_0. \tag{3.4}$$

In fact, denoting $D_{k,p} = \mathrm{diag}(B_{k,p})$ and $\widetilde{B}_{k,p} = B_{k,p} - D_{k,p}$, we will verify (3.4) in all cases.

First, when the ith element of z^*, $[z^*]_i$, is positive and $[B_{k,p}z^{k,p} + q_{k,p}]_i = 0$, it clearly holds that

$$[B_{k,p}(z^{k,p} - z^*) + C_{k,p}(z^p - z^*)]_i = 0$$

or equivalently

$$[D_{k,p}(z^{k,p} - z^*)]_i = [-\tilde{B}_{k,p}(z^{k,p} - z^*) - C_{k,p}(z^p - z^*)]_i.$$

Therefore,

$$[D_{k,p}|z^{k,p} - z^*|]_i \le [|\tilde{B}_{k,p}||z^{k,p} - z^*| + |C_{k,p}||z^p - z^*|]_i,$$

and hence

$$[\langle B_{k,p}\rangle|z^{k,p} - z^*|]_i \le [|C_{k,p}||z^p - z^*|]_i. \tag{3.5}$$

Second, when $[z^*]_i > 0$ and $[B_{k,p}z^{k,p} + q_{k,p}]_i > 0$, it holds that $[z^{k,p}]_i = 0$ and

$$\begin{cases} [\widetilde{B}_{k,p}z^{k,p} + C_{k,p}z^p + q]_i > 0, \\ [\widetilde{B}_{k,p}z^* + C_{k,p}z^* + q]_i = -[D_{k,p}z^*]_i. \end{cases}$$

By subtracting these two inequalities we obtain

$$[\widetilde{B}_{k,p}(z^{k,p}-z^*)+C_{k,p}(z^p-z^*)]_i > [D_{k,p}z^*]_i.$$

Therefore,

$$\begin{aligned}[D_{k,p}|z^{k,p}-z^*|]_i = [D_{k,p}z^*]_i &< |[\widetilde{B}_{k,p}(z^{k,p}-z^*)+C_{k,p}(z^p-z^*)]_i| \\ &\le [|\widetilde{B}_{k,p}||z^{k,p}-z^*|+|C_{k,p}||z^p-z^*|]_i,\end{aligned}$$

and hence (3.5) also holds.

Finally, noticing that when $[z^*]_i = 0$ and $[z^{k,p}]_i = 0$ (3.5) holds automatically, and that when $[z^*]_i = 0$ and $[z^{k,p}]_i > 0$, (3.5) can be demonstrated analogously to the second case, we conclude that (3.5) holds for all $i \in \{1, 2, \dots, n\}$, all $k \in \{1, 2, \dots, K\}$, and all $p \in N_0$.

Since $B_{k,p}(k = 1, 2, \dots, K, p = 0, 1, 2, \dots)$ are H-matrices, it holds that

$$|z^{k,p}-z^*| \le \langle B_{k,p}\rangle^{-1}|C_{k,p}||z^p-z^*|, \qquad k = 1, 2, \dots, K, \quad p = 0, 1, 2, \dots .$$

Because for each $k \in \{1, 2, \dots, K\}$ and $p \in N_0$, $B_{k,p}$ is an H_+-matrix and $(\widehat{B}_{k,p}, \widehat{C}_{k,p})$ is a majorizing sequence of $(B_{k,p}, C_{k,p})$, we have

$$\langle B_{k,p}\rangle^{-1}|C_{k,p}| \le \langle \widehat{B}_{k,p}\rangle^{-1}|\widehat{C}_{k,p}|, \qquad k = 1, 2, \dots, K, \quad p = 0, 1, 2, \dots ,$$

and hence

$$\begin{aligned}|z^{p+1}-z^*| = \left|\sum_{k=1}^{K} E_{k,p}(z^{k,p}-z^*)\right| &\le \sum_{k=1}^{K} E_{k,p}|z^{k,p}-z^*| \\ &\le \sum_{k=1}^{K} E_{k,p}\langle \widehat{B}_{k,p}\rangle^{-1}|\widehat{C}_{k,p}||z^p-z^*| := \widehat{H}_p|z^p-z^*|.\end{aligned}$$

Let $\delta > 0$ be such that $|z^0 - z^*| \le \delta u$. Then we have

$$|z^{p+1}-z^*| \le \widehat{H}_p\widehat{H}_{p-1}\cdots\widehat{H}_0\delta u \le \theta^{p+1}\delta u \longrightarrow 0 \qquad (p \to \infty).$$

That is to say, $\lim_{p\to\infty} z^p = z^*$. □

In Theorem 3.1 no restriction is imposed on the weighting matrices $E_{k,p}(k = 1, 2, \dots, K, p \in N_0)$. Hence, we can choose them suitably such that Method 2.1 achieves high parallel computational efficiency. On the other hand, the restriction on the matrix splittings $M = B_{k,p} - C_{k,p}(k = 1, 2, \dots, K, p \in N_0)$ in Theorem 3.1 is quite different from that in Theorem 2.1. Conditions imposed to ensure the existence of the multiple splittings satisfying (3.1) may be more restrictive than those in Theorem 2.1 and in [14] in some sense. However, because the nonsymmetric problem is considerably harder than the symmetric one, these conditions are necessary for guaranteeing the convergence of Method 2.1. As a matter of fact, they naturally originate from the standard conditions for the parallel matrix multisplitting iterative methods of the systems of linear equations [9, 16, 17]. Two concrete choices of the multiple splittings satisfying condition (3.1) are given in the following examples.

EXAMPLE 3.1. *Let $M = (m_{ij}) \in \mathbb{R}^{n\times n}$ be an H_+-matrix, $\overline{D} = \mathrm{diag}(M)$, and $\overline{B} = M - \overline{D}$. For a given positive constant $\beta \ge 1$, let $D_{k,p}(k = 1, 2, \dots, K, p \in N_0)$ be positive diagonal matrices satisfying $\overline{D} \le D_{k,p} \le \beta\overline{D}$ $(k = 1, 2, \dots, K, p \in N_0)$. Take*

$$B_{k,p} = D_{k,p}, \quad C_{k,p} = M - D_{k,p}, \qquad k = 1, 2, \dots, K, \quad p = 0, 1, 2, \dots .$$

Then condition (3.1) *is satisfied because*

$$\widehat{B}_{k,p} = D_{k,p}, \quad \widehat{C}_{k,p} = D_{k,p} - \overline{D} + |\overline{B}|, \qquad k = 1, 2, \dots, K, \quad p = 0, 1, 2, \dots,$$

$u = \langle M \rangle^{-1} e$ *with* $e = (1, 1, \dots, 1)^T \in \mathbb{R}^n$, *and* $\theta = \max\{0, 1 - \frac{1}{\max\{\beta m_{ii} u_i\}}\}$.

EXAMPLE 3.2. *Let* $M = (m_{ij}) \in \mathbb{R}^{n \times n}$ *be an* H_+*-matrix,* $\overline{D} = \operatorname{diag}(M)$, *and* $\overline{B} = M - \overline{D}$. *For each* $p \in N_0$ *and* $k \in \{1, 2, \dots, K\}$, *take* $B_{k,p} = (b_{ij}^{(k,p)}) \in \mathbb{R}^{n \times n}$ *and* $C_{k,p} = (c_{ij}^{(k,p)}) \in \mathbb{R}^{n \times n}$ *in accordance with the following rule:*

$$b_{ij}^{(k,p)} = \begin{cases} m_{ij} & \text{for } i = j, \\ b_{ij}^{(k,p)} \in [0, m_{ij}] & \text{for } i \neq j \text{ and } m_{ij} \geq 0, \\ b_{ij}^{(k,p)} \in [m_{ij}, 0] & \text{for } i \neq j \text{ and } m_{ij} < 0, \end{cases} \qquad i, j = 1, 2, \dots, n,$$

$$c_{ij}^{(k,p)} = \begin{cases} 0 & \text{for } i = j, \\ m_{ij} - b_{ij}^{(k,p)} & \text{for } i \neq j, \end{cases} \qquad i, j = 1, 2, \dots, n.$$

Then condition (3.1) *is satisfied because*

$$\widehat{B}_{k,p} = \langle B_{k,p} \rangle, \quad \widehat{C}_{k,p} = |C_{k,p}|, \qquad k = 1, 2, \dots, K, \quad p = 0, 1, 2, \dots,$$

$u = \langle M \rangle^{-1} e$ *with* $e = (1, 1, \dots, 1)^T \in \mathbb{R}^n$, *and* $\theta = \max\{0, 1 - \frac{1}{\max\{m_{ii} u_i\}}\}$.

Theorem 3.1 immediately implies the following convergence conclusion about Method 2.1.

THEOREM 3.2. *Let* $M \in \mathbb{R}^{n \times n}$ *be an* H_+*-matrix, and for every* $p \in N_0$ *let* $(B_{k,p}, C_{k,p}, E_{k,p})(k = 1, 2, \dots, K)$ *be a multisplitting of the matrix* M *such that*

(a) $B_{k,p}(k = 1, 2, \dots, K)$ *are* H_+*-matrices;*

(b) *for each* k, *there exists a majorizing pair* $(\widehat{B}_{k,p}, \widehat{C}_{k,p})$ *of the splitting* $(B_{k,p}, C_{k,p})$ *satisfying*

$$\begin{cases} \langle \widehat{B}_{k,p} \rangle - |\widehat{C}_{k,p}| \geq \widehat{M}_k, \\ u > 0, \quad \widehat{M}_k u > 0 \Longrightarrow \langle \widehat{B}_{k,p} \rangle u \leq \dfrac{1}{1 - \theta} \widehat{M}_k u \quad \text{for some} \quad \theta \in [0, 1), \end{cases} \tag{3.6}$$

for some matrices $\widehat{M}_k \in \mathbb{R}^{n \times n}(k = 1, 2, \dots, K)$, *for which there exists a positive vector* $u \in \mathbb{R}^n$ *satisfying* $\widehat{M}_k u > 0$.

Then the sequence $\{z^p\}_{p \in N_0}$ *generated by Method* 2.1 *converges to the unique solution of the LCP(M,q).*

Proof. Obviously, we need only to verify that (3.6) implies the validity of (3.1). As a matter of fact, because there exists a positive vector $u \in \mathbb{R}^n$ such that $\widehat{M}_k u > 0(k = 1, 2, \dots, K)$, in accordance with Lemma 2.1 we know that $\widehat{M}_k \in \mathbb{R}^{n \times n}$ are M-matrices. From (3.6) we have

$$\langle \widehat{B}_{k,p} \rangle u - |\widehat{C}_{k,p}| u \geq \widehat{M}_k u$$

or equivalently

$$\langle \widehat{B}_{k,p} \rangle^{-1} |\widehat{C}_{k,p}| u \leq u - \langle \widehat{B}_{k,p} \rangle^{-1} \widehat{M}_k u \leq u - (1 - \theta) u = \theta u.$$

Therefore,

$$\widehat{H}_p u = \sum_{k=1}^{K} E_{k,p}\langle \widehat{B}_{k,p}\rangle^{-1}|\widehat{C}_{k,p}|u \le \sum_{k=1}^{K} E_{k,p}\theta u = \theta u.$$

That is to say, (3.1) holds. □

Theorems 3.1 and 3.2 and Lemmas 2.1 and 2.2 directly give the following results.

THEOREM 3.3. *Let $M \in \mathbb{R}^{n\times n}$ be an H_+-matrix. Assume that for each $p \in N_0$ and $k \in \{1, 2, \dots, K\}$, $M = B_{k,p} + C_{k,p}$ is a splitting with $B_{k,p}$ having positive diagonal elements, and there exist matrices $M_k \in \mathbb{R}^{n\times n}(k = 1, 2, \dots, K)$, a constant $\theta \in [0, 1)$, and a positive vector $u \in \mathbb{R}^n$ such that $M_k u > 0$ and*

$$\begin{cases} \langle B_{k,p}\rangle - |C_{k,p}| \ge M_k, \\ \langle B_{k,p}\rangle u \le \dfrac{1}{1-\theta} M_k u. \end{cases}$$

Then the sequence $\{z^p\}_{p\in N_0}$ generated by Method 2.1 converges to the unique solution of the LCP(M,q).

Proof. Noticing that

$$\langle M\rangle \ge \langle B_{k,p}\rangle - |C_{k,p}| \ge M_k, \qquad k = 1, 2, \dots, K, \quad p = 0, 1, 2, \dots,$$

we see that $\langle B_{k,p}\rangle - |C_{k,p}|$ and $\langle B_{k,p}\rangle$ are M-matrices and hence $B_{k,p}$ are H_+-matrices. Now, taking

$$\widehat{B}_{k,p} = B_{k,p}, \qquad \widehat{C}_{k,p} = C_{k,p}, \qquad k = 1, 2, \dots, K, \quad p = 0, 1, 2, \dots,$$

we know that all the conditions of Theorem 3.2 are satisfied, and thus the conclusion of this theorem holds. □

THEOREM 3.4. *Let $M \in \mathbb{R}^{n\times n}$ be an H_+-matrix. Assume that for each $p \in N_0$ and $k \in \{1, 2, \dots, K\}$, $M = B_{k,p} + C_{k,p}$ is an H_+-compatible splitting. Then the sequence $\{z^p\}_{p\in N_0}$ generated by Method 2.1 converges to the unique solution of the LCP(M,q).*

4. Applications to multisplitting relaxation methods. Method 2.1 is an implicit one, and at every iterate each processor needs to solve a linear complementarity problem of much smaller size. This makes it less convenient in concrete applications. In this section, we will discuss several special explicit forms of Method 2.1, which are convenient for practical implementations.

For $M = (m_{ij}) \in \mathbb{R}^{n\times n}$, denote $D = \mathrm{diag}(M) = \mathrm{diag}(m_{11}, m_{22}, \dots, m_{nn})$. For $k = 1, 2, \dots, K$ and $p \in N_0$, let $L_{k,p} = (l_{ij}^{(k,p)}) \in \mathbb{R}^{n\times n}$ be strictly lower triangular matrices, let $W_{k,p} = (w_{ij}^{(k,p)}) \in \mathbb{R}^{n\times n}$ be zero-diagonal matrices, and let $E_{k,p} = \mathrm{diag}(e_i^{(k,p)}) \in \mathbb{R}^{n\times n}$ be nonnegative diagonal matrices, such that

(1) $M = D + L_{k,p} + W_{k,p}$, $k = 1, 2, \dots, K$, $p = 0, 1, 2, \dots$;
(2) $\det(D) \ne 0$;
(3) $\sum_{k=1}^{K} E_{k,p} = I$, $p = 0, 1, 2, \dots$.

Given real constants γ_k and $\omega_k (\ne 0)$, $k = 1, 2, \dots, K$. Take

$$\begin{cases} B_{k,p} = \dfrac{1}{\omega_k}(D + \gamma_k L_{k,p}), \\ C_{k,p} = \dfrac{1}{\omega_k}((1-\omega_k)D - (\omega_k - \gamma_k)L_{k,p} - \omega_k W_{k,p}), \end{cases} \qquad k = 1, 2, \dots, K, \quad p = 0, 1, 2, \dots$$

in Method 2.1. Then we obtain the following dynamic multisplitting AOR method for solving the LCP(M,q).

METHOD 4.1 (multisplitting AOR method for the LCP(M,q)).

Step 1. *Choose an arbitrary starting vector* $z^0 \in \mathbb{R}^n$. *Set* $p := 0$.

Step 2. *For each* $k \in \{1, 2, \ldots, K\}$ *and for* $i = 1, 2, \ldots, n$, *let* $z^{k,p}$ *be calculated according to*

$$[z^{k,p}]_i = \begin{cases} 0 & \text{if} \quad \gamma_k \sum_{j=1}^{i-1} l_{ij}^{(k,p)}([z^{k,p}]_j - [z^p]_j) + \omega_k[Mz^p + q]_i > m_{ii}[z^p]_i, \\ [z^p]_i + \frac{\gamma_k}{m_{ii}} \sum_{j=1}^{i-1} l_{ij}^{(k,p)}([z^p]_j - [z^{k,p}]_j) - \frac{\omega_k}{m_{ii}}[Mz^p + q]_i & \text{otherwise.} \end{cases}$$

Step 3. $z^{p+1} = \sum_{k=1}^{K} E_{k,p} z^{k,p}$.

Step 4. *If* $z^{p+1} = z^p$, *then stop. Otherwise, set* $p := p + 1$ *and return to Step* 2.

Evidently, Method 4.1 includes Algorithm 4.1 in Machida, Fukushima, and Ibaraki [14] as a special case. On the other hand, when we choose the parameter pairs (γ_k, ω_k), $k = 1, 2, \ldots, K$, to be (ω_k, ω_k), $k = 1, 2, \ldots, K$, $(1, 1)$ and $(0, 1)$, respectively, Method 4.1 gives the dynamic multisplitting SOR, Gauss–Seidel, and Jacobi methods, correspondingly, for the LCP(M,q). Hence, an extensive sequence of parallel multisplitting relaxation methods can be obtained, which is quite practical and efficient for solving the large sparse linear complementarity problems on the high-speed multiprocessor systems. Moreover, the parameters (γ_k, ω_k), $k = 1, 2, \ldots, K$, can be adjusted suitably so that the convergence property of the multisplitting AOR method can be substantially improved. We remark that Method 4.1 is stationary if the matrices $L_{k,p}$, $W_{k,p}$, and $E_{k,p}$ are chosen independently of p, the iterate index.

We have the following convergence theorem for Method 4.1.

THEOREM 4.1. *Let* $M \in \mathbb{R}^{n \times n}$ *be an* H_+*-matrix and* $D = \operatorname{diag}(M)$. *Assume that there exists a matrix* $B \in \mathbb{R}^{n \times n}$ *such that, for every* $p \in N_0$,

$$|L_{k,p}| + |W_{k,p}| \le |B|, \qquad k = 1, 2, \ldots, K. \tag{4.1}$$

If $\rho(D^{-1}|B|) < 1$, *then the sequence* $\{z^p\}_{p \in N_0}$ *generated by Method* 4.1 *converges to the unique solution of the LCP(M,q), provided the relaxation parameters* γ_k *and* ω_k, $k = 1, 2, \ldots, K$, *satisfy*

$$0 \le \gamma_k \le \omega_k, \qquad 0 < \omega_k < \frac{2}{1 + \rho(D^{-1}|B|)}, \qquad k = 1, 2, \ldots, K. \tag{4.2}$$

Proof. Because for each $k \in \{1, 2, \ldots, K\}$ and each $p \in N_0$

$$\begin{cases} \langle B_{k,p} \rangle = \dfrac{1}{\omega_k}(D - \gamma_k |L_{k,p}|) := \widehat{\mathcal{B}}_{k,p}, \\ |C_{k,p}| \le \dfrac{1}{\omega_k}[|1 - \omega_k| D + (\omega_k - \gamma_k)|L_{k,p}| + \omega_k |W_{k,p}|] := \widehat{\mathcal{C}}_{k,p}, \end{cases}$$

$\widehat{\mathcal{B}}_{k,p}$ is an M-matrix, and $\widehat{\mathcal{C}}_{k,p}$ is a nonnegative matrix, by Lemma 2.2 we see that $(\widehat{\mathcal{B}}_{k,p}, \widehat{\mathcal{C}}_{k,p})$ is a majorizing pair of the splitting $(B_{k,p}, C_{k,p})$. Moreover, it holds that

$$\begin{aligned}\widehat{\mathcal{B}}_{k,p} - \widehat{\mathcal{C}}_{k,p} &= \frac{1}{\omega_k}[(1-|1-\omega_k|)D - \omega_k(|L_{k,p}| + |W_{k,p}|)] \\ &\geq \frac{1}{\omega_k}[(1-|1-\omega_k|)D - \omega_k|B|] \\ &> \frac{1}{\omega_k}[(1-|1-\omega_k|)D - \omega_k(|B| + \varepsilon ee^T)] := \widehat{M}_k,\end{aligned}$$

where $\varepsilon > 0$ is an arbitrary number and $e = (1, 1, \ldots, 1)^T \in \mathbb{R}^n$.

Considering that $\rho(D^{-1}|B|) < 1$, by the continuity of the spectral radius we know that it has $\rho(D^{-1}(|B| + \varepsilon ee^T)) < 1$ for a sufficiently small ε. Hence, in accordance with the Perron–Frobenius theorem (see [20]), there exists a positive vector $u \in \mathbb{R}^n$ such that

$$D^{-1}(|B| + \varepsilon ee^T)u = \rho u, \qquad \rho := \rho(D^{-1}(|B| + \varepsilon ee^T)).$$

On the other hand, since (4.2) implies that the relaxation parameters ω_k, $k = 1, 2, \ldots, K$, satisfy

$$|1-\omega_k| + \omega_k\rho(D^{-1}|B|) < 1, \qquad k = 1, 2, \ldots, K,$$

in accordance with the continuity of the spectral radius we can again let ε be small enough such that

$$|1-\omega_k| + \omega_k\rho < 1, \qquad k = 1, 2, \ldots, K.$$

Now, by direct calculations we have

$$\widehat{M}_k u = \left(\frac{1-|1-\omega_k|}{\omega_k} - \rho\right) Du > 0, \qquad k = 1, 2, \ldots, K,$$

provided the relaxation parameters γ_k and $\omega_k (k = 1, 2, \ldots, K)$ are within the regions defined by (4.2). Moreover, since

$$\begin{aligned}\langle B_{k,p}\rangle u &= \frac{1}{\omega_k}(D - \gamma_k|L_{k,p}|)u \leq \frac{1}{\omega_k}Du \\ &= \frac{1}{1-(|1-\omega_k| + \omega_k\rho)}\widehat{M}_k u,\end{aligned}$$

by letting $\theta = \max_{1\leq k\leq K}\{|1-\omega_k| + \omega_k\rho\}$, we see that $\theta \in [0, 1)$ and

$$\langle B_{k,p}\rangle u \leq \frac{1}{1-\theta}\widehat{M}_k u.$$

By making use of Theorem 3.2, we immediately reach the conclusion of this theorem. □

The condition $\rho(D^{-1}|B|) < 1$ in Theorem 4.1 is a natural alternative to condition (3.1) for the multisplitting AOR Method 4.1. One concrete choice of the multiple splittings that can ensure the existence of the splitting matrices satisfying (4.1) and $\rho(D^{-1}|B|) < 1$ is given in the following example.

EXAMPLE 4.1. *Let $M = (m_{ij}) \in \mathbb{R}^{n\times n}$ be an H_+-matrix and $D = \mathrm{diag}(M)$, $B = M - D$. For each $p \in N_0$ and $k \in \{1, 2, \ldots, K\}$, take $L_{k,p} = (l_{ij}^{(k,p)}) \in \mathbb{R}^{n\times n}$ and*

$W_{k,p} = (w_{ij}^{(k,p)}) \in \mathbb{R}^{n\times n}$ *in accordance with the following rule:*

$$l_{ij}^{(k,p)} = \begin{cases} l_{ij}^{(k,p)} \in [0, m_{ij}] & \text{for } i > j \text{ and } m_{ij} \geq 0, \\ l_{ij}^{(k,p)} \in [m_{ij}, 0] & \text{for } i > j \text{ and } m_{ij} < 0, \\ 0 & \text{otherwise,} \end{cases} \qquad i, j = 1, 2, \ldots, n,$$

$$w_{ij}^{(k,p)} = \begin{cases} 0 & \text{for } i = j, \\ m_{ij} - l_{ij}^{(k,p)} & \text{for } i \neq j, \end{cases} \qquad i, j = 1, 2, \ldots, n.$$

Then (4.1) *and* $\rho(D^{-1}|B|) < 1$ *are satisfied.*

Theorem 4.1 directly results in the following convergence results for the multisplitting SOR, Gauss–Seidel, and Jacobi methods.

THEOREM 4.2. *Let the conditions of Theorem* 4.1 *be satisfied. Then*

(a) *the sequence* $\{z^p\}_{p\in N_0}$ *generated by the multisplitting SOR method converges to the unique solution of the LCP(M,q), provided the relaxation parameters* $\omega_k (k = 1, 2, \ldots, K)$ *satisfy*

$$0 < \omega_k < \frac{2}{1 + \rho(D^{-1}|B|)}, \qquad k = 1, 2, \ldots, K;$$

(b) *the sequence* $\{z^p\}_{p\in N_0}$ *generated by the multisplitting Gauss–Seidel method converges to the unique solution of the LCP(M,q);*

(c) *the sequence* $\{z^p\}_{p\in N_0}$ *generated by the multisplitting Jacobi method converges to the unique solution of the LCP(M,q).*

5. Conclusions. A new variant of the multisplitting methods in Machida, Fukushima, and Ibaraki [14] is presented for the linear complementarity problems. This new method inherits the good parallel computational properties of its predecessor and, in particular, includes a class of practical multisplitting relaxation methods. The convergence of these new methods is generally and systematically studied for the nonsymmetric linear complementarity problems. In addition, sufficient conditions and concrete examples about the multiple splittings that satisfy the general convergence criteria are investigated. Hence, the convergence theorems of the multisplitting methods in [14] for the symmetric linear complementarity problems are developed for the nonsymmetric linear complementarity problems, which completes the convergence theory of this class of parallel multisplitting iterative methods.

Acknowledgments. The author is very much indebted to Professor R. W. Cottle for his warm encouragement and valuable suggestions. He also wants to thank the referees for their comprehensive comments and suggestions, which greatly improved the original manuscript. Part of this work was fulfilled during the author's visit to Oxford University Computing Laboratory and Atlas Centre, Rutherford Appleton Laboratory in England. Therefore, he sincerely thanks his advisors, Dr. Andrew J. Wathen and Professor Iain S. Duff, for their constant direction and warm encouragement.

REFERENCES

[1] Z.-Z. BAI AND D.J. EVANS, *Matrix multisplitting relaxation methods for linear complementarity problems*, Internat. J. Comput. Math., 63 (1997), pp. 309–326.

[2] A. BERMAN AND R.J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[3] R.W. Cottle and G.B. Dantzig, *Complementarity pivot theory of mathematical programming*, Linear Algebra Appl., 1 (1968), pp. 103–125.
[4] R.W. Cottle, G.H. Golub, and R.S. Sacher, *On the solution of large, structured linear complementarity problems: The block partitioned case*, Appl. Math. Optim., 4 (1978), pp. 347–363.
[5] R.W. Cottle, J.S. Pang, and R.E. Stone, *The Linear Complementarity Problem*, Academic Press, San Diego, 1992.
[6] R.W. Cottle and R.S. Sacher, *On the solution of large, structured linear complementarity problems: The tridiagonal case*, Appl. Math. Optim., 4 (1977), pp. 321–340.
[7] C.W. Cryer, *The solution of a quadratic programming problem using systematic overrelaxation*, SIAM J. Control, 9 (1971), pp. 385–392.
[8] K. Fan, *Topological proofs for certain theorems on matrices with non-negative elements*, Monatsh. Math., 62 (1958), pp. 219–237.
[9] A. Frommer and G. Mayer, *Convergence of relaxed parallel multisplitting methods*, Linear Algebra Appl., 119 (1989), pp. 141–152.
[10] A.N. Iusem, *On the convergence of iterative methods for symmetric linear complementarity problems*, Math. Programming, 59 (1993), pp. 33–48.
[11] A.N. Iusem, *On the convergence of iterative methods for nonsymmetric linear complementarity problems*, Math. Appl. Comput., 10 (1991), pp. 27–41.
[12] C.E. Lemke, *Bimatrix equilibrium points and mathematical programming*, Management Sci., 11 (1965), pp. 681–689.
[13] Z.-Q. Luo and P. Tseng, *On the convergence of a matrix splitting algorithm for the symmetric monotone linear complementarity problem*, SIAM J. Control Optim., 29 (1991), pp. 1037–1060.
[14] N. Machida, M. Fukushima, and T. Ibaraki, *A multisplitting method for symmetric linear complementarity problems*, J. Comput. Appl. Math., 62 (1995), pp. 217–227.
[15] O.L. Mangasarian, *Solution of symmetric linear complementarity problems by iterative methods*, J. Optim. Theory Appl., 22 (1977), pp. 465–485.
[16] M. Neumann and R.J. Plemmons, *Convergence of parallel multisplitting iterative methods for M-matrices*, Linear Algebra Appl., 88/89 (1987), pp. 559–573.
[17] D.P. O'Leary and R.E. White, *Multisplittings of matrices and parallel solution of linear systems*, SIAM J. Alg. Discrete Methods, 6 (1985), pp. 630–640.
[18] J.M. Ortega and W.C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
[19] J.S. Pang, *Necessary and sufficient conditions for the convergence of iterative methods for the linear complementarity problem*, J. Optim. Theory Appl., 42 (1984), pp. 1–17.
[20] R.S. Varga, *Matrix Iterative Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1962.
[21] D.R. Wang, *On the convergence of the parallel multisplitting AOR algorithm*, Linear Algebra Appl., 154/156 (1991), pp. 473–486.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 1, pp. 79–92
© 1999 Society for Industrial and Applied Mathematics

SPECTRAL RADII OF FIXED FROBENIUS NORM PERTURBATIONS OF NONNEGATIVE MATRICES*

LIXING HAN†, MICHAEL NEUMANN†, AND MICHAEL TSATSOMEROS‡

Abstract. Let A be an $n \times n$ nonnegative matrix. In this paper we consider the problems of maximizing the spectral radii of (i) $A + X$ and (ii) $A + D$, where X is a real $n \times n$ matrix whose Frobenius norm is restricted to be 1 and where D is as X but is further constrained to be a diagonal matrix. For both problems the maximums occur at nonnegative X and D, and we use tools of nonnegative matrices, most notably due to Levinger and Fiedler, as well as the Kuhn–Tucker criterion for constrained optimization, to find upper and lower bounds on the maximums, and, when A is additionally assumed to be irreducible, to characterize cases of equalities in these bounds. In the case of the first problem, when A is irreducible, we characterize a matrix which gives the global maximum. A matrix which yields a global maximum to the second problem is more complicated to characterize as, depending on A, the problem admits local maximums within the nonnegative diagonal matrices of Frobenius norm 1.

Key words. nonnegative matrices, Perron eigenvalue, perturbation, maximization

AMS subject classifications. 15A48, 15A18

PII. S0895479897318241

1. Introduction.

Let A be an $n \times n$ nonnegative matrix. In this paper we consider optimization questions about the spectral radius of $A + X$, where X is a matrix whose *Frobenius norm* is constrained to be 1. After examining the case when X is any such matrix, we turn to the more restricted case of when X is diagonal. To be specific, we shall be concerned here with the following problems.

Maximization Problem I. Let A be an $n \times n$ nonnegative matrix. Then determine

$$\max_{\substack{X \in \mathbb{R}^{n \times n} \\ \|X\|_F = 1}} \rho(A + X). \tag{1.1}$$

Maximization Problem II. Let A be an $n \times n$ nonnegative matrix. Then determine

$$\max_{\substack{D \in \mathbb{R}^{n \times n}, \|D\|_F = 1 \\ D \text{ diagonal}}} \rho(A + D). \tag{1.2}$$

As the sets over which the maximizations are desired in both problems are compact and the spectral radius is a continuous function in the matrix entries, we are guaranteed that both problems have a solution within the restricted sets. Moreover, since $\|X\|_F = \||X|\|_F$ for any $X \in \mathbb{R}^{n,n}$, where $|X|$ is the nonnegative matrix whose entries are the absolute values of the corresponding entries in X, we have from the Perron–Frobenius theory that

$$\rho(A + X) \leq \rho(A + |X|).$$

Thus the solutions to the above maximization problems must occur at a nonnegative element whose Frobenius norm is 1.

*Received by the editors March 12, 1997; accepted for publication (in revised form) by M. Eiermann July 2, 1998; published electronically August 3, 1999.
http://www.siam.org/journals/simax/21-1/31824.html

†Department of Mathematics, University of Connecticut, Storrs, CT 06269-3009 (lih95001@uconnvm.uconn.edu, neumann@math.uconn.edu).

‡Department of Mathematics and Statistics, University of Regina, Regina, SK S4S 0A2, Canada (tsat@plato.math.uregina.ca). The work of this author was supported in part by NSERC.

In section 3 we analyze Maximization Problem I and prove a string of results that can be summarized as follows.

THEOREM 1.1. *Let A be an $n \times n$ nonnegative matrix and let*

$$\mu := \max_{\substack{X \in \mathbb{R}^{n\times n} \\ \|X\|_F = 1}} \rho(A + X). \tag{1.3}$$

Then

$$\rho(A) + 1 \leq \mu \leq \rho\left(\frac{A + A^T}{2}\right) + 1 \tag{1.4}$$

and μ is the unique value in the interval $(\rho(A), \infty)$ for which

$$\left\|(\mu I - A)^{-1}\right\|_2 = 1. \tag{1.5}$$

Moreover, if A is also irreducible, then μ is attained at a matrix W, $\|W\|_F = 1$, of rank 1. Furthermore, for the case when A is irreducible, equality holds in either of the inequalities in (1.4) *if and only if A has a common right and left Perron vector.*

In section 4 we examine the Maximization Problem II. We can summarize our observations there as follows.

THEOREM 1.2. *Let A be an $n \times n$ nonnegative matrix. Set*

$$\nu := \max_{\substack{D \in \mathbb{R}^{n\times n}, \|D\|_F = 1 \\ D \text{ diagonal}}} \rho(A + D). \tag{1.6}$$

Then

$$\rho(A) + \frac{1}{\sqrt{n}} \leq \nu \leq \rho(A) + 1. \tag{1.7}$$

Furthermore, suppose in addition that $n \geq 2$ and that A is irreducible and let $x = (x_1 \dots x_n)^T$ and $y = (y_1 \dots y_n)^T$ be, respectively, right and left Perron vectors of A normalized so that $y^T x = 1$. Then

$$\rho(A) + \frac{1}{\sqrt{n}} \leq \rho(A) + \sum_{i=1}^{n} (x_i y_i)^{3/2} \leq \nu < \rho(A) + 1. \tag{1.8}$$

Moreover, equality holds in the inequality

$$\rho(A) + \sum_{i=1}^{n} (x_i y_i)^{3/2} \leq \nu$$

if and only if

$$\rho(A) + \frac{1}{\sqrt{n}} = \rho(A) + \sum_{i=1}^{n} (x_i y_i)^{3/2} = \nu$$

and, under either equivalent conditions, we have that

$$x_i y_i = \frac{1}{n}, \quad i = 1, \dots, n. \tag{1.9}$$

Last, equality holds in the leftmost inequality in (1.8) *if and only if condition* (1.9) *holds.*

Our results in this paper are made possible with the help of three useful tools in analysis and in the theory of nonnegative matrices: the Kuhn–Tucker criterion for constrained optimization, an inequality for the spectral radius of an $n \times n$ nonnegative matrix due to Levinger, and a characterization for the spectral radius of an $n \times n$ nonnegative and irreducible matrix due to Fiedler. For convenience we quote these required results in section 2, where we also present other essential preliminaries.

We may view Theorems 1.1 and 1.2 as nonlocal in the sense that they tell us about the maxima of the spectral radii "far away" from A. In section 2 we also make the simple observation that the steepest ascent of the spectral radius at A occurs in the direction of the transpose of the eigenprojection onto the Perron eigenspace of A.

We comment that in section 3 it will be argued, using (1.5), that when A is irreducible, the Maximization Problem I has no local maxima in the nonnegative orthant other than, of course, the global maximum.[1] However, Maximization Problem II can have local maxima. We further mention that the results of Theorems 1.1 and 1.2 can be easily modified to solve the respective maximization problems subject to the perturbation matrices X or D having Frobenius norm $t > 0$ rather than 1. All that needs to be done is to replace that matrix A in (1.1) and (1.2) by the matrix A/t.

It should finally be remarked that in two recent papers, Johnson, Stanford, Olesky, and van den Driessche [7] and Johnson, Loewy, Olesky, and van den Driessche [8] consider optimization problems for the spectral radius of $A + D$, where A is an $n \times n$ nonnegative and irreducible matrix, but now D is a diagonal matrix whose *trace* is constrained in some fashion rather than its Frobenius norm.

2. Preliminaries. In this section we shall present some preliminary notation, some known results from the literature, as well as some observations which should aid in the understanding of the paper.

The following basic tenets of the Perron–Frobenius theory are taken from the book of Berman and Plemmons [1]. Let B be an $n \times n$ nonnegative irreducible matrix. The latter condition is equivalent to the existence of *no* permutation matrix P for which

$$PBP^T = \begin{pmatrix} \tilde{B}_{1,1} & \tilde{B}_{1,2} \\ 0 & \tilde{B}_{2,2} \end{pmatrix},$$

where both diagonal blocks are square and nonvacuous. The spectral radius of B, denoted by $\rho(B)$, is a simple eigenvalue, frequently called the *Perron root of* B, and it possesses right and left Perron eigenvectors having only positive entries. We shall use "$\gg 0$" to denote an array, square, or rectangular, whose entries are all positive and we shall use "$\circ$" to denote the entrywise product of two arrays whose dimensions coincide.

As a *function* of any of the matrix entries, the Perron root is strictly increasing at B. Moreover, as it is a *simple* eigenvalue we know, for instance from Wilkinson [11], that $\rho(\cdot)$ is an analytic function in each of the entries in an $\mathbb{R}^{n,n}$-neighborhood of B and hence the partial derivative of $\rho(\cdot)$ with respect to each of the entries exists at B.

Let x and y be positive right and left Perron vectors of B, respectively. Then

[1] Note the different sense in which the notion of "local" is used now compared with the sense in which it was used in the previous paragraph.

according to, e.g., Stewart [10, Ex. 1, p. 305], we have that

$$\frac{\partial \rho(B)}{\partial_{i,j}} = \frac{y_i x_j}{x^T y}, \quad i,j = 1, \ldots, n. \tag{2.1}$$

Thus, since the eigenprojection onto the Perron eigenspace is given by

$$E := \frac{xy^T}{y^T x},$$

it follows at once, relying on the said differentiability of the Perron root at B, that

$$\rho'(B) = \nabla \rho(B) = E^T = \frac{yx^T}{y^T x}. \tag{2.2}$$

With an expression for the gradient of the Perron root at our disposal, we can compute its directional derivatives at B. For that purpose let us recall that the inner product on $\mathbb{R}^{n,n}$ is given by

$$P \cdot Q = \operatorname{trace}(P^T Q)$$

which induces the matrix norm

$$\|P\|_F = \sqrt{\operatorname{trace}(P^T P)}.$$

Now let $X \in \mathbb{R}^{n,n}$ with $\|X\|_F = 1$. Then the directional derivative in the direction of X at B is given by

$$\rho'_X(B) = \lim_{t \to 0} \frac{\rho(B + tX) - \rho(B)}{t} = \nabla \rho(B) \cdot X = E^T \cdot X = \operatorname{trace}(EX). \tag{2.3}$$

Multivariate calculus tells us now that the steepest rate of increase at B is in the direction of the gradient and hence it is given by

$$\rho'_{E^T/\|E\|_F}(B) = \lim_{t \to 0} \frac{\rho\left(B + t\dfrac{E^T}{\|E\|_F}\right) - \rho(B)}{t} = E^T \cdot \frac{E^T}{\|E\|_F} = \|E\|_F. \tag{2.4}$$

By the Cauchy–Schwarz inequality we know that

$$\rho'_X(B) < \rho'_{E^T/\|E\|_F}(B) \tag{2.5}$$

unless for $X \in R^{n,n}$ with $\|X\|_F = 1$, $X = kE^T$ for some constant k. Thus we can conclude the following about the local behavior of the Perron root in the vicinity of B.

THEOREM 2.1. *Let B be an $n \times n$ nonnegative irreducible matrix and let E be the eigenprojection onto the eigenspace of $\rho(B)$. Then for any $X \in \mathbb{R}^{n \times n}$ with $\|X\|_F = 1$ and $X \neq kE^T$, k a constant, there exists $t_X > 0$ such that for all $0 < t \leq t_X$,*

$$\rho\left(B + t\frac{E^T}{\|E\|_F}\right) > \rho(B + tX). \tag{2.6}$$

Proof. Let X be as prescribed. The conclusion follows at once from (2.3), (2.5), and (2.4). ∎

If we regard for a moment the Perron root as a function of the diagonal entries only, then from (2.1) we see that

$$\rho'(B) = \nabla\rho(B) = \frac{1}{y^T x}\begin{pmatrix} x_1 y_1 \\ \vdots \\ x_n y_n \end{pmatrix} := E_D. \tag{2.7}$$

Moreover, we see that the rate of steepest ascent of the spectral radius as a function of the diagonal entries at B is given by

$$\rho'_{\mathrm{diag}(E_D)/\|\mathrm{diag}(E_D)\|_F}(B) = \frac{\|x \circ y\|_2}{y^T x}.$$

Now using similar argumentation which led us to Theorem 2.1, we have the following local behavior of the Perron root with regard to diagonal perturbations.

THEOREM 2.2. *Let B be an $n \times n$ nonnegative irreducible matrix. Then for any diagonal matrix $\Delta \in \mathbb{R}^{n\times n}$ with $\|\Delta\|_F = 1$ and $\Delta \neq kE_D^T = kE_D$, k a constant, there exists $t_\Delta > 0$ such that for all $0 < t \leq t_\Delta$,*

$$\rho\left(B + t\frac{\mathrm{diag}\,(E_D)}{\|\mathrm{diag}\,(E_D)\,\|_F}\right) \;>\; \rho(B + t\Delta). \tag{2.8}$$

To obtain our goal of proving Theorems 1.1 and 1.2, we shall rely on the following results on the spectral radius of $n \times n$ nonnegative matrices.

THEOREM LEVINGER (see [9]). *Let B be an $n \times n$ nonnegative matrix. Then*

$$\rho(B) \leq \rho\left(\frac{B + B^T}{2}\right) \tag{2.9}$$

with equality holding in the inequality if and only if B has common right and left Perron vectors.

THEOREM FIEDLER (see [4, Theorem]). *Let B be an irreducible $n \times n$ nonnegative matrix with x and y positive right and left Perron vectors, respectively, viz., $Bx = \rho(B)x$ and $B^T y = \rho(B)y$. Then*

(a)

$$\rho(B) = \max_{z\gg 0}\ \min_{\substack{u,v\gg 0\\ u\circ v = z}} \frac{v^T B u}{v^T u}. \tag{2.10}$$

(b) *$v^T Bu \geq y^T Bx = \rho(B)y^T x$ whenever $u, v \gg 0$ and $u \circ v = x \circ y$, and equality holds if and only if $u = kx$ for some positive scalar k.*

(c) *In particular, $x^T By \geq y^T Bx = \rho(B)y^T x$, with equality if and only if B has the same right and left Perron vectors.*

We mention that while no proof of Theorem Levinger is given in [9], a variety of proofs of his result can be found in the literature. Actually, Theorem Fiedler, whose statement and proof are given in [4, Theorem] comes from Fiedler [2]. For a more recent treatment of both the Levinger and Fiedler theorems see [3].

Finally, as both Maximization Problems I and II are of *constrained optimization problems*, we shall make use of the Kuhn–Tucker constrained optimization criterion subject to one equality constraint as in the following theorem.

THEOREM KUHN–TUCKER (see [6, p. 81]). *Consider the equality constrained maximization problem*

$$\begin{cases} \max_{t\in\mathbb{R}^k} F(t), \\ \\ \textit{subject to } c(t) = 0. \end{cases} \tag{2.11}$$

Suppose the functions F and c are differentiable in some $\mathbb{R}^k$-neighborhood of the optimal solution t_*. Then

$$\nabla F(t_*) = \lambda \nabla c(t_*) \tag{2.12}$$

for some real scalar λ.

3. Maximation Problem I. The goal of this section is to provide all claims and proofs that are necessary to establish Theorem 1.1.

To get an initial lower estimate on the value of μ in (1.3) is trivial. All we need to do is make the following observation.

Observation 3.1. *Let A be an $n \times n$ nonnegative matrix and let x be a right Perron vector of A with $\|x\|_2 = 1$. Then*

$$\rho(A + xx^T) = \rho(A) + 1. \tag{3.1}$$

Proof. Observe that $\|xx^T\|_F = 1$ and that

$$(A + xx^T)x = [\rho(A) + 1]x. \qquad \square$$

The above observation implies at once that

$$\max_{\substack{X \in \mathbb{R}^{n \times n} \\ \|X\|_F = 1}} \rho(A + X) \geq \rho(A) + 1. \tag{3.2}$$

For the symmetric case we can obtain with the aid of (3.2) that the following corollary holds.

COROLLARY 3.2. *Let A be an $n \times n$ symmetric nonnegative matrix. Then*

$$\max_{\substack{X \in \mathbb{R}^{n \times n} \\ \|X\|_F = 1}} \rho(A + X) = \rho(A) + 1. \tag{3.3}$$

Proof. On the one hand, we know by (3.2) that

$$\max_{\substack{X \in \mathbb{R}^{n \times n} \\ \|X\|_F = 1}} \rho(A + X) \geq \rho(A) + 1.$$

On the other hand, we can write that for any nonnegative $X \in \mathbb{R}^{n,n}$ with Frobenius norm 1,

$$\begin{aligned} \rho(A + X) &\leq \|A + X\|_2 \leq \|A\|_2 + \|X\|_2 \\ &\leq \rho\left(\frac{A + A^T}{2}\right) + \|X\|_F = \rho\left(\frac{A + A^T}{2}\right) + 1 \\ &= \rho(A) + 1. \end{aligned} \tag{3.4}$$

This completes the proof. $\square$

Now if A is *not necessarily symmetric*, then we can use Theorem Levinger (see section 2) to obtain the following upper estimate:

$$\rho(A + X) \leq \rho\left(\frac{A + A^T}{2} + \frac{X + X^T}{2}\right).$$

However,

$$\left\| \frac{X + X^T}{2} \right\|_2 \leq 1$$

so that we can certainly increase elements of $(X + X^T)/2$ if necessary to arrive at a matrix $\tilde{X}$ with $\|\tilde{X}\|_F = 1$, in which case we see from (3.4) that

$$\begin{aligned}\rho(A+X) &\leq \rho\left(\frac{A+A^T}{2}+\frac{X+X^T}{2}\right)\\ &\leq \rho\left(\frac{A+A^T}{2}+\tilde{X}\right) \leq \rho\left(\frac{A+A^T}{2}\right)+1.\end{aligned}$$

We have thus shown that Proposition 3.3 holds.

PROPOSITION 3.3. *Let A be an $n \times n$ nonnegative matrix. Then*

$$\max_{X\in\mathbb{R}^{n\times n},\|X\|_F=1} \rho(A+X) \leq \rho\left(\frac{A+A^T}{2}\right)+1. \tag{3.5}$$

We turn now to the problem of showing that the value of μ given in (1.3) is given by the unique value of μ which satisfies (1.5). To this end, we begin by assuming that A is also irreducible and we characterize a nonnegative matrix W with $\|W\|_F = 1$ which solves (1.1). We can use the remarks preceding Theorem 1.1 to reason as follows: As there exists a matrix W that furnishes a global solution and that is nonnegative and as A is irreducible and nonnegative, at such a W, $A + W$ is nonnegative and irreducible, thus securing the differentiability of the Perron root at $A + W$ in each of the matrix entries. This means that we can use the Kuhn–Tucker necessary conditions for the solution of the constrained optimization problem (see Theorem Kuhn–Tucker in section 2) to make an observation essential to this section.

LEMMA 3.4. *Let A be an $n \times n$ nonnegative and irreducible matrix and let W be a nonnegative global maximizer for* (1.1). *Then $W \gg 0$ and* $\operatorname{rank}(W) = 1$.

Proof. On the one hand, we know by (2.2) that the gradient of $\rho(\cdot)$ at $A + W$ is the eigenprojection onto the Perron eigenspace of $(A + W)^T$ and hence is a positive matrix of rank 1. On the other hand, the gradient of $c(X) = \|X\|_F^2 - 1$ at W is $2W$ from which our conclusion simply follows. ∎

Lemma 3.4 leads us to the following characterization for the maximal value attained in (1.1).

THEOREM 3.5. *Let A be an $n \times n$ nonnegative matrix. If*

$$\mu = \max_{\substack{X\in\mathbb{R}^{n,n}\\ \|X\|_F=1}} \rho(A+X), \tag{3.6}$$

then

$$\left\|(\mu I - A)^{-1}\right\|_2 = 1. \tag{3.7}$$

Proof. Let us begin by assuming that A is (also) irreducible and suppose that W is a maximizing nonnegative solution to (1.1). Then, by Lemma 3.4, W is positive, of rank 1, and of Frobenius norm 1 and therefore can be taken without loss of generality to be $W = qp^T$, where $\|q\|_2 = \|p\|_2 = 1$. Moreover, it follows from the arguments in the proof of Lemma 3.4 that W is a multiple of the eigenprojection of $(A+W)^T$ onto its Perron eigenspace and so

$$(A+W)^T W = \mu W \tag{3.8}$$

so that

$$(A^T + pq^T)q = \mu q. \tag{3.9}$$

It is easily realized that W^T is the global optimal solution to (1.1) when $A \to A^T$. Thus we also have that

$$(A + qp^T)p = \mu p. \tag{3.10}$$

However, (3.9) and (3.10) together give that

$$\left(\mu I - A^T\right)^{-1} (\mu I - A)^{-1} q = q.$$

As both matrices in the product above are positive, 1 is the spectral radius of this product and hence $\|(\mu I - A)^{-1}\|_2 = 1$.

Suppose now that A is reducible and let J be the $n \times n$ matrix of all 1's. For $\epsilon > 0$, set $A_\epsilon = A + \epsilon J$, and let

$$\mu_\epsilon := \max_{X \in \mathbb{R}^{n\times n} \|X\|_F = 1} \rho(A_\epsilon + X) = \rho\left(A_\epsilon + X_\epsilon\right). \tag{3.11}$$

Then by the initial part of the proof,

$$\left\|(\mu_\epsilon I - A)^{-1}\right\|_2 = 1. \tag{3.12}$$

Since the set

$$\{X \in \mathbb{R}^{n,n} \mid X \geq 0,\ \|X\|_F = 1\} \tag{3.13}$$

is compact, there exists a monotonically decreasing sequence $\{\epsilon_i\}_{i=1}^{\infty}$ such that $X_{\epsilon_i} \to \tilde{X}$ with $\|\tilde{X}\|_F = 1$ as $\epsilon_i \to 0$. Note that $A_{\epsilon_i} \to A$ and so, by the continuity of the spectral radius,

$$\mu_{\epsilon_i} = \rho\left(A_{\epsilon_i} + X_{\epsilon_i}\right) \ \to \rho\left(A + \tilde{X}\right) = \tilde{\mu} \tag{3.14}$$

with

$$\left\|(\tilde{\mu} I - A)^{-1}\right\|_2 = 1. \tag{3.15}$$

Next, clearly for some $\hat{X} \in \mathbb{R}^{n,n}$ with $\hat{X} \geq 0$ and with $\|\hat{X}\|_F = 1$,

$$\tilde{\mu} \leq \max_{\substack{X \in \mathbb{R}^{n\times n} \\ \|X\|_F = 1}} \rho(A + X) = \rho\left(A + \hat{X}\right) = \mu.$$

On the other hand,

$$\mu_\epsilon \geq \rho(A_\epsilon + \hat{X})$$

and so

$$\tilde{\mu} \geq \lim_{i \to \infty} \rho\left(A_{\epsilon_i} + \hat{X}\right) = \mu.$$

Thus $\mu = \tilde{\mu}$ and $\left\|(\mu I - A)^{-1}\right\|_2 = 1$ and our proof is complete. ∎

COROLLARY 3.6. *Let A be an $n \times n$ nonnegative matrix. Then*

$$\mu := \max_{\substack{X \in \mathbb{R}^{n\times n} \\ \|X\|_F = 1}} \rho(A + X)$$

if and only if

$$\left\|(\mu I - A)^{-1}\right\|_2 = 1.$$

Since for $\nu > \rho(A)$,

$$\left\|(\nu I - A)^{-1}\right\|_2^2 = \rho\left\{(\nu I - A)^{-1}\left(\nu I - A^T\right)^{-1}\right\},$$

and due to Neumann's expansion, as a function of ν in the interval $(\rho(A), \infty)$, $\left\|\left(\nu I - A\right)^{-1}\right)\right\|_2$ is strictly decreasing and hence for only one value of ν, namely, $\nu = \mu$, the function assumes the value 1.

Remark 3.7. If A is an $n \times n$ irreducible matrix, then on the set of nonnegative X's in $\mathbb{R}^{n,n}$ with Frobenius norm 1, $\rho(A + X)$ has no local maxima in addition to its global maximum. This is because, once again, the use of the Kuhn–Tucker conditions as in Lemma 3.4 would give that a local maximum has to occur at a rank 1 element. An analysis such as in the proof of Theorem 3.5 would then imply that the value of the local maximum has to be equal to the only value of μ that can render $\|(\mu I - A)^{-1}\|_2 = 1$.

As an example for the results of Lemma 3.4 and Theorem 3.5, consider that matrix

$$A = \begin{bmatrix} 12 & 1 \\ 99 & 2 \end{bmatrix}.$$

Our computations yield that the maximizing matrix is given by the rank 1 matrix

$$W \approx \begin{bmatrix} 0.1963 & 0.9756 \\ 0.01946 & 0.09675 \end{bmatrix}$$

and

$$\rho(A + W) - \rho(A) \approx 22.0166 - 18.1355 = 3.8811.$$

We remark that the upper bound $\rho((A + A)^T/2) + 1$ on $\mu = \rho(A + W)$ given on the right-hand side of (1.4) comes to approximately 53.2494. Finally, if we attempt to verify (3.8) for this example, then we get that

$$(A + W)^T W - \rho(A + W)W \approx 10^{-3} \begin{bmatrix} -0.09520 & -0.4730 \\ 0.01910 & 0.09520 \end{bmatrix}.$$

We computed W above using a function routine in conjunction with the MATLAB Optimization Toolbox. In our function, the search was done only over the rank 1 nonnegative matrices of Frobenius norm 1. However, to confirm the results, as well as the results for other examples, we used a version of the SQP method, (see Fletcher [5]) devised by M. J. D. Powell, FRS. We implemented the algorithm on an IBM mainframe with single precision arithmetic. This usually gave nonsingular solutions W to our various examples. A quick check showed that the computed W's had a high condition number.

In our next results we shall characterize precisely when equalities hold in (3.2) and (3.5) for the case when A is irreducible.

THEOREM 3.8. *Let A be an $n \times n$ nonnegative and irreducible matrix. Then equality holds in* (3.2) *if and only if A has a common left and right eigenvector.*

Proof. Since A is irreducible, there exist unique n-vectors $x \gg 0$ and $y \gg 0$ with $\|x\|_2 = \|y\|_2 = 1$ such that $Ax = \rho(A)x$ and $A^T y = \rho(A)y$. Suppose that $\mu = \rho(A)+1$. Then

$$(\mu I - A)^{-1} x = x \tag{3.16}$$

and

$$\left(\mu I - A^T\right)^{-1} y = y. \tag{3.17}$$

Now from the proof of Theorem 3.5, there exists a unique n-vector $z \gg 0$ with $\|z\|_2 = 1$ such that

$$\left(\mu I - A^T\right)^{-1} (\mu I - A)^{-1} z = z. \tag{3.18}$$

However, from (3.17), (3.18), and the irreducibility of $\left(\mu I - A^T\right)^{-1}$ and $\left(\mu I - A^T\right)^{-1}$ $(\mu I - A)^{-1}$ we get that $y = z$ and $y = (\mu I - A)^{-1} z$, whence $y = (\mu I - A)^{-1} y$ so that, by (3.16), $x = y$ from which we conclude that A has a common left and right Perron vector.

Conversely, if $x = y$, then

$$\left((1 + \rho(A))I - A^T\right)^{-1} ((1 + \rho(A))I - A)^{-1} x = x.$$

Hence

$$\|(1 + \rho(A)I - A)^{-1}\|_2 = 1$$

which implies that $\mu = \rho(A) + 1$. □

THEOREM 3.9. *Let A be an $n \times n$ nonnegative and irreducible matrix. Then equality holds in* (3.5) *if and only if A has a common right and left Perron eigenvector.*

Proof. Let W be a nonnegative global maximizer for the maximization problem (1.1). Then by the Theorem Levinger inequality, inequality (3.5) as applied to the matrix $(A + A^T)/2$, and because $\|(W + W^T)/2\|_F \leq 1$, we can write that

$$\rho(A + W) \leq \rho\left(\frac{A + A^T}{2} + \frac{W + W^T}{2}\right) \leq \rho\left(\frac{A + A^T}{2}\right) + 1. \tag{3.19}$$

However, by the hypothesis of the theorem,

$$\rho(A + W) = \rho\left(\frac{A + A^T}{2} + \frac{W + W^T}{2}\right). \tag{3.20}$$

It now follows by the equality case in Levinger's inequality that $A + W$ has a common right and left Perron vector. Now according to our characterization of the maximizing element using the Kuhn–Tucker condition, $W = qp^T$, with $(A + W)p = \rho(A + W)p$ and $(A^T + W^T)q = \rho(A + W)q$, so that we must have that $p = q$. However, again using the fact that $W = qp^T$, it easily follows that p (or $q = p$) has to be concurrently a right and left Perron vector of A.

Conversely, if A has a common right and left Perron vector, then by Levinger's inequality, $\rho(A) = \rho((A + A^T)/2)$. That equality holds in (3.5) now follows from Theorem 3.8 and (3.5). □

4. Diagonal perturbations. In this section we consider the solution to Maximization Problem II. We begin by establishing the results which give the lower bounds in (1.8) in the irreducible case.

LEMMA 4.1. *Let A be an $n \times n$ nonnegative and irreducible matrix and suppose that $x = (x_1 \dots x_n)^T$ and $y = (y_1 \dots y_n)^T$ are, respectively, right and left Perron vectors of A normalized so that $y^T x = 1$. Then*

$$\rho(A) + \frac{1}{\sqrt{n}} \leq \rho(A) + \sum_{i=1}^{n} (x_i y_i)^{3/2} \leq \nu. \tag{4.1}$$

Moreover, for equality to hold in the rightmost inequality it is necessary and sufficient that equalities hold throughout (4.1)*, that is,* $\nu = \rho(A) + 1/\sqrt{n}$*, and, under either equivalent conditions,*

$$x_i y_i = \frac{1}{n}, \quad i = 1, \dots, n. \tag{4.2}$$

Last, equality holds in the leftmost inequality in (4.1) *if and only if* (4.2) *holds.*

Proof. We begin by showing the validity of the rightmost inequality in (4.1). For that purpose let $z = x \circ y$ and let $D = \operatorname{diag}\left(\sqrt{z_1}, \dots, \sqrt{z_n}\right)$ so that D is a diagonal matrix whose Frobenius norm is 1. Then for any two positive vectors $\xi = (\xi_1 \dots \xi_n)^T$ and $\eta = (\eta_1 \dots \eta_n)^T$ with $\xi_i \eta_i = z_i$, $i = 1, \dots, n$, we have, by Theorem Fiedler part (b) of section 2 as applied to the matrix A, that

$$\xi^T (A + D)\eta \geq \rho(A) + \sum_{i=1}^{n} \xi_i \left(\sqrt{z_i}\right) \eta_i = \rho(A) + \sum_{i=1}^{n} (x_i y_i)^{3/2}. \tag{4.3}$$

Next, the proof of Theorem Fiedler given in [4], as applied now to the matrix $A + D$, assures us that for the vector $z \gg 0$, there exist positive vectors $\tilde{\xi}$ and $\tilde{\eta}$ such that $\tilde{\xi} \circ \tilde{\eta} = z$ and such that

$$\tilde{\xi}^T (A + D)\tilde{\eta} = \min_{\substack{\xi \circ \eta = z \\ \xi \gg 0, \eta \gg 0}} \xi^T (A + D)\eta. \tag{4.4}$$

Thus, from the maximization that occurs in Theorem Fiedler part (a) when applied to the matrix $A + D$ and by (4.3) we can write that

$$\nu \geq \rho(A + D) \geq \tilde{\xi}^T (A + D)\tilde{\eta} \geq \rho(A) + \sum_{i=1}^{n} (x_i y_i)^{3/2}. \tag{4.5}$$

We have thus proved that the rightmost inequality in (4.1) holds.

Suppose now that equality holds in the rightmost inequality in (4.1) so that

$$\nu = \rho(A) + \sum_{i=1}^{n} (x_i y_i)^{3/2}.$$

Then, by (4.4) and (4.5), we have that

$$\nu = \tilde{\xi}^T (A + D)\tilde{\eta} = \rho(A) + \sum_{i=1}^{n} (x_i y_i)^{3/2}. \tag{4.6}$$

As $\tilde{\xi}^T A \tilde{\eta} \geq \rho(A)$, we must actually have that $\tilde{\xi}^T A \tilde{\eta} = \rho(A)$. However, by part (b) of Theorem Fiedler, now as applied to the matrix A, it follows that $\tilde{\eta} = kx$ for some positive scalar k and therefore also that $\tilde{\xi} = (1/k)y$. Without loss of generality we can assume that $k = 1$. Hence, by (4.6) and part (b) of Theorem Fiedler as applied again to the matrix $A + D$, it follows that

$$(A + D)\tilde{\eta} = \nu \tilde{\eta} \quad \text{and} \quad (A^T + D)\tilde{\xi} = \nu \tilde{\xi}.$$

This means that x and y are also right and left Perron eigenvectors of $A + D$, respectively, and so it holds that

$$Dx = (\nu - \rho(A))x \quad \text{and} \quad Dy = (\nu - \rho(A))y.$$

As the diagonal entries of D are $\sqrt{x_i y_i}$, $i = 1, \ldots, n$, the above relations easily yield that

$$x_i y_i = (\nu - \rho(A))^2, \quad i = 1, \ldots, n, \tag{4.7}$$

whence

$$\sum_{i=1}^{n} (x_i y_i)^{3/2} = n(\nu - \rho(A))^3 \tag{4.8}$$

which easily yields, using (4.6), that $\nu - \rho(A) = 1/\sqrt{n}$. This shows that equality must hold also in the leftmost inequality in (4.1). Observe too that (4.7) validates (4.2) under the present assumptions.

Let us now show the leftmost inequality in (4.1). Consider the problem of minimizing the function $\sum_{i=1}^{n} u_i^{3/2}$ subject to the conditions that $u_i > 0$, $i = 1, \ldots, n$, and $\sum_{i=1}^{n} u_i = 1$. The method of the Lagrange multipliers easily gives that we must have the vector equality $(u_1^{1/2} \ldots u_n^{1/2})^T = \lambda(1 \ldots 1)^T$ for some nonzero scalar λ so that $u_1 = \cdots = u_n$. As $\sum_{i=1}^{n} u_i = 1$, it is necessary that $u_i = 1/n$, $i = 1, \ldots, n$, and so the leftmost inequality in (4.1) holds.

That (4.2) is a necessary and sufficient condition for equality to hold in the leftmost equality of (4.1) is an immediate consequence of the arguments in the preceding paragraph.

Finally, if $\nu = \rho(A) + 1/\sqrt{n}$, then by (4.3) and the last paragraph but one, we have that

$$\nu = \rho(A) + \sum_{i=1}^{n} (x_i y_i)^{3/2}$$

which implies, again by using the preceding paragraphs, that $x_i y_i = 1/n$, $i = 1, \ldots, n$. □

We remark that the lower bound

$$\nu \geq \rho(A) + \frac{1}{\sqrt{n}}$$

holds also when A is reducible by obvious continuity arguments.

We next provide two examples to illustrate the behavior of some matrices in the presence of equalities in some or all of the inequalities of (4.1). We begin with the case when $\nu = \rho(A) + (1/\sqrt{n})$; i.e., equalities hold throughout (4.1). This can occur without A being either symmetric or doubly stochastic. To see this let

$$A_1 = \begin{bmatrix} 2.0 & 1.0 \\ 2.0 & 2.0 \end{bmatrix}.$$

On computing right and left eigenvectors of A_1, say $x = (0.6124\ 0.8660)^T$ and $y = (0.8165\ 0.5774)^T$, we see $x \circ y = (1/2\ 1/2)^T$. In our second example we show that equality in the leftmost inequality of (4.1) *does not* imply equality also in the rightmost

inequality of (4.1). Take

$$A_2 = \begin{bmatrix} 0 & 0 & 0 & 0.6124 & 0.2494 & 0.1381 \\ 0 & 0 & 0 & 0.2494 & 0.3685 & 0.3821 \\ 0 & 0 & 0 & 0.1381 & 0.3821 & 0.4797 \\ 0.6124 & 0.2494 & 0.1381 & 0 & 0 & 0 \\ 0.2494 & 0.3685 & 0.3821 & 0 & 0 & 0 \\ 0.1381 & 0.3821 & 0.4797 & 0 & 0 & 0 \end{bmatrix}. \tag{4.9}$$

Then A_2 is symmetric, doubly stochastic, and even has a constant diagonal, but $\nu > \rho(A_2) + 1/\sqrt{6}$ as can be seen by selecting

$$D = \operatorname{diag}(0.07869, 0.1312, 0.8979, 0.08612, 0.2206, 0.3382)$$

for which $\|D\|_F = 1$ and $\rho(A + D) - \rho(A) = 0.4245 > 1/\sqrt{6}$.

Obtaining the upper bound in Theorem 1.2 is simple.

LEMMA 4.2. *Let A be an $n \times n$, $n \geq 2$, nonnegative matrix. Then, in the notation of Theorem* 1.2, *$\nu \leq \rho(A) + 1$. If, in addition, A is irreducible, then strict inequality holds.*

Proof. Recall that

$$\nu := \max_{\substack{D \in \mathbb{R}^{n \times n}, \|D\|_F = 1 \\ D \text{ diagonal}}} \rho(A + D). \tag{4.10}$$

Without loss of generality, we can assume that $D \geq 0$. As $D \leq I$, we have from the Perron–Frobenius theory that

$$\rho(A + D) \leq \rho(A + I) = 1 + \rho(A).$$

As $n \geq 2$, we must have that $D \neq I$ and hence strict inequality when A is irreducible. ∎

Two comments are in order as follows concerning our findings in this section:

(i) Consider for a moment the doubly stochastic and symmetric matrix A in (4.9) for which we showed that $\nu \geq 1.4245$. By Theorem Kuhn–Tucker of section 2 and by the preceding lemma, we know that there exists a positive diagonal matrix W, $\|W\|_F = 1$, for which

$$(A + W)\, p = \nu p \quad \text{and} \quad \left(A^T + W\right) q = \nu q$$

for some positive vectors n-vectors p and q. On the other hand, we know that the vector $e = (1, \ldots, 1)^T$ is both a right and left Perron eigenvector of A. Moreover, e is, in fact, both a left and right Perron vector of the matrix $A + \operatorname{diag}(e \circ e) := A + E$. Thus, for the Perron root viewed as a function of the diagonal entries, we observe via (2.7) that

$$\rho'(A + E) = \nabla \rho(A + E) = k_1 (e \circ e) = k_1 e$$

for some positive constant k. Now as $2e$ is the gradient of the function $\|x\|_2^2 - 1$ at $x = e$, we see not only that W satisfies the necessary condition (2.11) of Theorem Kuhn–Tucker but also the matrix E. This suggests that unlike Maximization Problem I for which we showed that in the nonnegative orthant it has no local maxima other than the global maximum, Maximization Problem II can have local maxima.

(ii) In Lemma 4.2 we show that for $n \geq 2$, the solution to Maximization Problem II always satisfies that $\nu < \rho(A)+1$ when A is irreducible, but that when A is reducible, equality can hold. As can be expected, we can show that there are irreducible A's for which ν is arbitrarily close to $\rho(A)+1$. This is borne out by analyzing nonnegative diagonal perturbations of Frobenius norm 1 in the example

$$A(\epsilon) = \begin{pmatrix} 1 & \epsilon \\ \epsilon & \epsilon \end{pmatrix}, \quad \epsilon > 0,$$

and by letting $\epsilon \to 0+$.

Acknowledgments. The authors wish to thank the referee for many helpful suggestions that in some places have led to improvements, simplifications, and economizations in our proofs. The authors also wish to acknowledge interesting conversations and exchange of ideas with Professor Paul van Dooren and Mr. Vincent Vermaut of the Université Catholique de Louvain.

REFERENCES

[1] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, SIAM, Philadelphia, 1994.
[2] M. Fiedler, *An Alternative Proof of Levinger's Theorem*, manuscript, 1970.
[3] M. Fiedler, *Numerical range of matrices and Levinger's theorem*, Linear Algebra Appl., 220 (1995), pp. 171–180.
[4] M. Fiedler, C. R. Johnson, T. L. Markham, and M. Neumann, *A trace inequality for M-matrices and symmetrizability of a real matrix by a positive diagonal matrix*, Linear Algebra Appl., 71 (1985), pp. 81–94.
[5] R. Fletcher, *Practical Methods of Optimization*, 2nd ed., Wiley, New York, 1987.
[6] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*, Academic Press, New York, 1981.
[7] C. R. Johnson, D. Stanford, D. D. Olesky, and P. van den Driessche, *Dominant eigenvalues under trace-preserving diagonal perturbations*, Linear Algebra Appl., 212/213 (1994), pp. 415–435.
[8] C. R. Johnson, R. Loewy, D. D. Olesky, and P. van den Driessche, *Maximizing the spectral radius of fixed trace diagonal perturbations of nonnegative matrices*, Linear Algebra Appl., 241–243 (1996), pp. 635–654.
[9] B. W. Levinger, *An inequality for nonnegative matrices*, Notices Amer. Math. Soc., 17 (1970), p. 260.
[10] G. W. Stewart, *Introduction to Matrix Computations*, Academic Press, New York, 1973.
[11] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 1, pp. 93–105
© 1999 Society for Industrial and Applied Mathematics

POINTWISE AND UNIFORMLY CONVERGENT SETS OF MATRICES*

ADAM L. COHEN†, LEIBA RODMAN‡, AND DAVID P. STANFORD‡

Abstract. Pointwise and uniform convergence properties of sets of real matrices are investigated. In particular, topological properties of uniformly and pointwise convergent sets are studied, as well as uniformly and pointwise convergent sets of matrices with special features: entrywise nonnegative or triangular. A detailed description of pointwise and uniformly convergent sets of two 2×2 matrices is presented.

Key words. pointwise convergent, uniformly convergent, matrix sets

AMS subject classification. 15A99

PII. S0895479898338433

1. Introduction.

Let $\mathcal{A} = \{A_j\}_{j \in J}$ be an indexed set of matrices, where each A_j is an $n \times n$ real matrix. The set $\mathcal{A}$ is said to be *pointwise convergent* if, for each $x \in \mathbb{R}^n$, there is an index sequence $\{p(x)_i\}_{i=1}^{\infty}$ such that $\lim_{k\to\infty}((\prod_{i=k}^{1} A_{p(x)_i})x) = 0$. $\mathcal{A}$ is said to be *uniformly convergent* if the index sequence can be taken independent of x; that is, there is an index sequence $\{p_i\}_{i=1}^{\infty}$ such that, for each $x \in \mathbb{R}^n$, $\lim_{k\to\infty}((\prod_{i=k}^{1} A_{p_i})x) = 0$.

The notions of pointwise and uniform convergence of sets of matrices arise naturally in multimodal control problems; see [BCS], [S]. To describe such control problems, it is convenient to assume that the set $\mathcal{A} = \{A_1, \ldots, A_m\}$ is finite. Consider m discrete linear time invariant systems $x_{k+1} = A_j x_k$, $k = 1, 2, \ldots$. A *control function* is a map $\sigma : \{1, 2, \ldots\} \to \{1, 2, \ldots, m\}$. Given a control function σ, we define the control system by $x_{k+1} = A_{\sigma(k)} x_k$, $k = 1, 2, \ldots$. Now, the set $\mathcal{A}$ is pointwise convergent if and only if for any initial state x_1 the state can be steered to zero using a suitable control function. The set $\mathcal{A}$ is uniformly convergent if and only if there exists a control function that will steer the state to zero from any initial position.

Recently, pointwise and uniform convergence have been studied in [GR], where it was proved that both types of convergence are equivalent for polynomially bounded semigroups, and in [SU], where criteria for pointwise and uniform convergence of sets $\mathcal{A}$ have been presented. In the next section, we summarize some results from [SU].

In the present paper we investigate further pointwise and uniform convergence properties of matrix sets, with emphasis on topological properties (description of the closure) and special classes of matrices: entrywise nonnegative or triangular. In section 5, we present a detailed description of pointwise and uniformly convergent sets of two 2×2 matrices; the description is partial in the case of pointwise convergence.

It should be noted that, in the absence of special properties such as the matrices being triangular or having nonnegative entries, the pointwise convergence property is

*Received by the editors May 6, 1998; accepted for publication (in revised form) by R. Brualdi, September 1, 1998; published electronically August 16, 1999.

http://www.siam.org/journals/simax/21-1/33843.html

†Oregon State University, Corvallis, OR 97330. The research of this author was supported by NSF grant DMS-9300395 and carried out at the College of William and Mary, Williamburg, VA.

‡College of William and Mary, Department of Mathematics, P.O. Box 8795, Williamsburg, VA 23187-8795 (lxrodm@math.wm.edu, stanford@math.wm.edu). The research of the second author was partially supported by NSF grant DMS-9500924.

often much harder to assess than the uniform convergence. Our results bear this out; compare Theorems 4.3 and 4.6 and compare Theorems 5.1 and 5.2.

We remark that other types of convergence of matrix sets have been studied in the literature as well; see for example [BY], [DL1], [DL2], motivated by the problem of constructing wavelets with compact support and other applications, and [G1],[G2], motivated by the problem of stability of linear inclusions. In contrast with the definitions given above, the other types of convergence require convergence (or convergence to zero) of *every* infinite product.

Preliminary versions of many results of the present paper have been reported in [CRS].

2. Previous results. The results presented in this section are found in [SU].

Given a norm $\|\ \|$ on $\mathbb{R}^n$ the set $\mathcal{A} = \{A_j\}_{j\in J}$ is *contractive relative to* $\|\ \|$ provided that if $x \in \mathbb{R}^n$, $x \neq 0$, then there is an index j such that $\|A_j x\| < \|x\|$. $\mathcal{A}$ is *precontractive relative to* $\|\ \|$ provided that if $x \in \mathbb{R}^n$, $x \neq 0$, then there is an index sequence $\{p(x)_i\}_{i=1}^r$ such that $\| \left(\prod_{i=1}^r A_{p(x)_i}\right) x\| < \|x\|$. It follows from the compactness of the unit sphere in $\mathbb{R}^n$ that $\mathcal{A}$ is precontractive relative to $\|\ \|$ if and only if some finite set of finite products of the members of $\mathcal{A}$ is contractive relative to $\|\ \|$.

THEOREM 2.1. *The set $\mathcal{A}$ is pointwise convergent if and only if $\mathcal{A}$ is precontractive relative to some norm on $\mathbb{R}^n$.*

THEOREM 2.2. *The set $\mathcal{A}$ is uniformly convergent if and only if there is a finite product $P = \prod_{i=1}^k A_{p_i}$ of members of $\mathcal{A}$ with $\rho(P) < 1$, where ρ denotes spectral radius.*

Theorem 2.1 implies that precontractiveness is norm independent, and so we may omit reference to a particular norm in using that term. Clearly, in applying the definition of pointwise convergence or that of precontractiveness, it is sufficient to consider only vectors x of norm 1 (any norm).

While it is obvious that every uniformly convergent set is pointwise convergent, there are sets $\mathcal{A}$ that are pointwise convergent and not uniformly convergent. In fact, we will see examples in section 4 of sets of two 2×2 matrices that are pointwise, but not uniformly, convergent.

LEMMA 2.3. *Pointwise convergence and uniform convergence are invariant under simultaneous similarity of the matrices in $\mathcal{A}$ and under the replacement of any of the matrices in $\mathcal{A}$ with its negative.*

LEMMA 2.4. *Pointwise and uniform convergence are equivalent for sets $\mathcal{A}$ of diagonal matrices.*

Thus, we may omit the modifier "pointwise" or "uniform" for convergence of sets of diagonal matrices.

A real $m \times n$ matrix A is said to be *semipositive* provided there is an $x \in \mathbb{R}^n$, $x \geq 0$, such that $Ax > 0$. Here, and throughout the paper, inequalities involving vectors or matrices are to be interpreted as entrywise inequalities. See [JKS] for a discussion of semipositivity.

Given a finite set $\mathcal{D} = \{D_1, D_2, \ldots, D_m\}$ of $n \times n$ diagonal matrices $D_j = \operatorname{diag}(d_{1j}, d_{2j}, \ldots, d_{nj})$, the $n \times m$ matrix $C(\mathcal{D}) = [c_{ij}]$ is defined by

$$c_{ij} = \begin{cases} 1 & \text{if } d_{ik} = 0 \text{ for some } k, \\ -\ln(|d_{ij}|) & \textit{otherwise.} \end{cases}$$

THEOREM 2.5. *The finite set $\mathcal{D}$ of diagonal matrices is convergent if and only if the matrix $C(\mathcal{D})$ is semipositive.*

If $\mathcal{A}$ is a set of square matrices, we denote by $\mathcal{D}(\mathcal{A})$ the set of diagonal matrices obtained by setting all off-diagonal entries of the matrices in $\mathcal{A}$ equal to 0. The following is then a corollary to Theorem 2.2.

COROLLARY 2.6. *A set $\mathcal{A}$ of upper triangular matrices (or of lower triangular matrices) is uniformly convergent if and only if the set $\mathcal{D}(\mathcal{A})$ is convergent.*

It follows from the results presented in this section that a set $\mathcal{A}$ is pointwise (resp., uniformly) convergent if and only if some finite subset of $\mathcal{A}$ is pointwise (resp., uniformly) convergent. Therefore, in what follows we often will focus attention on finite sets $\mathcal{A}$.

3. Sets of nonnegative matrices. In this section we focus on *nonnegative* matrices, i.e., matrices with nonnegative elements.

THEOREM 3.1. *Let $\mathcal{A}$ be a set of nonnegative $n \times n$ matrices. Then $\mathcal{A}$ is pointwise convergent if and only if it is uniformly convergent. Moreover, if $\mathcal{A}$ is uniformly convergent, then $\mathcal{D}(\mathcal{A})$ is convergent as well.*

Proof. Let $\mathcal{A}$ be a pointwise convergent set of nonnegative matrices. Then $\mathcal{A}$ is precontractive relative to any norm on $\mathbb{R}^n$ and in particular is relative to $\|x\|_\infty = \max\{|x_i| : 1 \le i \le n\}$. Hence there is a product P of matrices in $\mathcal{A}$ satisfying $\|Pu\|_\infty < 1$, where u denotes the vector of 1's in $\mathbb{R}^n$. Since P is nonnegative, the entries in the vector Pu are the sums of absolute values of entries along the rows of P, so $\|P\|_\infty < 1$, where for any $n \times n$ matrix A, $\|A\|_\infty$ denotes the matrix row sum norm, $\|A\|_\infty = \max\{\sum_{j=1}^n |a_{ij}| : 1 \le i \le n\}$. Since every matrix norm dominates the spectral radius, we have $\rho(P) < 1$; thus $\mathcal{A}$ is uniformly convergent by Theorem 2.2.

Now suppose that $\mathcal{A}$ is uniformly convergent but $\mathcal{D}(\mathcal{A}) = \{D_i\}_{i\in J}$ is not convergent. Then there is a vector $x \in \mathbb{R}^n$, $x \neq 0$, such that, for each index sequence $\{p_i\}_{i=1}^r$, we have $\|(\prod_{i=1}^r D_{p_i})\, x\|_\infty \ge \|x\|_\infty$. We let $|x|$ denote the entrywise absolute value of x. Since the A_i's are nonnegative and $\prod_{i=1}^r D_{p_i}$ is diagonal, we may write

$$\left\| \left(\prod_{i=1}^r A_{p_i} \right) |x| \right\|_\infty \ge \left\| \left(\prod_{i=1}^r D_{p_i} \right) |x| \right\|_\infty = \left\| \left| \left(\prod_{i=1}^r D_{p_i} \right) x \right| \right\|_\infty$$
$$= \left\| \left(\prod_{i=1}^r D_{p_i} \right) x \right\|_\infty \ge \|x\|_\infty = \| \, |x| \, \|_\infty.$$

This contradicts the pointwise convergence of $\mathcal{A}$, and therefore the theorem is proved. ∎

We note that in the absence of triangularity, the pointwise convergence of $\mathcal{D}(\mathcal{A})$ does not imply that of $\mathcal{A}$. The set $\mathcal{A} = \{\left[\begin{smallmatrix} .5 & 10 \\ 10 & .5 \end{smallmatrix}\right]\}$ is not uniformly convergent but $\mathcal{D}(\mathcal{A})$ is.

For any matrix $A = [a_{ij}]$, we denote $[|a_{ij}|]$ by $|A|$, and for $\mathcal{A} = \{A_j\}_{j\in J}$, the set $\{|A_1|\}_{j\in J}$ is denoted by $|\mathcal{A}|$.

COROLLARY 3.2. *If $\mathcal{A}$ is a set of $n \times n$ matrices and $|\mathcal{A}|$ is pointwise convergent, then $\mathcal{A}$ is uniformly convergent.*

Proof. By Theorem 3.1, $|\mathcal{A}|$ is uniformly convergent, so $\rho\,(\prod_{i=1}^r |A_{p_i}|) < 1$ for some index sequence $\{p_i\}_{i=1}^r$. Now $|\prod_{i=1}^r A_{p_i}| \le \prod_{i=1}^r |A_{p_i}|$ and therefore, using monotonicity properties of the spectral radius of real matrices (see [HJ, Thm. 8.1.18]), we have $\rho\,(\prod_{i=1}^r A_{p_i}) \le \rho\,(|\prod_{i=1}^r A_{p_i}|) \le \rho\,(\prod_{i=1}^r |A_{p_i}|) < 1$. Hence $\mathcal{A}$ is uniformly convergent. ∎

The converse fails because if $\mathcal{A} = \{A\}$ with $A = \left[\begin{smallmatrix} -2/3 & 2/3 \\ 2/3 & 2/3 \end{smallmatrix}\right]$, then $\rho(A) = \frac{2\sqrt{2}}{3}$. Thus $\mathcal{A}$ is uniformly convergent but $\rho(|A|) = \frac{4}{3}$, so $|\mathcal{A}|$ is not uniformly convergent

and hence is not pointwise convergent.

The property of uniform convergence for sets of nonnegative matrices is inherited by sets of principal submatrices in the following sense. If A is an $n \times n$ matrix and $\alpha \subset \{1, 2, \ldots, n\}$, then $A[\alpha]$ denotes the principal submatrix of A lying in the rows and columns indexed by α. If $\mathcal{A} = \{A_j\}_{j \in J}$ is a set of $n \times n$ matrices, then $\mathcal{A}[\alpha]$ denotes $\{A_j[\alpha]\}_{j \in J}$.

THEOREM 3.3. *Let $\mathcal{A}$ be a uniformly convergent set of nonnegative $n \times n$ matrices and let $\alpha \subset \{1, 2, \ldots, n\}$. Then $\mathcal{A}[\alpha]$ is uniformly convergent.*

Proof. By simultaneous permutation similarity, it suffices to prove the theorem for the case $\alpha = \{1, 2, \ldots, t\}$ for $1 \leq t \leq n$. Suppose first that $n = 2$ so that we need consider only $\alpha = \{1\}$. By Theorem 3.1, $\mathcal{D}(\mathcal{A})$ is uniformly convergent. Hence there is a matrix in $\mathcal{D}(\mathcal{A})$ with a $1,1$ entry less than 1. The 1×1 matrix with only this entry is a member of $\mathcal{A}[\{1\}]$, and so $\mathcal{A}[\{1\}]$ is uniformly convergent and the theorem holds when $n = 2$. Now let $n > 2$ and suppose the theorem holds for sets of nonnegative $(n-1) \times (n-1)$ matrices. Let $\mathcal{A}$ be a uniformly convergent set of $n \times n$ nonnegative matrices and let $\alpha = \{1, 2, \ldots, t\}$ with $t \leq n$. If $t = n$, there is nothing to prove. If $t < n - 1$, the uniform convergence of $\mathcal{A}[\alpha]$ will follow from that of $\mathcal{A}[\{1, 2, \ldots, n-1\}]$ by the inductive assumption. Hence we may assume $t = n - 1$ so that $\alpha = \{1, 2, \ldots, n-1\}$.

If $i, j \in J$, we have

$$A_i A_j = \begin{bmatrix} A_i[\alpha] & X \\ Y^t & a \end{bmatrix} \begin{bmatrix} A_j[\alpha] & Z \\ W^t & b \end{bmatrix} = \begin{bmatrix} A_i[\alpha]A_j[\alpha] + XW^t & * \\ * & * \end{bmatrix}$$

and so, since $XW^t \geq 0$, $(A_i A_j)[\alpha] \geq A_i[\alpha]A_j[\alpha]$. By induction on the number of factors, it follows that if $P = \prod_{i=1}^{r} A_{p_i}$ is any product of matrices in $\mathcal{A}$, then $P[\alpha] \geq \prod_{i=1}^{r} A_{p_i}[\alpha]$. Since $\mathcal{A}$ is uniformly convergent, we may take P to satisfy $\rho(P) < 1$. Again using [HJ, Thm. 8.1.18], we have $\rho\left(\prod_{i=1}^{r} A_{p_i}[\alpha]\right) \leq \rho(P[\alpha]) \leq \rho(P) < 1$ and so $\mathcal{A}[\alpha]$ is uniformly convergent. ∎

We note that if $\mathcal{A}$ is as given after the proof of Theorem 3.1, then $\mathcal{A}[\alpha]$ is uniformly convergent for all proper α but $\mathcal{A}$ is not uniformly convergent.

4. Topological properties of the class of convergent matrix sets. In this section we identify points in $\mathbb{R}^{mn^2}$ with the set of ordered m-tuples of $n \times n$ real matrices $\mathcal{A} = \{A_1, A_2, \ldots, A_m\}$ and discuss topological properties of the collections of pointwise and uniformly convergent sets as subsets of $\mathbb{R}^{mn^2}$. While the convergence properties of $\mathcal{A}$ are invariant under reordering of the matrices in the set, it is convenient to think of the matrix sets as ordered sets. We denote by $\mathcal{U}_{m,n}$ (resp., $\mathcal{P}_{m,n}$) the collection of all uniformly convergent (resp., pointwise convergent) sets in $\mathbb{R}^{mn^2}$.

Suppose that for each positive integer k and for each p with $1 \leq p \leq m$, $A_p^{(k)} = [a_{p,ij}^{(k)}]$ is a real $n \times n$ matrix and $A_p = [a_{p,ij}]$ is a real $n \times n$ matrix. Let $\mathcal{A}^{(k)} = \{A_1^{(k)}, A_2^{(k)}, \ldots, A_m^{(k)}\}$ and $\mathcal{A} = \{A_1, A_2, \ldots, A_m\}$. Then the sequence $\{\mathcal{A}^{(k)}\}_{k=1}^{\infty}$ converges to the set $\mathcal{A}$ provided that, for each i and j with $1 \leq i, j \leq n$ and each p with $1 \leq p \leq m$, we have $\lim_{k \to \infty} a_{p,ij}^{(k)} = a_{p,ij}$.

If $\mathcal{A} = \{A_1, A_2, \ldots, A_m\} \in \mathbb{R}^{mn^2}$ and if $c \in \mathbb{R}^m$, then $\mathcal{A}_c$ will denote the set $\{c_1 A_1, c_2 A_2, \ldots, c_m A_m\}$. In case $c_i = \epsilon$ for $1 \leq i \leq m$, $\mathcal{A}_c$ may be denoted by $\epsilon\mathcal{A}$.

LEMMA 4.1. *Let $\mathcal{A} = \{A_1, A_2, \ldots, A_m\} \in \mathbb{R}^{mn^2}$ and let $|c_p| \leq 1$ for $1 \leq p \leq m$. If $\mathcal{A} \in \mathcal{U}_{m,n}$, then $\mathcal{A}_c \in \mathcal{U}_{m,n}$. If $\mathcal{A} \in \mathcal{P}_{m,n}$, then $\mathcal{A}_c \in \mathcal{P}_{m,n}$.*

The proof of Lemma 4.1 is immediate.

THEOREM 4.2. *The sets $\mathcal{U}_{mn}$ and $\mathcal{P}_{mn}$ are open and star shaped about zero (in particular, connected) in $\mathbb{R}^{mn^2}$.*

Proof. Let $\lim_{k\to\infty} \mathcal{A}^{(k)} = \mathcal{A}$.

First suppose that no $\mathcal{A}^{(k)}$ is uniformly convergent. Any product P of matrices in $\mathcal{A}$ is the limit of corresponding products $P^{(k)}$ of matrices in $\mathcal{A}^{(k)}$, and $\rho(P^{(k)}) \geq 1$. Then $\rho(P) = \lim_{k\to\infty} \rho(P^{(k)}) \geq 1$; thus, $\mathcal{A}$ is not uniformly convergent. Hence the complement of $\mathcal{U}_{m,n}$ is closed and thus $\mathcal{U}_{m,n}$ is open.

Now suppose that no $\mathcal{A}^{(k)}$ is pointwise convergent. For each k, we may choose a vector $x_k \in \mathbb{R}^n$ with $\|x_k\|_\infty = 1$ such that no product of the matrices in $\mathcal{A}^{(k)}$ contracts x_k relative to $\| \ \|_\infty$. By compactness, we may assume that $\lim_{k\to\infty} x_k = x$, where $\|x\|_\infty = 1$. Any product P of matrices in $\mathcal{A}$ is the limit of corresponding products $P^{(k)}$ of matrices in $\mathcal{A}^{(k)}$, and $\|P^{(k)}x\|_\infty \geq 1$. Then $\|Px\|_\infty = \lim_{k\to\infty} \|P^{(k)}x_k\|_\infty \geq 1$, so P does not contract x relative to $\| \ \|_\infty$, and $\mathcal{A}$ is not precontractive. Thus $\mathcal{A}$ is not pointwise convergent, so the complement of $\mathcal{P}_{m,n}$ is closed and $\mathcal{P}_{m,n}$ is open.

The star-shapedness property follows from Lemma 4.1. ∎

We now present results concerning the closures of $\mathcal{U}_{m,n}$ and $\mathcal{P}_{m,n}$. For $\mathcal{A} \in \mathbb{R}^{mn^2}$, we denote by $\Pi_\mathcal{A}$ the set of all finite products of members of $\mathcal{A}$. The quantity $\inf\{\rho(P) : P \in \Pi_\mathcal{A}\}$ will be denoted by $\rho_\mathcal{A}$. Note that if $\rho_\mathcal{A} < 1$, then a consideration of P^k, where $P \in \Pi_\mathcal{A}$ and $\rho(P) < 1$, shows that $\rho_\mathcal{A} = 0$.

THEOREM 4.3. *Let $\mathcal{A} \in \mathbb{R}^{mn^2}$ and suppose that either* (1) *or* (2) *holds:*

(1) $\rho_\mathcal{A} \in \{0, 1\}$*;*

(2) $\rho(P) > \rho_\mathcal{A}$ *for all* $P \in \Pi_\mathcal{A}$.

Then $\mathcal{A}$ is in the closure of $\mathcal{U}_{m,n}$.

Proof. Suppose (1) holds. The result is clear if $\rho_\mathcal{A} = 0$, so let $\rho_\mathcal{A} = 1$. Assume first that there is a $P = \prod_{i=1}^r A_{q(i)} \in \Pi_\mathcal{A}$ such that $\rho(P) = 1$. For $0 < \epsilon < 1$, consider $\epsilon\mathcal{A}$. Then $\rho\left(\prod_{i=1}^r \epsilon A_{q(i)}\right) = \epsilon^r \rho(P) < 1$, so $\epsilon\mathcal{A} \in \mathcal{U}_{m,n}$. Clearly, $\lim_{\epsilon\to 1} \epsilon\mathcal{A} = \mathcal{A}$ and so $\mathcal{A}$ is in the closure of $\mathcal{U}_{m,n}$.

Now suppose $\rho_\mathcal{A} \geq 1$ and $\rho(P) > \rho_\mathcal{A}$ for all $P \in \Pi_\mathcal{A}$. For each positive integer t, let $\Pi_\mathcal{A}^t = \{\prod_{i=1}^r A_{q(i)} : 1 \leq r \leq t, \ 1 \leq q(i) \leq m\}$. $\Pi_\mathcal{A}^t$ is finite, so there is an index sequence $\{s_t(i)\}_{i=1}^{r_t}$, where $r_t \leq t$, such that $P_t = \prod_{i=1}^{r_t} A_{s_t(i)}$ satisfies $\rho_\mathcal{A} < \rho(P_t) \leq \rho(P)$ for all $P \in \Pi_\mathcal{A}^t$. Let $\delta_t = (2\rho(P_t))^{-\frac{1}{r_t}}$. Now $\rho\left(\prod_{i=1}^{r_t} \delta_t A_{s_t(i)}\right) = \delta_t^{r_t}\rho(P_t) = \frac{1}{2} < 1$, so $\delta_t\mathcal{A} \in \mathcal{U}_{m,n}$. If $\{r_t\}_{t=1}^\infty$ were bounded, then $\rho_\mathcal{A} = \inf\{\rho(P) : P \in \Pi_\mathcal{A}\}$ would be attained by some some P_t. Hence $\lim_{t\to\infty} r_t = \infty$. Since $A_1 \in \Pi_\mathcal{A}^1 \subset \Pi_\mathcal{A}^t$, we have $1 \leq \rho_\mathcal{A} < \rho(P_t) \leq \rho(A_1)$ for all t, and it follows that $\lim_{t\to\infty} \delta_t = 1$. Hence $\lim_{t\to\infty} \delta_t\mathcal{A} = \mathcal{A}$, and $\mathcal{A}$ is in the closure of $\mathcal{U}_{m,n}$. ∎

The following example shows a situation in which $\rho_\mathcal{A} = 1$ but $\rho(P) > 1$ for every $P \in \Pi_\mathcal{A}$. Let

$$\mathcal{A} = \{A_1, A_2\}, \quad A_1 = \begin{bmatrix} a & 0 \\ 0 & a^{-1} \end{bmatrix}, \quad A_2 = \begin{bmatrix} b & 0 \\ 0 & b^{-1} \end{bmatrix},$$

where $a > 1$ and $b > 1$ are real numbers such that their logarithms $\ln(a)$, $\ln(b)$ are not commensurable. Then there exists a sequence (m_q, n_q), $q = 1, 2, \ldots$, of pairs of integers such that $\lim_{q\to\infty}(m_q\ln(a) + n_q\ln(b)) = 0$. We can assume that all integers m_q are of the same sign and all integers n_q are of the same sign, say, $m_q > 0$ and $n_q < 0$ for all q. Then, replacing A_2 by A_2^{-1}, we obtain a set $\mathcal{A}$ with the desired properties.

While we have not obtained the converse of Theorem 4.3 in general, we can show that if $\mathcal{A}$ is a set of upper triangular matrices (or of lower triangular ones) and if $\mathcal{A}$ is in the closure of $\mathcal{U}_{m,n}$, then (1) must hold.

THEOREM 4.4. *Suppose $\mathcal{A} = \{A_1, A_2, \ldots, A_m\}$ is a set of upper triangular matrices, and suppose $\mathcal{A}$ is in the closure of $\mathcal{U}_{m,n}$. Then $\rho_{\mathcal{A}} \leq 1$.*

We need a lemma for the proof of Theorem 4.4.

LEMMA 4.5. *Let A be a $p \times q$ real matrix such that $Ax \geq 0$ (componentwise) for some nonzero vector $x \in \mathbb{R}^q$ with nonnegative components. Then for every positive ϵ there exists a nonzero vector y with nonnegative integer components such that no component of Ay is less than $-\epsilon$.*

Proof. We use induction on q. When $p = q = 1$, then either $A = 0$ (in which case every positive integer y will do), or $A > 0$, $x > 0$. In the latter case, approximate x by a rational number $\frac{u}{v}$, where u and v are positive integers, so that

$$\left| x - \frac{u}{v} \right| < \frac{1}{v^2}.$$

Now

$$Au = A(u - vx) + Avx \geq A(u - vx) \geq \frac{-A}{v},$$

which can be made not smaller than $-\epsilon$ for v large enough; thus, one can take $y = u$. If $q = 1$ but $p > 1$, then apply the foregoing argument with A replaced by the minimal positive element of A.

Consider now the general case. If a component of x is equal to zero, then we can omit the corresponding column of A, thereby reducing q by one, and use the induction hypothesis. Thus we can assume that all components of x are positive. Write $x = [x_1, \ldots, x_q]^T > 0$. By a result in geometric number theory (Sect. V.10 of [C]), for every sufficiently large positive integer Q there exist positive integers $u_0, u_1, \ldots, u_q$ such that

$$|u_0 x_j - u_j| < \frac{1}{\sqrt[q]{Q}} \qquad \text{for} \quad j = 1, \ldots, q.$$

We claim that $y = [u_1, \ldots, u_q]^T$ will do (for sufficiently large Q). Indeed, letting a_{ij} be the (i, j) element of the matrix A, we obtain

$$\sum_{j=1}^{q} a_{ij} u_j = \sum_{j=1}^{q} a_{ij}(u_j - u_0 x_j) + u_0 \sum_{j=1}^{q} a_{ij} x_j \geq \sum_{j=1}^{q} a_{ij}(u_j - u_0 x_j)$$

$$\geq -q \max_{1 \leq i \leq p; 1 \leq j \leq q} |a_{ij}| \frac{1}{\sqrt[q]{Q}},$$

and this expression tends to zero as $Q \to \infty$. ∎

Proof of Theorem 4.4. Let $\{\mathcal{A}^{(k)}\}_{k=1}^{\infty}$ be a sequence of matrix sets in $\mathcal{U}_{m,n}$ such that $\lim_{k \to \infty} \mathcal{A}^{(k)} = \mathcal{A}$. By Corollary 2.6, $\mathcal{D}^{(k)} = \mathcal{D}(\mathcal{A}^{(k)}) \in \mathcal{U}_{m,n}$ for each k and clearly $\lim_{k \to \infty} \mathcal{D}^{(k)} = \mathcal{D}(\mathcal{A})$. Let $\mathcal{D} = \mathcal{D}(\mathcal{A}) = \{D_1, D_2, \ldots, D_m\}$ with $D_j = diag(d_{1j}, d_{2j}, \ldots, d_{nj})$ for $1 \leq j \leq m$. By Theorem 2.5, the $n \times m$ matrix $C(\mathcal{D}^{(k)})$ is semipositive, so there is a nonnegative vector $x_k \in \mathbb{R}^m$ such that $C(\mathcal{D}^{(k)})x_k > 0$. We may assume that $\|x_k\|_2 = 1$, where $\| \ \|_2$ denotes the Euclidean norm on $\mathbb{R}^m$, and by compactness we may further assume that $\lim_{k \to \infty} x_k = x \geq 0$. It follows that $\lim_{k \to \infty} C(\mathcal{D}^{(k)})x_k = C(\mathcal{D})x \geq 0$.

We now choose $\epsilon > 0$ and construct $P_\epsilon \in \Pi_{\mathcal{A}}$ such that $\rho_{\mathcal{A}} \leq \rho(P_\epsilon) < e^\epsilon$. It will follow that $\rho_{\mathcal{A}} \leq \lim_{\epsilon \to 0} e^\epsilon = 1$.

Since $C(\mathcal{D})x \geq 0$, by Lemma 4.4 there is a nonzero nonnegative vector $p^{(\epsilon)} \in \mathbb{R}^m$ with integer entries such that $C(\mathcal{D})p^{(\epsilon)} > -\epsilon u$, where u denotes the vector of 1's in $\mathbb{R}^m$. Thus $\sum_{j=1}^m (-\ln|d_{ij}|)p_j^{(\epsilon)} > -\epsilon$ for $1 \leq i \leq m$. Hence, for $1 \leq i \leq m$, we have $\ln(\prod_{j=1}^m |d_{ij}|^{p_j^{(\epsilon)}}) < \epsilon$, and taking exponentials we have $\prod_{j=1}^m |d_{ij}|^{p_j^{(\epsilon)}} < e^\epsilon$. Let $P_\epsilon = \prod_{j=1}^m A_j^{p_j^{(\epsilon)}}$. Then $P_\epsilon \in \Pi_{\mathcal{A}}$ and, because of triangularity, we have

$$\rho(P_\epsilon) = \rho\left(\prod_{j=1}^m D_j^{p_j^{(\epsilon)}}\right) = \max\left\{\prod_{j=1}^m |d_{ij}|^{p_j^{(\epsilon)}} : 1 \leq i \leq n\right\} < e^\epsilon. \qquad \square$$

It is instructive to compare $\rho_{\mathcal{A}}$ with another measure of joint spectral radii, namely, the lower joint spectral radius. Given an ordered m-tuple $\mathcal{A}$ of $n \times n$ real matrices, denote by $\pi_{\mathcal{A}}^k$ the set of all products of exactly k matrices from $\mathcal{A}$, $k = 1, 2, \ldots$. The *lower joint spectral radius* of $\mathcal{A}$ is defined by

$$\underline{\rho}_{\mathcal{A}} = \liminf_{k\to\infty}\Big(\min\{\rho(P)^{1/k} \ : \ P \in \pi_{\mathcal{A}}^k\}\Big).$$

This definition was introduced in [G1]; a discussion of this notion in the context of Lyapunov exponents, as well as an extensive bibliography, is found in [TB]. It was proved in [G1] that for any (submultiplicative) matrix norm we have

$$\underline{\rho}_{\mathcal{A}} = \liminf_{k\to\infty}\Big(\min\{\|P\|^{1/k} \ : \ P \in \pi_{\mathcal{A}}^k\}\Big).$$

(The finiteness of the set $\mathcal{A}$ is essential for the proof of this equality to hold.) It turns out that if $\underline{\rho}_{\mathcal{A}} \leq 1$, then $\mathcal{A}$ belongs to the closure of $\mathcal{U}_{mn}$. Indeed,

$$\underline{\rho}_{c\mathcal{A}} = c\underline{\rho}_{\mathcal{A}} \leq c$$

for any positive constant c; if $c < 1$, then some finite product of matrices from $c\mathcal{A}$ has spectral radius less than one, and therefore $c\mathcal{A} \in \mathcal{U}_{mn}$.

We also observe that condition (2) of Theorem 4.3 implies $\underline{\rho}_{\mathcal{A}} \leq 1$. To verify that, note that if this condition holds, then there is an infinite sequence of positive integers $k_1 < k_2 < \cdots$ such that $\rho(P_j) < \rho_{\mathcal{A}} + 1$ for some $P_j \in \pi_{\mathcal{A}}^{k_j}$, $j = 1, 2, \ldots$. Therefore,

$$\underline{\rho}_{\mathcal{A}} \leq \liminf_{j\to\infty}\Big(\rho(P_j)^{\frac{1}{k_j}}\Big) \leq \lim_{j\to\infty} \ (\rho_{\mathcal{A}} + 1)^{\frac{1}{k_j}} = 1.$$

A pointwise analogy to Theorem 4.3 seems difficult to obtain. The following theorem is a partial result with rather restrictive hypotheses. Its formulation requires the following definitions. Let $\|\ \|$ be any norm on $\mathbb{R}^n$ and let S^{n-1} denote the unit sphere $\{x \in \mathbb{R}^n : \|x\| = 1\}$. For $\mathcal{A} \in \mathbb{R}^{mn^2}$ and $t \in \mathbb{N}$, the set of all positive integers, define $\pi_{\mathcal{A}}^t$ and $\Pi_{\mathcal{A}}^t$ as before, and define $\mu_{\mathcal{A}}^t$ on S^{n-1} by $\mu_{\mathcal{A}}^t(x) = \min\{\|Px\| : P \in \Pi_{\mathcal{A}}^t\}$. Let

$$\mu_{\mathcal{A}}(x) = \inf\{\mu_{\mathcal{A}}^t(x) : t \in \mathbb{N}\} = \inf\{\|Px\| : P \in \Pi_{\mathcal{A}}\}.$$

THEOREM 4.6. *Let $\mathcal{A} \in \mathbb{R}^{mn^2}$ and let $B = \{x \in S^{n-1} : \mu_{\mathcal{A}}(x) \geq 1\}$. Suppose that either*

(1) *for each $x \in S^{n-1}$, there is a $P_x \in \Pi_{\mathcal{A}}$ such that $\|P_x x\| \leq 1$, or*

(2) *there is a sequence $\{r_t\}_{t=1}^\infty$ of positive integers such that $\lim_{t\to\infty} r_t = \infty$ and, for each sufficiently large positive integer t and each $x \in B$, there is a $P_{t,x} \in \pi_{\mathcal{A}}^{r_{t,x}}$*

such that $r_t \le r_{t,x} \le t$ *and* $\mu_{\mathcal{A}}^t(x) = \|P_{t,x}x\|$.
Then $\mathcal{A}$ *is in the closure of* $\mathcal{P}_{m,n}$.

Proof. Suppose (1). For $x \in S^{n-1}$, let $P_x = \prod_{i=1}^{w_x} A_{s_x(i)}$. Then, for $0 < \epsilon < 1$ and for each $x \in S^{n-1}$,

$$\left\| \left(\prod_{i=1}^{w_x} \epsilon A_{s_x(i)} \right) x \right\| = \epsilon^{w_x} \|P_x x\| < 1.$$

Hence by Theorem 2.1 $\epsilon\mathcal{A} \in \mathcal{P}_{m,n}$ and $\mathcal{A}$ is in the closure of $\mathcal{P}_{m,n}$.

Now suppose (2). For each $t \in \mathbb{N}$, $\mu_{\mathcal{A}}^t$ is the minimum of finitely many continuous functions of the form $f_P : x \to \|Px\|$. Therefore on the compact set B, $\mu_{\mathcal{A}}^t$ is continuous and assumes a maximum value M_t. Let $\delta_t = (2M_t)^{-\frac{1}{r_t}}$. We claim that $\delta_t\mathcal{A} \in \mathcal{P}_{m,n}$. Clearly, $M_t \ge 1$, so $\delta_t < 1$ and so for each $x \in B$, $\delta_t^{r_{t,x}} \le \delta_t^{r_t}$. Let $P_{t,x} = \prod_{i=1}^{r_{t,x}} A_{s_{t,x}(i)}$. Then, for $x \in B$ we have

$$\left\| \left(\prod_{i=1}^{r_{t,x}} \delta_t A_{s_{t,x}(i)} \right) x \right\| = \|\delta_t^{r_{t,x}} P_{t,x}x\| \le \delta_t^{r_t} \|P_{t,x}x\| = \frac{1}{2M_t}\mu_{\mathcal{A}}^t(x) \le \frac{1}{2} < 1.$$

For $x \notin B$, any product of matrices in $\mathcal{A}$ that reduces the norm of x will provide a product of matrices in $\delta_t\mathcal{A}$ that does so. Hence $\delta_t\mathcal{A} \in \mathcal{P}_{m,n}$ by Theorem 2.1 again.

For each $t \in \mathbb{N}$ and each $x \in B$, $P_{1,x} \in \Pi_{\mathcal{A}}^t$, and so $\mu_{\mathcal{A}}^t(x) \le \mu_{\mathcal{A}}^1(x)$. It follows that $1 \le M_t \le M_1$ and since $\lim_{t\to\infty} r_t = \infty$, $\lim_{t\to\infty} \delta_t = 1$. Hence $\mathcal{A}$ is in the closure of $\mathcal{P}_{m,n}$. ∎

5. Upper triangular matrices of size two. Throughout this section, we consider sets of two real upper triangular 2×2 matrices:

$$\mathcal{A} = \left\{ A_1 = \begin{bmatrix} a_1 & b_1 \\ 0 & c_1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} a_2 & b_2 \\ 0 & c_2 \end{bmatrix} \right\}.$$

The uniform convergence of such $\mathcal{A}$ is easily decided using Theorem 2.5 and Corollary 2.6, as follows.

THEOREM 5.1. *Assume that* $a_1c_1a_2c_2 \neq 0$ *and* $|a_1| \le |a_2|$. *Then* $\mathcal{A}$ *is uniformly convergent if and only if either the spectral radius of at least one of the matrices* A_1 *and* A_2 *is less than one or the following conditions hold:* $\rho(A_j) \ge 1$ $(j = 1, 2)$, $\max\{|a_1|, |c_2|\} < 1$, *and* $\ln(|a_1|) \cdot \ln(|c_2|) > \ln(|a_2|) \cdot \ln(|c_1|)$.

Under the hypotheses of Theorem 5.1, the matrix $\mathcal{C}(\mathcal{D}(\mathcal{A}))$ (see the definitions immediately before and after Theorem 2.5) is given by

$$\mathcal{C}(\mathcal{D}(\mathcal{A})) = \begin{bmatrix} -\ln|a_1| & -\ln|a_2| \\ -\ln|c_1| & -\ln|c_2| \end{bmatrix}.$$

The uniform convergence of $\mathcal{A}$ in the case $a_1a_2c_1c_2 = 0$ is easily sorted out. We omit the proof of Theorem 5.1.

For the pointwise convergence the situation is much more subtle. We consider first the case when $|\det(A_1)| = |\det(A_2)| = 1$. Obviously, in this case the set $\mathcal{A}$ cannot be uniformly convergent. However, often it is pointwise convergent, as follows.

THEOREM 5.2. *Assume* $|a_1c_1| = |a_2c_2| = 1$. *Then* $\mathcal{A}$ *is pointwise convergent unless at least one of the three following conditions occur:* (1) $\min\{|a_1|, |a_2|\} \ge 1$ *or* $\min\{|c_1|, |c_2|\} \ge 1$; (2) $\det A_1 = \det A_2 = 1$; (3) A_1 *and* A_2 *are diagonalizable by simultaneous similarity, i.e.,* $S^{-1}A_1S$ *and* $S^{-1}A_2S$ *are diagonal for some invertible*

(real) matrix S. If one of the conditions (1)–(3) *occurs, then the set $\mathcal{A}$ is not pointwise convergent.*

Proof. We dispose of the conditions (1)–(3) first. If (1) holds and, say, min $\{|c_1|, |c_2|\} \geq 1$, then the vector $[0\ 1]^T \in \mathbb{R}^2$ is not contracted (relative to the Euclidean norm) by any finite product of the matrices A_1 and A_2. Hence by Theorem 2.1, $\mathcal{A}$ is not pointwise convergent.

For the rest of the proof we assume that (1) does not hold. In particular, the numbers a_j and c_j $(j = 1, 2)$ are different from ± 1. If (3) holds, then in view of Lemma 2.4 $\mathcal{A}$ is not pointwise convergent (because it is not uniformly convergent). If (2) holds, then (multiplying if necessary one or both matrices A_j by -1, which does not affect pointwise convergence) we can assume that a_1, c_1, a_2, c_2 are positive. Since $a_1 \neq c_1$, we also can assume (using a simultaneous similarity with an upper triangular similarity matrix S) that $b_1 = 0$. If $b_2 \geq 0$, then by Theorem 3.1 $\mathcal{A}$ cannot be pointwise convergent. If $b_2 < 0$, then the simultaneous similarity with $S = \left[\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right]$ replaces b_2 by $-b_2$, thereby reducing to the previous case.

We assume from now on in the proof that none of the conditions (1), (2), or (3) hold, and we will prove that $\mathcal{A}$ is pointwise convergent. Throughout the proof we will repeatedly use the following easily verified fact: if a subset of the multiplicative semigroup generated by $\mathcal{A}$ is pointwise convergent, then the set $\mathcal{A}$ itself is pointwise convergent.

Without loss of generality, we assume that $\det A_2 = -1$. Further, by reducing A_1 to a diagonal form using simultaneous similarity with a suitable upper triangular matrix S, we assume that

$$\mathcal{A} = \left\{ A_1 = \begin{bmatrix} \delta a & 0 \\ 0 & a^{-1} \end{bmatrix},\ A_2 = \begin{bmatrix} -b^{-1} & 1 \\ 0 & b \end{bmatrix} \right\}.$$

Here a and b are nonzero real numbers such that $|a|$ and $|b|$ are both greater than, or both less than, 1 and $\delta = \pm 1$.

First, consider the case when $a = b$:

$$A_1 = \begin{bmatrix} \delta a & 0 \\ 0 & a^{-1} \end{bmatrix}, \qquad A_2 = \begin{bmatrix} -a^{-1} & 1 \\ 0 & a \end{bmatrix},$$

where $a \neq \pm 1, 0$. Assume $a > 1$. If $\delta = 1$, a calculation reveals that, relative to the Euclidean norm, A_2 contracts $\left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$; A_1 contracts $\left[\begin{smallmatrix} x \\ 1 \end{smallmatrix}\right]$ if $|x| < 1/a$; A_2A_1 contracts $\left[\begin{smallmatrix} x \\ 1 \end{smallmatrix}\right]$ if $x > 1/(2a)$; and $A_2^2A_1^2$ contracts $\left[\begin{smallmatrix} x \\ 1 \end{smallmatrix}\right]$ if $x < -1/(2a)$. By Theorem 2.1, the set $\{A_1, A_2\}$ is pointwise convergent.

If $\delta = -1$ and $a > 1$, a calculation reveals that A_2 contracts $\left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$; A_1 contracts $\left[\begin{smallmatrix} x \\ 1 \end{smallmatrix}\right]$ if $|x| \leq 1/a$; and A_2A_1 contracts $\left[\begin{smallmatrix} x \\ 1 \end{smallmatrix}\right]$ if $x < -1/(2a)$. Let k be a positive integer such that $k < a^2 \leq k + 1$, and consider

$$B_k = A_1A_2(A_2A_1)^k = \begin{bmatrix} 1 & \frac{k}{a} - a \\ 0 & 1 \end{bmatrix}.$$

Then B_k contracts $\left[\begin{smallmatrix} x \\ 1 \end{smallmatrix}\right]$ if $x > 1/(2a)$. Again, by Theorem 2.1, $\{A_1, A_2\}$ is pointwise convergent.

If $0 < a < 1$ and $\delta = -1$, then by simultaneuos similarity we reduce A_2 to the form $\left[\begin{smallmatrix} -a^{-1} & 0 \\ 0 & a \end{smallmatrix}\right]$ and A_1 to the form $\left[\begin{smallmatrix} -a & 1 \\ 0 & a^{-1} \end{smallmatrix}\right]$. Now apply the already proved part to the new pair of matrices, with the roles of A_1 and A_2 interchanged and a^{-1} playing

the role of a. The pointwise convergence of $\{A_1, A_2\}$ follows. If $0 < a < 1$ and $\delta = 1$, then consider

$$B_1 = A_1^2 A_2 = \begin{bmatrix} -a & a^2 \\ 0 & a^{-1} \end{bmatrix}.$$

A simultaneous similarity reduces the matrices B_1 and A_2 to the forms $\left[\begin{smallmatrix} -a & 0 \\ 0 & a^{-1} \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} -a^{-1} & 1 \\ 0 & a \end{smallmatrix}\right]$. But the pair of matrices in this form is already proved to be pointwise convergent (the case $0 < a < 1$, $\delta = -1$, considered above), therefore the pair B_1, A_2 is also pointwise convergent. Since B_1 belongs to the semigroup generated by A_1 and A_2, the same property holds for A_1 and A_2.

Finally, if $a < 0$, then by multiplying the matrices A_1 and A_2 by -1 and applying a simultaneous similarity with an upper triangular similarity matrix S, we reduce the proof of pointwise convergence to the previously considered case $a > 0$. This concludes the consideration of the case when $a = b$.

We now assume $a \neq b$ and consider first the case when $a, b > 1$. We already know that the two sets

$$\mathcal{A}_1 = \left\{ \begin{bmatrix} \pm a & 0 \\ 0 & a^{-1} \end{bmatrix}, \begin{bmatrix} -a^{-1} & 1 \\ 0 & a \end{bmatrix} \right\}, \qquad \mathcal{A}_2 = \left\{ \begin{bmatrix} \pm a/b & 0 \\ 0 & b/a \end{bmatrix}, \begin{bmatrix} -b/a & 1 \\ 0 & a/b \end{bmatrix} \right\}$$

are pointwise convergent. By Theorem 4.2, there exist $\epsilon_1, \epsilon_2 > 0$ such that

$$|c - a| < \epsilon_1 \Rightarrow \left\{ \begin{bmatrix} \pm a & 0 \\ 0 & a^{-1} \end{bmatrix}, \begin{bmatrix} -c^{-1} & 1 \\ 0 & c \end{bmatrix} \right\} \text{ is pointwise convergent} \tag{5.1}$$

and

$$|c - a/b| < \epsilon_2 \Rightarrow \left\{ \begin{bmatrix} \pm a/b & 0 \\ 0 & b/a \end{bmatrix}, \begin{bmatrix} -c^{-1} & 1 \\ 0 & c \end{bmatrix} \right\} \text{ is pointwise convergent.} \tag{5.2}$$

The 1 in the $(1,2)$ position of the second matrices in (5.1) and (5.2) can be changed to any nonzero real number without affecting either implication by using simultaneous similarity.

Let $\epsilon = \min\{\epsilon_1, \epsilon_2\}$. By Theorem 6.8 in [NZ] there exist positive integers x and $y + 1$ such that

$$\left| \frac{x}{y+1} - \log(a) \right| < \frac{\ln(\epsilon/a + 1)}{(y+1)\ln(b)},$$

where $\log(a) = \frac{\ln(a)}{\ln(b)}$ is the base b logarithm. Multiplying by $(y+1)\ln(b)$ and then taking the exponential on both sides, we get

$$1 \leq b^{|x - (\log(a))(y+1)|} < \epsilon/a + 1. \tag{5.3}$$

If the expression inside the absolute value sign in (5.3) is positive, then

$$1 \leq \frac{b^x}{b^{(\log(a))(y+1)}} < 1 + \epsilon/a,$$

and therefore

$$a \leq \frac{b^x}{a^y} < a + \epsilon. \tag{5.4}$$

Similarly, if the expression inside the absolute value sign in (5.3) is negative, we obtain

$$1 \leq \frac{b^{(\log(a))(y+1)}}{b^x} < \epsilon/a + 1;$$

therefore

$$\frac{b^x}{a^y} \leq a < \frac{b^x}{a^y}(\epsilon/a + 1) = \frac{b^x}{a^y}\left(\frac{\epsilon}{a}\right) + \frac{b^x}{a^y} \leq a\left(\frac{1}{a}\right)\epsilon + \frac{b^x}{a^y} = \frac{b^x}{a^y} + \epsilon. \tag{5.5}$$

Inequalities (5.4) and (5.5) imply that $\left|\frac{b^x}{a^y} - a\right| < \epsilon$. If $\delta^y(-1)^x = -1$, then the set $\{A_1, A_1^y A_2^x\}$ is pointwise convergent by the implication (5.1). Hence the set $\mathcal{A}$ itself is pointwise convergent. If $\delta^y(-1)^x = 1$, then

$$\left|\frac{b^{x-1}}{a^y} - \frac{a}{b}\right| < \epsilon/b < \epsilon.$$

Now,

$$A_1A_2 = \begin{bmatrix} \delta a/(-b) & \delta a \\ 0 & b/a \end{bmatrix}, \qquad A_2A_1 = \begin{bmatrix} \delta a/(-b) & 1/a \\ 0 & b/a \end{bmatrix},$$
$$A_1^y A_2^{x-1} = \begin{bmatrix} -a^y/b^{x-1} & q \\ 0 & b^{x-1}/a^y \end{bmatrix}$$

for some real number q. Since $\delta a \neq 1/a$, reducing by simultaneous similarity (with an upper triangular similarity matrix) at least one of the two pairs $\{A_1A_2, A_1^y A_2^{x-1}\}$ and $\{A_2A_1, A_1^y A_2^{x-1}\}$ to the form

$$\left\{\begin{bmatrix} \delta a/(-b) & 0 \\ 0 & b/a \end{bmatrix}, \begin{bmatrix} -a^y/b^{x-1} & q' \\ 0 & b^{x-1}/a^y \end{bmatrix}\right\}$$

will result in a nonzero number q', which can be made to equal 1 by a subsequent simultaneous similarity. By the implication (5.2) it follows that at least one of the two pairs $\{A_1A_2, A_1^y A_2^{x-1}\}$ and $\{A_2A_1, A_1^y A_2^{x-1}\}$ is pointwise convergent, and then the pointwise convergence holds for the pair $\mathcal{A}$ as well.

Now assume $0 < a < 1$ and $0 < b < 1$. There are a positive even integer k such that $a/(-b)^k > 1$ and a positive integer j such that $0 > a^j/(-b) > -1$. The set

$$\left\{\begin{bmatrix} \frac{a}{b^k} & 0 \\ 0 & \frac{b^k}{a} \end{bmatrix}, \quad \begin{bmatrix} -\frac{a^j}{b} & 1 \\ 0 & \frac{b}{a^j} \end{bmatrix}\right\} \tag{5.6}$$

is pointwise convergent by the already proved part of the theorem. If $\delta = 1$, then at least one of the two pairs $(A_1A_2^k, A_1^jA_2)$ and $(A_1A_2^k, A_2A_1^j)$ can be reduced to (5.6) by simultaneous similarity (an analogous argument was used in the preceding paragraph) and therefore itself is pointwise convergent. However, the pointwise convergence of $\mathcal{A}$ then follows. If $0 < a < 1$ and $0 < b < 1$, but $\delta = -1$, then (A_1^2, A_2) is already proved to be pointwise convergent, and therefore $\mathcal{A}$ also has this property.

Finally, if a and/or b are negative, then negating one or both matrices in $\mathcal{A}$ and performing (if necessary) a simultaneous diagonal similarity to make the $(1,2)$ entry of A_2 equal to 1 will produce a set of a form already proved to be pointwise convergent. Therefore, $\mathcal{A}$ is itself pointwise convergent. □

It follows from Theorem 4.2 that given $N > 0$, there exists positive ϵ such that the set

$$\left\{ A_1 = \begin{bmatrix} a_1 & b_1 \\ 0 & c_1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} a_2 & b_2 \\ 0 & c_2 \end{bmatrix} \right\}$$

is pointwise convergent, provided

$$|a_1c_1 - 1| < \epsilon, \ |a_2c_2 - 1| < \epsilon, \ \max\{|a_1|, |b_1|, |c_1|, |a_2|, |b_2|, |c_2|\} \leq N, \tag{5.7}$$

and none of the conditions (1), (2), (3) of Theorem 5.2 occurs when $|a_1c_1| = |a_2c_2| = 1$. However, the thickness of this "fringe" ϵ is unknown.

The following example illustrates Theorems 5.1 and 5.2.

EXAMPLE 5.3. *Fix $a > 1$ and consider the pair*

$$\left\{ A_1 = \begin{bmatrix} a & 0 \\ 0 & 1/a \end{bmatrix}, \quad A_2 = \begin{bmatrix} c & 1 \\ 0 & d \end{bmatrix} \right\}, \tag{5.8}$$

where c and d are real numbers. The set (5.8) *is uniformly convergent if and only if $|c| < 1$ and $|cd| < 1$. If $|c| \geq 1$ or if $|c| < 1$ and $cd \geq 1$, then the set* (5.8) *is not pointwise convergent. If $|c| < 1$ and $cd = -1$, then the set* (5.8) *is pointwise convergent. By Theorem* 4.2, (5.8) *is also pointwise convergent in a sufficiently small neighborhood of each of the two hyperbolic arcs $cd = -1, -1 < c < 0$ and $cd = -1, 0 < c < 1$. It is not known whether the set* (5.8) *is pointwise convergent for every pair (c, d) in the region $0 < |c| < 1, cd < -1$.*

Acknowledgments. We thank I. M. Spitkovsky and L. Gurvits for several useful discussions concerning the subject matter of this paper.

REFERENCES

[BCS] G. P. BARKER, L. T. CONNER, AND D. P. STANFORD, *Complete controllability and contractability in multi-modal systems*, Linear Algebra Appl., 110 (1988), pp. 55–74.

[BY] M. A. BERGER AND Y. WANG, *Bounded semigroups of matrices*, Linear Algebra Appl., 166 (1992), pp. 21–27.

[C] J. W. C. CASSELS, *An Introduction to the Geometry of Numbers*, Springer, Berlin, New York, 1959.

[CRS] A. L. COHEN, L. RODMAN, AND D. P. STANFORD, *Convergent Sets of Matrices*, REU Report, The College of William and Mary, Williamsburg, VA, 1994.

[DL1] I. DAUBECHIES AND J. C. LAGARIAS, *Sets of matrices all infinite products of which converge*, Linear Algebra Appl., 161 (1992), pp. 227–283.

[DL2] I. DAUBECHIES AND J. C. LAGARIAS, *Two-scale difference equations* II. *Local regularity, infinite products of matrices and fractals*, SIAM J. Math. Anal., 23 (1992), pp. 1031–1079.

[G1] L. GURVITS, *Stability of discrete linear inclusion*, Linear Algebra Appl., 231 (1995), pp. 47–85.

[G2] L. N. GURVITS, *Stability of discrete linear inclusions* II, preprint.

[GR] L. GURVITS AND L. RODMAN, *Convergence of polynomially bounded semigroups of matrices*, SIAM J. Matrix Anal. and Appl., 18 (1997), pp. 360–368.

[HJ] R. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, London, UK, 1985.

[JKS] C. R. JOHNSON, M. K. KERR, AND D. P. STANFORD, *Semipositivity of matrices*, Linear and Multilinear Algebra, 27 (1994), pp. 265–271.

[NZ] I. NIVEN AND H. S. ZUCKERMAN, *Introduction to the Theory of Numbers*, John Wiley and Sons, New York, 1980.

[S] D. P. Stanford, *Stability for a multi-rate sampled-data system*, SIAM J. Control Optim. 17 (1979), pp. 390–399.

[SU] D. P. Stanford and J. M. Urbano, *Some convergence properties of matrix sets* , SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1132–1140.

[TB] J. N. Tsitsiklis and V. D. Blondel, *The Lyapunov exponent and joint spectral radius of pairs of matrices are hard—when not impossible—to compute and to approximate*, Math. Control Signals Systems, 10 (1997), pp. 31–40.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 1, pp. 106–128
© 1999 Society for Industrial and Applied Mathematics

ON SINGULARITIES OF A BOUNDARY OF THE STABILITY DOMAIN*

ALEXEI A. MAILYBAEV† AND ALEXANDER P. SEYRANIAN†

This paper is dedicated to V. I. Arnold on the occasion of his 60th birthday.

Abstract. This paper deals with the study of generic singularities of a boundary of the stability domain in a parameter space for systems governed by autonomous linear differential equations $\dot{y} = Ay$ or $x^{(m)} + a_1 x^{(m-1)} + \cdots + a_m x = 0$. It is assumed that elements of the matrix A and coefficients of the differential equation of mth order smoothly depend on one, two, or three real parameters. A constructive approach allowing the geometry of singularities (orientation in space, magnitudes of angles, etc.) to be determined with the use of tangent cones to the stability domain is suggested. The approach allows the geometry of singularities to be described using only first derivatives of the coefficients a_i of the differential equation or first derivatives of the elements of the matrix A with respect to problem parameters with its eigenvectors and associated vectors calculated at the singular points of the boundary. Two methods of study of singularities are suggested. It is shown that they are constructive and can be applied to investigate more complicated singularities for multiparameter families of matrices or polynomials. Two physical examples are presented and discussed in detail.

Key words. stability boundary, generic singularity, tangent cone, collapse of the Jordan block, versal deformation

AMS subject classifications. 93D99, 34D20

PII. S0895479897326675

Introduction. We consider a system of autonomous linear differential equations $\dot{y} = Ay$ assuming that the real matrix operator A of dimension $m \times m$ smoothly depends on n real parameters. Stability of the trivial solution $y \equiv 0$ of the system is considered. It is well known that the trivial solution is asymptotically stable, if all eigenvalues of A have negative real part, and unstable if at least one of the eigenvalues of A has positive real part. According to this definition the parameter space $\mathbf{R}^n$ is divided into the stability and instability domains. Boundary between these domains corresponds to the cases when some of the eigenvalues have zero real part while other eigenvalues have negative real part.

Arnold [3, 4, 5] listed all the generic singularities arising at the stability boundary in two- and three-dimensional space of parameters and gave their description up to a smooth change of problem parameters (diffeomorphism). In this paper we suggest a constructive approach allowing one to determine the geometry of singularities (orientation in space, magnitudes of angles, etc.) using only first derivatives of the matrix A with respect to parameters and left and right eigenvectors and associated vectors of A, corresponding to the Jordan structure of the matrix A at the singular points of the boundary. Our study is essentially based on the perturbation theory of eigenvalues and eigenvectors, developed by Vishik and Lyusternik [16] and Lidskii [11] and applied by Seyranian [12, 13] to the case of multiple parameters, and the theory of

*Received by the editors August 27, 1997; accepted for publication (in revised form) by A. Edelman September 8, 1998; published electronically August 16, 1999. A preliminary version of this paper appeared in *Proceedings of the Seventh AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, St. Louis, MO, 1998.

http://www.siam.org/journals/simax /21-1/32667.html

†Institute of Mechanics, Moscow State Lomonosov University, Michurinsky pr. 1, 117192 Moscow, Russia (mailybaev@inmech.msu.su, seyran@inmech.msu.su).

normal forms of families of matrices by Arnold [2] and Galin [7]. Investigation of the boundary of the stability domain is closely related to construction of tangent cones to the stability domain at the boundary point, introduced by Levantovskii [9], and the problem of finding stable perturbations of nonsymmetric matrices considered by Burke and Overton [6].

We investigate generic singularities of a boundary of the stability domain for linear autonomous differential equation of mth order $x^{(m)} + a_1 x^{(m-1)} + \cdots + a_m x = 0$, assuming that the coefficients a_i smoothly depend on one, two, or three real parameters. Explicit formulae to describe the geometry of singularities in the parameter space are derived.

As examples, two physical problems are considered: stability of equilibrium of a voltaic arc in an electric circuit and stability of Ziegler's double pendulum with two different damping parameters. In the first problem at the singular point "double zero" we find the angle of the corner of the stability boundary and its orientation in parameter plane. In the second problem it is shown that the singularity, arising at the critical load of the system without damping, represents, according to Arnold's terminology [3, 4], the "deadlock of an edge." This leads to the effects of destabilization due to small damping and absence of a limit of the critical load when damping parameters tend to zero. Similar effects could be expected for systems with singularities like "break of an edge."

The main result of the paper is that the Jordan structure of the matrix A and its first derivatives with respect to problem parameters at any point of the stability boundary define a linear approximation of the stability domain in the vicinity of the considered point. Similar results are valid for stability problems governed by a linear differential equation of mth order: to determine the geometry of the stability domain in the vicinity of a singular point of the boundary we need only to know the multiplicity of the root of the characteristic polynomial and the first derivatives of the coefficients of the differential equation with respect to problem parameters at this point.

1. Collapse of Jordan blocks. Let us consider an eigenvalue problem

$$A\,u = \lambda\,u. \tag{1.1}$$

Here A is a real nonsymmetric square $m \times m$ matrix, the elements of which, $a_{ij}(p)$, $i, j = 1, 2, \ldots, m$, are smooth functions of a real vector of parameters $p = (p_1, p_2, \ldots, p_n)^T$; λ is an eigenvalue; and u is a corresponding eigenvector of dimension m.

It is assumed that at fixed $p = p_0$, λ_0 is an eigenvalue of $A(p_0)$, and a change of the eigenvalue λ_0 is sought that depends on a change of the vector of parameters p. For this purpose let us consider a perturbation of the vector p_0 in the form $p = p(\varepsilon)$, $p(0) = p_0$, where ε is a small positive number and $p(\varepsilon)$ is a smooth function of ε. Determine a real vector of direction $e = (e_1, e_2, \ldots, e_n) = dp/d\varepsilon \neq 0$, where the derivative is calculated at $\varepsilon = 0$. As a result the matrix A takes the increment

$$A(p(\varepsilon)) = A_0 + \varepsilon A_1 + \varepsilon^2 A_2 + \cdots, \tag{1.2}$$

where the matrices A_0 and A_1 are given by the relations

$$A_0 = A(p_0), \quad A_1 = \sum_{s=1}^{n} \frac{\partial A}{\partial p_s}\, e_s. \tag{1.3}$$

The derivatives in (1.3) are taken at $p = p_0$.

Due to the perturbation of the vector p_0, the eigenvalue λ_0 and the eigenvector u_0 take increments. According to the perturbation theory of non-self-adjoint operators, developed in [16, 11], these increments can be expressed as series in integer or fractional powers of ε, depending on the Jordan structure corresponding to the eigenvalue λ_0. A multiple eigenvalue λ_0 generally splits into l simple eigenvalues under the perturbation of parameters $p = p(\varepsilon)$. Expansions for eigenvalues and eigenvectors contain terms with fractional powers $\varepsilon^{j/l}$, $j = 0, 1, 2 \ldots$, where l is the length of the Jordan chain [16, 11].

1.1. Simple eigenvalue. Assume that λ_0 is a simple eigenvalue of the matrix A_0 and u_0 is the corresponding eigenvector. In this case expansions of λ and u take the form [16, 11]

$$\begin{aligned} \lambda &= \lambda_0 + \varepsilon\,\lambda_1 + \varepsilon^2\lambda_2 + \cdots, \\ u &= u_0 + \varepsilon\,w_1 + \varepsilon^2 w_2 + \cdots. \end{aligned} \tag{1.4}$$

For the following presentation we also need the left eigenvector v_0, corresponding to λ_0,

$$v_0^T A_0 = \lambda_0\, v_0^T. \tag{1.5}$$

The eigenvectors u_0 and v_0 in the case of simple eigenvalue λ_0 are related by the condition $v_0^T u_0 \neq 0$. Substituting (1.2), (1.4) into (1.1) and using (1.5) we find [16, 11]

$$\lambda_1 = \frac{v_0^T A_1 u_0}{v_0^T u_0}. \tag{1.6}$$

This expression with the use of (1.3) can be given in the form

$$\lambda_1 = (r,\, e) + i\,(k,\, e), \tag{1.7}$$

where brackets denote the scalar product in $\mathbf{R}^n$, i.e., $(a,\, b) = \sum_{s=1}^n a_s b_s$. Vectors $r = (r^1,\, r^2,\, \ldots,\, r^n)^T$ and $k = (k^1,\, k^2,\, \ldots,\, k^n)^T$ are the gradient vectors of real and imaginary parts of λ at $p = p_0$, given by

$$r^s + i\,k^s = \frac{v_0^T \dfrac{\partial A}{\partial p_s}\, u_0}{v_0^T u_0}\,, \quad s = 1,\, 2,\, \ldots\,,\, n. \tag{1.8}$$

There are two complex-conjugate quantities λ_1, $\overline{\lambda}_1 = (r,\, e) \pm i\,(k,\, e)$ corresponding to a complex-conjugate pair of simple eigenvalues λ_0, $\overline{\lambda}_0 = \alpha_0 \pm i\,\omega_0$. The increments of these eigenvalues are given in the form

$$\lambda,\, \overline{\lambda} = \alpha_0 + (r,\, e)\,\varepsilon \pm i\,[\,\omega_0 + (k,\, e)\,\varepsilon\,] + o(\varepsilon). \tag{1.9}$$

In the case of a real eigenvalue $\lambda_0 = \alpha_0$, the vector $k = 0$.

1.2. Double eigenvalue. Let us consider the case of a double eigenvalue λ_0 with the length of the Jordan chain equal to 2. This means that at $p = p_0$ the eigenvalue λ_0 corresponds to an eigenvector u_0 and an associated vector u_1 governed by the equations

$$\begin{aligned} A_0\, u_0 &= \lambda_0\, u_0, \\ A_0\, u_1 &= \lambda_0\, u_1 + u_0. \end{aligned} \tag{1.10}$$

For a left eigenvector v_0 and an associated vector v_1 we have

$$\begin{aligned} v_0^T A_0 &= \lambda_0 v_0^T, \\ v_1^T A_0 &= \lambda_0 v_1^T + v_0^T. \end{aligned} \tag{1.11}$$

From (1.10) and (1.11) it directly follows that the vectors u_0, u_1, v_0, and v_1 are related by the conditions [16]

$$v_0^T u_0 = 0, \quad v_1^T u_0 = v_0^T u_1 \neq 0. \tag{1.12}$$

In the case of a double eigenvalue we have expansions [16, 11]

$$\begin{aligned} \lambda &= \lambda_0 + \varepsilon^{1/2}\lambda_1 + \varepsilon\,\lambda_2 + \varepsilon^{3/2}\lambda_3 + \cdots, \\ u &= u_0 + \varepsilon^{1/2} w_1 + \varepsilon\, w_2 + \varepsilon^{3/2} w_3 + \cdots. \end{aligned} \tag{1.13}$$

Substituting (1.13) and (1.2) into (1.1) and using (1.10)–(1.12) we obtain expressions for determining λ_1 and λ_2:

$$\begin{aligned} \lambda_1 &= \pm\sqrt{\frac{v_0^T A_1 u_0}{v_0^T u_1}}, \\ \lambda_2 &= \frac{v_0^T A_1 u_1 + v_1^T A_1 u_0 - \lambda_1^2\, v_1^T u_1}{2\, v_0^T u_1}. \end{aligned} \tag{1.14}$$

Expressions (1.13), (1.14) are correct if $v_0^T A_1 u_0 \neq 0$ (the condition Γ in [16]).

Note that the vectors u_0 and v_0 are defined up to arbitrary nonzero multipliers; the vectors u_1 and v_1 are defined up to additive terms αu_0 and βv_0, respectively, where α and β are arbitrary constants. However, the values of λ_1 and λ_2 in (1.14) don't depend on the way that the vectors u_0, u_1, v_0, and v_1 are chosen.

Assuming that the vectors u_0, u_1 are fixed we use the following normalization conditions for v_0 and v_1:

$$v_0^T u_1 = 1, \quad v_1^T u_1 = 0. \tag{1.15}$$

Combining the expressions (1.14) with (1.3), (1.13) and the normalization conditions (1.15) gives

$$\lambda = \lambda_0 \pm \sqrt{[\,(f_1,\, e) + i\,(q_1,\, e)\,]\,\varepsilon} + \frac{1}{2}\,[\,(f_2,\, e) + i\,(q_2,\, e)\,]\,\varepsilon + o(\varepsilon), \tag{1.16}$$

where components of the vectors $f_j = (f_j^1,\, f_j^2,\, \ldots,\, f_j^n)$, $q_j = (q_j^1,\, q_j^2,\, \ldots,\, q_j^n)$, $j = 1,\, 2$, are real and imaginary parts of quantities defined by

$$\begin{aligned} f_1^s + i\, q_1^s &= v_0^T \frac{\partial A}{\partial p_s}\, u_0, \\ f_2^s + i\, q_2^s &= v_0^T \frac{\partial A}{\partial p_s}\, u_1 + v_1^T \frac{\partial A}{\partial p_s}\, u_0, \quad s = 1,\, 2,\, \ldots,\, n. \end{aligned} \tag{1.17}$$

If λ_0 is a real number, then the vectors $q_1 = q_2 = 0$.

1.3. Triple eigenvalue. Consider the case of a triple eigenvalue that is characterized by a Jordan chain of the length 3. This means that there are an eigenvector u_0 and associated vectors u_1, u_2 satisfying the equations

$$\begin{aligned} A_0\,u_0 &= \lambda_0\,u_0,\\ A_0\,u_1 &= \lambda_0\,u_1 + u_0,\\ A_0\,u_2 &= \lambda_0\,u_2 + u_1. \end{aligned} \tag{1.18}$$

For a left eigenvector v_0 and associated vectors v_1 and v_2 we have

$$\begin{aligned} v_0^T A_0 &= \lambda_0\,v_0^T,\\ v_1^T A_0 &= \lambda_0\,v_1^T + v_0^T,\\ v_2^T A_0 &= \lambda_0\,v_2^T + v_1^T. \end{aligned} \tag{1.19}$$

The vectors u_j, v_j, $j = 1, 2, 3$, are related by the conditions

$$\begin{gathered} v_0^T u_0 = v_0^T u_1 = v_1^T u_0 = 0,\\ v_0^T u_2 = v_1^T u_1 = v_2^T u_0 \neq 0,\\ v_1^T u_2 = v_2^T u_1. \end{gathered} \tag{1.20}$$

These conditions can be proved by means of (1.18), (1.19). Assuming that the vectors u_j, $j = 1, 2, 3$, are fixed we use the following normalization conditions for the vectors v_j, $j = 1, 2, 3$:

$$v_0^T u_2 = 1, \quad v_1^T u_2 = v_2^T u_2 = 0. \tag{1.21}$$

These conditions define the vectors v_j, $j = 1, 2, 3$, uniquely.

The eigenvalue λ_0 generally splits to three simple eigenvalues due to perturbation of parameters. Then eigenvalues and eigenvectors can be given in the form [16, 11]

$$\begin{aligned} \lambda &= \lambda_0 + \varepsilon^{1/3}\lambda_1 + \varepsilon^{2/3}\lambda_2 + \varepsilon\,\lambda_3 + +\varepsilon^{4/3}\lambda_4 + \cdots,\\ u &= u_0 + \varepsilon^{1/3}w_1 + \varepsilon^{2/3}w_2 + \varepsilon\,w_3 + \varepsilon^{4/3}w_4 + \cdots. \end{aligned} \tag{1.22}$$

Substituting expansions (1.22) and (1.2) into (1.1) and using (1.18)–(1.21) we obtain expressions for the first three coefficients λ_j, $j = 1, 2, 3$:

$$\begin{aligned} \lambda_1 &= \sqrt[3]{v_0^T A_1 u_0},\\ \lambda_2 &= \frac{v_0^T A_1 u_1 + v_1^T A_1 u_0}{3\,\lambda_1},\\ \lambda_3 &= \frac{1}{3}\left(v_0^T A_1 u_2 + v_1^T A_1 u_1 + v_2^T A_1 u_0\right). \end{aligned} \tag{1.23}$$

These expressions are correct if $v_0^T A_1 u_0 \neq 0$ (the condition Γ in [16]). In this case the first expression of (1.23) defines three different complex roots λ_1. Then values of λ_2 and λ_3 are determined for each root λ_1 from the second and third expressions of (1.23).

Combining expressions (1.23) with (1.22) and (1.3) gives

$$\begin{aligned} \lambda = \lambda_0 &+ \sqrt[3]{[\,(h_1,\,e) + i\,(t_1,\,e)\,]\,\varepsilon}\\ &+ \frac{(h_2,\,e) + i\,(t_2,\,e)}{3\sqrt[3]{(h_1,\,e) + i\,(t_1,\,e)}}\,\varepsilon^{2/3} + \frac{1}{3}\,[\,(h_3,\,e) + i\,(t_3,\,e)\,]\,\varepsilon + o(\varepsilon), \end{aligned} \tag{1.24}$$

where components of the vectors $h_j = (h_j^1, h_j^2, \ldots, h_j^n)$, $t_j = (t_j^1, t_j^2, \ldots, t_j^n)$, $j = 1, 2, 3$, are real and imaginary parts of the quantities

$$
\begin{aligned}
h_1^s + i\,t_1^s &= v_0^T \frac{\partial A}{\partial p_s}\, u_0, \\
h_2^s + i\,t_2^s &= v_0^T \frac{\partial A}{\partial p_s}\, u_1 + v_1^T \frac{\partial A}{\partial p_s}\, u_0, \\
h_3^s + i\,t_3^s &= v_0^T \frac{\partial A}{\partial p_s}\, u_2 + v_1^T \frac{\partial A}{\partial p_s}\, u_1 + v_2^T \frac{\partial A}{\partial p_s}\, u_0, \\
s &= 1, 2, \ldots, n.
\end{aligned} \tag{1.25}
$$

The cubic roots in the second and third terms of the right-hand side of (1.24) are the same and take three different complex values.

If λ_0 is a real number, then the vectors $t_j = 0$, $j = 1, 2, 3$.

2. One- and two-parameter families of matrices. Let us consider a linear evolutionary system of the form

$$\dot{y} = A\,y, \tag{2.1}$$

where A is a real autonomous $m \times m$ matrix and y is a real vector of dimension m. The system is stable (asymptotically stable) if all eigenvalues λ of the matrix A have negative real parts, $\operatorname{Re}\lambda < 0$. If there exists at least one eigenvalue such that $\operatorname{Re}\lambda > 0$, the system is unstable. If there are some eigenvalues with $\operatorname{Re}\lambda = 0$ while for all others $\operatorname{Re}\lambda < 0$, we have a boundary point.

A *family* of matrices is a mapping $A : \Lambda \longrightarrow \mathbf{R}^{m^2}$ of the parameter space into the space of matrices. The set of values $p \in \Lambda$, such that $A(p)$ is a stable matrix, is called the stability domain. First, let us consider a one-parameter family $A(p)$, $p \in \mathbf{R}$. The stability domain boundary of a *generic* one-parameter family is characterized by one simple eigenvalue $\lambda = 0$ or by one pair of complex-conjugate simple eigenvalues $\lambda = \pm i\,\omega$ of the matrix A [3, 4]. In the technical literature these cases are called divergence and flutter boundaries, respectively.

Using (1.9) we have

$$\operatorname{Re}\lambda = r\,(p - p_0) + o(p - p_0) \tag{2.2}$$

for a simple eigenvalue λ in the neighborhood of the stability boundary point p_0, $\operatorname{Re}\lambda_0 = 0$. Hence, location of the stability and instability domains is determined by the sign of the quantity

$$r = \operatorname{Re}\left[\left(v_0^T\, dA/dp\, u_0\right)/\left(v_0^T u_0\right)\right].$$

For example, if $r > 0$, then the system is stable ($\operatorname{Re}\lambda < 0$) at $p < p_0$ and unstable ($\operatorname{Re}\lambda > 0$) at $p > p_0$ for p sufficiently close to p_0. Note that $r \neq 0$ in the generic case, i.e., the case $r = 0$ can be removed by an arbitrarily small shift of the family.

In the case of a two-parameter generic family $A(p)$, $p \in \mathbf{R}^2$, the stability boundary is a smooth curve whose only singularities are corners. The curve in nonsingular points is characterized by a simple eigenvalue $\lambda = 0$ or a simple pair $\lambda = \pm i\,\omega$ and has a normal vector r defined in (1.8). From (1.9) it follows that the normal r lies in the instability domain. The corners correspond to matrices $A(p)$ of the three following types (strata) [3, 4, 5]: $F_1(0^2)$—double eigenvalue $\lambda = 0$ with the

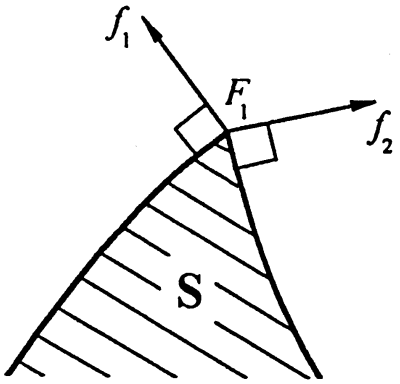


FIG. 2.1. *A corner of the stability domain.*

corresponding Jordan block of the order 2; $F_2(0,\, \pm i\,\omega)$— simple eigenvalues 0, $\pm i\,\omega$; $F_3(\pm i\,\omega_1,\, \pm i\,\omega_2)$—two pairs of simple different eigenvalues $\pm i\,\omega_1$, $\pm i\,\omega_2$. Other types of singularities can be destroyed by an arbitrarily small shift of the family.

Using expansions (1.16) for a double eigenvalue $\lambda_0 = 0$ we have

$$\lambda = \pm\sqrt{(f_1,\, e)\,\varepsilon} + \frac{1}{2}\,(f_2,\, e)\,\varepsilon + o(\varepsilon), \tag{2.3}$$

where vectors f_1 and f_2 are defined by means of (1.17) and e is a vector of variation (direction). In the generic case the vectors f_1 and f_2 are linearly independent. For an arbitrary fixed vector e, such that $(f_1,\, e) < 0$ and $(f_2,\, e) < 0$, we have $\operatorname{Re}\lambda < 0$ (stability) for sufficiently small ε. If at least one of these inequalities has the opposite sign, we have $\operatorname{Re}\lambda > 0$ (instability).

For the following presentation we need a concept of *tangent cone*. A tangent cone to the stability domain at the boundary point is a set of direction vectors of the curves starting at this point and lying in the stability domain [9]. A tangent cone is nondegenerate if it cuts out on a sphere a set of nonzero measure. Otherwise, the cone is called degenerate. A tangent cone can be considered as a linear approximation of the stability domain.

In accordance with (2.3) a tangent cone at the boundary point corresponding to the stratum $F_1(0^2)$ takes the form

$$K_{F_1} = \big\{\, e : (f_1,\, e) \le 0,\ \ (f_2,\, e) \le 0 \,\big\}. \tag{2.4}$$

Using expression (2.2) for a simple eigenvalue we similarly obtain tangent cones at boundary points corresponding to the strata $F_2(0,\, \pm i\,\omega)$ and $F_3(\pm i\,\omega_1,\, \pm i\,\omega_2)$:

$$K_{F_2} = \big\{\, e : (r_0,\, e) \le 0,\ \ (r,\, e) \le 0 \,\big\}, \tag{2.5}$$

$$K_{F_3} = \big\{\, e : (r_1,\, e) \le 0,\ \ (r_2,\, e) \le 0 \,\big\}. \tag{2.6}$$

Here the vectors r_0, r, r_1, r_2 correspond to the simple eigenvalues 0, $\pm i\,\omega$, $\pm i\,\omega_1$, $\pm i\,\omega_2$, respectively, and are calculated using (1.8). In the generic case the vectors r, r_0 and also r_1, r_2 are linearly independent.

Using the relations (2.4)–(2.6) for tangent cones we can find tangent vectors to the stability boundary. For example, in the case of a singular point of the type F_1

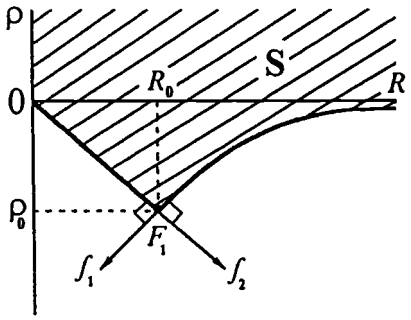


FIG. 3.1. *The stability domain of equilibrium of the voltaic arc in electric circuit.*

tangent vectors g_1 and g_2 can be found from the following system of linear equations (see Figure 2.1):

$$(f_i,\, g_j) = -\delta_{ij}, \quad i,\, j = 1,\, 2, \tag{2.7}$$

where δ_{ij} is Kronecker's delta and f_1, f_2 are the vectors from (2.4).

The inequalities in (2.4)–(2.6) define an intersection of two halfplanes. Hence, the stability domain wedges into the instability domain with the angle of wedge less than π; see Figure 2.1. This is a quantitative justification of the Arnold's principle of "fragility of all good things" [3, 4, 5] and quasi convexity of the stability domain [9].

3. Example: Stability of equilibrium of a voltaic arc in electric circuit. As an example let us consider a stability problem of equilibrium of a circuit consisting of a voltaic arc, resister R, inductance L, and shunting capacitor C connected in series. Linearized differential equations near the equilibrium of the system have the form [1]

$$\begin{aligned} \frac{d\xi}{dt} &= -\frac{\rho\xi}{L} + \frac{\eta}{L}, \\ \frac{d\eta}{dt} &= -\frac{\xi}{C} - \frac{\eta}{RC}, \end{aligned} \tag{3.1}$$

where $\xi(t)$, $\eta(t)$ are, respectively, an electric current and a voltage in the voltaic arc, and ρ is a resistance of the arc.

The system (3.1) depends on four parameters: three positive quantities L, C, R and parameter ρ, which can take both positive and negative values. Assuming that the parameters L and C are fixed, we consider the stability problem on the plane of two parameters: $p_1 = R$ and $p_2 = \rho$. The matrix A corresponding to the system (3.1) is

$$A = \begin{pmatrix} -\dfrac{\rho}{L} & \dfrac{1}{L} \\ -\dfrac{1}{C} & -\dfrac{1}{CR} \end{pmatrix}. \tag{3.2}$$

The characteristic equation of the system takes the form

$$\lambda^2 + \left(\frac{1}{RC} + \frac{\rho}{L}\right)\lambda + \frac{1}{LC}\left(\frac{\rho}{R} + 1\right) = 0. \tag{3.3}$$

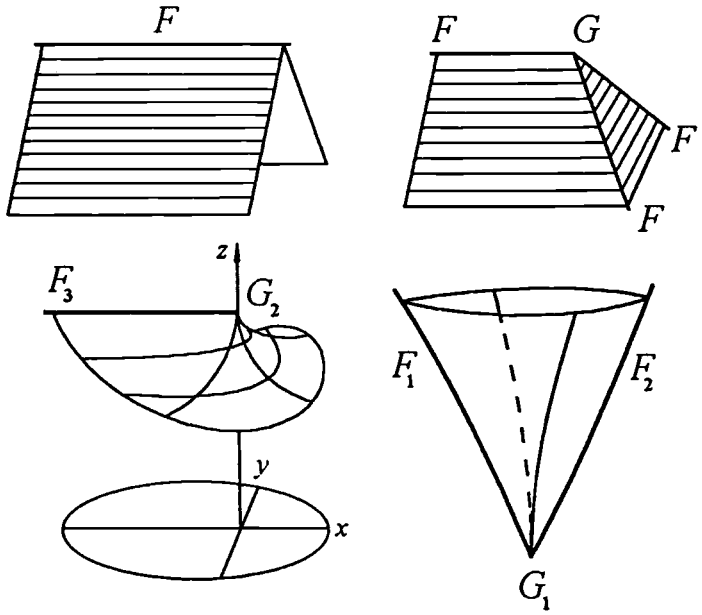


FIG. 4.1. *Singularities of the stability boundary of a three-parameter generic family.*

At the point $R_0 = \sqrt{L/C}$, $\rho_0 = -\sqrt{L/C}$ the characteristic equation (3.3) has a double root $\lambda_0 = 0$ with the length of the Jordan chain equal to 2. The equations for Jordan chains (1.10), (1.11) yield at this point

$$u_0 = \begin{pmatrix} 1 \\ -\sqrt{L/C} \end{pmatrix}, \quad u_1 = \begin{pmatrix} 0 \\ L \end{pmatrix}, \quad v_0 = \begin{pmatrix} 1/\sqrt{LC} \\ 1/L \end{pmatrix}, \quad v_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Using these vectors and the matrix A from (3.2) we calculate the vectors f_1 and f_2 according to (1.17):

$$f_1 = -\frac{1}{L\sqrt{LC}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad f_2 = \frac{1}{L} \begin{pmatrix} 1 \\ -1 \end{pmatrix}. \tag{3.4}$$

Thus, we have found the tangent cone (2.4) to the stability domain at the point $R = R_0$, $\rho = \rho_0$; see Figure 3.1. The tangent vectors to the stability boundary (2.7) up to a positive factor are $g_1 = (1, 1)^T$, $g_2 = (-1, 1)^T$. Hence, the angle of the wedge of the stability domain is equal to $\pi/2$. This result is in accordance with [1], where it has been shown that the stability boundary consists of the line $\rho = -R$, $0 \le R \le \sqrt{L/C}$ and the hyperbola $\rho = -L/(CR)$, $\sqrt{L/C} \le R$; see Figure 3.1.

4. Three-parameter family of matrices. Consider a generic three-parameter family of matrices $A(p)$, $p \in \mathbf{R}^3$. The stability domain boundary of the family is a smooth surface characterized by one simple eigenvalue $\lambda = 0$ or a pair of simple eigenvalues $\lambda = \pm i\,\omega$ [3, 4]. The normal vector r to this surface is defined by the relation (1.8) in the same way as in the two-parameter case. The vector r lies in the instability domain. According to [3, 4] the only singularities of the stability boundary of a generic three-parameter family are of four types: dihedral angle, trihedral angle, deadlock of an edge, and break of an edge; see Figure 4.1.

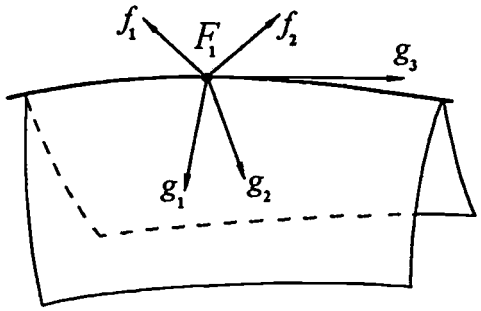


FIG. 4.2. *An edge of the stability domain.*

The dihedral angle singularity is connected with the strata $F_1(0^2)$, $F_2(0, \pm i\,\omega)$, $F_3(\pm i\,\omega_1, \pm i\,\omega_2)$ examined in section 2. Therefore tangent cones to the stability domain at these singular points are determined by the relations (2.4)–(2.6), respectively.

The tangent cone K_{F_1} can be written in a more suitable form introducing vectors g_1, g_2, g_3 by the equations

$$(f_i, g_j) = -\delta_{ij}, \quad i = 1, 2, \quad j = 1, 2, 3, \tag{4.1}$$

where f_1, f_2 are the vectors from (2.4). The equations (4.1) are solvable because the vectors f_1 and f_2 are linearly independent for a generic family. The vector g_3 is directed along the edge and the vectors g_1, g_2 are tangent to the sides of the dihedral angle; see Figure 4.2. Using these vectors the set (2.4) can be written in the form

$$K_{F_1} = \{\, e :\ e = \alpha\, g_1 + \beta\, g_2 + \gamma\, g_3;\ \ \alpha, \beta, \gamma \in \mathbf{R},\ \alpha \geq 0,\ \beta \geq 0 \,\}. \tag{4.2}$$

Substituting the expression $e = \alpha\, g_1 + \beta\, g_2 + \gamma\, g_3$ into (2.4) and using (4.1) we find $(f_1, e) = -\alpha \leq 0$, $(f_2, e) = -\beta \leq 0$. It proves the representation (4.2). Similar representations can be obtained for the tangent cones K_{F_2} and K_{F_3} using in (4.1) the vectors r_0, r and r_1, r_2 instead of f_1, f_2, respectively.

The trihedral angle singularity is characterized by the following strata [3, 4]: $G_3(0^2, \pm i\,\omega)$—a double eigenvalue $\lambda = 0$ with the length of the Jordan chain equal to 2 and a pair of simple pure imaginary eigenvalues; $G_4(0, \pm i\,\omega_1, \pm i\,\omega_2)$—simple $\lambda = 0$ and two different pairs of simple pure imaginary eigenvalues; $G_5(\pm i\,\omega_1, \pm i\,\omega_2, \pm i\,\omega_3)$—three different pairs of simple pure imaginary eigenvalues.

Note that these strata differ from the strata F_1, F_2, F_3 by the presence of an additional pair of simple eigenvalues of the type $\lambda = \pm i\,\omega$. Therefore tangent cones at the boundary points of these types, similarly to (2.4)–(2.6), can be written in the form

$$K_{G_3} = \{\, e : (f_1, e) \leq 0,\ \ (f_2, e) \leq 0,\ \ (r, e) \leq 0 \,\}, \tag{4.3}$$

$$K_{G_4} = \{\, e : (r_0, e) \leq 0,\ \ (r_1, e) \leq 0,\ \ (r_2, e) \leq 0 \,\}, \tag{4.4}$$

$$K_{G_5} = \{\, e : (r_1, e) \leq 0,\ \ (r_2, e) \leq 0,\ \ (r_3, e) \leq 0 \,\}, \tag{4.5}$$

where the vectors r_0, r, r_j, $j = 1, 2, 3$ correspond to the eigenvalues 0, $\pm i\,\omega$, $\pm i\,\omega_j$, $j = 1, 2, 3$, respectively, and are defined in (1.8). The vectors f_1 and f_2 correspond

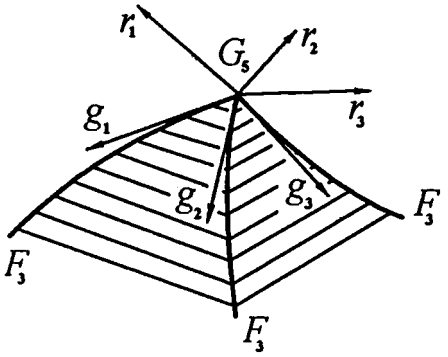


FIG. 4.3. *The singularity trihedral angle.*

to the double eigenvalue $\lambda = 0$ and are defined in (1.17). The sets (4.3)–(4.5) describe a trihedral angle, which is located entirely in a closed half-space; see Figure 4.3.

As in (4.1), using the vectors determining the cones, we can find vectors g_1, g_2, g_3 tangential to the edges of the trihedral angle; see Figure 4.3. For example, for the singular point of the type G_5 we have

$$(r_i,\, g_j) = -\delta_{ij}, \quad i,\, j = 1,\, 2,\, 3, \tag{4.6}$$

where r_j, $j = 1, 2, 3$ are the vectors from (4.5). With the use of these vectors the tangent cone K_{G_5} can be described in the following way:

$$K_{G_5} = \big\{\, e :\ e = \alpha\, g_1 + \beta\, g_2 + \gamma\, g_3;\ \ \alpha,\, \beta,\, \gamma \geq 0 \,\big\}. \tag{4.7}$$

Similar representations can be deduced for the cones K_{G_3} and K_{G_4}.

Note that the vectors determining dihedral and trihedral angles are linearly independent for a generic family.

Singularity deadlock of an edge is characterized by the stratum $G_2\big((\pm i\,\omega)^2\big)$—a pair of double pure imaginary eigenvalues $\lambda = \pm i\,\omega$ with the length of the Jordan chain equal to 2. It is well known that the stability domain in the neighborhood of this singularity up to a smooth change of parameters (diffeomorphism) is given by [3, 4]

$$z + \left|\, \mathrm{Re}\sqrt{x + i\,y}\,\right| < 0. \tag{4.8}$$

The stability boundary of (4.8) is a half of the so-called Witney–Cayley umbrella surface [4, 5]. The tangent cone to the domain (4.8) at the singular point G_2, i.e., at $x = y = z = 0$, is degenerate and represents a plane angle

$$K^0_{G_2} = \big\{\, e = (e_1,\, e_2,\, e_3) :\ e_1 \leq 0,\ \ e_2 = 0,\ \ e_3 \leq 0 \,\big\}.$$

Note that the singularity G_2 is formed by a collision of two different simple pure imaginary eigenvalues $i\,\omega_1$ and $i\,\omega_2$ at the singular point when they move along the edge of the type F_3.

Let us calculate the tangent cone for G_2 in the generic case. Using the expansions (1.16) for a double eigenvalue $\lambda = i\,\omega$ we have

$$\lambda = i\,\omega \pm \sqrt{[\,(f_1,\, e) + i\,(q_1,\, e)\,]\,\varepsilon} + \frac{1}{2}\,[\,(f_2,\, e) + i\,(q_2,\, e)\,]\,\varepsilon + o(\varepsilon), \tag{4.9}$$

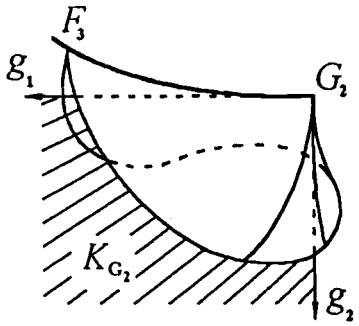


FIG. 4.4. *The singularity deadlock of an edge.*

where the vectors f_1, q_1, f_2, q_2 correspond to $\lambda = i\,\omega$ and are defined in (1.17). The expansions for a complex-conjugate double eigenvalue $\overline{\lambda} = -i\,\omega$ can be found by taking complex conjugation of (4.9).

If either $(f_1,\, e) > 0$ or $(q_1,\, e) \neq 0$ under the radical in (4.9), then one of the eigenvalues has positive real part for sufficiently small ε (instability). In the case when $(f_1,\, e) < 0,\ (q_1,\, e) = 0$, the second (with the radical) term is a pure imaginary number. Hence, for $(f_2,\, e) < 0$ and small ε we have $\operatorname{Re}\lambda < 0$ (stability) and for $(f_2,\, e) > 0$ we have $\operatorname{Re}\lambda > 0$ (instability). Therefore, the tangent cone to the stability domain at a singular point G_2 is the plane angle of the form

$$K_{G_2} = \big\{\, e : (f_1,\, e) \le 0,\ (f_2,\, e) \le 0,\ (q_1,\, e) = 0 \,\big\}. \tag{4.10}$$

From (4.9) it directly follows that all smooth curves, emitted from the singular point along the direction e, satisfying the conditions $(f_1,\, e) < 0,\ (f_2,\, e) < 0,\ (q_1,\, e) = 0$, lie in the stability domain for rather small ε.

Determining the vectors g_1, g_2 by

$$(f_i,\, g_j) = -\delta_{ij}, \quad (q_1,\, g_j) = 0, \quad i,\, j = 1,\, 2,$$

we can write (4.10) in the form

$$K_{G_2} = \big\{\, e :\ e = \alpha\, g_1 + \beta\, g_2;\ \ \alpha,\, \beta \ge 0 \,\big\}. \tag{4.11}$$

The vectors g_1 and g_2 are directed along the sides of the plane angle K_{G_2}. The vector g_1 is tangent to the edge F_3 of the stability domain characterized by two different pairs of simple eigenvalues $\pm i\,\omega_1$, $\pm i\,\omega_2$; see Figure 4.4.

Note, that the vectors f_1, f_2, q_1 are linearly independent for a generic family of matrices.

5. Singularity break of an edge. The singularity break of an edge is characterized by the stratum $G_1(0^3)$—by one triple zero eigenvalue of the matrix $A(p_0)$ with the Jordan chain of the length equal to 3. The expansion of a triple eigenvalue is described by (1.24). The cubic root in (1.24) takes three different complex values. This means that if $(h_1, e) \neq 0$ (note that $t_1 = 0$ since $\lambda_0 = 0$), then at least one eigenvalue has positive real part (instability). Hence, the tangent cone to the stability domain lies in the plane $(h_1, e) = 0$, where the expansion (1.24) is not valid due to violation of the Γ condition $v_0^T A_1 u_0 = (h_1, e) \neq 0$. Therefore the tangent cone in this case cannot be found with the method used for the investigation of the previous singularities. For this reason we take another approach to study this singularity.

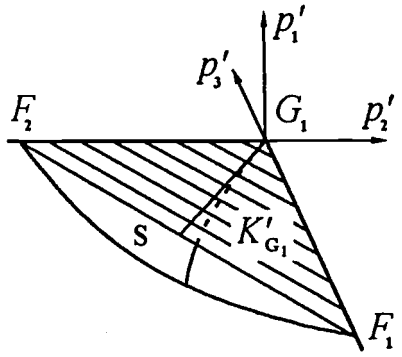


FIG. 5.1. *The singularity break of an edge.*

Let us construct a versal deformation of the matrix $A_0 = A(p_0)$ that is a smooth matrix family $A'(p')$, $p' \in \mathbf{R}^d$, such that any smooth family $A(p)$, $A(p_0) = A_0$, can be represented in a neighborhood of $p = p_0$ in the form

$$A(p) = C(p)A'(\varphi(p))C^{-1}(p), \tag{5.1}$$

where C(p) is a smooth family of nonsingular matrices, $p' = \varphi(p)$ is a smooth mapping from a neighborhood of the point p_0 in $\mathbf{R}^3$ into a neighborhood of the origin of coordinate system in $\mathbf{R}^d$, and $\varphi_1(p_0) = \varphi_2(p_0) = \cdots = \varphi_d(p_0) = 0$. A versal deformation with the minimum possible number of parameters d is called miniversal one. The miniversal deformation of the matrix A_0 can be chosen in the block diagonal form [2, 7]

$$A'(p') = A'(0) + B(p'). \tag{5.2}$$

Here $A'(0)$ is the Jordan form of A_0 and $B(p')$ is a family of block diagonal matrices whose blocks are determined in accordance with the structure of $A'(0)$. The first block of $A'(p')$ corresponding to the triple zero eigenvalue (0^3) can be taken in the form

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ p_1' & p_2' & p_3' \end{pmatrix}. \tag{5.3}$$

The other blocks correspond to eigenvalues with negative real parts. Due to (5.1) the characteristic equations for the matrices $A(p)$ and $A'(p')$, $p' = \varphi(p)$, coincide identically. Stability of the matrix $A'(p')$ in a neighborhood of the point $p' = 0$ is determined by the first block (5.3) due to its block diagonal structure. The characteristic equation of (5.3) takes the form $\lambda^3 - p_3'\lambda^2 - p_2'\lambda - p_1' = 0$. The stability domain of this equation is found using the Routh–Hurwitz conditions

$$\begin{gathered} p_1' + p_2'p_3' > 0, \\ p_1' < 0, \quad p_2' < 0, \quad p_3' < 0. \end{gathered} \tag{5.4}$$

This domain in 3-parameter space p_1', p_2', p_3' is shown in Figure 5.1. Directly from (5.4) we find that the tangent cone to the stability domain at $p' = 0$ is degenerate. It is defined by the relations

$$e_1' = 0, \quad e_2' \le 0, \quad e_3' \le 0. \tag{5.5}$$

Let us calculate the vectors h'_j, $j = 1, 2, 3$, defining collapse of a triple zero eigenvalue of the matrix $A'(0)$. Finding eigenvectors and associated vectors u'_j, v'_j, $j = 1, 2, 3$, satisfying the normalization conditions (1.21), and using (1.25) we obtain

$$h'_1 = (1, 0, 0, 0, \dots, 0)^T,$$
$$h'_2 = (0, 1, 0, 0, \dots, 0)^T,$$
$$h'_3 = (0, 0, 1, 0, \dots, 0)^T.$$

By means of these vectors the tangent cone (5.5) can be given in the form

$$K'_{G_1} = \left\{ e' \in \mathbf{R}^d : (h'_1, e') = 0,\ (h'_2, e') \le 0,\ (h'_3, e') \le 0 \right\}. \tag{5.6}$$

Let us determine a tangent cone for the family $A(p)$. For this purpose we have to find a relation between the vectors h'_j and h_j, $j = 1, 2, 3$. Let u'_j, v'_j, $j = 0, 1, 2$, be left and right eigenvectors and associated vectors of the matrix $A'(0)$, corresponding to the triple eigenvalue $\lambda_0 = 0$ and satisfying the normalization conditions (1.21). Then, using (5.1) we get relations between u'_j, v'_j, $j = 1, 2, 3$, and eigenvectors and associated vectors u_j, v_j, $j = 1, 2, 3$, of the matrix A_0

$$\begin{aligned} u_j &= C(p_0)u'_j, \\ v_j^T &= v_j'^T C^{-1}(p_0)\ , \quad j = 0, 1, 2. \end{aligned} \tag{5.7}$$

We differentiate the expression (5.1) with respect to p_j and find the value of the derivative at $p = p_0$, $p' = \varphi(p_0) = 0$,

$$\frac{\partial A}{\partial p_j} = \frac{\partial C}{\partial p_j} A' C^{-1} + C A' \frac{\partial C^{-1}}{\partial p_j} + \sum_{s=1}^{d} \left(C \frac{\partial A'}{\partial p'_s} C^{-1} \right) \frac{\partial \varphi_s}{\partial p_j}\ , \quad j = 1, 2, 3. \tag{5.8}$$

Multiplying (5.8) by v_0^T and u_0 from left and right, respectively, we have

$$\begin{aligned} h_1^j = v_0^T \frac{\partial A}{\partial p_j} u_0 &= v_0^T \frac{\partial C}{\partial p_j} A' C^{-1} u_0 \\ + v_0^T C A' \frac{\partial C^{-1}}{\partial p_j} u_0 &+ \sum_{s=1}^{d} v_0^T C \frac{\partial A'}{\partial p'_s} \frac{\partial \varphi_s}{\partial p_j} C^{-1} u_0 \\ &= \sum_{s=1}^{d} \frac{\partial \varphi_s}{\partial p_j} \left[v_0'^T \frac{\partial A'}{\partial p'_s} u'_0 \right], \quad j = 1, 2, 3. \end{aligned} \tag{5.9}$$

Here we have used the relations (5.7) and $A'(0)u'_0 = 0$, $v_0'^T A'(0) = 0$. Thus, from (5.9) we get the relation between the vectors h_1 and h'_1:

$$h_1^T = h_1'^T D_\varphi, \quad D_\varphi = \left[\frac{\partial \varphi_i}{\partial p_j} \right], \quad i = 1, 2, \dots, d, \quad j = 1, 2, 3.$$

Analogously, we can prove this relation for the vectors h_2, h'_2 and h_3, h'_3. In this proof expressions (5.7), (5.8), and the equalities

$$v_s^T \frac{\partial C}{\partial p_i} C^{-1} u_j + v_s^T C \frac{\partial C^{-1}}{\partial p_i} u_j = v_s^T \frac{\partial C C^{-1}}{\partial p_i} u_j = 0,$$
$$i = 1, 2, \dots, n, \quad s, j = 0, 1,$$

are used. Thus, the vectors h_s and h'_s are related by

$$h_s^T = h_s'^T D_\varphi, \quad s = 1, 2, 3. \tag{5.10}$$

Now we find a relation between direction vectors e and e' from the tangent cones in $\mathbf{R}^3$ and $\mathbf{R}^d$, respectively:

$$e'_i = \frac{\partial p'_i}{d\varepsilon} = \sum_{j=1}^{d} \frac{\partial \varphi_i}{\partial p_j} \frac{dp_j}{d\varepsilon} = \sum_{j=1}^{d} \frac{\partial \varphi_i}{\partial p_j} e_j, \quad i = 1, 2, \ldots, d\,, \quad j = 1, 2, 3.$$

Consequently,

$$e' = D_\varphi e. \tag{5.11}$$

Any curve $p(\varepsilon)$, $p(0) = p_0$, with the direction $e = dp/d\varepsilon$, lying in the stability domain, corresponds to a curve $p'(\varepsilon) = \varphi(p(\varepsilon))$ with the direction $e' = D_\varphi e$ lying in the stability domain in $\mathbf{R}^d$. Similarly, for any curve $p'(\varepsilon)$, $p'(0) = 0$, $dp'/d\varepsilon = e'$, lying in the stability domain in $\mathbf{R}^d$, in the case of linearly independent vectors h_s, $s = 1, 2, 3$, there exist curves $p(\varepsilon)$, $p(0) = p_0$, with directions e, related to e' by (5.11), and lying in the stability domain. In the case of the generic family of matrices $A(p)$ the vectors h_s, $s = 1, 2, 3$, are linearly independent.

Multiplying (5.10) by e and using (5.11) we obtain

$$h_s^T e = h_s'^T e', \quad s = 1, 2, 3. \tag{5.12}$$

Using (5.6) and (5.12) we find the tangent cone to the stability domain at the singular point $G_1(0^3)$ in the form

$$K_{G_1} = \big\{\, e : (h_1, e) = 0,\ (h_2, e) \le 0,\ (h_3, e) \le 0 \,\big\}. \tag{5.13}$$

The tangent cone K_{G_1} is degenerate and represents a plane angle. Recall that the vectors h_s, $s = 1, 2, 3$, determining the cone, are defined by (1.25) and need only eigenvectors and associated vectors corresponding to the triple zero eigenvalue and the derivatives of A with respect to p_j, $j = 1, 2, 3$, at the point under consideration.

Introducing the vectors g_1, g_2 by formulae

$$(g_j, h_{4-s}) = -\delta_{js}, \quad j = 1, 2, \quad s = 1, 2, 3,$$

the tangent cone K_{G_1} can be written in the form

$$K_{G_1} = \big\{\, e : e = \alpha g_1 + \beta g_2, \quad \alpha, \beta \ge 0 \,\big\},$$

where the vectors g_1, g_2 are tangent to the edges of the singularity.

6. A simple model of a damped follower force column. A simple, two-degrees-of-freedom pendulum loaded by a follower force P has been studied by Ziegler [17] and in the present extended version with two different damping parameters by Herrmann and Jong [8]. Boundary surface of the stability domain of this system was plotted and studied by Seyranian and Pedersen [15]. We consider this example from the point of view of singularities of the stability boundary and show that the effects known as destabilization due to damping [8] and uncertainty of the critical load when damping parameters tend to zero [14, 15] are closely related to the deadlock of an

edge singularity, which takes place at the point of the critical load of the system with no damping, and the dihedral angle singularity at lower values of the load.

The linearized equations of free motion of the pendulum in nondimensional variables are [8]

$$\begin{pmatrix} 3 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \ddot{\varphi}_1 \\ \ddot{\varphi}_2 \end{pmatrix} + \begin{pmatrix} \gamma_1+\gamma_2 & -\gamma_2 \\ -\gamma_2 & \gamma_2 \end{pmatrix} \begin{pmatrix} \dot{\varphi}_1 \\ \dot{\varphi}_2 \end{pmatrix} + \begin{pmatrix} 2-p & -1+p \\ -1 & 1 \end{pmatrix} \begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{6.1}$$

where γ_1 and γ_2 are independent nonnegative damping parameters and p is a magnitude of the follower force.

Introducing variables $\varphi_3 = \dot{\varphi}_1$ and $\varphi_4 = \dot{\varphi}_2$, (6.1) takes the form

$$\dot{\varphi} = A\varphi, \qquad \varphi = (\varphi_1, \varphi_2, \varphi_3, \varphi_4)^T, \tag{6.2}$$

$$A = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ p/2-3/2 & 1-p/2 & -\gamma_1/2-\gamma_2 & \gamma_2 \\ 5/2-p/2 & p/2-2 & \gamma_1/2+2\gamma_2 & -2\gamma_2 \end{pmatrix}. \tag{6.3}$$

We investigate the stability domain of the system in the space of three parameters (γ_1, γ_2, p). The characteristic equation of the system (6.2), (6.3) is

$$2\lambda^4 + (\gamma_1 + 6\gamma_2)\lambda^3 + (7 - 2p + \gamma_1\gamma_2)\lambda^2 + (\gamma_1+\gamma_2)\lambda + 1 = 0. \tag{6.4}$$

At $\gamma_1 = \gamma_2 = 0$ (the system without damping) we find

$$\lambda^2 = \frac{1}{2}\left(p - 7/2 \pm \sqrt{(p-7/2)^2 - 2}\right). \tag{6.5}$$

Hence at $p \in [0,\, 7/2 - \sqrt{2})$ we have two different pairs of simple complex conjugate imaginary eigenvalues corresponding to the dihedral angle singularity (F_3). At $p_0 = 7/2 - \sqrt{2}$ there exists a pair of double complex conjugate imaginary eigenvalues with the Jordan chain (1.10), corresponding to the singularity deadlock of an edge (G_2). Thus, the segment $\gamma_1 = \gamma_2 = 0,\ p \in [0,\, p_0]$ is an edge of the stability boundary with the deadlock at the point $p = p_0$; see Figure 6.1.

At the point $\gamma_1 = \gamma_2 = 0,\ p \in [0,\, p_0)$ the tangent cone $K_{F_3}(p)$ has been determined in (2.6). The vectors r_1 and r_2, given by (1.8), for the matrix A from (6.3) take the form

$$r_{1,2}(p) = \frac{1}{8}\begin{pmatrix} \pm\dfrac{3/2-p}{\sqrt{(p-7/2)^2-2}} - 1 \\ \pm\dfrac{19-6p}{\sqrt{(p-7/2)^2-2}} - 6 \\ 0 \end{pmatrix}, \tag{6.6}$$

where plus corresponds to r_1 and minus to r_2. The angle between r_1 and r_2 (equal to the difference of π and the angle of the dihedron) increases with the increase of p from zero and reaches π at $p = p_0$. But at $p = p_0$ the vectors r_1 and r_2 become infinite because the radicand in (6.6) is equal to zero. Therefore the tangent cone K_{F_3}

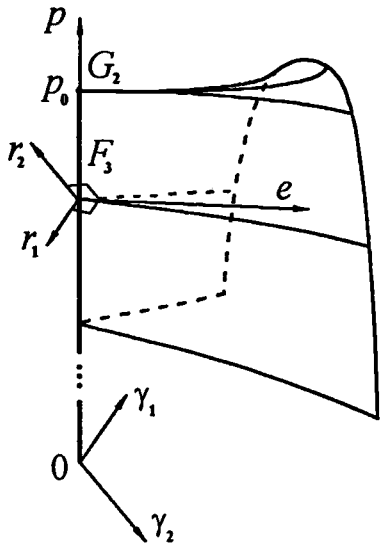


FIG. 6.1. *The stability domain of Ziegler's double pendulum.*

degenerates into the cone K_{G_2} of the deadlock of an edge singularity at $p = p_0$; see Figure 6.1.

The cone K_{G_2} has been found in (4.10), where the vectors f_1, q_1, f_2 are given by (1.17). For the matrix (6.3) they are, up to a positive factor, equal to

$$\begin{aligned} f_1 &= (0, 0, 1), \\ q_1 &= (1, -4 - 5\sqrt{2}, 0), \\ f_2 &= (-1, -6, 0). \end{aligned} \tag{6.7}$$

This cone also can be written in the form

$$K_{G_2} = \left\{ (e_1, e_2, e_3) : \ e_1 = (4 + 5\sqrt{2})e_2, \ \ e_2 \geq 0, \ \ e_3 \leq 0 \right\}. \tag{6.8}$$

In the parameter space (γ_1, γ_2, p) it represents a plane angle.

At fixed values of damping parameters γ_1, γ_2 a critical load p_{cr} is defined as the smallest value of p at which the system becomes unstable. Consider damping in the form $\gamma_1 = e_1\varepsilon$, $\gamma_2 = e_2\varepsilon$, where ε is a small positive number. Since the segment $\gamma_1 = \gamma_2 = 0$, $p \in [0, p_0]$ is the edge of the stability boundary, the limit of the critical load, when damping tends to zero, $p_0^e = \lim_{\varepsilon \to 0} p_{cr}(\gamma_1, \gamma_2)$ for a fixed direction (e_1, e_2) is equal to the value of p at which the vector $e = (e_1, e_2, 0)$ leaves the tangent cone $K_{F_3}(p)$ with the increase of p from zero. In this case either the condition $(r_1(p_0^e), e) = 0$ or $(r_2(p_0^e), e) = 0$ is fulfilled. For example, at $\gamma_1 = \varepsilon$, $\gamma_2 = 0$ we have $e = (1, 0, 0)$, $p_0^e = 2$, $r_2(2) = (0, -5/2, 0)$, $(r_2(2), e) = 0$. From this it is seen that the limit of the critical load p_0^e is different for various directions (e_1, e_2). For all $(e_1, e_2) \neq c(4+5\sqrt{2}, 1)$, $c > 0$, this limit is less then p_0. At $(e_1, e_2) = c(4+5\sqrt{2}, 1)$, $c > 0$, we have $p_0^e = p_0$. It is connected with the fact that the directions $c(4 + 5\sqrt{2}, 1, \alpha)$, $\alpha \leq 0$, $c > 0$, belong to the tangent cone K_{G_2} from (6.8).

Degeneration of the dihedral angle at the deadlock of an edge singular point geometrically illustrates the effects of destabilization of a nonconservative system by small dissipative forces [8] and uncertainty of the critical load when damping

parameters tend to zero [14]. Similar effects should be expected for other systems with deadlock of an edge and break of an edge singularities of the stability boundary.

7. Family of polynomials. Consider a linear homogeneous differential equation of the order m

$$x^{(m)} + a_1 x^{(m-1)} + \cdots + a_m x = 0 \tag{7.1}$$

whose coefficients $a_j \in \mathbf{R},\ j = 1, 2, \ldots, m$ smoothly depend on a vector of parameters $p \in \mathbf{R}^n$. The characteristic equation for (7.1) is

$$\lambda^m + a_1 \lambda^{m-1} + \cdots + a_m = 0. \tag{7.2}$$

The trivial solution of (7.1) is asymptotically stable if and only if every root of (7.2) has a negative real part $\mathrm{Re}\lambda < 0$.

The stability domain boundary of a generic one-parameter family of polynomials ($n = 1$) is characterized by a simple root $\lambda = 0$ or a pair of simple imaginary roots $\lambda = \pm i\,\omega$. The stability boundary of a generic two-parameter (three-parameter) family consists of smooth curves (surfaces), corresponding to simple roots $\lambda = 0$ or $\lambda = \pm i\,\omega$, whose only singularities are characterized by the following strata [10]:

$$\begin{aligned} n = 2: &\quad \widehat{F}_1(0^2),\ \widehat{F}_2(0, \pm i\,\omega),\ \widehat{F}_3(\pm i\,\omega_1, \pm i\,\omega_2), \\ n = 3: &\quad \widehat{F}_1(0^2),\ \widehat{F}_2(0, \pm i\,\omega),\ \widehat{F}_3(\pm i\,\omega_1, \pm i\,\omega_2),\ \widehat{G}_1(0^3), \\ &\quad \widehat{G}_2\big((\pm i\,\omega)^2\big), \widehat{G}_3(0^2, \pm i\,\omega),\ \widehat{G}_4(0, \pm i\,\omega_1, \pm i\,\omega_2), \\ &\quad \widehat{G}_5(\pm i\,\omega_1, \pm i\,\omega_2, \pm i\,\omega_3), \end{aligned} \tag{7.3}$$

where all imaginary roots at the singular point are taken in brackets with a power denoting the multiplicity of a root. Other singularities disappear under an arbitrary small deformation of the family.

Introducing the vector $y \in \mathbf{R}^m$, with the components $y_i = x^{(i-1)},\ i = 1, 2, \ldots, m$, (7.1) takes the form

$$\dot{y} = Ay, \tag{7.4}$$

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_m & -a_{m-1} & -a_{m-2} & \cdots & -a_1 \end{pmatrix}. \tag{7.5}$$

The characteristic equation for the matrix (7.5) is identically equal to (7.2); hence the stability domains for (7.1) and (7.4) coincide. For every root λ_0 of (7.2) there exists precisely one corresponding eigenvector u_0,

$$u_0 = c(1, \lambda_0, \lambda_0^2, \ldots, \lambda_0^{m-1}), \quad c = \mathrm{const}.$$

For every multiple root λ_0 there exists one corresponding Jordan chain with the length equal to the multiplicity of λ_0. It is easy to see that in the case of generic one-, two-, and three-parameter families the singularities of the stability boundary $\widehat{F}_j$,

$\widehat{G}_s$, $j = 1, 2, 3$, $s = 1, 2, \ldots, 5$, for polynomials coincide with the singularities F_j, G_s, $j = 1, 2, 3$, $s = 1, 2, \ldots, 5$, for matrices, studied above. All the results for tangent cones and other geometric characteristics of singularities, obtained for families of matrices, are directly transferred to families of polynomials. The tangent cones to the stability domain of (7.1) at singular points (7.3) are determined in (2.4)–(2.6), (4.3)–(4.5), (4.10), (5.13), where the vectors r_j, f_j, q_j, h_j are calculated by the formulae (1.8), (1.17), (1.25) for the matrix (7.5).

Let us find the expressions for these vectors by means of the coefficients a_j, $j = 1, 2, \ldots, m$, of the equation (7.1). Denoting $Q(\lambda, p) = \lambda^m + a_1\lambda^{m-1} + \cdots + a_m$ and differentiating the equation $Q(\lambda, p) = 0$ for a simple root λ ($Q = 0$, $\partial Q/\partial\lambda \neq 0$) we have

$$\frac{\partial Q}{\partial \lambda} d\lambda + (\nabla Q, dp) = 0,$$

$$\nabla\lambda = -\nabla Q \Big/ \frac{\partial Q}{\partial \lambda},$$

where ∇ is the gradient operator

$$\nabla = \left(\frac{\partial}{\partial p_1}, \frac{\partial}{\partial p_2}, \ldots, \frac{\partial}{\partial p_n}\right)^T.$$

Recall that the vector r, determined in (1.8), is the gradient of the real part of a simple eigenvalue with respect to p. Hence,

$$r = -\mathrm{Re}\left(\nabla Q \Big/ \frac{\partial Q}{\partial \lambda}\right). \tag{7.6}$$

Substituting the expression for $Q(\lambda, p)$ into (7.6) we obtain

$$r = -\mathrm{Re}\,\frac{\displaystyle\sum_{j=1}^{m} \nabla a_j \lambda^{m-j}}{m\lambda^{m-1} + \displaystyle\sum_{j=1}^{m-1} (m-j) a_j \lambda^{m-j-1}}. \tag{7.7}$$

For the simple root $\lambda = 0$ ($a_m = 0$, $a_{m-1} \neq 0$) it gives

$$r = -\frac{\nabla a_m}{a_{m-1}}. \tag{7.8}$$

Consider a double root λ_0. The vectors f_j, q_j, $j = 1, 2$, determined by (1.17), describe collapse of a double eigenvalue with the Jordan chain (1.10). Expanding $Q(\lambda, p)$ in Taylor series in the neighborhood of $\lambda = \lambda_0$, $p = p_0$ we have

$$\begin{aligned} Q(\lambda, p) = (\nabla Q, \Delta p) + \frac{1}{2}\frac{\partial^2 Q}{\partial \lambda^2}\Delta\lambda^2 + \left(\nabla \frac{\partial Q}{\partial \lambda}, \Delta p\right)\Delta\lambda \\ + \frac{1}{2}\Delta p^T \nabla\nabla^T Q \Delta p + \frac{1}{6}\frac{\partial^3 Q}{\partial \lambda^3}\Delta\lambda^3 + \cdots, \end{aligned} \tag{7.9}$$

$$\Delta\lambda = \lambda - \lambda_0, \quad \Delta p = p - p_0.$$

Here the equalities $Q = 0,\ \partial Q/\partial\lambda = 0$ at $\lambda = \lambda_0,\ p = p_0$, determining a double root, were used. We substitute the perturbation of p in the form $\Delta p = \varepsilon e + O(\varepsilon^2)$ and the expansion for λ (1.13) into (7.9) and then equate the coefficients at ε and $\varepsilon^{3/2}$ zero. As the result, for the coefficient at ε we get

$$(\nabla Q, e) + \frac{1}{2}\frac{\partial^2 Q}{\partial\lambda^2}\lambda_1^2 = 0,$$

$$\lambda_1^2 = -\frac{2(\nabla Q, e)}{\dfrac{\partial^2 Q}{\partial\lambda^2}}.$$

If $\lambda_1^2 \neq 0$, then equating the coefficient at $\varepsilon^{3/2}$ zero we obtain

$$\lambda_2 = -\frac{\left(\nabla\dfrac{\partial Q}{\partial\lambda}, e\right) + \dfrac{1}{6}\dfrac{\partial^3 Q}{\partial\lambda^3}\lambda_1^2}{\dfrac{\partial^2 Q}{\partial\lambda^2}}$$

$$= \frac{\dfrac{1}{3}\dfrac{\partial^3 Q}{\partial\lambda^3}(\nabla Q, e) - \dfrac{\partial^2 Q}{\partial\lambda^2}\left(\nabla\dfrac{\partial Q}{\partial\lambda}, e\right)}{\left(\dfrac{\partial^2 Q}{\partial\lambda^2}\right)^2}.$$

Note that $\partial^2 Q/\partial\lambda^2 \neq 0$ for a double root. Thus, for the vectors $f_j,\ q_j,\ \ j = 1, 2$, we get the following expressions:

$$f_1 + i\,q_1 = -\frac{2\nabla Q}{\dfrac{\partial^2 Q}{\partial\lambda^2}},$$

$$f_2 + i\,q_2 = \frac{\dfrac{2}{3}\dfrac{\partial^3 Q}{\partial\lambda^3}\nabla Q - 2\dfrac{\partial^2 Q}{\partial\lambda^2}\nabla\dfrac{\partial Q}{\partial\lambda}}{\left(\dfrac{\partial^2 Q}{\partial\lambda^2}\right)^2}. \tag{7.10}$$

Substituting the explicit form of $Q(\lambda, p)$ into (7.10), by analogy with (7.7), the expression (7.10) can be written by means of the coefficients $a_j(p),\ \ j = 1, 2, \ldots, m$, and λ_0. In the case of the double root $\lambda_0 = 0$ we get

$$f_1 = -\frac{\nabla a_m}{a_{m-2}},$$

$$f_2 = \frac{a_{m-3}\nabla a_m - a_{m-2}\nabla a_{m-1}}{a_{m-2}^2}. \tag{7.11}$$

The vectors $h_j,\ \ j = 1, 2, 3$, used for determining the tangent cone K_{G_1}, are needed only for $\lambda_0 = 0$. In this case $a_m = a_{m-1} = a_{m-2} = 0,\ \ a_{m-3} \neq 0$. The left and right eigenvectors and the associated vectors of the matrix (7.5), corresponding to the triple zero root and satisfying the normalization conditions (1.21), are

$$u_0 = (1, 0, 0, 0, \ldots, 0)^T,$$

$$u_1 = (0, 1, 0, 0, \dots, 0)^T,$$

$$u_2 = (0, 0, 1, 0, \dots, 0)^T,$$

$$v_0 = \left(0, 0, 1, *, \dots, *, \frac{1}{a_{m-3}}\right)^T,$$

$$v_1 = \left(0, 1, 0, *, \dots, *, -\frac{a_{m-4}}{a_{m-3}^2}\right)^T,$$

$$v_2 = \left(1, 0, 0, *, \dots, *, \frac{a_{m-4}^2 - a_{m-3}a_{m-5}}{a_{m-3}^3}\right)^T,$$

where asterisks denote the components, which don't affect the resultant expressions. Substitution of these vectors into (1.25) gives

$$h_1 = -\frac{\nabla a_m}{a_{m-3}},$$

$$h_2 = \frac{a_{m-4}\nabla a_m - a_{m-3}\nabla a_{m-1}}{a_{m-3}^2}, \tag{7.12}$$

$$h_3 = \frac{(a_{m-3}a_{m-5} - a_{m-4}^2)\nabla a_m + a_{m-3}a_{m-4}\nabla a_{m-1} - a_{m-3}^2\nabla a_{m-2}}{a_{m-3}^3}.$$

Note that in the case of the generic family of polynomials, vectors determining tangent cones to the stability domain at a singular point are linearly independent.

Example. As an example let us consider the stability problem from section 3. The characteristic equation has the form (3.3). The system is considered as dependent on two parameters R and ρ. At the point $R_0 = \sqrt{L/C}$, $\rho_0 = -\sqrt{L/C}$ the characteristic equation (3.3) has the double zero root corresponding to the singularity $\widehat{F}_1(0^2)$. The vectors f_1 and f_2, calculated with the use of (7.11), are

$$f_1 = \frac{1}{L\sqrt{LC}}\begin{pmatrix}-1\\-1\end{pmatrix}, \quad f_2 = \frac{1}{L}\begin{pmatrix}1\\-1\end{pmatrix}.$$

These expressions coincide with (3.4), where these vectors were calculated using (1.17). The tangent cone to the stability domain $K_{\widehat{F}_1}$ at the point under consideration has the form (2.4); see Figure 3.1.

Also, the results (6.6)–(6.8) on singularities of the stability boundary for Ziegler's pendulum can be derived using polynomial formulation (6.4).

8. Concluding remarks. Two methods of investigation of singularities are developed in this paper. They are constructive and convenient for numerical implementation. These methods can be applied for studying other types of singularities (also nongeneric) of the stability domain of systems depending on an arbitrary number of parameters. The first method connected with expansions of eigenvalues can be applied

if the tangent cone of the singularity does not belong to the set of directions violating the condition Γ [16] (for nonderogatory eigenvalue of the matrix A this condition is ${v_0}^T A_1 u_0 \neq 0$):

$$K \not\subset \{e : \Gamma \text{ is not satisfied}\}.$$

In this set of directions the expansions of eigenvalues in powers of $\varepsilon^{1/l}$ (l is a multiplicity) are not valid. For example, this method can be applied to the singularities, where there are only simple and double eigenvalues with zero real part. Singularities investigated in sections 2 and 4 are just of this type.

The second method connected with versal deformations can be useful for investigation of singularities determined by pure imaginary eigenvalues with higher multiplicities. The main point here is to find appropriate vectors, like h_1, h_2, h_3 in section 5, which connect parameter space of the problem with parameter space of the versal deformation.

As an example of application of the second method let us consider the singular point of the stability domain determined by zero eigenvalue with one Jordan block of the order k. Let $u_0, u_1, \ldots, u_{k-1}$ and $v_0, v_1, \ldots, v_{k-1}$ be corresponding Jordan chains of right and left eigenvectors and associated vectors satisfying normalization conditions $v_0^T u_{k-1} = 1$, $v_i^T u_{k-1} = 0$, $i = 1, \ldots, k-1$. Then introducing the vectors h_i, $i = 1, 2, \ldots, k$, with components defined by the formulae

$$h_i^j = \sum_{s=0}^{i-1} v_s^T \frac{\partial A}{\partial p_j} u_{i-s-1}, \quad i = 1, 2, ..., k, \;\; j = 1, 2, ..., n,$$

and making the same steps as in the section 5, in the case of linearly independent vectors h_i, $i = 1, 2, \ldots, k$, we get the expression for the tangent cone K_0 to the stability domain at this singular point:

$$K_0 = \{\, e : (h_1, e) = \cdots = (h_{k-2}, e) = 0,\ (h_{k-1}, e) \leq 0,\ (h_k, e) \leq 0 \,\}.$$

Note that for $k = 3$ these expressions are the same as the expressions which have been found in section 5 for the singularity G_1. Evidently, to fulfill the linear independence condition we need $n \geq k$, i.e., the dimension of the parameter space must be greater than or equal to the multiplicity of zero eigenvalue. If $n \geq k$, then the vectors h_i, $i = 1, 2, \ldots, k$, are linearly independent for the generic family of matrices.

To investigate singularities of a boundary of the stability domain of a family of polynomials, first we have to consider the singularity of the corresponding family of matrices (7.5) and then express the result in terms of the coefficients of the polynomial and their derivatives with respect to parameters.

REFERENCES

[1] A. A. Andronov, A. A. Vitt and S. E. Khaikin, *Vibration Theory*, 2d ed., Nauka, Moscow, 1981 (in Russian).

[2] V. I. Arnold, *On matrices depending on parameters*, Russian Math. Surveys, 26 (1971), pp. 29–43.

[3] V. I. Arnold, *Lectures on bifurcations in versal families*, Russian Math. Surveys, 27 (1972), pp. 119–184.

[4] V. I. Arnold, *Geometrical methods in the theory of ordinary differential equations*, Springer Verlag, New York, Berlin, 1983.

[5] V. I. Arnold, *Bifurcations and singularities in mathematics and mechanics*, in Theoretical and Applied Mechanics, P. Germain, M. Piau, and D. Caillerie, eds., North–Holland, Amsterdam, New York, 1989, pp. 1–25.

[6] J. V. Burke and M. L. Overton, *Stable perturbations of nonsymmetric matrices*, Linear Algebra Appl., 171 (1992), pp. 249–273.
[7] D. M. Galin, *On real matrices depending on parameters*, Uspekhi Math. Nauk, 27 (1972), pp. 241–242 (in Russian).
[8] G. Herrmann and I.-C. Jong, *On the destabilizing effect of damping in nonconservative elastic systems*, Trans. ASME J. Appl. Mech. Ser. E, 32 (1965), pp. 592–597.
[9] L. V. Levantovskii, *The boundary of a set of stable matrices*, Russian Math. Surveys, 35 (1980), pp. 249–250.
[10] L. V. Levantovskii, *Singularities of the boundary of the stability domain*, Funct. Anal. Appl., 16 (1982), pp. 34–37.
[11] V. B. Lidskii, *Perturbation theory of nonconjugate operators*, USSR Comput. Math. Math. Phys., 1 (1966), pp. 73–85.
[12] A. P. Seyranian, *Sensitivity analysis of multiple eigenvalues*, Mech. Structures Mach., 21 (1993), pp. 261–284.
[13] A. P. Seyranian, *Collision of eigenvalues in linear oscillatory systems*, J. Appl. Math. Mech. 58 (1994), pp. 805–813.
[14] A. P. Seyranian, *Stabilization of nonconservative systems by dissipative forces and uncertainties in the critical load*, Phys. Dokl., 41 (1996), pp. 214–217.
[15] A. P. Seyranian and P. Pedersen, *On two effects in fluid/structure interaction theory*, Proceedings of the Sixth International Conference on Fluid-Induced Vibration, P. Bearman, ed., London, 1995, pp. 565–576.
[16] M. I. Vishik and L. A. Lyusternik, *The solution of some perturbation problems for matrices and selfadjoint or non-selfadjoint differential equations*, Russian Math. Surveys, 15 (1960), pp. 1–74.
[17] H. Ziegler, *Die Stabilitätskriterien der Elastomechanik*, Ing.-Arch., 20 (1952), pp. 49–56.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 1, pp. 129–144
© 1999 Society for Industrial and Applied Mathematics

IMMANANT INEQUALITIES FOR LAPLACIANS OF TREES*

ONN CHAN† AND T. K. LAM†

Abstract. Let $\mathcal{T}_n$ be the collection of trees on n vertices. Let $\mathcal{T}_n(b;p,q)$, $\mathcal{T}_n(m;k)$, and $\mathcal{T}_n(d;k)$ be subsets of $\mathcal{T}_n$ comprising trees, each whose vertex set has bipartition (p,q), trees whose maximum matching has size k, and trees of diameter k, respectively. Brualdi and Goldwasser [*Discrete Math.*, 48 (1984), pp. 1–21] obtained lower bounds on the permanent of the Laplacian matrix of a tree from each of these subsets. They characterized the tree in each of these subsets whose Laplacian matrix has the smallest permanent as the "double star" in $\mathcal{T}_n(b;p,q)$, the "spur" in $\mathcal{T}_n(m;k)$, and the "broom" in $\mathcal{T}_n(d;k)$. In this work, the concept of vertex orientations and a new interpretation of the matching numbers in a tree allow us to formulate a unified approach to extending these results to all other immanant functions besides the permanent. It turns out that the "double star" and the "spur" remain the tree in $\mathcal{T}_n(b;p,q)$ and the tree in $\mathcal{T}_n(m;k)$, respectively, whose Laplacian matrix has the smallest immanant value for all immanants. For $\mathcal{T}_n(d;k)$ the tree that has the smallest immanant value varies with the immanant function, but it belongs to a small family of "caterpillars" whose legs are all concentrated on a single vertex.

Key words. immanants, Laplacians of trees, matchings

AMS subject classifications. 05C05, 15A15

PII. S0895479897318423

1. Introduction.

Let S_n be the permutation group on n symbols and χ_λ be an irreducible character of the symmetric group S_n, indexed by a partition λ of n. The *immanant function*, d_λ associated with the character χ_λ acting on an $n \times n$ matrix $A = [a_{ij}]$, is defined as

$$d_\lambda(A) = \sum_{\sigma \in S_n} \chi_\lambda(\sigma) \prod_{i=1}^{n} a_{i\sigma(i)}.$$

Both the determinant and permanent functions are special cases of the immanant function corresponding to the alternating and trivial characters, respectively.

Here we consider immanant functions over the family of Laplacian matrices of graphs. Much is known about the permanent of the Laplacian of a graph, and its polynomial analogues have been studied by several researchers (see, for example, [2, 9, 10, 17, 22]). Some results pertaining to the study of other immanants may be found in [1, 6, 18, 19, 26]. For example, in [19], it is shown that if $L(G)$ is the Laplacian matrix of a graph G, then

$$d_{(2,1^{n-2})}(L(G)) = 2e(G)\tau(G),$$

where $e(G)$ is the number of edges in G and $\tau(G)$ is the number of spanning trees in G. Another result of a similar flavor links a hook immanant to the sum of all distances in a tree. If T is a tree on n vertices with vertex set $V(T)$, then

$$d_{(3,1^{n-3})}(L(T)) = 4 \sum_{\substack{u,v \in V(T) \\ u \neq v}} (d(u,v) - 1),$$

*Received by the editors March 12, 1997; accepted for publication (in revised form) by R. Brualdi June 9, 1998; published electronically August 16, 1999.
http://www.siam.org/journals/simax/21-1/31842.html
†Department of Mathematics, National University of Singapore, Lower Kent Ridge Road, S119260, Republic of Singapore (matchano@leonis.nus.edu.sg, matlamtk@math.nus.sg).

where $d(u, v)$ denotes the distance between u and v, measured by the number of edges on the unique path joining u and v (see [6, Theorem 4.3]). The interested reader may consult [8, 13, 12, 14, 15, 16, 20, 21, 23, 24] for other works in this area.

Let $\mathcal{T}_n$ denote the collection of trees on n vertices. Let $\mathcal{T}_n(b; p, q)$ denote the subset of $\mathcal{T}_n$ that contains all the trees whose vertex set admits a bipartition of sizes p and q where $p + q = n$. Let $\mathcal{T}_n(m; k)$ and $\mathcal{T}_n(d; k)$ be subsets of $\mathcal{T}_n$ comprising trees whose maximum matching has size k and trees of diameter k, respectively. Brualdi and Goldwasser [2] studied the permanent of the Laplacian matrix $L(G)$ of a graph G. They obtained bounds for the permanent per $L(G)$ when G is connected and bipartite and lower bounds for per $L(T)$ in each of the three cases where $T \in \mathcal{T}_n(b; p, q)$, $T \in \mathcal{T}_n(m; k)$, and $T \in \mathcal{T}_n(d; k)$. In each of these families, there is a unique tree whose Laplacian matrix attains the smallest permanent, and in [2] they are called the "double star," the "spur," and the "broom," respectively. As the permanent is an immanant function, it is natural to seek lower bounds for all the other immanant functions for each of these three families of trees and to characterize the trees that attain the lower bounds whenever possible.

To this end, we first give a combinatorial description to the term $\prod_{i=1}^{n} a_{i\sigma(i)}$ in the expansion of the immanant of the Laplacian matrix of a graph. This is done by introducing the notion of vertex orientation, which is based on the idea of a functional digraph defined by Goulden and Jackson in [11, 12]. It turns out that the "double star" and the "spur" are still the "smallest" in $\mathcal{T}_n(b; p, q)$ and $\mathcal{T}_n(m; k)$, respectively. In the case of $\mathcal{T}_n(d; k)$, this is no longer true. The tree that attains the smallest immanant value depends on the immanant function. However we manage to show that it must come from a small collection of "caterpillars" whose legs are all concentrated on a single vertex.

It is well known that $d_{(1^n)}(L(T)) = \det L(T) = 0$. From [19, Theorem 2], we also know that $d_{(2,1^{n-2})}(L(T)) = 2(n-1)$ for all trees T with n vertices. These two immanants are constants on all trees of the same size. In this paper, we will consider only immanants which are different from these two unless otherwise stated.

2. Weighted matching and vertex orientation. Let T be a tree with vertex set $V(T) = \{v_1, v_2, \ldots, v_n\}$ and edge set $E(T)$. Its *Laplacian matrix* $L(T) = [l_{ij}]$ is defined by

$$l_{ij} := \begin{cases} \deg_T(v_i) & \text{if } i = j, \\ -1 & \text{if } \{v_i, v_j\} \in E(T), \\ 0 & \text{otherwise,} \end{cases}$$

where $\deg_T(v_i)$ is the degree of the vertex v_i in T. We shall write $\deg(v)$ in place of $\deg_T(v)$ when the tree T being discussed is obvious from the context.

As observed in [1], in the expansion of $d_\lambda(L(T)) = \sum_{\sigma \in S_n} \chi_\lambda(\sigma) \prod_{i=1}^{n} l_{i\sigma(i)}$, many of the terms are zero. The only instances when the term $\prod_{i=1}^{n} l_{i\sigma(i)} \neq 0$ are when $\{v_i, v_{\sigma(i)}\} \in E(T)$ for every i such that $i \neq \sigma(i)$. This induces us to define the weighted j-matching numbers.

A *j-matching* $M \subseteq E(T)$ is a subset of j independent edges in T. Two edges are said to be independent if they are not incident with a common vertex. In Figure 2.1 below, $\{\{v_1, v_6\}, \{v_3, v_4\}\}$ is a 2-matching, while $\{\{v_0, v_1\}, \{v_2, v_3\}, \{v_4, v_5\}\}$ is a 3-matching.

We use the notation $v \in M$ to mean that the vertex v is incident with some edge in M and $v \notin M$ to mean otherwise. For $j = 0, 1, \ldots, \lfloor \frac{n}{2} \rfloor$, the *weighted j-matching*

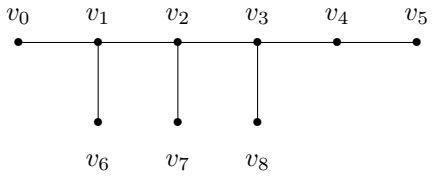


FIG. 2.1.

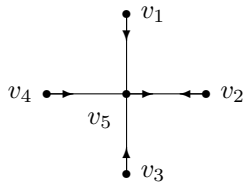


FIG. 2.2.

number of T is defined to be

$$m_T(j) = \sum_M \prod_{v \notin M} \deg(v),$$

where the sum is taken over all possible j-matchings M of the tree T. Note that $m_T(0) = \prod_{v \in V(T)} \deg(v)$ and, if M is a total matching, $m_T(j)$ is taken to be 1. Set $\chi_\lambda(j) := \chi_\lambda(\sigma)$ where σ has cycle type $(2^j, 1^{n-2j})$. It is straightforward to check that

$$d_\lambda(L(T)) = \sum_{j=0}^{\lfloor \frac{n}{2} \rfloor} \chi_\lambda(j) m_T(j).$$

This derivation in terms of matchings was used for immanantal polynomials [1, equation (4)].

We now describe another graph-theoretic invariant that is closely related to the j-matching numbers. Let T be a tree. For each vertex v, we assign an arrow pointing away from the vertex along one of its adjacent edges. There are $\deg(v)$ ways to do this. This is essentially the same as a functional digraph defined in [12, section 4] and [11, Proposition 3.3.11]. However, in our context, the resulting graph is not a digraph, as not all the edges will have arrows on them, and there are some edges with two arrows on them. We call such an arrangement with exactly j edges having two arrows on them a *j-vertex orientation*. These edges can be viewed as directed cycles of length 2. For $j = 1, 2, \ldots, \lfloor \frac{n}{2} \rfloor$, define $a_T(j)$ as the number j-vertex orientations. We also define $a_T(0) = 0$.

Example. Consider the star $S(5)$ on five vertices as shown in Figure 2.2.

There are four possible arrangements of arrows for $S(5)$, but they all resemble the 1-vertex orientation shown above. So, $a_{S(5)}(1) = 4$ and $a_{S(5)}(j) = 0$ for all $j \neq 1$.

In general, the star $S(n)$ is the tree on n vertices with one central vertex of degree $n-1$ and the remaining vertices of degree 1 joined to the central vertex. A simple calculation gives

$$m_{S(n)}(0) = m_{S(n)}(1) = n - 1 \quad \text{and} \quad m_{S(n)}(j) = 0 \text{ for } j > 1$$

and

$$a_{S(n)}(1) = n-1 \quad \text{and} \quad a_{S(n)}(j) = 0 \text{ for } j \neq 1. \tag{2.1}$$

Observe that there are $\deg(v)$ ways to assign an arrow for the vertex v. Thus the product of the degrees of the vertices in a tree gives the total number of all the vertex orientations of the tree. For a tree T with n vertices, we have

$$\sum_{j=0}^{\lfloor \frac{n}{2} \rfloor} a_T(j) = \prod_{v \in V(T)} \deg(v) = m_T(0).$$

The relation between $m_T(j)$ and $a_T(j)$ is made more explicit below.

THEOREM 2.1. *Let T be a tree with n vertices. Then*

$$m_T(j) = \sum_{k=j}^{\lfloor \frac{n}{2} \rfloor} \binom{k}{j} a_T(k), \tag{2.2}$$

$$a_T(j) = \sum_{k=j}^{\lfloor \frac{n}{2} \rfloor} (-1)^{k-j} \binom{k}{j} m_T(k) \tag{2.3}$$

for $j = 0, 1, \ldots, \lfloor \frac{n}{2} \rfloor$.

Proof. We will prove only (2.2) since (2.3) follows easily from it. This is done by counting the set

$$Q_j = \{(M, O) : M \text{ is a } j\text{-matching}, O \text{ is a } k\text{-vertex orientation with } k \geq j \text{ such that the edges in } M \text{ all have two arrows in } O\}.$$

Given a j-matching M, we can form a k-vertex orientation O by directing the vertices $v \in M$ along the edges $e \in M$. The other vertices are allowed to point in any direction. Clearly, $(M, O) \in Q_j$. Therefore,

$$|Q_j| = \sum_M \prod_{v \notin M} \deg(v) = m_T(j),$$

where the sum is over all j-matchings.

Given a k-vertex orientation O where $j \leq k \leq \lfloor \frac{n}{2} \rfloor$, we note that the k edges which have two arrows on them are independent. We can form an ordered pair (M, O) by choosing j edges from them. This gives

$$m_T(j) = |Q_j| = \sum_{k=j}^{\lfloor \frac{n}{2} \rfloor} \binom{k}{j} a_T(k),$$

proving (2.2). ∎

This relation allows us to express $d_\lambda(L(T))$ in terms of $a_T(j)$. So,

$$\begin{aligned} d_\lambda(L(T)) &= \sum_{j=0}^{\lfloor \frac{n}{2} \rfloor} \chi_\lambda(j) m_T(j) \\ &= \sum_{j=0}^{\lfloor \frac{n}{2} \rfloor} \chi_\lambda(j) \sum_{k=j}^{\lfloor \frac{n}{2} \rfloor} \binom{k}{j} a_T(k) \\ &= \sum_{k=0}^{\lfloor \frac{n}{2} \rfloor} a_T(k) \sum_{j=0}^{k} \chi_\lambda(j) \binom{k}{j}. \end{aligned} \tag{2.4}$$

It is natural to examine the properties of the sum $\sum_{j=0}^{k} \chi_\lambda(j)\binom{k}{j}$ since $d_\lambda(L(T))$ depends on them. We do this in the next section.

3. Binomial sums of characters. It is well known that $\sum_{j=0}^{n}\binom{n}{j} = 2^n$ and $\sum_{j=0}^{n}(-1)^j\binom{n}{j} = 0$. These two sums are special cases of the more general sum

$$\sum_{j=0}^{n} \chi_\lambda(j)\binom{n}{j}.$$

It turns out that these sums are always nonnegative.

LEMMA 3.1 (see [4, Theorem 2.2]). *Let χ_λ be an irreducible character of S_m where $m \geq 2n$. The sum*

$$\sum_{j=0}^{n} \chi_\lambda(j)\binom{n}{j} \geq 0.$$

Define $l(\lambda)$ to be the number of parts in the partition λ. The sum is exactly 0 if and only if $m - l(\lambda) < n$.

Proof. We induct on n. When $n = 1$, the sum on the left is $\chi_\lambda(0) + \chi_\lambda(1)$. An application of the Murnaghan–Nakayama rule (see, for example, page 177 in [25]) shows that this sum is nonnegative since

$$\chi_\lambda(0) + \chi_\lambda(1) \;=\; 2\sum_{\mu\in I(\lambda)} \chi_\mu(0) \;+\; 2\sum_{\mu\in K(\lambda)} \chi_\mu(0) \;\geq\; 0,$$

where $I(\lambda)$ and $K(\lambda)$ are the sets of partitions of $m-2$ that are obtained from λ by removing a rim hook of the form $\square\square$ and two nonadjacent rim hooks of size 1, respectively, from the Young tableau associated with λ. Furthermore, the sum is zero if and only if $\lambda = (1^m)$ where $m \geq 2$. Note that $\lambda = (1^m)$ is the only partition with $l(\lambda) = m$, thus satisfying the stated condition.

For bigger values of n, we use the Murnaghan–Nakayama rule again to get

$$\begin{aligned}
&\sum_{j=0}^{n} \chi_\lambda(j)\binom{n}{j} \\
=\;& \sum_{j=0}^{n} \chi_\lambda(j)\left(\binom{n-1}{j-1} + \binom{n-1}{j}\right) \\
=\;& \sum_{j=0}^{n-1} (\chi_\lambda(j) + \chi_\lambda(j+1))\binom{n-1}{j} \\
=\;& \sum_{j=0}^{n-1} \left(\sum_{\mu\in I(\lambda)} \chi_\mu(j) + \sum_{\mu\in J(\lambda)} \chi_\mu(j) + 2\sum_{\mu\in K(\lambda)} \chi_\mu(j) + \sum_{\mu\in I(\lambda)} \chi_\mu(j) - \sum_{\mu\in J(\lambda)} \chi_\mu(j)\right) \\
\times\;& \binom{n-1}{j} \\
=\;& 2\sum_{\mu\in I(\lambda)} \sum_{j=0}^{n-1} \chi_\mu(j)\binom{n-1}{j} + 2\sum_{\mu\in K(\lambda)} \sum_{j=0}^{n-1} \chi_\mu(j)\binom{n-1}{j},
\end{aligned}$$

where $I(\lambda), J(\lambda), K(\lambda)$ are the sets of partitions of size $m-2$ that are obtained from λ by removing a rim hook of the form $\square\square$, a rim hook of the form $\begin{smallmatrix}\square\\\square\end{smallmatrix}$, and two nonadjacent rim hooks of size 1, respectively. In the penultimate line above, the first

three terms come from the term $\chi_\lambda(j)$ of the previous line, while the last two terms come from the term $\chi_\lambda(j+1)$ of the previous line. Since all the irreducible characters χ_μ that appear in the final expression are characters of S_{m-2} and $m-2 \geq 2(n-1)$, we can apply the induction hypothesis to the sums in the expression above and obtain

$$\sum_{j=0}^{n} \chi_\lambda(j) \binom{n}{j} \geq 0.$$

For the second half of the lemma, note that for all $\mu \in I(\lambda)$ and $K(\lambda)$, $l(\mu) = l(\lambda) - 1$ or $l(\lambda)$. If λ is such that $m - l(\lambda) < n$, then $m - 2 - l(\mu) < n - 1$. By the induction hypothesis, $\sum \chi_\mu(j)\binom{n-1}{j} = 0$ for all such $\mu \in I(\lambda) \cup K(\lambda)$. Therefore, $\sum \chi_\lambda(j)\binom{n}{j} = 0$ for those λ where $m - l(\lambda) < n$.

In the case when λ satisfies the condition $m - l(\lambda) \geq n$, we claim that we can always find some $\mu \in I(\lambda) \cup K(\lambda)$ such that $l(\mu) \leq m - n - 1$. Clearly, this is true for all $\mu \in I(\lambda) \cup K(\lambda)$ if $l(\lambda) \leq m - n - 1$. If $l(\lambda) = m - n$, then $l(\lambda) \geq \frac{m}{2}$. This forces λ to have at least $\lceil \frac{m}{2} \rceil$ parts and smallest part of size 1 or 2. However, this would mean that $I(\lambda)$ or $K(\lambda)$ contains a partition μ such that $l(\mu) = l(\lambda) - 1$, hence using the recurrence relation

$$\sum_{j=0}^{n} \chi_\lambda(j) \binom{n}{j} \geq 2 \sum_{j=0}^{n} \chi_\mu(j) \binom{n-1}{j} > 0$$

by the induction hypothesis. This proves the lemma. □

We state the special case of the theorem for $n = 2$ below, as we will need it later.

COROLLARY 3.2. *For irreducible characters of $m \geq 4$, the sum $\chi_\lambda(0) + 2\chi_\lambda(1) + \chi_\lambda(2)$ is zero if and only if $\lambda = (1^m), (2, 1^{m-2})$.*

Improved lower bounds on the quantity $\sum_{j=0}^{n} \chi_\lambda(j)\binom{n}{j}$ including a combinatorial formula for the exact value of $\chi_\lambda(0) + 2\chi_\lambda(1) + \chi_\lambda(2)$ may be found on page 48 in [26]; see also [5].

4. The family of trees $\mathcal{T}_n$. In this section, we show that the lower bound on the immanant of the Laplacian matrix of a tree is attained by the star. In [8], the proof of this result is based on a reduction argument on the diameter of the tree. Here we present a simpler proof using vertex orientations.

LEMMA 4.1. *Let T be a tree with n vertices. Then*

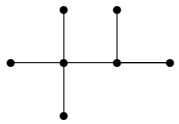
$$a_T(1) = n - 1.$$

Proof. Given an edge e of T, we point all arrows toward e. This gives an arrangement of arrows with exactly one edge, namely, e, with two arrows. Clearly, any arrangement that has this property must be of this form. Since there are $n-1$ edges, $a_T(1) = n - 1$. □

THEOREM 4.2 (see [8, Theorem 5]). *Let $S(n)$ denote the star on n vertices and let T be any tree on n vertices which is not the star. Then*

$$d_\lambda(L(T)) > d_\lambda(L(S(n))).$$

Proof. From Lemma 4.1, $a_T(1) = n - 1$ and since T is not the star, T admits a 2-vertex orientation. From Corollary 3.2, we also know that $\sum_{j=0}^{2} \chi_\lambda(j)\binom{2}{j} > 0$ when



$D(4,3)$

FIG. 5.1.

$\lambda \neq (1^n), (2,1^{n-2})$. Hence

$$\begin{aligned} d_\lambda(L(T)) &= \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} a_T(k) \sum_{j=0}^{k} \chi_\lambda(j) \binom{k}{j} \\ &> (n-1)(\chi_\lambda(0) + \chi_\lambda(1)) \\ &= d_\lambda(L(S(n))) \end{aligned}$$

using the values of $a_{S(n)}(j)$ given by (2.1). This completes the proof. □

When $\lambda = (n)$, both $\chi_\lambda(0)$ and $\chi_\lambda(1)$ are equal to 1. We obtain as a corollary the lower bound on the permanent of all trees given in [17].

COROLLARY 4.3 (see [17]). *Let T be a tree with n vertices. Then*

$$\operatorname{per} L(T) \geq 2(n-1)$$

with equality if and only if T is the star $S(n)$.

5. Trees with given bipartition $\mathcal{T}_n(b; p, q)$. Since a tree is also a bipartite graph, we can divide the vertices into two sets V_1 and V_2 so that every edge has one end in V_1 and the other in V_2. If we let $p = |V_1|$ and $q = |V_2|$, we then say that the tree T has a (p,q)-*bipartition*.

Take the star $S(p+1)$ and append $q-1$ vertices to one of the leaves. Call this new tree the *double star* and denote it by $D(p,q)$. For example, the double star $D(4,3)$ is shown in Figure 5.1.

It can be shown that $a_{D(p,q)}(1) = n-1$, $a_{D(p,q)}(2) = (p-1)(q-1)$, and $a_{D(p,q)}(j) = 0$ for all $j > 2$.

LEMMA 5.1. *Let T be a tree on n vertices which has a (p,q)-bipartition. Then*

$$a_T(2) \geq (p-1)(q-1).$$

Furthermore, equality holds if and only if $T = D(p,q)$.

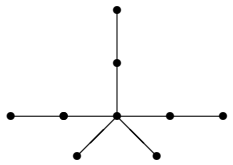
Proof. Let V_1, V_2 be the partition of the vertices of the tree T corresponding to the (p,q) bipartition. It is obvious that $\sum_{v \in V_1} \deg(v) = \sum_{v \in V_2} \deg(v) = n-1$.

For any tree T, the number of 2-matchings is given by the formula

$$\binom{n-1}{2} - \sum_{v \in V_1} \binom{\deg(v)}{2} - \sum_{v \in V_2} \binom{\deg(v)}{2}.$$

The number of 2-vertex orientations is greater than or equal to the number of 2-matchings. A simple exercise shows that

$$a_T(2) \geq \binom{n-1}{2} - \sum_{v \in V_1} \binom{\deg(v)}{2} - \sum_{v \in V_2} \binom{\deg(v)}{2}$$



$S(9,4)$

FIG. 6.1.

$$\geq \binom{n-1}{2} - \binom{n-1-p+1}{2} - \binom{n-1-q+1}{2}$$
$$= (p-1)(q-1)$$

with equality if and only if the degrees of the vertices in V_1 are $q, 1, 1, \ldots, 1$ and the degrees of the vertices in V_2 are $p, 1, 1, \ldots, 1$. This means equality holds if and only if T is the double star $D(p,q)$. □

We are now ready to determine the lower bound of $d_\lambda(L(T))$ when $T \in \mathcal{T}_n(b; p, q)$.

THEOREM 5.2. *Let T be a tree with n vertices which has a (p,q)-bipartition, $n = p+q$, and T is not the double star $D(p,q)$. Then*

$$d_\lambda(L(T)) > d_\lambda(L(D(p,q))).$$

Proof. The proof is similar to Theorem 4.2. We use Corollary 3.2 and Lemma 5.1:

$$d_\lambda(L(T)) = \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} a_T(k) \sum_{j=0}^{k} \chi_\lambda(j) \binom{k}{j}$$
$$> (n-1)(\chi_\lambda(0) + \chi_\lambda(1)) + (p-1)(q-1)(\chi_\lambda(0) + 2\chi_\lambda(1) + \chi_\lambda(2))$$
$$= d_\lambda(L(D(p,q))). \quad \square$$

In the special case of the permanent, a simple calculation gives

$$\begin{aligned} \operatorname{per} L(D(p,q)) &= (n-1)2 + (p-1)(q-1)4 \\ &= (2p-1)(2q-1) + 1. \end{aligned}$$

This gives a lower bound on the permanent of the Laplacian matrix of a tree with a (p,q)-bipartition.

COROLLARY 5.3 (see [2, Theorem 2.16]). *Let T be a tree with a (p,q)-bipartition. Then*

$$\operatorname{per} L(T) \geq (2p-1)(2q-1) + 1$$

with equality if and only if T is a double star $D(p,q)$.

6. Trees with maximum matching of size k, $\mathcal{T}_n(m; k)$. Take the star $S(n-k+1)$ and add a leaf to each $k-1$ of its leaves. The resulting tree is called a *spur* and we denote it by $S(n,k)$. The spur $S(9,4)$ is shown in Figure 6.1.

Let $\alpha'(T)$ denote the size of a maximum matching in T. Note that $S(n,1)$ is just the star $S(n)$ and $\alpha'(S(n,k)) = k$.

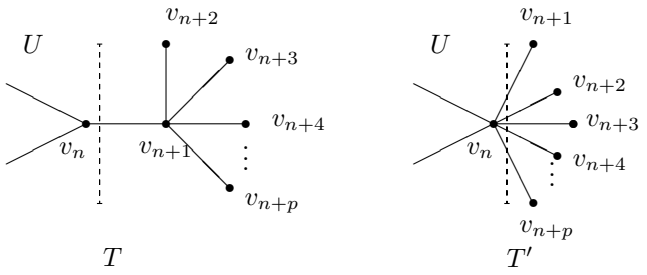


FIG. 6.2.

It can be shown that for $j = 1, 2, \ldots, k$,

$$a_{S(n,k)}(j) = (n + 1 - 2j)\binom{k-1}{j-1}$$

and $a_{S(n,k)}(j) = 0$ for $j > k$. The set of j-orientations of $S(n, k)$ can be partitioned into two types: those in which the central vertex is not incident with an edge having two arrows and those in which the central vertex is incident with an edge having two arrows. The number of j-orientations of the former type is $j\binom{k-1}{j}$ and the number of the latter type is $((n-k)-(j-1))\binom{k-1}{j-1}$. The sum of these two numbers equals $a_{S(n,k)}(j)$.

Again, we will use the same idea of showing that these are lower bounds for $a_T(j)$ for trees T such that $\alpha'(T) = k$. We will first prove a recursive relation on the $a_T(j)$'s. The analogue of this result for weighted j-matching numbers may be found in [8, Lemma 4] and [6, Lemma 5.4].

LEMMA 6.1. *Let U be a tree with n vertices. Let T and T' be trees with $n + p$ vertices where $p \geq 2$ and both trees contain U as a common subtree as shown in Figure* 6.2.

Then for $j = 1, 2, \ldots, \lfloor \frac{n}{2} \rfloor + 1$,

$$a_T(j) = a_{T'}(j) + (p-1)a_U(j-1).$$

Proof. We give a combinatorial proof of this result. Consider the j-vertex orientations of T.

Case 1. The arrow from v_n points toward v_{n+1} and for some $i = 2, 3, \ldots, p$, the edge $\{v_{n+1}, v_{n+i}\}$ has two arrows. We associate the j-vertex orientation with a j-vertex orientation of T' where $\{v_n, v_{n+i}\}$ has two arrows and all the arrows in the common subtree U remain the same.

Case 2. The edge $\{v_n, v_{n+1}\}$ has two arrows. We associate this j-vertex orientation of T with a j-vertex orientation of T' where $\{v_n, v_{n+1}\}$ has two arrows and the rest remains the same.

Case 3. The arrow from v_n points away from v_{n+1} and none of the edges $\{v_{n+1}, v_{n+i}\}$ has two arrows for some $i = 2, \ldots, p$. Here, the arrow from v_{n+1} points toward v_n. We associate this j-vertex orientation with a j-vertex orientation of T' where the arrows point from v_{n+i} to v_n for all $i = 1, 2, \ldots, p$, and the arrows in the common subtree U remain the same.

Case 4. The arrow from v_n points away from v_{n+1} and for some $i = 2, 3, \ldots, p$, the edge $\{v_{n+1}, v_{n+i}\}$ has two arrows. We associate this j-vertex orientation with a $(j-1)$-vertex orientation of U and take note of the number i.

The number of j-vertex orientations in Cases 1, 2, and 3 combined is exactly $a_{T'}(j)$, and the number in Case 4 is $(p-1)a_U(j-1)$. Therefore,

$$a_T(j) = a_{T'}(j) + (p-1)a_U(j-1). \qquad \square$$

Remarks.

1. If the subtree U has one or more edges, that is, if T is not the star, then $a_T(j) > a_{T'}(j)$ for all $j = 2, 3, \ldots, \alpha'(T)$. By Lemma 3.1 and formula (2.4), it follows that $d_\lambda(L(T)) > d_\lambda(L(T'))$ whenever T is not the star.
2. The construction of T' from T is a useful one and we shall give it a name to facilitate discussion later on. A vertex v in T attached to $\deg(v) - 1$ leaves is called a *penultimate vertex*. In the construction we shall say that T' is the contraction of T via the penultimate vertex v_{n+1} and write $T' = \mathrm{co}(T, v_{n+1})$.

LEMMA 6.2. *Let T be a tree with a matching of maximum size $\alpha'(T) = k$. Then*

$$a_T(j) \geq (n+1-2j)\binom{k-1}{j-1} = a_{S(n,k)}(j)$$

for $j = 1, 2, \ldots, k$.

Furthermore for $j = 2, \ldots, k$, equality holds if and only if T is the spur $S(n, k)$.

Proof. We prove this by induction on $i(T)$ where $i(T)$ is the number of vertices in T of degree greater than 1. The "i" in $i(T)$ stands for "interior vertices" of T. The base case is when $i(T) = 1$, that is, when T is the star $S(n)$. In this case $k = 1$ and the inequality is true since $a_{S(n)} = n-1$. Recall that the star $S(n)$ is the spur $S(n, 1)$.

If $i(T) > 1$, then T is not the star $S(n)$. We can find a penultimate vertex v in T. That is to say, T has the form given in Lemma 6.1. Let $T' = \mathrm{co}(T, v)$ as in Lemma 6.1. By Lemma 6.1

$$a_T(j) = a_{T'}(j) + (p-1)a_U(j-1),$$

where U is the common subtree obtained from T by deleting v and all its leaves. Note that T' has the same number of vertices as T but $i(T') = i(T) - 1$. We split the induction step into two cases.

Case 1. $\alpha'(T') = k = \alpha'(T)$. Then by the remarks following Lemma 6.1 and the induction hypothesis,

$$a_T(j) > a_{T'}(j) \geq (n+1-2j)\binom{k-1}{j-1} = a_{S(n,k)}(j).$$

Note that the inequality is strict for $j = 2, 3, \ldots, k$. This shows that T is not the spur $S(n, k)$.

Case 2. $\alpha'(T') = k - 1$. It follows that $\alpha'(U) = k - 1$ as well. Also $i(U) \leq i(T')$. So,

$$\begin{aligned}
a_T(j) &= a_{T'}(j) + (p-1)a_U(j-1) \\
&\geq (n+1-2j)\binom{k-2}{j-1} + (p-1)(n-p+1-2(j-1))\binom{k-2}{j-2} \\
&= (n+1-2j)\binom{k-2}{j-1} + (n+1-2j)\binom{k-2}{j-2} + (p-2)(n-p-2(j-1))\binom{k-2}{j-2} \\
&\geq (n+1-2j)\binom{k-2}{j-1} + (n+1-2j)\binom{k-2}{j-2} \\
&= (n+1-2j)\binom{k-1}{j-1}.
\end{aligned}$$

In this case, equality holds if and only if $p = 2$ and both T' and U are the spurs $S(n, k-1)$ and $S(n-2, k-1)$, respectively. This means that T has to be the spur $S(n, k)$. □

Basically, Lemma 6.2 says that among all trees in $\mathcal{T}_n(m; k)$, $a_T(j)$ is smallest when T is $S(n, k)$.

THEOREM 6.3. *Let T be a tree on n vertices with a maximum matching of size k and where T is different from the spur $S(n, k)$. Then*

$$d_\lambda(L(T)) > d_\lambda(L(S(n,k))).$$

Proof. Using Lemma 6.2 and Corollary 3.2, we see that if T is not the spur $S(n, k)$,

$$\begin{aligned} d_\lambda(L(T)) &= \sum_{h=1}^{\lfloor \frac{n}{2} \rfloor} a_T(h) \sum_{j=0}^{h} \chi_\lambda(j) \binom{h}{j} \\ &> \sum_{h=1}^{\lfloor \frac{n}{2} \rfloor} a_{S(n,k)}(h) \sum_{j=0}^{h} \chi_\lambda(j) \binom{h}{j} \\ &= d_\lambda(L(S(n,k))). \end{aligned}$$ □

The permanent for the spur $S(n, k)$ can be easily computed using the values for $a_{S(n,k)}(j)$:

$$\begin{aligned} \text{per } L(S(n,k)) &= \sum_{j=1}^{k} (n+1-2j) \binom{k-1}{j-1} 2^j \\ &= 3^{k-2}(6n-8k+2). \end{aligned}$$

The lower bound for the permanent in this case is thus attained by the spur $S(n, k)$.

COROLLARY 6.4 (see [2, Theorem 2.9]). *Let T be a tree with n vertices and a k-matching where $k \geq 1$. Then*

$$\text{per } L(T) \geq 3^{k-2}(6n-8k+2),$$

with equality if and only if T is the spur $S(n, k)$.

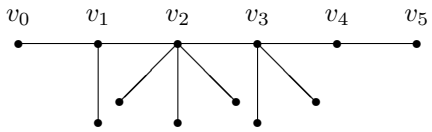
7. Trees with diameter k, $\mathcal{T}_n(d; k)$. From the previous sections, we see that in $\mathcal{T}_n$, $\mathcal{T}_n(b; p, q)$, and $\mathcal{T}_n(m, k)$ the minimum immanant value is attained by the same tree independent of the character chosen. However, this property no longer holds for $\mathcal{T}_n(d; k)$. This makes the characterization of the tree having the minimum immanant value more challenging.

Let P be a path of length k. We label the vertices of P by $v_0, v_1, v_2, \ldots, v_k$. Then a *caterpillar* of diameter k can be described as the path P with leaves (or legs) attached to the vertices $v_1, v_2, \ldots, v_{k-1}$.

Note that in the caterpillar, $\deg(v_0) = \deg(v_k) = 1$ and $\deg(v_i) \geq 2$ for $i = 1, 2, \ldots, k-1$. Since it depends only on $\deg(v_i)$ where $i = 1, 2, \ldots, k-1$, we denote it by

$$C(k; l_1, l_2, \ldots, l_{k-1}),$$

where $l_i = \deg(v_i) - 2$ is the number of legs at vertex v_i. In Figure 7.1, we show the caterpillar $C(5; 1, 3, 2, 0)$.



$C(5; 1, 3, 2, 0)$

FIG. 7.1.

Let $T \in \mathcal{T}_n(d; k)$ and P be a path $v_0, v_1, \ldots, v_k$ in T of length k. Suppose T is not a caterpillar. Then it has a penultimate vertex, say, v, that is neither v_1 nor v_{k-1}. Let $T' = \mathrm{co}(T, v)$ be the contraction of T via v as in Lemma 6.1. Then $d_\lambda(L(T')) < d_\lambda(L(T))$ and T' still has diameter k. By using the contraction process repeatedly, we see that the tree in $\mathcal{T}_n(d; k)$ that has the smallest immanant value must be a caterpillar. We summarize this in a lemma.

LEMMA 7.1. *Let $T \in \mathcal{T}_n(d; k)$. If T is not a caterpillar, then there exists a caterpillar C such that*

$$d_\lambda(L(T)) > d_\lambda(L(C)).$$

In fact, as we shall see later, the caterpillar that is minimum must have all its legs concentrated at a vertex.

For a caterpillar $C = C(k; l_1, l_2, \ldots, l_{k-1})$, the numbers $a_C(j)$'s can be expressed as polynomials in l_i's. For example, a direct calculation gives

$$\begin{aligned} a_C(1) &= \sum_{i=1}^{k-1} l_i + k, \\ a_C(2) &= \sum_{j=2}^{k-2} (l_1+1) l_j (j-1) + \sum_{1<i<j<k-1} l_i l_j (j-i) + \sum_{i=2}^{k-2} l_i (l_{k-1}+1)(k-1-i) \\ &\quad + (l_1+1)(l_{k-1}+1)(k-2) + (l_1+1)\binom{k-1}{2} + \sum_{i=2}^{k-2} l_i \left(\binom{i}{2} + \binom{k-1-i}{2} \right) \\ &\quad + (l_{k-1}+1)\binom{k-1}{2} + \binom{k-2}{3}. \end{aligned} \tag{7.1}$$

We say that a polynomial is *square-free* if the highest power of any of its variables is not greater than 1. So both $a_C(1)$ and $a_C(2)$ are square-free polynomials in the variables l_i. This is also true of other $a_C(j)$.

LEMMA 7.2. *Let C be the caterpillar $C(k; l_1, l_2, \ldots, l_{k-1})$. Then we have that for $j = 1, 2, \ldots, k-1$, $a_C(j)$ is a square-free polynomial in the variables $l_1, l_2, \ldots, l_{k-1}$ with nonnegative integer coefficients.*

Proof. We will prove a stronger statement that $a_C(j)$ is a square-free polynomial in the variables $l_1+1, l_2, \ldots, l_{k-2}, l_{k-1}+1$. We prove this by induction on the diameter k. When $k = 2$, $C(2; l_1)$ is just the star with $l_1 + 3$ vertices and $a_C(1) = l_1 + 2$.

For larger values of k, let C' denote the caterpillar $C(k-1; l_1, \ldots, l_{k-3}, l_{k-2} + (l_{k-1}+1))$ and C'' denote the caterpillar $C(k-1; l_1, \ldots, l_{k-3}, l_{k-2}-1)$. In the case $l_{k-2} = 0$, we let C'' be the caterpillar $C(k-2; l_1, \ldots, l_{k-3})$. Then we can apply Lemma 6.1 to C and C' to obtain

$$a_C(j) = a_{C'}(j) + (l_{k-1}+1) a_{C''}(j-1).$$

Since C' and C'' have shorter diameters, by the induction hypothesis, $a_{C'}(j)$ and $a_{C''}(j)$ can be expressed in terms of $l_1+1, l_2, \ldots, l_{k-2}, l_{k-1}+1$ and $l_1+1, l_2, \ldots, l_{k-2}$, respectively, with nonnegative integer coefficients. Both these expressions are square-free. Hence, $a_C(j)$ is a square-free polynomial in the variables $l_1+1, l_2, \ldots, l_{k-2}, l_{k-1}+1$ with nonnegative integer coefficients. □

We will make use of this to prove Theorem 7.3.

THEOREM 7.3. *Let* $C[n;k;i]$ *denote the caterpillar* $C(k;\underbrace{0,\ldots,0}_{i-1},n-k-1,\underbrace{0,\ldots,0}_{k-1-i})$. *For any tree* T *with order* n *and diameter* k,

$$d_\lambda(L(T)) \geq d_\lambda(L(C[n;k;i]))$$

for some $i = 1, 2, \ldots, \lfloor \frac{k}{2} \rfloor$ *where* i *depends on* λ. *Furthermore, the inequality is strict if* T *is not* $C[n;k;j]$ *for* $j = 1, 2, \ldots, k-1$.

Proof. Let $C[n;k;i]$ denote the caterpillar $C(k;\underbrace{0,\ldots,0}_{i-1},n-k-1,\underbrace{0,\ldots,0}_{k-1-i})$. From Lemma 7.1, it suffices to consider the smaller set of caterpillars C with diameter k and $C \neq C[n;k;i]$ for all i. Using Lemma 7.2 and (2.4), we can express $d_\lambda(L(C))$ as a square-free polynomial in the variables l_i with nonnegative coefficients. Let $t_{j;i_1,i_2,\ldots,i_h}$ denote the coefficient of $l_{i_1}l_{i_2}\cdots l_{i_h}$ in $a_C(j)$ where $1 \leq i_1 < i_2 < \cdots < i_h \leq k-1$. The constant will be denoted by t_j. Note that $t_{j;i_1,i_2,\ldots,i_h}$ is independent of the values of l_i. Also from (7.1), $t_{2;i_1,i_2} > 0$ for all i_1, i_2. Then

$$\begin{aligned}
d_\lambda(L(C)) &= \sum_g a_C(g) \sum_{j=0}^{g} \chi_\lambda(j) \binom{g}{j} \\
&= \sum_g \left(t_g + \sum_{1 \leq i_1 < \cdots < i_h \leq k-1} l_{i_1} l_{i_2} \cdots l_{i_h} t_{g;i_1,i_2,\ldots,i_h} \right) \sum_{j=0}^{g} \chi_\lambda(j) \binom{g}{j} \\
&= \sum_g t_g \sum_j \chi_\lambda(j) \binom{g}{j} \\
&\quad + \sum_{1 \leq i_1 < \cdots < i_h \leq k-1} l_{i_1} l_{i_2} \cdots l_{i_h} \left(\sum_g t_{g;i_1,i_2,\ldots,i_h} \sum_j \chi_\lambda(j) \binom{g}{j} \right) \\
&> \sum_g t_g \sum_j \chi_\lambda(j) \binom{g}{j} + \sum_{i=1}^{k-1} l_i \sum_g t_{g;i} \sum_j \chi_\lambda(j) \binom{g}{j}
\end{aligned}$$

since for C, $l_i l_j > 0$ for some i, j and by Corollary 3.2, $\sum_j \chi_\lambda(j)\binom{2}{j} > 0$. Now, recall that the total number of legs is $\sum_{i=1}^{k-1} l_i = n-k-1$. So, we can pick i such that the sum $\sum_g t_{g;i} \sum_j \chi_\lambda(j)\binom{g}{j}$ is minimum. Then the inequality becomes

$$\begin{aligned}
d_\lambda(L(C)) &> \sum_g t_g \sum_j \chi_\lambda(j) \binom{g}{j} + (n-k-1) \sum_g t_{g;i} \sum_j \chi_\lambda(j) \binom{g}{j} \\
&= d_\lambda(L(C[n;k;i])). \qquad \square
\end{aligned}$$

From this theorem, we see that to find the tree in $\mathcal{T}(d;k)$ that attains the minimum for some particular immanant it suffices to compare the values $d_\lambda(L(C[n;k;i]))$ where $i = 1, 2, \ldots, \lfloor \frac{k}{2} \rfloor$. As an example, we illustrate this for the permanent.

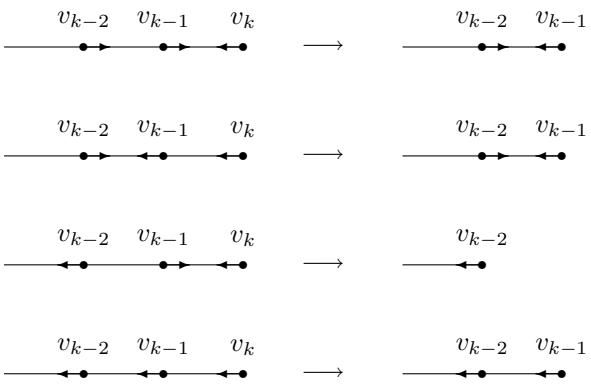
LEMMA 7.4. *Let $C[n;k;i]$ denote the caterpillar with n vertices, diameter k, and all its legs concentrated at the vertex v_i. If $i < k-2$, then for $j = 1, 2, \ldots, \alpha'(C[n;k;i])$,*

$$a_{C[n;k;i]}(j) = 2a_{C[n-1;k-1;i]}(j) + a_{C[n-2;k-2;i]}(j-1) - a_{C[n-2;k-2;i]}(j). \tag{7.2}$$

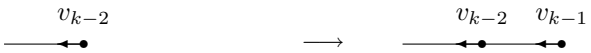
Proof. We give a combinatorial proof of the equivalent statement

$$a_{C[n;k;i]}(j) + a_{C[n-2;k-2;i]}(j) = 2a_{C[n-1;k-1;i]}(j) + a_{C[n-2;k-2;i]}(j-1).$$

We divide the j-vertex orientations of $C[n;k;i]$ into four cases as shown below. We map these to the vertex orientations of $C[n-1;k-1;i]$ and $C[n-2;k-2;i]$, changing only the arrows shown.



For j-vertex orientations of $C[n-2;k-2;i]$, we map them to j-vertex orientations of $C[n-1;k-1;i]$.



Clearly, this gives a bijection. Therefore, the formula holds. □

THEOREM 7.5 (see [2, Theorem 2.14]). *Let T be a tree from the set $\mathcal{T}(d;k)$. Then*

$$\text{per } L(T) \geq \text{per } L(C[n;k;1])$$

with equality if and only if $T = C[n;k;1]$.

Proof. It suffices to show that per $L(C[n;k;i]) >$ per $L(C[n;k;1])$ for $i = 2, \ldots, \lfloor \frac{k}{2} \rfloor$. We prove this by induction on k. The base case is when $k = 4$. There are only two caterpillars to compare. A computation gives

$$\text{per } L(C[n;4;1]) = 14n - 46 \text{ and per } L(C[n;4;2]) = 18n - 66$$

, and since $n > 5$, $C[n;4;1]$ attains the unique permanent value.

Assume $k \geq 5$. In this case, $1 < i \leq \lfloor \frac{k}{2} \rfloor < k-2$. So, we can use (7.2) to obtain

$$\begin{aligned}
& \text{per } L(C[n;k;i]) \\
= \;& 2 \text{ per } L(C[n-1;k-1;i]) + 2 \text{ per } L(C[n-2;k-2;i]) - \text{per } L(C[n-2;k-2;i]) \\
> \;& 2 \text{ per } L(C[n-1;k-1;1]) + \text{per } L(C[n-2;k-2;1]) \\
= \;& \text{per } L(C[n;k;1]).
\end{aligned}$$

Thus per $L(T) >$ per $L(C[n;k;1])$ if $T \neq C[n;k;1]$. □

TABLE 7.1

λ	(1^8)	$(2,1^6)$	$(3,1^5)$	$(4,1^4)$	$(5,1^3)$	$(6,1^2)$	$(7,1)$	(8)
$C[8;6;1]$	0	14	52	270	600	660	372	82
$C[8;6;2]$	0	14	40	226	560	680	424	102
$C[8;6;3]$	0	14	26	222	568	684	420	98

The caterpillar $C[n;k;1]$ is called the *broom* in [2] and is denoted by $B(n,k,n-k)$. Just for the record,

$$\text{per } L(C[n;k;1]) = \left(n-k+\frac{1}{\sqrt{2}}\right)(1+\sqrt{2})^{k-1} + \left(n-k-\frac{1}{\sqrt{2}}\right)(1-\sqrt{2})^{k-1}.$$

Lest it be thought that $C[n;k;1]$ attains the smallest immanant value for all immanants, we quote a result to dispel this notion.

THEOREM 7.6 (see [6, Theorem 4.3]). *When* $\lambda = (3,1^{n-3})$,

$$d_\lambda(L(T)) = 4 \sum_{v_i,v_j \in V(T)} \left(d(v_i,v_j)-1\right),$$

where $d(v_i,v_j)$ *is the distance between the vertices* v_i *and* v_j.

Using this, it is easy to see that $d_\lambda(L(T))$ attains a minimum when $T = C[n;k;\lfloor \frac{k}{2} \rfloor]$.

8. Open problems. To conclude, we list some interesting problems. It is natural to propose the following problems.

PROBLEM 1. *For each partition* λ *of* n *determine the tree or trees in* $\mathcal{T}_n(d;k)$ *such that* $d_\lambda(\cdot)$ *is minimized.*

Example. Consider the set of trees of order 8 and diameter 6. By symmetry, we only need to calculate the immanant values for the three caterpillars, $C[8;6;1]$, $C[8;6;2]$, and $C[8;6;3]$, where the explicit values corresponding to the hook characters are listed below.

From Table 7.1, we can see that there is a gradual movement of the minimum from $C[8;6;3]$ to $C[8;6;1]$ when the partition changes.

PROBLEM 2. *For every* $C[n;k;i]$, *does there exist a partition* λ *such that* $C[n;k;i]$ *is the tree in* $\mathcal{T}(d;k)$ *that attains the minimum immanant value corresponding to* λ*?*

The methods in this article can be used to show that among all the trees in $\mathcal{T}_n$, the path attains the maximum for all immanant values. However, if we restrict the discussion to the subsets of trees $\mathcal{T}_n(b;p,q)$, $\mathcal{T}_n(m;k)$, and $\mathcal{T}_n(d;k)$ our approach fails to provide a satisfactory answer.

PROBLEM 3. *Characterize the tree that attains the maximum immanantal value in each of* $\mathcal{T}_n(b;p,q), \mathcal{T}_n(m;k)$, *and* $\mathcal{T}_n(d;k)$.

Finally, we note that these problems can also be extended to bipartite graphs and graphs in general.

REFERENCES

[1] P. BOTTI AND R. MERRIS, *Almost all trees share a complete set of immanantal polynomials*, J. Graph Theory, 17 (1993), pp. 467–476.

[2] R. BRUALDI AND J. GOLDWASSER, *Permanent of Laplacian matrix of trees and bipartite graphs*, Discrete Math., 48 (1984), pp. 1–21.

[3] S. CHAIKEN, *A combinatorial proof of the all minors matrix tree theorem*, SIAM J. Alg. Discrete Methods, 3 (1982), pp. 319–329.

[4] Onn Chan and T. K. Lam, *Binomial Coefficients and Characters of the Symmetric Group*, Technical Report 693, National University of Singapore, November 1996.
[5] Onn Chan and T. K. Lam, *Certain Sums of Binomial Coefficients: A Symmetric Function Viewpoint*, Technical Report 696, National University of Singapore, December 1996.
[6] Onn Chan and T. K. Lam, *Hook immanantal inequalities for Laplacians of trees*, Linear Algebra Appl., 261 (1997), pp. 23–47.
[7] Onn Chan and T. K. Lam, *Immanant of Laplacian Matrix of Trees*, manuscript, 1997.
[8] Onn Chan, T. K. Lam, and H.-K. Tang, *A Partial Order on Trees Based on Immanants*, manuscript, 1996.
[9] I. Faria, *Permanental roots and the star degree of a graph*, Linear Algebra Appl., 64 (1985), pp. 255–265.
[10] J. L. Goldwasser, *Permanent of the Laplacian matrix of trees with a given matching*, Discrete Math., 61 (1986), pp. 197–212.
[11] I. P. Goulden and D. M. Jackson, *Combinatorial Enumeration*, Wiley, New York, 1983.
[12] I. P. Goulden and D. M. Jackson, *Immanants of combinatorial matrices*, J. Algebra, 148 (1992), pp. 305–324.
[13] P. Heyfron, *Immanant dominance orderings for hook partitions*, Linear and Multilinear Algebra, 24 (1988), pp. 65–78.
[14] P. Heyfron, *Some inequalities concerning immanants*, Math. Proc. Cambridge Philos. Soc., 109 (1991), pp. 15–30.
[15] G. D. James, *Hecke algebras and immanants*, Linear Algebra Appl., 197/198 (1994), pp. 659–670.
[16] G. D. James and M. Liebeck, *Permanents and immanants of Hermitian matrices*, Proc. London Math. Soc. (3), 55 (1987), pp. 243–265.
[17] R. Merris, *The Laplacian permanental polynomial for trees*, Czechoslovak Math. J., 32 (1982), pp. 397–403.
[18] R. Merris, *Single-hook characters and Hamiltonian circuits*, Linear and Multilinear Algebra, 14 (1983), pp. 21–35.
[19] R. Merris, *The second immanantal polynomial and the centroid of a graph*, SIAM J. Alg. Discrete Methods, 7 (1986), pp. 484–503.
[20] R. Merris, *Laplacian matrices of graphs: A survey*, Linear Algebra Appl., 197/198 (1994), pp. 143–176.
[21] R. Merris, *A survey of graph Laplacians*, Linear and Multilinear Algebra, 39 (1995), pp. 19–13.
[22] R. Merris, K. R. Rebman, and W. Watkins, *Permanental polynomials of graphs*, Linear Algebra Appl., 38 (1981), pp. 273–288.
[23] T. H. Pate, *Immanant inequalities and partition node diagrams*, J. London Math. Soc. (2), 46 (1992), pp. 65–80.
[24] T. H. Pate, *Immanant inequalities, induced characters and rank two partitions*, J. London Math. Soc. (2), 49 (1994), pp. 40–60.
[25] B. Sagan, *The Symmetric Group. Representations, Combinatorial Algorithms, and Symmetric Functions*, Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1991.
[26] H.-K. Tang, *Generalized Matrix Functions*, Honours Thesis, National University of Singapore, Singapore, 1995.
[27] W. T. Tutte, *The dissection of equilateral triangles into equilateral triangles*, Proc. Cambridge Philos. Soc., 44 (1948), pp. 463–482.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 1, pp. 145–165
© 1999 Society for Industrial and Applied Mathematics

THE KREISS MATRIX THEOREM ON A GENERAL COMPLEX DOMAIN*

KIM-CHUAN TOH† AND LLOYD N. TREFETHEN‡

Abstract. Let A be a bounded linear operator in a Hilbert space $\mathcal{H}$ with spectrum $\Lambda(A)$. The Kreiss matrix theorem gives bounds based on the resolvent norm $\|(zI-A)^{-1}\|$ for $\|A^n\|$ if $\Lambda(A)$ is in the unit disk or for $\|e^{tA}\|$ if $\Lambda(A)$ is in the left half-plane. We generalize these results to a complex domain Ω, giving bounds for $\|F_n(A)\|$ if $\Lambda(A) \subset \Omega$, where F_n denotes the nth Faber polynomial associated with Ω. One of our bounds takes the form

$$\tilde{\mathcal{K}}(\Omega) \;\le\; 2 \sup_n \|F_n(A)\|, \qquad \|F_n(A)\| \;\le\; 2\,e\,(n+1)\,\tilde{\mathcal{K}}(\Omega),$$

where $\tilde{\mathcal{K}}(\Omega)$ is the "Kreiss constant" defined by

$$\tilde{\mathcal{K}}(\Omega) = \inf\left\{ C \,:\, \|(zI-A)^{-1}\| \;\le\; C/\mathrm{dist}(z,\Omega)\ \forall\ z \notin \Omega \right\}.$$

By means of an inequality due originally to Bernstein, the second inequality can be extended to general polynomials p_n. In the case where $\mathcal{H}$ is finite-dimensional, say, $\dim(\mathcal{H}) = N$, analogous results are also established in which $\|F_n(A)\|$ is bounded in terms of N instead of n when the boundary of Ω is twice continuously differentiable.

Key words. Kreiss matrix theorem, conformal mapping, Faber polynomials, polynomials of a matrix, Krylov subspaces

AMS subject classifications. 65F35, 65F10, 15A60, 30E10

PII. S0895479897324020

1. Introduction.

Let A be a bounded linear operator in a Hilbert space $\mathcal{H}$ with spectrum $\Lambda(A)$, and let $\|\cdot\|$ denote the operator 2-norm. The Kreiss matrix theorem, originally published in 1962 [9], concerns the problem of characterizing families of bounded linear operators that are uniformly power-bounded, with spectra contained in the closed unit disk D. Let us define

$$\mathcal{K}(D) = \inf\left\{ C \,:\, \|(zI-A)^{-1}\| \;\le\; \frac{C}{\mathrm{dist}(z,D)}\ \ \forall\ z \notin D \right\},$$

the *Kreiss constant* of A with respect to D. If $\mathcal{H}$ is finite-dimensional, say, $\dim(\mathcal{H}) = N$, the current, sharp form of the theorem reads as follows [17], [20]:

$$\mathcal{K}(D) \;\le\; \sup_n \|A^n\| \;\le\; e\,N\,\mathcal{K}(D). \tag{1}$$

Since its original appearance, the Kreiss matrix theorem has been one of the fundamental results for establishing numerical stability of discrete evolution process $\{\|A^n\|\}$.

When $\mathcal{H}$ is infinite-dimensional, unfortunately, the upper bound of (1) becomes vacuous. But there is another upper bound in the following (more elementary) variant of the Kreiss matrix theorem:

$$\|A^n\| \le e\,(n+1)\,\mathcal{K}(D). \tag{2}$$

*Received by the editors July 7, 1997; accepted for publication (in revised form) by L. Reichel August 31, 1998; published electronically August 26, 1999.
http://www.siam.org/journals/simax/21-1/32402.html

†Department of Mathematics, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260 (mattohkc@math.nus.edu.sg).

‡Oxford University Computing Laboratory, Wolfson Bldg., Pashs Road, Oxford OX1 3QD, UK (lnt@comlab.ox.ac.uk).

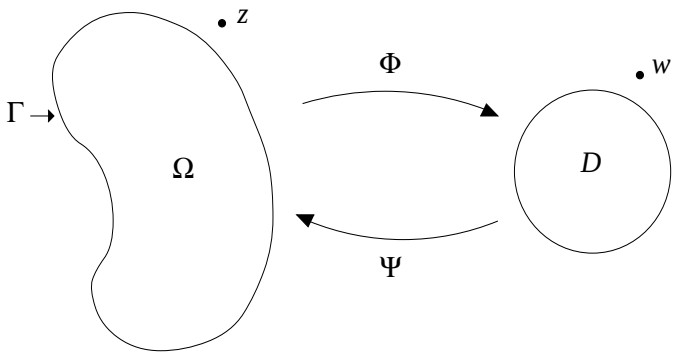


FIG. 1. *Conformal maps between the exterior of* Ω *and the exterior of the closed unit disk* D. Γ *is the boundary of* Ω.

In this paper, we generalize both (1) and (2) to a general complex domain Ω, giving bounds for $\|F_n(A)\|$ if $\Lambda(A) \subset \Omega$, where F_n is the nth Faber polynomial associated with Ω. On the unit disk D, the Faber polynomials are just the monomials z^n, and on the interval $[-1, 1]$, or on any ellipse with foci ± 1, they are twice the Chebyshev polynomials $T_n(z)$.

Let Ω be a compact set (possibly having empty interior) with boundary Γ in the complex plane $\mathbb{C}$ that contains $\Lambda(A)$, and assume that its complement Ω^c is simply connected in the extended complex plane. (This will be our assumption throughout this paper, except section 2.) By the Riemann mapping theorem [6], there exists a unique conformal map normalized by $\Phi(\infty) = \infty$ and $\Phi'(\infty) > 0$,

$$w \;=\; \Phi(z) \;=\; dz + d_0 + \sum_{k=1}^{\infty} \frac{d_k}{z^k} \quad (d > 0), \quad z \in \Omega^c,$$

from Ω^c to the exterior of the closed unit disk D (Figure 1). If $\Psi : D^c \longrightarrow \Omega^c$ is the inverse map of Φ, then Ψ has a similar expansion

$$z \;=\; \Psi(w) \;=\; cw + c_0 + \sum_{k=1}^{\infty} \frac{c_k}{w^k}, \qquad w \in D^c,$$

where $c = 1/d$. The positive constant c is known as the *logarithmic capacity* of Ω. For each nonnegative integer n, the polynomial part of $[\Phi(z)]^n$ is a polynomial of degree n. This polynomial is known as the nth Faber polynomial associated with Ω and is generally denoted by $F_n(z)$. For more details on Faber polynomials, see [2], [6], [16].

We define the Kreiss constant of A with respect to the region Ω by the following formula:

$$\mathcal{K}(\Omega) = \inf \left\{ C \,:\, \|(zI - A)^{-1}\| \;\le\; \frac{C\,|\,\Phi'(z)\,|}{|\,\Phi(z)\,| - 1} \;\forall\, z \in \Omega^c \right\}. \tag{3}$$

Our generalizations of the Kreiss matrix theorem read as follows.

THEOREM 1.1 (Kreiss matrix theorem I). *Suppose* A *is a bounded linear operator in a Hilbert space with spectrum* $\Lambda(A) \subset \Omega$ *and* $\mathcal{K}(\Omega) < \infty$. *Then* $\forall n \ge 0$,

$$\|F_n(A)\| \le e\,(n+1)\,\mathcal{K}(\Omega). \tag{4}$$

Conversely, if $\sup_{n\geq 0}\|F_n(A)\| < \infty$, *then* $\Lambda(A) \subset \Omega$, $\mathcal{K}(\Omega)$ *is finite, and*

$$\mathcal{K}(\Omega) \leq \sup_{n\geq 0}\|F_n(A)\|. \tag{5}$$

THEOREM 1.2 (Kreiss matrix theorem II). *Suppose* A *is bounded linear operator in a Hilbert space* $\mathcal{H}$ *with* $\Lambda(A) \subset \Omega$ *and* $\mathcal{K}(\Omega) < \infty$. *If* $\mathcal{H}$ *is finite-dimensional, say,* $\dim(\mathcal{H}) = N$, *and the boundary of* Ω *is twice continuously differentiable, then* $\forall n \geq 0$,

$$\|F_n(A)\| \leq C_\Omega\, e\, N\, \mathcal{K}(\Omega), \tag{6}$$

where the constant C_Ω *depends only on* Ω.

Remark. If the boundary of Ω is not smooth, then (6) holds with the constant C_Ω replaced by a constant of the form $C_\Omega\,(1 + \log n/N)$. We do not know if this latter bound is sharp.

Note that if Ω is the unit disk, with $F_n(A)$ equal to A^n, then we get back the standard results (1) and (2) from Theorems 1.1 and 1.2, aside from the introduction of an unspecified constant in Theorem 1.2.

In applications such as Krylov subspace iterations [4], we may not know the Faber polynomials for any particular region Ω of interest. Fortunately, this is not an essential restriction. By means of an inequality due to Bernstein in 1912, Theorem 1.1 can be extended to general polynomials p_n. We have the following theorem.

THEOREM 1.3 (Kreiss matrix theorem for p_n). *Suppose* A *is a bounded linear operator in a Hilbert space with spectrum* $\Lambda(A) \subset \Omega$ *and* $\mathcal{K}(\Omega) < \infty$. *Then for any polynomial* p_n *of any degree* n,

$$\|p_n(A)\| \leq e\,(n+1)\,\mathcal{K}(\Omega)\,\|p_n\|_\Omega. \tag{7}$$

Given the simplicity of the definition of the Kreiss constant with respect to the unit disk, one may wonder why we did not define the Kreiss constant for A with respect to Ω simply by

$$\tilde{\mathcal{K}}(\Omega) = \inf\left\{C \,:\, \|(zI-A)^{-1}\| \,\leq\, \frac{C}{\operatorname{dist}(z,\Omega)}\ \forall\, z \in \Omega^c\right\}. \tag{8}$$

The answer is that, as we shall see in sections 4 and 5, it is the constant $\mathcal{K}(\Omega)$ that is mathematically the most natural generalization of $\mathcal{K}(D)$. On the other hand, the constant $\tilde{\mathcal{K}}(\Omega)$ has undeniable appeal because of its geometric simplicity. In particular, $\tilde{\mathcal{K}}(\Omega)$ does not involve the conformal map Φ. We shall see that although we have two different versions of the Kreiss constant, they are equivalent in the sense that they differ by at most a factor of 2 (in either direction; see Theorem 3.2). If Ω is the unit disk itself, or any other disk, then $\tilde{\mathcal{K}}(\Omega) = \mathcal{K}(\Omega)$ for any A.

Besides being more memorable and intuitive, for most practical purposes, it is the constant $\tilde{\mathcal{K}}(\Omega)$ that is easier to use. Thus it is worthwhile for the reader to bear in mind that the Kreiss matrix theorems 1.1, 1.2, and 1.3 have a more memorable version where the constant $\mathcal{K}(\Omega)$ can be replaced by $2\tilde{\mathcal{K}}(\Omega)$ or $\frac{1}{2}\tilde{\mathcal{K}}(\Omega)$, depending on whether it appears as part of an upper bound or lower bound.

The paper is organized as follows. In section 2, we discuss some properties of $\tilde{\mathcal{K}}(\Omega)$. In section 3, we show the equivalence of the Kreiss constants $\mathcal{K}(\Omega)$ and $\tilde{\mathcal{K}}(\Omega)$. In section 4, we prove Theorems 1.1 and 1.3. In section 5, we prove Theorem 1.2. In section 6, we consider the special case $\Omega = [-1, 1]$, obtaining a Kreiss matrix theorem for the Chebyshev polynomials T_n. An example is also given to show that the linear dependence on N in the upper bound of this Kreiss matrix theorem is sharp.

Throughout, we write $\mathcal{K}(\Omega)$ simply as $\mathcal{K}$ when there is no danger of confusion.

2. Some properties of the Kreiss constant $\tilde{\mathcal{K}}(\Omega)$. In this section, Ω will denote any compact subset of the complex plane. Note that $\tilde{\mathcal{K}}(\Omega)$ remains well defined without the assumption that Ω^c is simply connected in the extended complex plane. In particular, it is well defined for $\Omega = \Lambda(A)$.

The Kreiss constant $\tilde{\mathcal{K}}(\Omega)$, being defined geometrically, can be computed without the knowledge of the conformal map Φ. This is a great advantage over $\mathcal{K}(\Omega)$ since, except for special domains such as ellipses or polygons, Φ cannot be computed easily. In practice, one might want to choose Ω to be an ϵ-pseudospectrum [19] of A for some ϵ, in which case Ω is rarely one of these special domains.

Besides this computational advantage, we shall see in this section that $\tilde{\mathcal{K}}(\Omega)$ has some appealing geometrical properties. The most important of these is that it has a geometrical interpretation: It measures the largest amount by which some ϵ-pseudospectrum of A protrudes outside the region Ω, relative to ϵ. We discuss this in Proposition 2.1. Also, $\tilde{\mathcal{K}}(\Omega) \geq 1$ for all Ω and A, since in the limit $z \to \infty$, $\|(zI - A)^{-1}\|$ and $\mathrm{dist}(z, \Omega)$ tend to $|z|^{-1}$ and $|z|$, respectively.

PROPOSITION 2.1. *For each $\epsilon \geq 0$, let $\Lambda_\epsilon(A)$ be the ϵ-pseudospectrum of A defined by*

$$\Lambda_\epsilon(A) \;=\; \{z \in \mathbb{C} \,:\, \|(zI - A)^{-1}\| \geq 1/\epsilon\}.$$

Then the following statements are equivalent:

1. $\|(zI - A)^{-1}\| \;\leq\; C/\mathrm{dist}(z, \Omega)\ \forall\ z$ *such that* $\mathrm{dist}(z, \Omega) > 0$;
2. $\mathrm{dist}(z, \Omega) \;\leq\; C\,\epsilon\ \forall\ z \in \Lambda_\epsilon(A),\ \epsilon \geq 0$;
3. $\Lambda_\epsilon(A) \;\subset\; \Omega + C\,D_\epsilon\ \forall\ \epsilon \geq 0$, *where D_ϵ is the closed disk with radius ϵ.*

Therefore, an equivalent definition for $\tilde{\mathcal{K}}(\Omega)$ is

$$\tilde{\mathcal{K}}(\Omega) = \sup_{\epsilon > 0} \frac{\mathrm{dist}(\Lambda_\epsilon(A), \Omega)}{\epsilon}, \tag{9}$$

where

$$\mathrm{dist}(\Lambda_\epsilon(A), \Omega) \;=\; \max_{z \in \Lambda_\epsilon(A)} \mathrm{dist}(z, \Omega).$$

Proof. This is an easy consequence of the definitions of $\tilde{\mathcal{K}}(\Omega)$ and $\Lambda_\epsilon(A)$. ∎

Other appealing properties of $\tilde{\mathcal{K}}(\Omega)$ include monotonicity and the fact that in the special case when A is normal and $\Omega = \Lambda(A)$ or Ω is the field of values $\mathcal{F}(A)$, $\tilde{\mathcal{K}}(\Omega)$ is equal to 1. (The field of values $\mathcal{F}(A)$ is the set $\{x^*Ax/x^*x : x \in \mathbb{C}^N, x \neq 0\}$.)

PROPOSITION 2.2. *Suppose Ω_1 and Ω_2 are subsets of $\mathbb{C}$ such that $\Lambda(A) \subset \Omega_1 \subset \Omega_2$. Then*

$$\tilde{\mathcal{K}}(\Omega_2) \leq \tilde{\mathcal{K}}(\Omega_1).$$

Proof. This result follows from the definitions of $\tilde{\mathcal{K}}(\Omega_1)$ and $\tilde{\mathcal{K}}(\Omega_2)$:

$$\begin{aligned}
\tilde{\mathcal{K}}(\Omega_2) &= \sup_{z \notin \Omega_2} \mathrm{dist}(z, \Omega_2)\, \|(zI - A)^{-1}\| \\
&\leq \sup_{z \notin \Omega_2} \mathrm{dist}(z, \Omega_1)\, \|(zI - A)^{-1}\| \\
&\leq \sup_{z \notin \Omega_1} \mathrm{dist}(z, \Omega_1)\, \|(zI - A)^{-1}\| \\
&= \tilde{\mathcal{K}}(\Omega_1). \qquad \blacksquare
\end{aligned}$$

PROPOSITION 2.3. *If A is a diagonalizable matrix with an eigenvalue decomposition $A = VDV^{-1}$, then*

$$\tilde{\mathcal{K}}(\Lambda(A)) \leq \kappa(V),$$

where $\kappa(V)$ is the condition number of V.

Proof. It is readily shown that

$$(zI - A)^{-1} = V\,(zI - D)^{-1}\,V^{-1}.$$

Taking norms on both sides gives

$$\|(zI - A)^{-1}\| \leq \frac{\kappa(V)}{\operatorname{dist}(z, \Lambda(A))}.$$

This implies that $\Lambda_\epsilon(A) \subset \Lambda(A) + \kappa(V)\,D_\epsilon$ for all $\epsilon \geq 0$, where D_ϵ is the closed disk of radius ϵ. Thus by Proposition 2.1, $\tilde{\mathcal{K}}(\Lambda(A)) \leq \kappa(V)$. ∎

PROPOSITION 2.4. *If A is a normal matrix and $\Lambda(A) \subset \Omega$, then*

$$\tilde{\mathcal{K}}(\Omega) = 1.$$

Proof. Since $\tilde{\mathcal{K}}(\Omega) \geq 1$ for any Ω and A, we need only show $\tilde{\mathcal{K}}(\Omega) \leq 1$. This follows from Propositions 2.2 and 2.3. ∎

PROPOSITION 2.5. *If $\mathcal{F}(A)$ is the field of values of A, then*

$$\tilde{\mathcal{K}}(\mathcal{F}(A)) = 1.$$

Proof. This is a restatement in a nonstandard language of mathematically standard material that is, for example, essentially the Hille–Yosida theorem of functional analysis [13]. ∎

3. Equivalence of the Kreiss constants $\tilde{\mathcal{K}}(\Omega)$ and $\mathcal{K}(\Omega)$. We have defined two different Kreiss constants: $\tilde{\mathcal{K}}(\Omega)$, which is defined geometrically, and $\mathcal{K}(\Omega)$, whose definition depends on the conformal maps Φ and Ψ. Thanks to results in geometric function theory stemming from the Koebe one-quarter theorem, these two constants are equivalent in the sense that they differ by at most a factor of 2.

THEOREM 3.1 (R. Kühnau [11]). *Suppose that Ψ is a conformal map of the exterior of the closed unit disk D to the exterior of a compact set Ω with boundary Γ and Ω^c is simply connected, with $\Psi(\infty) = \infty$. Then for any $|\,w_0\,| > 1$,*

$$\frac{1}{2}\,(|\,w_0\,| - 1) \;\leq\; \frac{\operatorname{dist}(z_0, \Gamma)}{|\,\Psi'(w_0)\,|} \;\leq\; 2\,(|\,w_0\,| - 1), \tag{10}$$

where $z_0 = \Psi(w_0)$. If Ψ is an "interval map" (defined below), then the factor $1/2$ on the left is attained in the limit $|\,w_0\,| \to 1$, and if Ψ is a "circular arc map" (likewise), then the factor 2 on the right is attained in the limit $|\,w_0\,| \to 1$.

Proof. We establish the right-hand inequality first. Let w_0 be fixed, and consider the "circular arc map"

$$g(w) = w\frac{w - w_0}{1 - \bar{w}_0 w}, \quad w \in D^c.$$

This function maps the exterior of D onto the exterior of an arc S of the unit circle, and w_0 to 0, as shown in Figure 2. The angle subtended by S at the origin is $4 \arctan 1/(|\,w_0\,|^2 - 1)^{1/2}$.

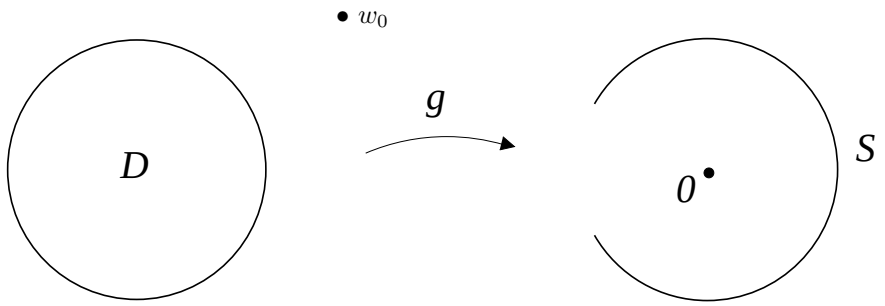


FIG. 2. *"Circular arc map," mapping the exterior of D onto the exterior of an arc S of the unit circle.*

On the contour $|w|=1$, we have $|\Psi(w)-\Psi(w_0)| = |\Psi(w)-z_0| \geq \operatorname{dist}(z_0,\Gamma)$. Thus

$$\frac{|\Psi(w)-\Psi(w_0)|}{|g(w)|} \geq \operatorname{dist}(z_0,\Gamma) \quad \forall \ |w|=1. \tag{11}$$

By the minimum principle applied to the analytic function $(\Psi(\cdot)-\Psi(w_0))/g(\cdot)$ defined on D^c, (11) holds $\forall\, |w| \geq 1$. In particular,

$$\lim_{w\to w_0} \left| \frac{\Psi(w)-\Psi(w_0)}{g(w)} \right| \ = \ \frac{|\Psi'(w_0)|}{|g'(w_0)|} \ \geq \ \operatorname{dist}(z_0,\Gamma). \tag{12}$$

Substituting $g'(w_0) = w_0/(1-|w_0|^2)$ into (12) gives

$$\frac{\operatorname{dist}(z_0,\Gamma)}{|\Psi'(w_0)|} \ \leq \ \frac{|w_0|^2-1}{|w_0|} \ \leq \ 2\operatorname{dist}(w_0,D). \tag{13}$$

If Ψ is a circular arc map, the left-hand inequality of (13) is actually an equality for all $|w_0|>1$. In this case, the factor 2 is attained in the limit as $|w_0|\to 1$.

To be precise in the above proof, we actually have to consider contours which are inside D^c, say, $|w|=r$ for $1<r<|w_0|$, and then take the limit as r tends to 1. We omit the details.

Next we show the left-hand inequality in (10). Again let w_0 be fixed, and this time, consider the "interval map"

$$h(w) = \frac{(w-w_0)(1-\bar{w}_0 w)}{4w\,(|w_0|-1)^2}, \quad w \in D^c.$$

This function maps the exterior of D onto the exterior of an interval I on the real line, and w_0 to 0, as shown in Figure 3. It is easily shown that

$$I = \left[\frac{1}{4}, \frac{(|w_0|+1)^2}{4(|w_0|-1)^2}\right] \ = \ \partial I, \tag{14}$$

$$h'(w_0) = \frac{1+|w_0|}{4w_0\,(1-|w_0|)}. \tag{15}$$

Now consider the conformal map $f = \Psi \circ h^{-1} \circ \kappa$, where $\kappa(\xi) = -\xi/(1-\xi)^2$ is the Koebe function defined on the open unit disk $\overset{\circ}{D}$ (a conformal map of $\overset{\circ}{D}$ onto

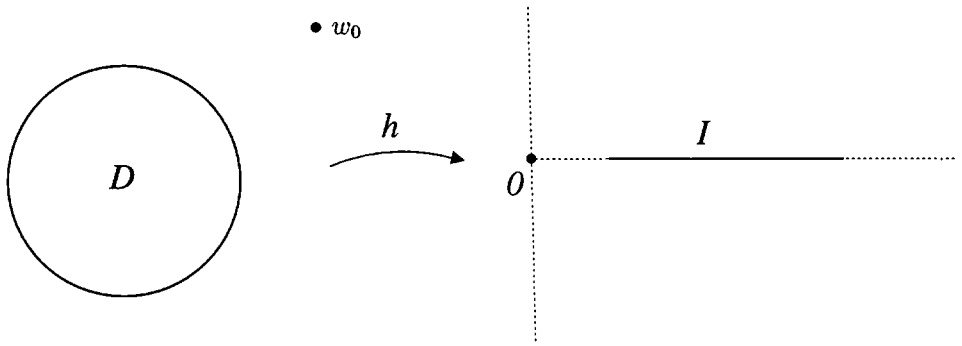


FIG. 3. *"Interval map," mapping the exterior of D onto the exterior of the interval* $I = \left[1/4,\ (|\,w_0\,|+1)^2/4(|\,w_0\,|-1)^2\right]$.

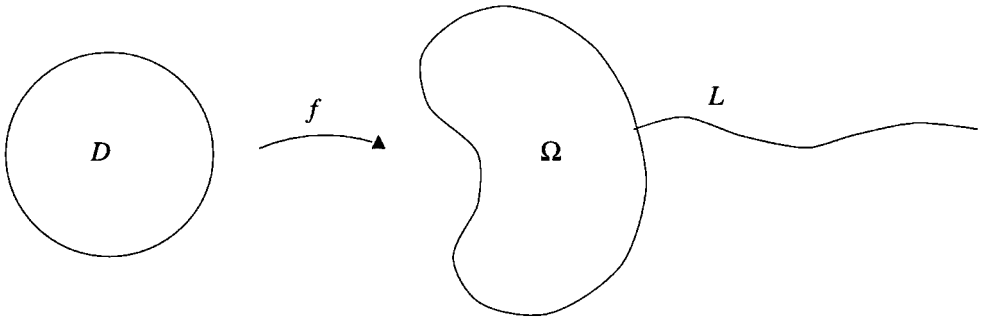


FIG. 4. *The function f maps the open unit disk $\overset{\circ}{D}$ onto the exterior of the set $\Omega \cup L$.*

$\mathbb{C} \setminus [1/4,\ \infty))$. The function f maps the open unit disk onto the exterior of $\Omega \cup L$, where L is a curve extending from the boundary of Ω to ∞, as shown in Figure 4.

By the Koebe one-quarter theorem [1, p. 29] applied to the function f, we have

$$\operatorname{dist}(f(0), \partial f(\overset{\circ}{D})) \geq \frac{1}{4}\,|\,f'(0)\,|\,. \tag{16}$$

Now

$$f(0) \;=\; \Psi(w_0), \quad f'(0) \;=\; \frac{\Psi'(w_0)}{h'(w_0)}, \quad \partial f(\overset{\circ}{D}) \;=\; \Gamma \cup L, \tag{17}$$

where

$$L = \Psi \circ h^{-1}\left(\frac{(|\,w_0\,|+1)^2}{4(|\,w_0\,|-1)^2}, \infty\right).$$

Substituting (17) into (16) gives

$$\operatorname{dist}(\Psi(w_0), \Gamma) \;\geq\; \operatorname{dist}(\Psi(w_0), \Gamma \cup L) \;\geq\; \frac{1}{4}\,\frac{|\,\Psi'(w_0)\,|}{|\,h'(w_0)\,|}. \tag{18}$$

Next, substituting (15) into (18) gives

$$\operatorname{dist}(\Psi(w_0), \Gamma) \geq \frac{|\,\Psi'(w_0)\,|}{1 + 1/\,|\,w_0\,|}\operatorname{dist}(w_0, D) \tag{19}$$

$$\geq \frac{1}{2}\ |\,\Psi'(w_0)\,|\ \mathrm{dist}(w_0, D).$$

If Ψ is an interval map, the inequality of (19) is actually an equality. In this case, the factor $1/2$ is attained in the limit as $|\,w_0\,| \to 1$. ∎

With the above theorem, it is now easy to show the equivalence of $\tilde{\mathcal{K}}(\Omega)$ and $\mathcal{K}(\Omega)$.

THEOREM 3.2. *Suppose A is a bounded linear operator in a Hilbert space with spectrum $\Lambda(A) \subset \Omega$. Then the Kreiss constants* (3) *and* (8) *are related by*

$$\frac{1}{2} \ \leq\ \frac{\tilde{\mathcal{K}}(\Omega)}{\mathcal{K}(\Omega)}\ \leq\ 2. \tag{20}$$

If Ω is an interval, then the factor $1/2$ on the left is attained in the limit as Ω becomes an infinitely long interval in the complex plane. If Ω is an arc of a circle, then the factor 2 on the right is attained in the limit as Ω approaches the full circle.

Proof. The result follows readily from Theorem 3.1 and the definitions of $\mathcal{K}(\Omega)$ and $\tilde{\mathcal{K}}(\Omega)$. ∎

4. The Kreiss matrix theorem (I). In this section, our goal is to establish Theorems 1.1 and 1.3 for a bounded linear operator A in a Hilbert space. The tools that we need are Faber series, the matrix analogue of the Cauchy integral formula, and Bernstein's lemma.

We shall begin with some basic results on Faber series. The following notation will be used in this and later sections:

- C_r: For each $r \geq 1$, C_r is the circle of radius r centered at the origin. The unit circle C_1 is also denoted by C.
- Γ_r: For each $r \geq 1$, Γ_r denotes the curve $\Psi(C_r)$. In the special case $r = 1$, Γ_1 is the boundary of Ω, and we denote this boundary also by Γ.
- Ω_r: For each $r \geq 1$, Ω_r denotes the closed region in the z-plane enclosed by $\Psi(C_r)$.
- $\|f\|_E$: For any continuous complex-valued function f defined on a compact subset E of the complex plane, $\|f\|_E$ denotes the maximum absolute value of f over E, i.e., $\|f\|_E = \max_{z\in E}\ |\,f(z)\,|$.
- P_n: The symbol P_n denotes the set of polynomials of degree at most n.

Suppose f is a function that is analytic in the interior of Ω and continuous on Ω. The *Faber series* associated with f is the formal series [2], [6, p. 44], [16]

$$f(\xi) \sim \sum_{n=0}^{\infty} a_n F_n(\xi), \quad \xi \in \Omega,$$

where

$$a_n = \frac{1}{2\pi i} \oint_C \frac{f(\Psi(w))}{w^{n+1}} dw.$$

Under suitable conditions, the Faber series actually converges to f in Ω. One sufficient condition for this is that the series converges uniformly in Ω [6, pp. 51–52]. For example, the function $(z-\xi)^{-1}$, where $z \in \Omega^c$ is fixed and $\xi \in \Omega$, can be expressed in terms of its Faber series because this series converges uniformly in Ω.

LEMMA 4.1. *For a fixed $z \in \Omega^c$, we have*

$$(z-\xi)^{-1} = \sum_{n=0}^{\infty} F_n(\xi) \frac{\Phi'(z)}{\Phi^{n+1}(z)} \quad \textit{for}\ \ \xi \in \Omega. \tag{21}$$

Proof. We omit the proof. ∎

We shall establish Theorem 1.3 first and then (4) of Theorem 1.1. Before doing so, let us state the Bernstein lemma.

LEMMA 4.2 (Bernstein). *Let $p_n \in P_n$ be arbitrary. Then*

$$\|p_n\|_{\Omega_r} \leq r^n \, \|p_n\|_{\Omega} \quad \forall \quad r \geq 1.$$

Proof. See [6, p. 27]. ∎

Now we can prove Theorem 1.3.

Proof of Theorem 1.3. Under the assumption that $\Lambda(A) \subset \Omega$, the matrix $p_n(A)$ can be represented in terms of the resolvent of A via a Cauchy integral (see [5, pp. 555–577])

$$p_n(A) = \frac{1}{2\pi i} \oint_{\Gamma_r} (zI - A)^{-1} \, p_n(z) dz \tag{22}$$

for any $r > 1$. Taking norms on both sides, we obtain the inequalities

$$\begin{aligned} \|p_n(A)\| &\leq \frac{1}{2\pi} \oint_{\Gamma_r} \|(zI - A)^{-1}\| \, |p_n(z)| \, |dz| \\ &\leq \frac{\|p_n\|_{\Omega_r}}{2\pi} \oint_{\Gamma_r} \|(zI - A)^{-1}\| \, |dz| \, . \end{aligned} \tag{23}$$

On the contour Γ_r, the resolvent norm $\|(zI - A)^{-1}\|$ is bounded above by $\mathcal{K} \, |\Phi'(z)| \, /(r-1)$, and $\|p_n\|_{\Omega_r}$ is bounded above by $r^n \|p_n\|_{\Omega}$ (by the Bernstein lemma). The inequality (23) thus reduces to

$$\|p_n(A)\| \;\leq\; \mathcal{K} \, \frac{\|p_n\|_{\Omega}}{2\pi} \, \frac{r^n}{r-1} \oint_{\Gamma_r} |\Phi'(z)| \, |dz| \;\leq\; \mathcal{K} \, \|p_n\|_{\Omega} \, \frac{r^{n+1}}{r-1}. \tag{24}$$

In deriving (24), we have used the equality

$$\oint_{\Gamma_r} |\Phi'(z)| \, |dz| \;=\; \oint_{C_r} |dw| \;=\; 2\pi r.$$

Taking $r = 1 + 1/(n+1)$ in (24) and noting that $r^{n+1} \leq e$ completes the proof. ∎

The proof of Theorem 1.3 is essentially based on ideas taken from [12], with the additional application of the Bernstein lemma. The Cauchy integral formula (22) is the standard tool that provides a link between the matrix or operator $p_n(A)$ and the resolvent $(zI - A)^{-1}$ of A, and this is the main tool that we use in this paper.

Proof of (4) *of Theorem* 1.1. The result follows immediately by applying the proof of Theorem 1.3 to the formula

$$F_n(A) = \frac{1}{2\pi i} \oint_{C_r} w^n \, \Psi'(w) \, (\Psi(w) I - A)^{-1} \, dw, \quad r > 1. \qquad ∎$$

Next we complete the proof of Theorem 1.1 by establishing the inequality (5), which states that the Kreiss constant $\mathcal{K}(\Omega)$ is bounded by the supremum of the norms $\{\|F_n(A)\|\}$. In the case of the standard Kreiss matrix theorem, a proof of such a result depends on the power series expansion of $(zI - A)^{-1}$. Here, the Faber series expansion of $(zI - A)^{-1}$ is required.

LEMMA 4.3. *Suppose* $\Lambda(A) \subset \Omega$. *Then*

$$(zI - A)^{-1} = \sum_{n=0}^{\infty} \frac{\Phi'(z)}{\Phi^{n+1}(z)} F_n(A) \quad \forall\ z \in \Omega^c. \tag{25}$$

Proof. By Theorem 16 in [5, p. 571], (25) is an easy consequence of Lemma 4.1. □

Proof of (5) *of Theorem* 1.1. Suppose $\sup_{n\geq 0} \|F_n(A)\| < \infty$. First we prove that $\Lambda(A) \subset \Omega$. Suppose there exists $\lambda \in \Lambda(A) \cap \Omega^c$. Then

$$\sup_n |F_n(\lambda)| \leq \sup_n \|F_n(A)\| < \infty.$$

But we also have (see [16])

$$\lim_{n\to\infty} |F_n(\lambda)|^{1/n} = |\Phi(\lambda)|,$$

which implies that $\sup_n |F_n(\lambda)| = \infty$ since $|\Phi(\lambda)| > 1$. Thus we have a contradiction. Hence $\Lambda(A) \subset \Omega$.

Now we prove the inequality (5). Let $z \in \Omega^c$ be fixed. Taking norms on both sides of (25) gives

$$\begin{aligned} \|(zI - A)^{-1}\| &\leq \sup_{n\geq 0} \|F_n(A)\| \sum_{k=0}^{\infty} \left| \frac{\Phi'(z)}{\Phi^{n+1}(z)} \right| \\ &= \sup_{n\geq 0} \|F_n(A)\| \frac{|\Phi'(z)|}{|\Phi(z)| - 1}, \end{aligned} \tag{26}$$

and (5) follows from (26). □

An analogue of (5) for a sequence of polynomials $\{p_n : p_n \in P_n\}$ can be derived from (5), but it is much less elegant. We give the result in the next corollary, for the sake of completeness.

COROLLARY 4.1. *Suppose* $\{p_n : p_n \in P_n,\ n \geq 0\}$ *is a sequence of polynomials,* $\Lambda(A) \subset \Omega$, *and* $\sup_{n\geq 0} \|p_n(A)\| < \infty$. *Let* $\mathbf{a}$ *denote the upper triangular array of Faber coefficients associated with* $\{p_n\}$, *i.e., for* $0 \leq k \leq n,\ n = 0, 1, 2, \ldots,$

$$a_{kn} = \frac{1}{2\pi i} \oint_C \frac{p_n(\Psi(w))}{w^{k+1}} dw.$$

If the inverse $\mathbf{b}$ *of* $\mathbf{a}$ *exists and* $\|\mathbf{b}\|_1 := \sup_{n\geq 0} \sum_{k=0}^{n} |b_{kn}| < \infty$, *then*

$$\mathcal{K}(\Omega) \leq \|\mathbf{b}\|_1 \sup_{n\geq 0} \|p_n(A)\|. \tag{27}$$

Proof. Let $\mathcal{P} = \sup\{\|p_n(A)\| : n \geq 0\}$. From the Faber series representation of p_n, we have

$$p_n(A) = \sum_{k=0}^{n} a_{kn} F_k(A), \quad n = 0, 1, 2, \ldots.$$

Thus

$$F_n(A) = \sum_{k=0}^{n} b_{kn}\, p_k(A), \quad n = 0, 1, 2, \ldots,$$
$$\|F_n(A)\| \le \sum_{k=0}^{n} |\, b_{kn}\,|\ \|p_k(A)\| \ \le\ \mathcal{P} \sum_{k=0}^{n} |\, b_{kn}\,|\,.$$

This implies that

$$\sup_{n\ge 0} \|F_n(A)\| \le \mathcal{P}\ \sup_{n\ge 0} \sum_{k=0}^{n} |\, b_{kn}\,|\ =\ \mathcal{P}\ \|\mathbf{b}\|_1.$$

By Theorem 1.1, we have established (27). □

The Kreiss matrix theorem as formulated so far is scaled to the domain Ω. However, it is an easy matter to extend this to the domain Ω_r for any fixed $r \ge 1$ by scaling the conformal map Φ of Ω by r. We conclude this section with a theorem summarizing this extension.

THEOREM 4.1. *Suppose $\Lambda(A) \subset \Omega$ and $\mathcal{K}(\Omega) < \infty$. For each fixed $r \ge 1$, let $\mathcal{K}_r$ be the Kreiss constant with respect to the domain Ω_r. If p_n is a polynomial of degree n and F_n is the nth Faber polynomial associated with Ω, then the following bounds hold for any $n \ge 0$:*

$$\|p_n(A)\| \le e\,(n+1)\, r^n\, \mathcal{K}_r\, \|p_n\|_\Omega; \tag{28}$$

$$\|F_n(A)\| \le e\,(n+1)\, r^n\, \mathcal{K}_r; \tag{29}$$

$$\mathcal{K}_r \le \sup_{n\ge 0}\ \|F_n(A)\|/r^n. \tag{30}$$

In addition,

$$\mathcal{K}_r = \inf\left\{ C\ :\ \|(zI - A)^{-1}\| \ \le\ \frac{C\,|\,\Phi'(z)\,|}{|\,\Phi(z)\,| - r} \quad \forall\ z \notin \Omega_r \right\}, \tag{31}$$

where Φ is the conformal map associated with Ω.

Proof. First, we note that (28) follows from Theorem 1.3 with an application of the Bernstein lemma.

Next we note that the conformal map associated with the domain Ω_r is simply $\Phi(\cdot)/r$. Hence the nth Faber polynomial associated with Ω_r is $F_n(\cdot)/r^n$. With this observation, (29) and (30) follow from Theorem 1.1.

Finally, (31) follows from the definition of $\mathcal{K}_r$. □

5. The Kreiss matrix theorem (II). Let $N = \dim(\mathcal{H})$. In this section, we shall assume that N is finite and that the boundary Γ of Ω is a rectifiable Jordan curve or a slit in the complex plane. We assume further that Γ is of bounded total rotation. The total rotation of Γ, denoted by V, is the total variation in the argument of the tangent to the curve Γ as the curve is described once; see [6, p. 45] for details. If Ω is a convex domain, then V is 2π; if it is a polygon, then V is the total exterior turning angle. A region Ω for which Γ is of bounded total rotation need not have an interior; it might be, for example, the interval $[-1, 1]$ or a circular arc.

The Kreiss matrix theorem of the last section gives an upper bound for $\|F_n(A)\|$ that depends linearly on n. Here we establish another upper bound for $\|F_n(A)\|$ that

depends linearly on N and at most logarithmically on n; if Γ is smooth, it depends only on N. The proof of these results depend on a sequence of lemmas established in Appendix A.

THEOREM 5.1. *Suppose A is a bounded linear operator in a finite-dimensional Hilbert space $\mathcal{H}$ with $\dim(\mathcal{H}) = N$, $\Lambda(A) \subset \Omega$, and $\mathcal{K}(\Omega) < \infty$. Let V be the total rotation of Γ. Then $\forall n \geq 0$,*

$$\|F_n(A)\| \leq \mathcal{K}\, e\, \frac{V}{2\pi}\,(4N\, +\, 1\, +\, \alpha_n), \tag{32}$$

where

$$\alpha_n\ :=\ \frac{2}{V}\int_{|w|=1+1/n} \frac{|\,\Psi''\,|}{|\,\Psi'\,|}|dw|\ \leq\ \frac{4}{\pi}\ [1\, +\, \operatorname{arcsinh}(2n+1)]\,. \tag{33}$$

Proof. As in the proof of Theorem 1.1, we can express $F_n(A)$ in terms of the resolvent of A via the Cauchy integral formula,

$$F_n(A) = \frac{1}{2\pi i}\int_{C_r} w^n\ (\Psi(w)I - A)^{-1}\,\Psi'(w)\,dw \quad \forall\ r > 1.$$

Let u, v be N-vectors such that $\|u\| = \|v\| = 1$. Then

$$v^* F_n(A) u = \frac{1}{2\pi i}\int_{C_r} w^n R(\Psi(w))\,\Psi'(w)\,dw \quad \forall\ \ r > 1, \tag{34}$$

where $R(z) = v^*(zI - A)^{-1}u$. It can be shown that $R(z)$ is a rational function of order $(N-1, N)$ with the characteristic polynomial $\det(zI - A)$ as its denominator (hence $R(z)$ has poles only inside Ω).

Performing integration by parts in (34) gives

$$2\pi i\,(v^*\,F_n(A)\,u) = \frac{-1}{n+1}\int_{C_r} w^{n+1}\frac{d}{dw}\,[R(\Psi(w))\Psi']\ dw,$$

implying that

$$\begin{aligned}|\,2\pi\, v^* F_n(A) u\,| &\leq \frac{r^{n+1}}{n+1}\int_{C_r}\left|\frac{d}{dw}[R(\Psi(w))\Psi']\right|\ |dw| \\ &\leq \frac{r^{n+1}}{n+1}\left[\int_{C_r}|\,\Psi'^2 R'(\Psi)\,|\ |dw|\ +\ \int_{C_r}|\,\Psi'' R(\Psi)\,|\ |dw|\right].\end{aligned} \tag{35}$$

Using the inequality of (58) in Lemma A.1, we get

$$\begin{aligned}|\,2\pi\, v^* F_n(A) u\,| &\leq \frac{r^{n+1}}{n+1}\left[\int_{C_r}|\,h'\,|\,|\,\Psi' R(\Psi)\,|\ |dw|\ +\ 2\int_{C_r}|\,\Psi'' R(\Psi)\,|\ |dw|\right] \\ &\leq \frac{r^{n+1}}{n+1}\,\|R(\Psi)\Psi'\|_{C_r}\left[\int_{C_r}|\,h'\,|\ |dw|\ +\ 2\int_{C_r}\frac{|\,\Psi''\,|}{|\,\Psi'\,|}\,|dw|\right],\end{aligned}$$

where $h(w) = \arg(w\,\Psi'(w)^2\,R'(\Psi(w)))$.

On the contour C_r, $|\,R(\Psi(w))\Psi'(w)\,| \leq \mathcal{K}/(r-1)$. Thus we have

$$|\,2\pi\, v^* F_n(A) u\,| \leq \frac{\mathcal{K}\, r^{n+1}}{(n+1)(r-1)}\left[\int_{C_r}|\,h'\,|\ |dw|\ +\ 2\int_{C_r}\frac{|\,\Psi''\,|}{|\,\Psi'\,|}\,|dw|\right].$$

Taking $r = 1 + 1/n$ and using the inequality of (59) in Lemma A.2, we get

$$|\, v^* F_n(A) u \,| \le \mathcal{K}\, e \left[\frac{V}{2\pi} (4N+1) \;+\; \frac{1}{\pi} \int_{|w|=1+1/n} \frac{|\,\Psi''\,|}{|\,\Psi'\,|} |dw| \right]$$

$$\le \mathcal{K}\, e\, \frac{V}{2\pi} (4N \;+\; 1 \;+\; \alpha_n). \tag{36}$$

Since $\|F_n(A)\|$ is the supremum of $|\, v^* F_n(A) u \,|$ over all unit vectors u and v, (32) follows from (36), and by Lemma A.3, it is easy to show that α_n satisfies the inequality of (33). This completes the proof. □

Ideally, one would like to have a bound that depends only on N in Theorem 5.1, as is the case with the standard Kreiss matrix theorem. However we have been unable to prove such a bound without making a stronger assumption on Γ. If we assume that Γ is *twice continuously differentiable*, then indeed we have a bound that depends only on N. This is the content of Theorem 1.2, whose proof follows easily from Theorem 5.1.

Proof of Theorem 1.2. From Theorem 5.1, we have

$$\|F_n(A)\| \le \mathcal{K}\, e\, \frac{V}{2\pi} (4N + 1 + \alpha), \tag{37}$$

where α is a constant independent of n, defined by

$$\alpha \;:=\; \sup_{n \ge 0} \alpha_n \;=\; \frac{2}{V} \int_C \frac{|\,\Psi''(w)\,|}{|\,\Psi'(w)\,|} \;|\, dw \,|\,. \tag{38}$$

(Note that α_n increases with n.) Under the assumption that Γ is twice continuously differentiable, α is finite, and therefore Theorem 1.2 is proved with

$$C_\Omega = \frac{V}{2\pi} \left(4 \;+\; \frac{1+\alpha}{N} \right). \qquad \square$$

Again, some of the ideas in the proof of Theorem 5.1 are taken from [12].

Like the first version of the Kreiss matrix theorem in the last section, analogous results of Theorem 4.1 scaled to the domain Ω_r can also be established for Theorem 5.1. This goes exactly as before, so we omit stating the results.

6. The Kreiss matrix theorem for the unit interval $[-1, 1]$. We have shown in section 5 that if the boundary of Ω is smooth, then $\|F_n(A)\|$ can be bounded in terms of N, independently of n. For a domain Ω such that the associated conformal map Ψ has the form

$$\Psi(w) = c\, w \;+\; \sum_{k=0}^{p} \frac{c_k}{w^k}, \tag{39}$$

where p is some nonnegative integer, it is also possible to bound $\|F_n(A)\|$ in terms of N, independently of n, even if $\partial\Omega$ is not smooth. In this section, we illustrate how this can be done for the special case $\Omega = [-1, 1]$. With this special case, it should be clear to the reader how an analogous result can be established for the general case.

For the unit interval $\Omega = [-1, 1]$, $\Psi(w)$ and $\Psi'(w)$ are given by

$$\Psi(w) = \frac{1}{2} \left(w + \frac{1}{w} \right), \qquad |w| > 1,$$

$$\Psi'(w) = \frac{1}{2} \left(1 - \frac{1}{w^2} \right).$$

It is well known that for $n \geq 1$, the nth Faber polynomial $F_n(z)$ associated with $[-1,1]$ is given by $F_n(z) = 2T_n(z)$, where $T_n(z) := \cos(n \arccos(z))$ is the nth Chebyshev polynomial. By means of Spijker's lemma, we shall establish a Kreiss matrix theorem for $[-1,1]$, giving a bound for $\|T_n(A)\|$ in terms of N. We state Spijker's lemma first and then prove the Kreiss matrix theorem.

LEMMA 6.1 (Spijker's lemma). *Suppose R is a rational function of order N. For any $r > 0$ such that R has no poles on C_r, there holds*

$$\int_{C_r} |R'(w)|\,|dw| \leq 2\pi N \,\|R\|_{C_r}.$$

Proof. See [17] or [20]. □

THEOREM 6.1. *Suppose A is a bounded linear operator in a finite-dimensional Hilbert space H with $\dim(\mathcal{H}) = N$, $\Lambda(A) \subset [-1,1]$, and $\mathcal{K}(\Omega) < \infty$. Then*

$$\|T_n(A)\| \leq e\,(N+1)\,\mathcal{K}(\Omega), \tag{40}$$

where $T_n(z) = \cos(n \arccos(z))$, for $n = 1, 2, \ldots$. Consequently,

$$\sup_{n \geq 0} \|T_n(A)\| \leq e\,(N+1)\,\mathcal{K}(\Omega). \tag{41}$$

Proof. For any unit N-vectors u and v, we have, as in (35),

$$\begin{aligned} |\,2\pi v^*(2T_n(A))u\,| &\leq \frac{r^{n+1}}{n+1} \int_{C_r} \left| \frac{d}{dw}[R(\Psi(w))\Psi'(w)] \right| |dw| \\ &= \frac{r^{n+1}}{n+1} \int_{C_r} |q'(w)|\,|dw|, \end{aligned}$$

where R is a rational function of order $(N-1, N)$ and $q(w) = R(\Psi(w))\Psi'(w)$. Since Ψ and Ψ' are rational functions of order $(2,1)$ and $(2,2)$, respectively, it is readily shown that q is a rational function of order $(2N+1, 2N+2)$. Applying Spijker's lemma to the rational function q and noting that $\|q\|_{C_r} \leq \mathcal{K}/(r-1)$ on the contour $C_r = \{w : |w| = r\}$, we obtain

$$|\,v^* T_n(A) u\,| \leq \frac{r^{n+1}}{(n+1)(r-1)}\,(N+1)\,\mathcal{K}. \tag{42}$$

Now (40) follows by taking $r = 1 + 1/(n+1)$ in (42). □

Next we give an example, analogous to a similar example in [12] for the standard Kreiss matrix theorem, to show that the linear dependence N in (41) is sharp.

Consider the $N \times N$ matrix

$$A = \begin{bmatrix} x_1 & \gamma & & & \\ & \ddots & \ddots & & \\ & & & x_{N-1} & \gamma \\ & & & & x_N \end{bmatrix}, \tag{43}$$

where γ is a positive real number greater than 3 and

$$x_i = \cos\left(\frac{i-1}{N-1}\pi\right), \quad i = 1, \ldots, N. \tag{44}$$

For this particular matrix, we have

$$(45)\quad (zI-A)^{-1} = \begin{bmatrix} \dfrac{1}{z-x_1} & \dfrac{\gamma}{\prod_{i=1}^{2}(z-x_i)} & \cdots & \dfrac{\gamma^{N-1}}{\prod_{i=1}^{N}(z-x_i)} \\ & \ddots & & \vdots \\ & & \dfrac{1}{z-x_{N-1}} & \dfrac{\gamma}{\prod_{i=N-1}^{N}(z-x_i)} \\ & & & \dfrac{1}{z-x_N} \end{bmatrix}$$

and

$$T_n(A) = \frac{1}{2\pi i}\int T_n(z)(zI-A)^{-1}dz$$

$$= \frac{1}{2\pi i}\begin{bmatrix} \displaystyle\int \frac{T_n(z)}{(z-x_1)}dz & \cdots & \displaystyle\int \frac{\gamma^{N-1}\,T_n(z)}{\prod_{i=1}^{N}(z-x_i)}dz \\ & \ddots & \vdots \\ & & \displaystyle\int \frac{T_n(z)}{(z-x_N)}dz \end{bmatrix}$$

$$(46)\qquad = \begin{bmatrix} T_n[x_1] & \gamma T_n[x_1,x_2] & \cdots & \gamma^{N-1}\,T_n[x_1,\cdots,x_N] \\ & \ddots & & \vdots \\ & & T_n[x_{N-1}] & \gamma\,T_n[x_{N-1},x_N] \\ & & & T_n[x_N] \end{bmatrix},$$

where $T_n[x_i,\ldots,x_j]$ denotes the divided difference of the polynomial T_n with respect to the points $x_i,\ldots,x_j$.

To show that the linear dependence on N in (41) is sharp, we need to establish a result of the form $\sup_{n\geq 0}\|T_n(A)\| \geq c\,N\,\mathcal{K}$ for some constant c independent of N.

THEOREM 6.2. *For the matrix A of* (43)*, we have*

$$(47)\qquad \frac{\gamma-3}{2\pi\gamma}\,(N-1)\,\mathcal{K} \leq \sup_{n\geq 0}\,\|T_n(A)\|,$$

where $T_n(z)=\cos(n\cos^{-1}(z))$, $z\in[-1,1]^c$.

Proof. For readability, we divide the proof into three parts.

(a) It is clear that

$$\sup_{n\geq 0}\,\|T_n(A)\| \geq \|T_{N-1}(A)\|$$

$$(48)\qquad \geq \gamma^{N-1}\,T_{N-1}[x_1,\ldots,x_N] \;=\; \gamma^{N-1}\,2^{N-2}.$$

(b) We need an upper bound for $\mathcal{K}$. But first we must obtain an upper bound for $\|(zI-A)^{-1}\|_2$, where $z\in[-1,1]^c$. For this purpose, we shall use the inequality

$$\|(zI-A)^{-1}\|_2^2 \leq \|(zI-A)^{-1}\|_1\,\|(zI-A)^{-1}\|_\infty.$$

It is easily shown that

$$\|(zI-A)^{-1}\|_1 = \frac{1}{\gamma}\sum_{k=1}^{N}\prod_{i=N-k+1}^{N}\frac{\gamma}{|z-x_i|}, \tag{49}$$

$$\|(zI-A)^{-1}\|_\infty = \frac{1}{\gamma}\sum_{k=1}^{N}\prod_{i=1}^{k}\frac{\gamma}{|z-x_i|}. \tag{50}$$

Now consider the region E defined by

$$E = \{z \in \mathbb{C} : |z-x_i| \le 3 \quad \forall\, i = 1,\ldots,N\}.$$

Noting that if $z \in E^c$, then $|z-x_i| \ge 1$ for all $i = 1,\ldots,N$, and with some algebraic manipulations in (49) and (50), we obtain

$$\|(zI-A)^{-1}\|_1\,,\ \|(zI-A)^{-1}\|_\infty \ \le\ \begin{cases} \dfrac{\gamma^N}{(\gamma-1)\,\mathrm{dist}(z,\Omega)} & \text{if } z \in E^c, \\ \dfrac{1}{(\gamma-3)}\ \dfrac{\gamma^N}{\prod_{i=1}^{N}|z-x_i|} & \text{if } z \in E\cap[-1,1]^c. \end{cases}$$

Therefore

$$\|(zI-A)^{-1}\|_2 \ \le\ \begin{cases} \dfrac{\gamma^N}{(\gamma-1)\,\mathrm{dist}(z,\Omega)} & \text{if } z \in E^c, \\ \dfrac{1}{(\gamma-3)}\ \dfrac{\gamma^N}{\prod_{i=1}^{N}|z-x_i|} & \text{if } z \in E\cap[-1,1]^c. \end{cases} \tag{51}$$

(c) Finally, recall that $\tilde{\mathcal{K}}$ is the supremum of the function $\mathrm{dist}(z,\Omega)\,\|(zI-A)^{-1}\|_2$ over the region $[-1,1]^c$. From (51), we get

$$\tilde{\mathcal{K}} \ \le\ \frac{\gamma^N}{\gamma-3}\ \max\left\{\sup_{z\in[-1,1]^c}\ \frac{\mathrm{dist}(z,\Omega)}{\prod_{i=1}^{N}|z-x_i|},\ 1\right\} \ =\ \pi\,\frac{\gamma^N}{\gamma-3}\,\frac{2^{N-2}}{N-1}, \tag{52}$$

where (52) follows from Lemma 6.2. Comparing (48) and (52), we get

$$\tilde{\mathcal{K}}\,(N-1)\,\frac{\gamma-3}{\pi\gamma} \le \sup_{n\ge 0}\|T_n(A)\|. \tag{53}$$

By Theorem 3.2, $\mathcal{K} \le 2\,\tilde{\mathcal{K}}$, and thus (47) follows from (53). ∎

The next lemma is a technical result used in the proof of Theorem 6.2. The reader is encouraged to skip the proof.

LEMMA 6.2. *Suppose* $\Omega = [-1,1]$ *and* x_i, $i = 1,\ldots,N$, *are given as in* (44). *If* $N \ge 3$, *then* $\forall z \in \Omega^c$,

$$\frac{\mathrm{dist}(z,\Omega)}{\prod_{i=1}^{N}|z-x_i|} \le \pi\,\frac{2^{N-2}}{N-1}. \tag{54}$$

Proof. First we note that the zeros of $T'_{N-1}(z)$ are the extrema $x_2,\ldots,x_{N-1}$ of $T_{N-1}(z)$ in the open interval $(-1,1)$. Thus

$$T'_{N-1}(z) = (N-1)\,2^{N-2}\prod_{i=2}^{N-1}(z-x_i).$$

We shall make use of this equation without explicitly referring to it in this proof. Note also that $|T'_{N-1}(1)| = (N-1)^2$.

We shall prove (54) for z in two separate regions, E and F, to be defined below. Note that by symmetry, it is sufficient to consider these two regions.

(a) Let E be the region defined by

$$E = \{z = x + iy \,:\, x \geq 1, y \in \mathbb{R}\}.$$

For $z \in E$, $\operatorname{dist}(z, \Omega) = |z - 1|$, and thus

$$\frac{\operatorname{dist}(z,\Omega)}{\prod_{k=1}^{N} |z - x_k|} = \frac{1}{\prod_{k=2}^{N} |z - x_k|}. \tag{55}$$

Since $1/\prod_{k=2}^{N}(z - x_k)$ is an analytic function in the interior of E and continuous in E, it is sufficient to show that for $z \in \partial E = \{1 + iy \,:\, y \in \mathbb{R}\}$, the function of (55) satisfies the bound of (54). Let $f(y)$ be the function defined by

$$f(y) = \frac{1}{\prod_{k=2}^{N} |(1 - x_k) + iy|}, \qquad y \in \mathbb{R}.$$

Elementary calculus shows that $y = 0$ is the maximizer of f and

$$f(0) \;=\; \frac{(N-1)\,2^{N-3}}{|T'_{N-1}(1)|} \;=\; \frac{2^{N-3}}{(N-1)}. \tag{56}$$

This establishes (54) for $z \in E$.

(b) Let $\theta = \pi/(N-1)$ and $I_j = [\xi_j,\ \xi_{j-1}]$, $j = 1, \ldots, N$, where $\xi_0 = 1$, $\xi_N = -1$, $\xi_j = \cos(j - \frac{1}{2})\theta$ for $j = 1, \ldots, N-1$.

Now consider the region $F = \{x + iy \,:\, -1 \leq x \leq 1,\ y > 0\}$. It is clear that $F = \cup_{j=1}^{N} F_j$, where

$$F_j \;=\; \{x + iy \,:\, x \in I_j,\ y > 0\}.$$

For $z = x + iy \in F_j$, we have

$$\frac{\operatorname{dist}(z,\Omega)}{\prod_{k=1}^{N} |z - x_k|} \;\leq\; \frac{1}{g_j(x)} \;\leq\; \frac{1}{\min_{x \in I_j} g_j(x)},$$

where $g_j(x)$, $x \in (x_{j+1},\ x_{j-1})$, is the function defined by

$$\begin{aligned} g_j(x) &:= \prod_{k=1,\, k \neq j}^{N} |x - x_k| \\ &= \frac{T'_{N-1}(x)}{(N-1)\,2^{N-1}} \, \frac{|1 - x^2|}{|x - x_j|}. \end{aligned}$$

It is easy to see that $g_j(x)$ has the form shown in Figure 5 in the interval (x_{j+1}, x_{j-1}) and the minimizer of g_j in the interval I_j is located at the end points of I_j. With some tedious algebra, it can be shown that

$$\min_{x \in I_j} g_j(x) = \min(g_j(\xi_{j-1}),\ g_j(\xi_j)) \;=\; g_j(\xi_{j-1})$$

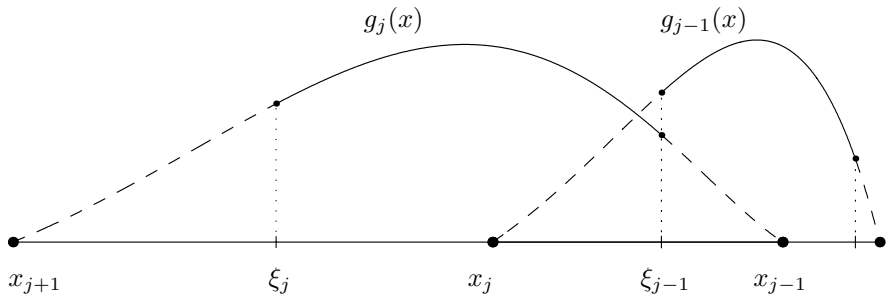


FIG. 5. *Proof of Lemma* 6.2.

$$
\begin{aligned}
&= \frac{1}{2^{N-1}} \, \frac{\sin(j-\frac{3}{2})\theta}{\sin\theta/4 \, \sin(j-\frac{5}{4})\theta} \\
&\geq \frac{1}{2^{N-1}} \, \frac{\sin\theta/2}{\sin\theta/4 \, \sin 3\theta/4} \\
&\geq \frac{1}{2^{N-1}} \, \frac{8}{3\theta} \;\geq\; \frac{(N-1)}{2^{N-2}} \, \frac{1}{\pi}. \qquad (57)
\end{aligned}
$$

By (57), we have established (54) for $z \in F_j$, $j = 1, \ldots, N$. This completes the proof. ∎

7. Discussion. We have generalized the Kreiss matrix theorem from the unit disk to a general complex domain, giving bounds for the associated Faber polynomials as well as for arbitrary polynomials. There are other generalizations of the Kreiss matrix theorem in the literature, in particular for the purpose of linear stability analysis of semidiscrete methods for time-dependent PDEs. In these problems, one is concerned with the boundedness of $\{\|\phi(\delta t\, A)^n\|\}$, where ϕ is a fixed rational function determined by the time-discretization scheme, A is determined by the space discretization scheme, and δt is the time step. The Kreiss matrix theorem has been generalized to the stability region S associated with the time-discretization scheme by a transplantation of the Kreiss matrix theorem from the unit disk to S [14]. Such a result is prototypical of the kind of generalizations of the Kreiss matrix theorem to regions in the complex plane for the purpose of linear stability analysis. For more details in this subject, see the survey article by van Dorsselaer, Kraaijevanger, and Spijker [3]. In the special case when ϕ is a fixed polynomial, our generalized Kreiss matrix theorems give the same results as those obtained from the transplantation technique, except for some technical differences.

We end with a note on the case when Ω consists of several connected components each bounded by a Jordan curve. We believe that analogues of our Kreiss matrix theorems can be established for this case. The main reason is that results such as the Bernstein lemma and the Cauchy integral formula that we have used in establishing our Kreiss matrix theorems continue to hold for this more general Ω; see [5], [21], [22].

Appendix A.

In this appendix, we present a sequence of lemmas whose results are applied in the proof of Theorem 5.1. The reader may skip these lemmas without any loss of understanding.

LEMMA A.1. *Suppose f is a analytic function defined in Ω^c. Then for each*

$r > 1$,

$$\int_{C_r} \left|\Psi'^2 f'(\Psi)\right| |dw| \le \int_{C_r} |h'|\,|\Psi' f(\Psi)|\ |dw| \ + \ \int_{C_r} |\Psi'' f(\Psi)|\ |dw|, \tag{58}$$

where $h(w) = \arg\left[w\Psi'^2(w) f'(\Psi(w))\right]$.

Proof. It is readily shown that

$$\begin{aligned}\int_{C_r} \left|\Psi'^2 f'(\Psi)\right| |dw| &= -i \int_{C_r} \Psi'(w) \frac{df(\Psi(w))}{dw} e^{-ih(w)} dw \\ &= -i \int_{C_r} \Psi'(w) e^{-ih(w)}\, df(\Psi(w)).\end{aligned}$$

Integration by parts gives

$$\int_{C_r} \left|\Psi'^2 f'(\Psi)\right| |dw| = \int_{|w|=r} h'\Psi' f(\Psi) e^{-ih} dw \ + \ i \int_{|w|=r} \Psi'' f(\Psi) e^{-ih} dw.$$

From this equation, (58) follows readily. □

LEMMA A.2. *Suppose R is a rational function of order $(N-1, N)$ with no poles inside Ω^c. Then*

$$\int_{C_r} |h'(w)|\,|dw| \le (4N+1)V \quad \forall\ r > 1, \tag{59}$$

where $h(w) = \arg\left[w\Psi'^2(w) R'(\Psi(w))\right]$.

Proof. Since R is a rational function of order $(N-1, N)$, this implies that R' is a rational function of order $(2N-2, 2N)$. Hence $R'(\Psi)$ can be written as a product

$$R'(\Psi(w)) = \prod_{k=1}^{2N-2} (a_k \Psi(w) + b_k) \prod_{k=2N-1}^{4N-2} \frac{1}{(a_k \Psi(w) + b_k)}.$$

This implies that

$$h(w) \ = \ 2\arg(w\Psi'(w)) \ - \ \arg w \ + \ \sum_{k=1}^{2N-2} \arg\left(a_k \Psi(w) + b_k\right) \ - \ \sum_{k=2N-1}^{4N-2} \arg\left(a_k \Psi(w) + b_k\right).$$

Therefore the total variation $\mathrm{TV}[h]$ of h around the circle C_r satisfies

$$\begin{aligned}\mathrm{TV}[h] &\le 2\,\mathrm{TV}[\arg(w\Psi'(w))] \ + \ \mathrm{TV}[\arg w] \ + \ \sum_{k=1}^{2N-2} \mathrm{TV}\left[\arg\left(a_k \Psi(w) + b_k\right)\right] \\ &\quad + \sum_{k=2N-1}^{4N-2} \mathrm{TV}\left[\arg\left(a_k \Psi(w) + b_k\right)\right] \\ &= 2\,\mathrm{TV}[\arg(w\Psi'(w))] \ + \ 2\pi \ + \ \sum_{k=1}^{4N-2} \mathrm{TV}\left[\arg\left(\Psi(w) + \gamma_k\right)\right],\end{aligned}$$

where $\gamma_k = b_k/a_k$.

Now by Lemma 1 of [8],

$$\mathrm{TV}[\arg(\Psi(w) + \gamma)] \le \mathrm{TV}[\arg(w\Psi'(w))] \quad \forall\ \ \gamma \in \mathbb{C}.$$

Thus

$$\mathrm{TV}[h] \le 4N\ \mathrm{TV}[\arg(w\Psi'(w))]\ +\ 2\pi. \tag{60}$$

Noting that $\mathrm{TV}[\arg(w\Psi'(w))]$ is the total variation in the argument of the tangent to the curve Γ_r as the curve is described counterclockwise once, and by definition, this is the total rotation $V(\Gamma_r)$ of Γ_r. It follows from (60) that

$$\mathrm{TV}[h]\ \le\ 4N\,V(\Gamma_r)\ +\ 2\pi\ \le\ (4N\ +\ 1)\,V(\Gamma_r).$$

It is known that $V(\Gamma_r) \le V\ \forall\ r > 1$ (see [10]). Thus we have established (59). ∎

LEMMA A.3. *For each $r > 1$,*

$$\int_{C_r} \frac{|\,\Psi''(w)\,|}{|\,\Psi'(w)\,|}|dw| \le \frac{2V}{\pi}\left[1 + \operatorname{arcsinh}\left(\frac{r+1}{r-1}\right)\right]. \tag{61}$$

Proof. From Lemma 5 of [10, p. 202], we have

$$\log\Psi'(w) = \frac{1}{\pi}\int_0^{2\pi} \log\left(1 - \frac{e^{i\theta}}{w}\right) du(\theta) \quad \forall\ \ |\,w\,| > 1, \tag{62}$$

where $u(\theta)$ is a function of bounded variation such that $\int_0^{2\pi}\ |\,du(\theta)\,| = V$. Differentiating (62) with respect to w gives

$$\frac{\Psi''(w)}{\Psi'(w)} = \frac{1}{\pi w}\int_0^{2\pi} \frac{e^{i\theta}}{w - e^{i\theta}}\, du(\theta).$$

Thus

$$\begin{aligned}\int_{C_r} \frac{|\,\Psi''(w)\,|}{|\,\Psi'(w)\,|}|dw| &\le \frac{1}{\pi r}\int_0^{2\pi}\int_{|w|=r} \frac{|dw|}{|\,w - e^{i\theta}\,|}\ |\,du(\theta)\,| \\ &\le \frac{4V}{\pi(r+1)}\left[1 + \operatorname{arcsinh}\left(\frac{r+1}{r-1}\right)\right], \end{aligned} \tag{63}$$

because

$$\int_{C_r} \frac{|dw|}{|\,w - e^{i\theta}\,|} = \frac{4r}{r+1}\left[1 + \operatorname{arcsinh}\left(\frac{r+1}{r-1}\right)\right].$$

Since $r > 1$, (61) follows from (63). ∎

Acknowledgments. We are grateful to Professor R. Kühnau for providing us the proof of Theorem 3.1 and to D. Gaier for helpful comments on a draft of this paper.

REFERENCES

[1] L.V. AHLFORS, *Conformal Invariants: Topics in Geometric Function Theory*, McGraw-Hill, New York, 1973.

[2] J.H. CURTISS, *Faber polynomials and the Faber series*, Amer. Math. Monthly, 78 (1971), pp. 577–596.

[3] J.L.M. VAN DORSSELAER, J.F.B.M. KRAAIJEVANGER, AND M.N. SPIJKER, *Linear stability analysis in the numerical solution of initial value problems*, Acta Numer. 1993, pp. 199–237.

[4] T.A. Driscoll, K.-C. Toh, and L.N. Trefethen, *From potential theory to matrix iterations in six steps*, SIAM Rev., 40 (1998), pp. 547–578.
[5] N. Dunford and J. Schwartz, *Linear Operators* I, Wiley–Interscience, New York, 1957.
[6] D. Gaier, *Lectures on Complex Approximation*, Birkhäuser, Boston, 1985.
[7] A. Greenbaum and L.N. Trefethen, *GMRES/CR and Arnoldi/Lanczos as matrix approximation problems*, SIAM J. Sci. Comp., 15 (1994), pp. 359–368.
[8] R.R. Hall and St. Ruscheweyh, *On transforms of functions with bounded boundary rotation*, Indian J. Pure Appl. Math., 16 (1985), pp. 1317–1325.
[9] H.O. Kreiss, *Über die Stabilitätsdefinition für Differenzengleichungen die partielle Differenzialgleichungen approximieren*, BIT, 2 (1962), pp. 153–181.
[10] T. Kövari and Ch. Pommerenke, *On Faber polynomials and Faber expansions*, Math. Zeitschr., 99 (1967), pp. 193–206.
[11] R. Kühnau, *private communication*, University of Halle, Germany, 1994.
[12] R.J. LeVeque and L.N. Trefethen, *On the resolvent condition in the Kreiss matrix theorem*, BIT, 24 (1984), pp. 584–591.
[13] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
[14] S.C. Reddy and L.N. Trefethen, *Stability of the method of lines*, Numer. Math., 62 (1992), pp. 235–267.
[15] E.B. Saff, *Polynomial and rational approximation in the complex domain*, Proc. Sympos. Appl. Math., 36 (1986), pp. 21-49.
[16] V.I. Smirnov and N.A. Lebedev, *Functions of a Complex Variable: Constructive Theory*, MIT Press, Cambridge, MA, 1968.
[17] M.N. Spijker, *On a conjecture by LeVeque and Trefethen related to the Kreiss matrix theorem*, BIT, 31 (1991), pp. 551–555.
[18] P.K. Suetin, *The basic properties of Faber polynomials*, Russian Math. Surveys, 19 (1964), pp. 121–149.
[19] L.N. Trefethen, *Spectra and Pseudospectra: The Behavior of Non-Normal Matrices and Operators*, book in preparation.
[20] E. Wegert and L.N. Trefethen, *From the Buffon needle problem to the Kreiss matrix theorem*, Amer. Math. Monthly, 101 (1994), pp. 132–139.
[21] J.L. Walsh, *Interpolation and Approximation*, Amer. Math. Soc. Colloq. Publ., New York, 1935.
[22] J.L. Walsh, *A generalization of Faber's polynomials*, Math. Ann., 136 (1958), pp. 23–33.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 1, pp. 166–177
© 1999 Society for Industrial and Applied Mathematics

SOLVING GENERALIZED LEAST-SQUARES PROBLEMS WITH LSQR*

STEVEN J. BENBOW†

Abstract. An iterative method for solving augmented linear systems in a generalized least-squares sense is given. The method, LSQR(A^{-1}), is shown to be a natural extension of the LSQR algorithm of Paige and Saunders [*ACM Trans. Math. Software*, 8 (1982), pp. 43–71], with generalized orthogonality properties so that the Cholesky factor of A is not required. Instead it is only assumed that some method of calculating the effect of A^{-1} on a vector is available. Numerical experiments comparing LSQR(A^{-1}) with similar preconditioned Krylov methods are described which demonstrate that the new method exhibits superior numerical properties when the Schur complement $B^TA^{-1}B$ is ill conditioned.

Key words. linear least-squares problem, generalized least-squares problem, Schur complement, sparse matrix, LSQR, MINRES

AMS subject classifications. 65F10, 65F20, 65F50

PII. S0895479897321830

1. Introduction.

Linear systems of the form

$$\begin{bmatrix} A & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}, \tag{1.1}$$

where the matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ $(n \geq m)$ are large and sparse, A is symmetric positive definite, and B is of full column rank, arise frequently in the numerical solution of problems in applied mathematics. Two classic examples from computational fluid dynamics which give rise to such systems are Darcy's law for incompressible flow in a saturated medium,

$$\frac{\mu}{k}u + \nabla p = 0, \qquad \nabla \cdot u = 0,$$

and the steady Stokes' equations for incompressible flow,

$$-\mu\triangle u + \nabla p = f, \qquad \nabla \cdot u = 0,$$

both in 2- or 3D together with some appropriate boundary conditions. Here the variable x of (1.1) represents the discretized velocity u, and y represents the discretized pressure p. The constant μ is the fluid viscosity and k is a discontinuous permeability function that is assumed to be piecewise constant. The equations are typically discretized using finite difference or finite element methods.

It can be seen (section 2) that the y component of the solution of (1.1) is the solution of the *generalized least-squares* problem

$$\textit{minimize}_{\ p\in\mathbb{R}^m} \|Bp - b\|_{A^{-1}}. \tag{1.2}$$

*Received by the editors May 23, 1997; accepted for publication (in revised form) by D. P. O'Leary November 11, 1998; published electronically August 26, 1999. The research of the author was supported by an Engineering and Physical Sciences Research Council of the United Kingdom CASE studentship with AEA Technology, Harwell, Oxfordshire, UK, and was carried out at the School of Mathematics, University of Bath, Claverton Down, Bath, BA2 7AY, UK.

http://www.siam.org/journals/simax/21-1/32183.html

†Quintessa Ltd., Bix Manor, Broadplat Lane, Bix, Henley-on-Thames, Oxfordshire, RG9 4RS, UK (stevenbenbow@quintessa.ltd.uk).

The LSQR algorithm [9] introduced in section 3 is a method for solving (linear) least-squares problems of the form *minimize* $_{p\in\mathbb{R}^m}\|Cp-c\|$, where $\|\cdot\|$ is used to denote the Euclidean norm. LSQR could be used naïvely to solve the generalized least-squares problem above, if the Cholesky factorization $A = LL^T$ were known, upon setting $C = L^{-1}B$ and $c = L^{-1}b$. This approach would require one operation with each of L^{-1} and L^{-T} per iteration step since C and C^T are both involved in LSQR. In section 4, it is shown that a variation of the LSQR algorithm exists that can be applied to the generalized least-squares problem in which the solves with L and L^T are replaced with a single solve with A. In cases where the level of fill-in in L is high, operations involving L can be expensive whereas an iterative method which calculates the effect of A^{-1} using only operations involving A can be relatively inexpensive. It is essential that L be computed accurately if it is to be used to form a linear least-squares problem as described above. This can be impractical if A is large. However, if a good preconditioner for A is known it may still be reasonable to expect to be able to operate with A^{-1}.

For example, consider the case when $A := A_M$ is the mass matrix associated with a piecewise linear finite element basis on an arbitrary triangular grid. Wathen [11] has shown that if $D = \mathrm{diag}(A)$, then $\kappa := \kappa(D^{-1}A) = 4$, so that the diagonally preconditioned conjugate gradient method applied to systems with coefficient matrix A will converge quickly at a rate independent of the mesh parameter. The matrix A representing the discretization of the μ/k operator is often badly scaled because the discontinuous permeability function can vary over several orders of magnitude. However, it is easy to show that $\kappa(D^{-1}A)$ (where $D = \mathrm{diag}(A)$) satisfies the same bound as the equivalent diagonally preconditioned mass matrix given in [11]. Hence the matrix A appearing in discretizations of Darcy's law can be preconditioned to be independent of k (and independent of the mesh parameter for suitable choices of finite elements). For a Stokes problem it may be the case that A^{-1} can be performed cheaply using FFT methods.

If it is assumed that it is possible to calculate the effect of A^{-1}, an obvious method of solving (1.1) is to solve the Schur complement system

$$B^TA^{-1}By = B^TA^{-1}b \tag{1.3}$$

and then recover x. In section 5, this approach is compared with the generalized least-squares approach outlined above.

2. The generalized least-squares connection. Notice that x can be eliminated from (1.1) to form (1.3), an equation in y only. From (1.3),

$$q^TB^TA^{-1}By = q^TB^TA^{-1}b \quad \forall q \in \mathbb{R}^m,$$

so that, on defining the inner-product $\langle\cdot\ ,\ \cdot\rangle$ on $\mathbb{R}^n$ by $\langle u\ ,\ v\rangle = u^TA^{-1}v$, (1.3) can be reformulated as

$$\textit{find } y \in \mathbb{R}^m \textit{ such that} \quad \langle Bq\ ,\ By\rangle = \langle Bq\ ,\ b\rangle \quad \forall q \in \mathbb{R}^m. \tag{2.1}$$

Setting $r = b - By$ implies that if y solves (2.1), $\langle Bq\ ,\ r\rangle = 0 \quad \forall q \in \mathbb{R}^m$, so that the residual r is orthogonal to the span of the columns of B, denoted span(B), with respect to $\langle\cdot\ ,\ \cdot\rangle$. This condition implies that By is the closest representation of b in span(B) with respect to the norm induced by $\langle\cdot\ ,\ \cdot\rangle$, i.e., that y is determined by the generalized least-squares problem

$$\textit{find } y \in \mathbb{R}^m \textit{ such that } \min_{q\in\mathbb{R}^m}\|b-Bq\|_{A^{-1}} = \|b-By\|_{A^{-1}}, \tag{2.2}$$

where $\|u\|_{A^{-1}} = \langle u\ ,\ u\rangle^{\frac{1}{2}} = (u^T A^{-1} u)^{\frac{1}{2}}$. The reverse argument, that (2.2) is equivalent to (1.3), can be seen by minimizing the functional $F(q) = \frac{1}{2}\|b - Bq\|^2_{A^{-1}}$.

3. Golub–Kahan bidiagonalization and LSQR. Before considering solving generalized least-squares problems, an introduction to the LSQR algorithm [9] is given. The LSQR algorithm can be seen to have its roots buried in the Lanczos process [6]. In section 3.1 the Lanczos process is reintroduced, and in section 3.2 it is explained how Golub and Kahan used the Lanczos process to reduce a given matrix to bidiagonal form [3] and how Paige and Saunders extended this bidiagonalization process to form an iterative solution method for least-squares problems [9].

3.1. The Lanczos Process. Given a symmetric matrix M and a starting vector b, the Lanczos process [6] reduces M to tridiagonal form as follows.

The Lanczos Process.
Set $\hat{\beta}_1 \hat{v}_1 = b$,
For $i = 1, 2, \ldots,$

$$\begin{aligned} w_i &= M\hat{v}_i - \hat{\beta}_i \hat{v}_{i-1}, \\ \hat{\alpha}_i &= \hat{v}_i^T w_i, \\ \hat{\beta}_{i+1}\hat{v}_{i+1} &= w_i - \hat{\alpha}_i \hat{v}_i, \qquad \text{where } \hat{\beta}_{j+1} > 0 \text{ and } \|\hat{v}_{j+1}\| = 1. \end{aligned}$$

Here $\hat{v}_0 := 0$. These recurrences can be written more succinctly as

$$M\hat{V}_k = \hat{V}_k T_k + \hat{\beta}_{k+1}\hat{v}_{k+1} e_k^T = \hat{V}_{k+1} H_k, \tag{3.1}$$

where $\hat{V}_k = [\hat{v}_1, \ldots, \hat{v}_k]$,

$$T_k = \begin{bmatrix} \hat{\alpha}_1 & \hat{\beta}_2 & & & \\ \hat{\beta}_2 & \hat{\alpha}_2 & \hat{\beta}_3 & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \hat{\beta}_k \\ & & & \hat{\beta}_k & \hat{\alpha}_k \end{bmatrix}, \text{ and } H_k = \begin{bmatrix} T_k \\ \hat{\beta}_{k+1} e_k^T \end{bmatrix}.$$

The set $\{\hat{v}_i\}_{i=1}^k$ can be seen to be constructed so that $\hat{v}_i^T \hat{v}_j = \delta_{ij}$ (in the absence of rounding error) so that $\hat{V}_k^T \hat{V}_k = I_k$.

Many iterative solution methods for the system $Mz = f$ can be derived from the Lanczos process. For example, if $\hat{p}_k$ is the solution of the system

$$T_k \hat{p}_k = \hat{\beta}_1 e_1 \tag{3.2}$$

and if $z_k := \hat{V}_k \hat{p}_k$, then (3.1) gives $Mz_k = f + \hat{\beta}_{k+1}\hat{\eta}_k \hat{v}_{k+1}$, where η_k is the last entry in $\hat{p}_k$. Hence, z_k will solve $Mz = f$ whenever $\hat{\beta}_{k+1}\hat{\eta}_k$ is negligible. When M (and therefore T_k) is positive definite, using Cholesky decompositions to solve (3.2) is equivalent to the conjugate gradient method [5]. Solving subproblems with H_k as coefficient matrix are the basis of the SYMMLQ and MINRES algorithms (see [8, 10]).

3.2. Golub–Kahan bidiagonalization and LSQR. In [3], the Lanczos process is applied to the coefficient matrix, denoted $\mathcal{A}_{LS}$, in the problem

$$\begin{bmatrix} I & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix} \tag{3.3}$$

to generate a bidiagonal matrix, E, having the same singular values as B. Applying the Lanczos process to $\mathcal{A}_{LS}$ with starting vector $[b^T \;\; 0^T]^T$, it can be seen that the orthonormal vectors $\hat{v}_i$ are of the form

$$\hat{v}_i =: \begin{cases} \begin{bmatrix} u_j \\ 0 \end{bmatrix} & i = 2j-1, \quad j \in \mathbb{N}, \\ \\ \begin{bmatrix} 0 \\ v_j \end{bmatrix} & i = 2j, \quad j \in \mathbb{N}, \end{cases} \tag{3.4a}$$

and that

$$\hat{\alpha}_i = \begin{cases} 1 & i = 2j-1, \quad j \in \mathbb{N}, \\ 0 & i = 2j, \quad j \in \mathbb{N}. \end{cases} \tag{3.4b}$$

Notice that since $\{\hat{v}_i\}_{i=1}^{k}$ is an orthonormal set, both $\{u_j\}_{j=1}^{[\frac{k+1}{2}]}$ and $\{v_j\}_{j=1}^{[\frac{k}{2}]}$ are orthonormal sets. Renaming the $\hat{\beta}_i$'s as

$$\hat{\beta}_i =: \begin{cases} \beta_j & i = 2j-1, \quad j \in \mathbb{N}, \\ \alpha_j & i = 2j, \quad j \in \mathbb{N} \end{cases} \tag{3.4c}$$

allows the Lanczos process on $\mathcal{A}_{LS}$ to be expressed in the following form, to be referred to as the Golub–Kahan lower-bidiagonalization procedure (GKLB).

The Golub–Kahan Lower-Bidiagonalization Procedure (GKLB).
Set $\beta_1 u_1 = b, \quad \alpha_1 v_1 = B^T u_1$,
For $j = 1, 2, \ldots,$

$$\begin{aligned} \beta_{j+1} u_{j+1} &= B v_j - \alpha_j u_j, && \text{where } \beta_{j+1} > 0 \text{ and } \|u_{j+1}\| = 1, \\ \alpha_{j+1} v_{j+1} &= B^T u_{j+1} - \beta_{j+1} v_j, && \text{where } \alpha_{j+1} > 0 \text{ and } \|v_{j+1}\| = 1. \end{aligned}$$

One step of GKLB is equivalent to performing two Lanczos steps on $\mathcal{A}_{LS}$ but requires only two multiplications with B, as opposed to four, because of relations (3.4a–3.4c). Defining $U_l = [u_1, \ldots, u_l]$ and $V_l = [v_1, \ldots, v_l]$ allows the recurrence relations of GKLB to be expressed as

$$\begin{aligned} U_{l+1}(\beta_1 e_1) &= b, \\ B V_l &= U_{l+1} E_l, \\ B^T U_{l+1} &= V_l E_l^T + \alpha_{l+1} v_{l+1} e_{l+1}^T, \end{aligned} \qquad \text{where} \quad E_l = \begin{bmatrix} \alpha_1 & & & \\ \beta_2 & \alpha_2 & & \\ & \cdot & \cdot & \\ & & \cdot & \alpha_l \\ & & & \beta_{l+1} \end{bmatrix}.$$

The lower-bidiagonal matrix E_l represents the restriction of B to the space spanned by the columns of V_l with respect to the basis formed by the columns of U_{l+1}.

The solution of the system (3.3) is now considered. Permuting rows and columns of the tridiagonal matrix T_k, it can be seen that the Lanczos subproblem (3.2) with $k = 2l + 1$ corresponds to

$$\begin{bmatrix} I & E_l \\ E_l^T & 0 \end{bmatrix} \begin{bmatrix} s_{l+1} \\ t_l \end{bmatrix} = \beta_1 \begin{bmatrix} e_1 \\ 0 \end{bmatrix},$$

where $\begin{bmatrix} s_{l+1}^T & t_l^T \end{bmatrix}^T$ is the permutation of $\hat{p}_{2l+1}$. Hence t_l is the solution of the least-squares problem

$$\textit{minimize}_{\; t \in \mathbb{R}^l} \|E_l t - \beta_1 e_1\|, \tag{3.5}$$

and $y_l = V_l t_l$, $x_l = U_{l+1} s_{l+1}$ are approximations to y and x from the spans of V_l and U_{l+1}, respectively. Paige and Saunders [9] showed that (3.5) can be easily solved at negligible cost using (updateable) QR decompositions of the matrices E_l. (See section 4.2 for a listing of the algorithm.)

4. Solution of generalized least-squares problems. In section 4.1, the bidiagonalization procedure GKLB is used to obtain a bidiagonalization procedure for B with new orthogonality properties. In the same way that LSQR was constructed from GKLB, an iterative solution method for the generalized least-squares problem (2.2) will then be derived from the new bidiagonalization procedure in section 4.2.

4.1. A bidiagonalization procedure with a different inner product. Typically, to solve (2.2) it is first assumed that the Cholesky factorization $A = LL^T$, where L is lower-triangular, is available. Then the problem (2.2) can be recast as the linear least-squares problem

$$\text{find } y \in \mathbb{R}^m \text{ such that } \min_{p \in \mathbb{R}^m} \|L^{-1}(b - Bp)\| = \|L^{-1}(b - By)\|, \tag{4.1}$$

which is then solved using any favorite method for linear least-squares. An example of solving (2.2) in this way is given in [7]. From section 2, it is obvious that (4.1) is equivalent to the linear system (3.3) with B and b replaced by $L^{-1}B$ and $L^{-1}b$, respectively, so clearly LSQR could be the method used to solve (4.1). However there is another way in which an LSQR-type method can be applied to solve (2.2) without using the Cholesky factor of A, which sacrifices some of the Euclidean orthonormality properties of the LSQR vectors for a similar orthonormality condition.

The GKLB procedure can be applied with the operator $L^{-1}B$ and with the transformed vector $w_i = Mu_i$ (where the choice of the nonsingular matrix M will be discussed shortly). Since the vectors $\{u_i\}$ form an orthonormal set,

$$w_i^T \left(MM^T\right)^{-1} w_j = \delta_{ij}. \tag{4.2}$$

Judicious choices for M are $M \in \{L, L^{-T}\}$. Choosing $M = L$ will cancel all applications of L^{-1}, and all occurrences of $L^{-T}M^{-1}$ become applications of A^{-1}. Similarly choosing $M = L^{-T}$ cancels all L^{-T} operations, and ML^{-1} becomes A^{-1}. In both cases, the Cholesky factor L is completely removed from the picture. The lower-bidiagonalization procedure resulting from the choice $M = L$ will be referred to as GKLB(A^{-1}) (since the vectors w_j are A^{-1}-orthonormal by (4.2)), and the procedure resulting from the choice $M = L^{-T}$ will be referred to as GKLB(A) for an analogous reason. The GKLB(A^{-1}) algorithm is presented below.

GKLB(A^{-1})
Set $\beta_1 w_1 = b$, $\alpha_1 v_1 = B^T A^{-1} w_1$,
For $j = 1, 2, \ldots$,

$$\begin{aligned} \beta_{j+1} w_{j+1} &= Bv_j - \alpha_j w_j, && \text{where } \beta_{j+1} > 0 \text{ and } \|w_{j+1}\|_{A^{-1}} = 1, \\ \alpha_{j+1} v_{j+1} &= B^T A^{-1} w_{j+1} - \beta_{j+1} v_j, && \text{where } \alpha_{j+1} > 0 \text{ and } \|v_{j+1}\| = 1. \end{aligned}$$

The GKLB(A) algorithm is similar, with A^{-1} acting on b and B, B^T appearing in place of $B^T A^{-1}$, and with $\beta_j \geq 0$ chosen so that $\|w_j\|_A = 1$. The recurrences of GKLB(A^{-1}) are more nicely expressed as

$$\begin{aligned} W_{k+1}(\beta_1 e_1) &= b, \\ BV_k &= W_{k+1} E_k, \\ B^T A^{-1} W_{k+1} &= V_k E_k^T + \alpha_{k+1} v_{k+1} e_{k+1}^T. \end{aligned}$$

The procedure for carrying out the calculation of $\beta_{j+1}w_{j+1}$ in GKLB(A^{-1}) is

$$\begin{aligned} \hat{w}_{j+1} &= Bv_j - \alpha_j w_j, \\ t &= A^{-1}\hat{w_{j+1}}, \\ \beta_{j+1} &= (\hat{w_{j+1}}^T t)^{\frac{1}{2}}, \\ w_{j+1} &= \hat{w_{j+1}}/\beta_{j+1}, \\ A^{-1}w_{j+1} &= t/\beta_{j+1}, \end{aligned}$$

so that $A^{-1}w_{j+1}$ is already known before it is required in the calculation of v_{j+1}; this does not require an operation with A^{-1}. The calculation for GKLB(A) is similar.

Since both bidiagonalization procedures GKLB(A^{-1}) and GKLB(A) are computationally equivalent, only GKLB(A^{-1}) will be considered further, although all the results given are equally true of both methods. The last note on GKLB(A) is the following: At each iteration step, one solve of the form $Af = g$ is required. If $f_k = W_k h_k$ is an approximation to f from span(W_k), then it may be hoped that $AW_k h_k \approx g$ (in some sense). If the matrix W_k is that obtained from GKLB(A), $h_k = W_k^T g$ and so $f_k = (W_k W_k^T)g$. This observation has not proved useful so far.

4.2. Extension of the bidiagonalization procedure to an iterative solution method. In exactly the same way that LSQR was derived from GKLB, GKLB(A^{-1}) forms the basis for an iterative solution method for generalized least-squares problems.

Let $y_k = V_k z_k$ be an approximation to the solution of (2.2). The optimal z_k is then the solution of $\min_{z\in\mathbb{R}^k} \|b - BV_k z\|_{A^{-1}}$, and, using the recurrences of GKLB(A^{-1}), it can be shown that the optimal z_k also solves $\min_{z\in\mathbb{R}^k} \|E_k z - \beta_1 e_1\|$, which is exactly the same subproblem which is solved by LSQR (cf. (3.5)), so that only the bidiagonalization step of LSQR needs to be swapped for GKLB(A^{-1}) in order to apply an LSQR-type algorithm to the generalized least-squares problem (2.2). The algorithm described above will be referred to as LSQR(A^{-1}) and is given below for completeness.

THE LSQR(A^{-1}) ALGORITHM.

Set $y_0 = 0$, $\beta_1 w_1 = b$, $\alpha_1 v_1 = B^T A^{-1} w_1$, $d_1 = v_1$, $\overline{\varphi}_1 = \beta_1$, $\overline{\rho}_1 = \alpha_1$,

For $j = 1, 2, \ldots$,

(Bidiagonalization step)

$$\begin{aligned} \beta_{j+1}w_{j+1} &= Bv_j - \alpha_j w_j, && \text{where } \beta_{j+1} > 0 \text{ and } \|w_{j+1}\|_{A^{-1}} = 1, \\ \alpha_{j+1}v_{j+1} &= B^T A^{-1} w_{j+1} - \beta_{j+1} v_j, && \text{where } \alpha_{j+1} > 0 \text{ and } \|v_{j+1}\| = 1; \end{aligned}$$

(Orthogonal transformation step)

$$\begin{aligned} \rho_j &= \left(\overline{\rho}_j^2 + \beta_{j+1}^2\right)^{\frac{1}{2}}, \\ c_j &= \overline{\rho}_j/\rho j, \\ s_j &= \beta_{j+1}/\rho_j, \\ \theta_{j+1} &= s_j \alpha_{j+1}, \\ \overline{\rho}_{j+1} &= -c_j \alpha_{j+1}, \\ \varphi_j &= c_j \overline{\varphi}_j, \\ \overline{\varphi}_{j+1} &= s_j \overline{\varphi}_j; \end{aligned}$$

(Solution update)

$$\begin{aligned} y_j &= y_{j-1} + (\varphi_j/\rho_j) d_j, \\ d_{j+1} &= v_{j+1} - (\theta_{j+1}/\rho_j) d_j. \end{aligned}$$

A computationally equivalent algorithm to that above, LSQR(A), can be derived from the GKLB(A) bidiagonalization process.

The LSQR(A^{-1}) and LSQR(A) algorithms can also provide estimates to the x component of the solution of (1.1) for a little more computational expense. If x_j is defined by $Ax_j = b - By_j$, then it can be seen that x_j can be updated with the rules

$$\begin{aligned} A^{-1}Bd_j &= A^{-1}Bv_j - (\theta_j/\rho_{j-1})A^{-1}Bd_{j-1}, \\ x_j &= x_{j-1} - (\varphi_j/\rho_j)A^{-1}Bd_j. \end{aligned}$$

Since $A^{-1}Bv_j$ is available in both algorithms from the first stage of the bidiagonalization step (provided one extra VAXPY operation (see [4]) is performed in the calculation of β_{j+1}), it is possible to update the quantity $A^{-1}Bd_j$ and hence the quantity x_j without performing any extra A^{-1} operations, the total extra cost being three VAXPY operations and storage of three n-vectors. Alternatively, once y_l is deemed sufficiently accurate, x_l can be recovered at a cost of less than one LSQR(A^{-1}) step using the rule $x_l = A^{-1}(b - By_l)$, which may be more numerically stable.

4.3. Preconditioning. When solving systems of the form

$$Mz = f,$$

where $M \in \mathbb{R}^{n\times n}$, it is usual to introduce a nonsingular preconditioning matrix N, or if N is positive definite, its Cholesky factor L, and instead solve one of the systems

$$\begin{aligned} N^{-1}Mz_L &= N^{-1}f \quad \text{(left preconditioning)}, \\ MN^{-1}z_R &= f \quad \text{(right preconditioning)}, \\ L^{-1}ML^{-T}z_C &= L^{-1}f \quad \text{(central preconditioning)}, \end{aligned}$$

where $z_L = z$, $z_R = Nz$, and $z_C = L^T z$; the idea in all cases being that the matrix $N^{-1}M$ should have better condition than M.

It is clear that of these preconditioning strategies, only right preconditioning is applicable to (generalized) least-squares problems since any other form of preconditioning affects the norm in which the minimization is taking place. Hence, preconditioned generalized least-squares problems,

$$\min \|BN^{-T}p - b\|_{A^{-1}}, \tag{4.3}$$

can be considered when solving (1.1), where $p = N^T y$. Analogous to the extension from least-squares to generalized least-squares, an iterative method for (4.3) can be devised from the Golub–Kahan bidiagonalization procedures by altering the inner product in which the m-vectors, $\{v_i\}$, are orthonormal. GKLB(A^{-1}) can be applied with the operator BN^{-T} with $z_i := Sv_i$. If $S = N$ is chosen and $H = NN^T$, the resulting algorithm, to be referred to as GKLB(A^{-1}, H^{-1}), is

GKLB(A^{-1}, H^{-1})
Set $\beta_1 w_1 = b$, $\quad \alpha_1 z_1 = B^T A^{-1} w_1$,
For $j = 1, 2, \ldots$,

$$\begin{aligned} \beta_{j+1}w_{j+1} &= BH^{-1}z_j - \alpha_j w_j, && \text{where } \beta_{j+1} > 0 \text{ and } \|w_{j+1}\|_{A^{-1}} = 1, \\ \alpha_{j+1}z_{j+1} &= B^T A^{-1} w_{j+1} - \beta_{j+1} z_j, && \text{where } \alpha_{j+1} > 0 \text{ and } \|z_{j+1}\|_{H^{-1}} = 1. \end{aligned}$$

The procedure for carrying out the calculation of $\alpha_{j+1}z_{j+1}$ is similar to that given for calculating $\beta_{j+1}w_{j+1}$ in GKLB(A^{-1}).

A similar algorithm, GKLB(A^{-1}, H), can be obtained on taking $S = N^{-T}$ above. Algorithms GKLB(A, H^{-1}) and GKLB(A, H) are also defined. Any of these bidiagonalization procedures can be plugged into the bidiagonalization step of LSQR to form an algorithm for the preconditioned system (4.3), the one corresponding to GKLB(A^{-1}, H^{-1}) being denoted LSQR(A^{-1}, H^{-1}), etc.

5. Comparing LSQR(A^{-1}) with preconditioned Krylov methods. The main factor influencing whether the LSQR(A^{-1}) method is applicable to a particular system of form (1.1) is that the action of A^{-1} must be easy to perform. Hence, it would be fair to compare the LSQR(A^{-1}) method with preconditioned Krylov methods, where the preconditioner contains the matrix A. In [2], preconditioners of the form

$$\begin{bmatrix} A & 0 \\ 0 & \pm H \end{bmatrix} \tag{5.1}$$

are considered for (1.1), where H is symmetric positive definite and is typically an approximation to the Schur complement $B^T A^{-1} B$. The coefficient matrix $\hat{\mathcal{A}}_{\pm}$ in the preconditioned system is symmetric if $+H$ is chosen, and positive semidefinite if $-H$ is chosen. Some elegant results depending on the choice of $\pm H$ are presented in [2], the most relevant being that if MINRES is used to solve the symmetrically preconditioned version of (1.1) with $\hat{\mathcal{A}}_{+}$ as coefficient matrix, then MINRES makes no progress every second step, in the sense that $\|r_{2k+1}^{\text{MINRES}}\| = \|r_{2k}^{\text{MINRES}}\|$, where r_j^{MINRES} denotes the residual in the preconditioned system.

The preconditioned system is the same as that solved by LSQR(A^{-1}, H^{-1}). The matrix H will certainly affect the convergence of both the least-squares and the MINRES methods of solution, but since only a comparison of these two approaches is required, the matrix H plays no part. Therefore it will be assumed that $H = I$. The MINRES approach then corresponds to solving the system

$$\begin{bmatrix} I & L^{-1}B \\ B^T L^{-T} & 0 \end{bmatrix} \begin{bmatrix} L^T x \\ y \end{bmatrix} = \begin{bmatrix} L^{-1} b \\ 0 \end{bmatrix}. \tag{5.2}$$

In the following table, the operation counts for one step of MINRES on (5.2) (taken from [1]) and LSQR(A^{-1}) are given. Here $(\cdot)$ indicates the dimension of the operation.

	MINRES	LSQR(A^{-1})	
Matrix-vector products	$2(n \times m)$	$2(n \times m)$	
A^{-1} operations	1	1	
Vector products	$2(n)$, $2(m)$	$1(n)$, $1(m)$	
Additional flops	$12n + 12m$	$4n + 8m$	($+6n$ with x update)
Stored vectors	$7(n)$, $7(m)$	$1(n)$, $4(m)$	($+3n$ with x update)

After $2l + 1$ steps of MINRES on (5.2), the approximation y_{2l+1}^{MINRES} is made from span (V_l) (see section 3.1), whereas LSQR(A^{-1}) provides an approximation from span(V_l) after only l steps, i.e., LSQR(A^{-1}) takes half the number of iterations (and hence half the number of matrix operations) of MINRES on (5.2) to produce an approximation to y from a subspace of the same size. Obviously the two approximations $y_l^{\text{LSQR}(A^{-1})}$ and y_{2l+1}^{MINRES} will not necessarily be the same since each method uses a different minimization subproblem. Hence the LSQR(A^{-1}) method provides a way of stepping over the redundant MINRES iterations found in [2] by using the GKLB process to step over the redundancies in the Lanczos process on the coefficient matrix in (5.2).

In [2], the authors concluded that the most efficient way to solve (1.1) (if the preconditioning (5.1) is viable) is to apply MINRES (or CG if B is of full column rank) to the Schur complement equations. Then an approximation y_l^{SCHUR} to y, from the same Krylov subspace as $y_l^{\text{LSQR}(A^{-1})}$, is obtained after l iterations. However, since the Schur complement equations are the normal equations of (1.1), the MINRES–

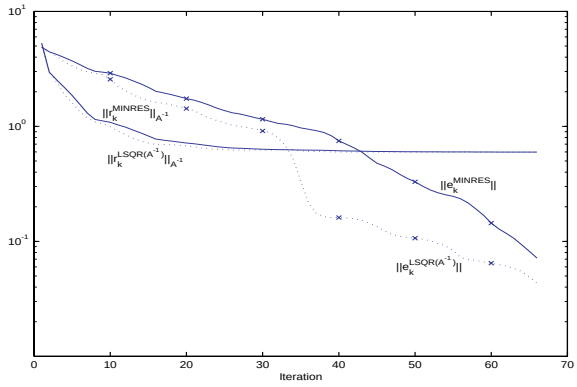


FIG. 1. *Graphs comparing the performance of MINRES on the Schur complement equations and LSQR(A^{-1}) for Example* 1. *LSQR(A^{-1}) is faster to reduce the error and slightly faster at reducing the residual norms.*

Schur complement approach is likely to be numerically inferior to LSQR(A^{-1}). This is investigated in the following section.

6. Numerical experiments and implementation. Following from the final comments of section 5, little difference between applying MINRES to (1.3) and LSQR(A^{-1}) to (1.2) is to be expected when the Schur complement is well conditioned. Experiments have shown this to be the case, with LSQR(A^{-1}) proving only slightly better at minimizing the error $\|y - y_k\|$ and the residual $\|By_k - b\|_{A^{-1}}$. The advantage of LSQR(A^{-1}) over the MINRES approach is more obvious when considering more ill-conditioned problems. The computations were performed on machines with precision $\epsilon \approx 10^{-16}$ using MATLAB 4.2*c*.

Example 1. Figure 1 illustrates the performance of LSQR(A^{-1}) and MINRES at minimizing the error $\|e_k\| = \|y - y_k\|$ and residual $\|r_k\|_{A^{-1}} = \|By_k - b\|_{A^{-1}}$ for a finite element discretization of pressure-driven Darcy flow in a porous region containing nonporous "fingers." Then (1.2) is incompatible ($By \neq b$), since $x = 0$ would correspond to a flow with zero velocity. After a similar initial performance, LSQR(A^{-1}) reduces the error faster than MINRES. Both methods perform similarly in reducing the residual, with LSQR(A^{-1}) being only slightly faster. The fact that (1.2) is incompatible makes a more precise comparison of each method's residual reducing properties difficult.

For the purpose of these experiments, $\|r_k\|_{A^{-1}}$ has been calculated exactly using an A^{-1} solve at each step of the iteration. In practice, however, the estimate

$$\|r_k\|_{A^{-1}} = \overline{\varphi}_{k+1} = \beta_1 s_k s_{k-1} \dots s_1$$

should be used since it is virtually free and has been observed to agree closely with the calculated value of $\|r_k\|_{A^{-1}}$. This estimate is analogous to that given in [9] for $\|r_k\|$ in LSQR and its derivation is the same.

Example 2. The system described in Example 1 was used to construct a compatible problem by setting $b = Bq$, where q is a vector all of whose entries equal 1. Since the system is compatible, comparison of the abilities of LSQR(A^{-1}) and MINRES at reducing the residual is easier. Figure 2 shows the performance of both methods on

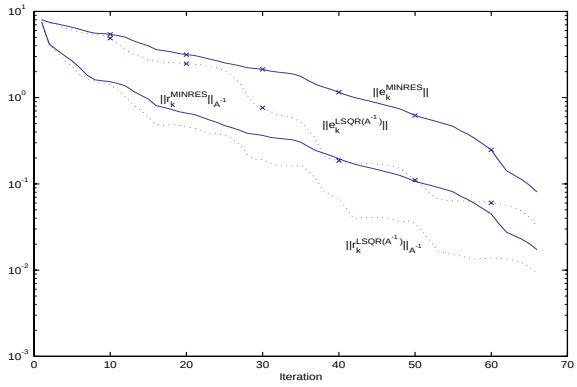


FIG. 2. *Graphs comparing the performance of MINRES on the Schur complement system and LSQR*(A^{-1}) *for Example* 2. *LSQR*(A^{-1}) *is faster at reducing the error and residual norms.*

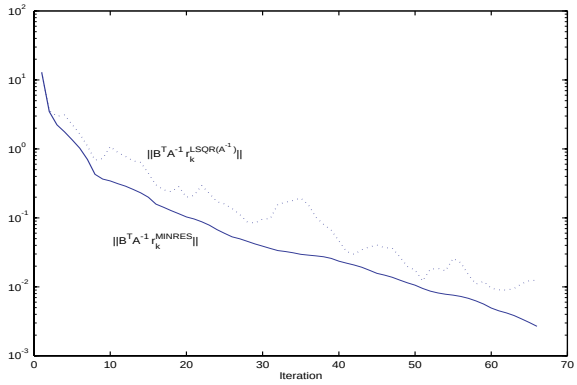


FIG. 3. *Graphs comparing* $\|B^T A^{-1} r_k\|$ *for Example* 3. *MINRES performs better at minimizing* $\|B^T A^{-1} r_k\|$ *as expected from its minimization property.*

this system. Again the advantage of LSQR(A^{-1}) is clear after only a small number of iterations.

Example 3. A similar estimate to that above for $\|r_k\|_{A^{-1}}$ is available for the residual of the Schur complement equations,

$$\|B^T A^{-1} r_k\| = \overline{\varphi}_{k+1} \alpha_{k+1} |c_k|.$$

This estimate is a useful stopping criterion since $\|B^T A^{-1} r_k\|$ tends to zero as y_k tends to y. MINRES on (1.3) is expected to be the superior of the two methods at minimizing $\|B^T A^{-1} r_k\|$ since this is precisely the quantity that is minimized at each step. This can be observed in Figure 3, which shows $\|B^T A^{-1} r_k\|$ for the groundwater flow problem in Example 1. Hence, care should be taken when using the residual of the Schur complement equations as a stopping criterion, since in Figure 1 MINRES was seen to be inferior when minimizing the error.

Example 4. As a final example, it is shown that LSQR(A^{-1}) can succeed in solving a moderately ill-conditioned problem upon which MINRES fails. Figure 4 depicts

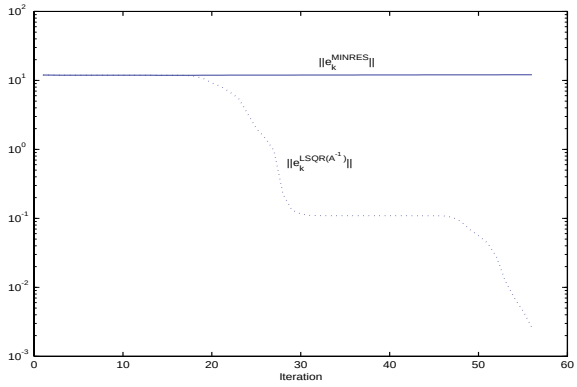


FIG. 4. *Graphs showing* $\|y - y_k\|$ *for Example* 4. *MINRES fails to reduce the norm of the error,* $LSQR(A^{-1})$ *succeeds eventually.*

what happens when both methods are applied to a system with $B \in \mathbb{R}^{32\times 16}$ and with $\kappa(B^T A^{-1} B) = 5.3 \times 10^4$. Although LSQR($A^{-1}$) requires a greater number of iterates than expected, it successfully minimizes $\|y - y_k\|$, whereas the MINRES approach fails to noticeably reduce the norm of the error.

7. Summary. Considering augmented systems as generalized least-squares problems can be advantageous when constructing iterative solution methods. LSQR can be extended to the algorithm LSQR(A^{-1}) for generalized least-squares problems without assuming availability of the Cholesky decomposition of the matrix A. LSQR(A^{-1}) is essentially twice as fast as the corresponding preconditioned MINRES method and is numerically superior to MINRES applied to the Schur complement system.

In conclusion, if it is reasonable to operate with A^{-1}, the LSQR(A^{-1}) algorithm should be used to solve (1.2) in favor of MINRES on the Schur complement equations.

Acknowledgments. The author would like to thank Professor Michael Saunders and Dr. Andrew Wathen for useful discussions and pointers to the literature, and Professor Alastair Spence, Andrew Cliffe, and two anonymous referees for suggesting improvements to the manuscript.

REFERENCES

[1] B. FISCHER, *Polynomial Based Iteration Methods for Symmetric Linear Systems*, Wiley–Teubner Series Advances in Numerical Mathematics, John Wiley & Sons, Chichester; Teubner, Stuttgart, 1996.

[2] B. FISCHER, A. RAMAGE, D. J. SILVESTER, AND A. J. WATHEN, *Minimum residual methods for augmented systems*, BIT, 38 (1998), pp. 527–543.

[3] G. GOLUB AND W. KAHAN, *Calculating the singular values and pseudo-inverse of a matrix*, SIAM J. Numer. Anal., 2 (1965), pp. 205–224.

[4] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 2nd ed., John Hopkins University Press, Baltimore, 1989.

[5] M. HESTENES AND E. STIEFEL, *Methods of conjugate gradients for solving linear systems*, J. Research Natl. Bur. Standards, 49 (1952), pp. 409–436.

[6] C. LANCZOS, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Research Natl. Bur. Standards, 45 (1950), pp. 255–282.

[7] C. PAIGE, *Fast numerically stable computations for generalized linear least squares problems*, SIAM J. Numer. Anal., 16 (1979), pp. 165–171.

[8] C. Paige and M. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.
[9] C. Paige and M. Saunders, *LSQR: An algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Software, 8 (1982), pp. 43–71.
[10] M. A. Saunders, *Solution of sparse rectangular systems using LSQR and CRAIG*, BIT, 7 (1995), pp. 588–604.
[11] A. Wathen, *Realistic eigenvalue bounds for the Galerkin mass matrix*, IMA J. Numer. Anal., 7 (1987), pp. 449–457.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 1, pp. 178–184
© 1999 Society for Industrial and Applied Mathematics

A REALIZATION ALGORITHM FOR $\mathrm{SL}_2(R[x_1,\ldots,x_m])$ OVER THE EUCLIDEAN DOMAIN*

HYUNGJU PARK†

Abstract. Let R be an arbitrary Euclidean domain and $\mathbf{A}$ be an $n\times n$ matrix of determinant 1 whose entries are elements of R. Gaussian elimination process based on Euclidean division algorithm (on R) allows one to write $\mathbf{A}$ as a product of elementary matrices over R. Suslin's stability theorem states that any $n\times n$ multivariate polynomial matrix of determinant 1 with $n\geq 3$ can be written as a product of elementary matrices. This result fails when $n=2$, and a counter-example was constructed by P. M. Cohn [*Inst. Hautes Études Sci. Publ. Math.* 30 (1996), pp. 365–413]. In this paper, an algorithm is developed that determines precisely when a given matrix in $\mathrm{SL}_2(R[x_1,\ldots,x_m])$ allows a factorization into elementary matrices and, if it does, expresses it as a product of elementary matrices. This algorithm has potential applications to signal processing, in which case the coefficient ring R is usually taken as a field or a ring $\mathbb{Z}$ of integers. Extending the idea used in this algorithm, we disprove a conjecture of Tolhuizen–Hollmann–Kalker regarding the realizability of a certain FIR filter.

Key words. polynomial matrix, elementary matrix factorization, Suslin's stability theorem

AMS subject classifications. 13P10, 15A54, 68Q40, 94A12

PII. S0895479897331096

1. Introduction.

Throughout this paper, the symbol k will denote a field and the symbol R will denote a Euclidean domain which is an integral domain equipped with an analogue of Euclidean division algorithm. Examples of such R are a field k, a univariate polynomial ring $k[x]$, or a ring $\mathbb{Z}$ of integers. Although many results used in this paper are valid over more general rings, we will exclusively work on Euclidean domains.

DEFINITION 1.1.

1. $\mathrm{SL}_n(R)$ *is the group of all the* $n\times n$ *matrices of determinant* 1 *whose entries are elements of* R.
2. *An* $n\times n$ *elementary matrix* $\mathbf{E}_{ij}(a)$ *over* R *is* $\mathbf{I}+a\cdot\mathbf{e}_{ij}$, *where* $i\neq j, a\in R$, *and* $\mathbf{e}_{ij}$ *is the* $n\times n$ *matrix whose* (i,j) *component is* 1 *and all other components are zero.*
3. *If a square matrix* $\mathbf{A}$ *can be written as a product of elementary matrices, then* $\mathbf{A}$ *is called* realizable.

Suslin's stability theorem asserts that any matrix in $\mathrm{SL}_n(R[x_1,\ldots,x_m])$ with $n\geq 3$ can be written as a product of elementary matrices, i.e.,

$$\mathrm{SL}_n(R[x_1,\ldots,x_m])=\mathrm{E}_n(R[x_1,\ldots,x_m]) \quad \text{for} \quad n\geq 3,$$

where $\mathrm{E}_n(R)$ is defined as the subgroup of $\mathrm{SL}_n(R)$ generated by the elementary matrices.

The above theorem was first proved by A. A. Suslin in [7], and an algorithmic proof for the case $R=k$ was found recently by C. Woodburn and H. Park in [6].

*Received by the editors December 3, 1997; accepted for publication (in revised form) by L. Reichel October 28, 1998; published electronically August 26, 1999.
http://www.siam.org/journals/simax/21-1/33109.html
†Department of Mathematics and Statistics, Oakland University, Rochester, MI 48309 (park@oakland.edu).

Suslin's stability theorem, however, fails when $n = 2$, and a counterexample was constructed by P. M. Cohn in [1], i.e., the *Cohn matrix*,

$$\begin{pmatrix} 1+xy & x^2 \\ -y^2 & 1-xy \end{pmatrix} \in \mathrm{SL}_2(k[x,y]),$$

was shown to be nonrealizable.

Recently, it was noted by L. Tolhuizen, H. Hollmann, and A. Kalker in [8] that the problem of decomposing a causal biorthogonal multidimensional 2-band filter bank into elementary ladder steps is essentially equivalent to developing an algorithm that determines if a given matrix in $\mathrm{SL}_2(k[x_1,\dots,x_m])$ is realizable, and if it is, expresses it as a product of elementary matrices. A reader interested in this application to signal processing should see [8] or [5].

In this paper, we develop a new algorithm for this task based on a monomial order on the polynomial ring $R[x_1,\dots,x_m]$, and we show that our algorithm terminates in a finite number of steps.

In [8, p. 647], the importance and the difficulty of extending a realization algorithm to matrices over a Laurent polynomial ring $k[x_1^{\pm},\dots,x_m^{\pm}]$ were mentioned, and a particular matrix in $\mathrm{SL}_2(k[x^{\pm},y^{\pm}])$ was conjectured to be nonrealizable. We prove that this conjecture is false and provide an explicit factorization of that matrix into elementary matrices.

2. Main theorem.

2.1. Preliminaries. For any monomial $x_1^{i_1}x_2^{i_2}\cdots x_m^{i_m}$ in $R[x_1,\dots,x_m]$, we use the shorthand notation

$$\boldsymbol{x}^{\mathbf{i}} := x_1^{i_1}x_2^{i_2}\cdots x_m^{i_m},$$

where $\mathbf{i} = (i_1,\dots,i_m) \in \mathbb{Z}_{\geq 0}^m$. Also, the polynomial ring $R[x_1,\dots,x_m]$ will be abbreviated as $R[\boldsymbol{x}]$.

DEFINITION 2.1. *A* monomial order *on $R[x_1,\dots,x_m]$ is any relation $>$ on $\mathbb{Z}_{\geq 0}^m$, or equivalently, any relation on the set of monomials $\boldsymbol{x}^{\mathbf{i}}$, $\mathbf{i} \in \mathbb{Z}_{\geq 0}^m$, satisfying the following:*

1. *$>$ is a total (or linear) order on $\mathbb{Z}_{\geq 0}^m$.*
2. *If $\mathbf{i} > \mathbf{j}$ and $\mathbf{h} \in \mathbb{Z}_{\geq 0}^m$, then $\mathbf{i}+\mathbf{h} > \mathbf{j}+\mathbf{h}$.*
3. *$>$ is a well-ordering on $\mathbb{Z}_{\geq 0}^m$. This means that every nonempty subset of $\mathbb{Z}_{\geq 0}^m$ has a smallest element under $>$.*

The following lemma plays a crucial role in proving the finite termination of our algorithm.

LEMMA 2.2 (see [2, Lemma 2, p. 53]). *An order relation $>$ on $\mathbb{Z}_{\geq 0}^m$ is a well-ordering if and only if every strictly decreasing sequence in $\mathbb{Z}_{\geq 0}^m$*

$$\mathbf{i}(1) > \mathbf{i}(2) > \mathbf{i}(3) > \cdots$$

eventually terminates.

A simple example of a monomial order is the usual numerical order

$$\cdots > m+1 > m > \cdots 3 > 2 > 1 > 0$$

on the elements of $\mathbb{Z}_{\geq 0}$. We list two popular monomial orders below.

DEFINITION 2.3 (lexicographic order). *For* $\mathbf{i}, \mathbf{j} \in \mathbb{Z}_{\geq 0}^m$, *we say* $\mathbf{i} >_{lex} \mathbf{j}$ *if the leftmost nonzero entry of the vector difference* $\mathbf{i} - \mathbf{j}$ *is positive. We write* $\boldsymbol{x}^{\mathbf{i}} >_{lex} \boldsymbol{x}^{\mathbf{j}}$ *if* $\mathbf{i} >_{lex} \mathbf{j}$.

Lexicographic order is usually abbreviated as lex order. As an example, consider the two monomials $xy^2, y^3z^4 \in k[x, y, z]$. Then we have $xy^2 >_{lex} y^3z^4$ since $(1, 2, 0) - (0, 3, 4) = (1, -1, -4)$.

DEFINITION 2.4 (degree lexicographic order). *For* $\mathbf{i}, \mathbf{j} \in \mathbb{Z}_{\geq 0}^m$, *we say* $\mathbf{i} >_{dlex} \mathbf{j}$ *if*

$$|\mathbf{i}| := \sum_{k=1}^{m} i_k > |\mathbf{j}| := \sum_{k=1}^{m} j_k, \qquad \text{or} \quad |\mathbf{i}| = |\mathbf{j}| \quad \text{and} \quad \mathbf{i} >_{lex} \mathbf{j}. \tag{2.1}$$

Note that degree lex order compares total degrees first, and then breaks ties using lex order.

Suppose that a fixed monomial order $>$ on $R[\boldsymbol{x}]$ is given. For any polynomial $f \in R[\boldsymbol{x}]$, its *leading term* or *highest degree term* with respect to $>$ is denoted as $\mathrm{lt}(f)$. Also, for a matrix $\mathbf{A} \in \mathrm{M}_n(R[\boldsymbol{x}])$, we define the matrix of its leading terms as $\mathrm{lt}(\mathbf{A}) := (\mathrm{lt}(a_{ij}))$.

2.2. A necessary condition for realizability. If a realizable matrix $\mathbf{A} \in \mathrm{SL}_2(R[\boldsymbol{x}])$ is a constant matrix, then the standard Gaussian elimination (based on Euclidean division algorithm on R) produces an explicit factorization of $\mathbf{A}$ into elementary matrices over R. Now the following theorem gives an important property for the *nonconstant* realizable 2×2 matrices.

THEOREM 2.1. *Suppose that* $\mathbf{A} = \left(\begin{smallmatrix} p & q \\ r & s \end{smallmatrix}\right) \in \mathrm{SL}_2(R[\boldsymbol{x}])$ *is a nonconstant realizable matrix. Then with respect to an arbitrary fixed monomial order* $>$, *either* $\mathbf{A}$ *has a zero entry or one of the row vectors of* $\mathrm{lt}(\mathbf{A})$ *is a monomial multiple of the other row.*

Remark 2.1. In the case of the Cohn matrix

$$\mathbf{A} = \begin{pmatrix} 1 + xy & x^2 \\ -y^2 & 1 - xy \end{pmatrix},$$

note that, with respect to the degree lex order, $\mathrm{lt}(\mathbf{A}) = \left(\begin{smallmatrix} xy & x^2 \\ -y^2 & -xy \end{smallmatrix}\right)$ is of rank 1, but neither of the two row vectors $(xy, x^2) = x \cdot (y, x)$ or $(-y^2, -xy) = -y \cdot (y, x)$ is a monomial multiple of the other. Therefore, the above theorem shows that the Cohn matrix is not realizable.

Remark 2.2. Note that the above theorem is valid with respect to an *arbitrary* monomial order. Since $\mathrm{lt}(\mathbf{A})$ usually changes as monomial order changes, we may get several relationships among different monomial vectors.

Remark 2.3. Theorem 2.1 is not valid for *constant* realizable matrices. For example, consider $\mathbf{A} := \left(\begin{smallmatrix} 5 & 2 \\ 2 & 1 \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z}[x, y])$.

Proof. Since $\mathbf{A}$ is realizable, we can write $\mathbf{A}$ as a product of elementary matrices. Let $\mathbf{A} = \mathbf{E}_1 \cdots \mathbf{E}_l$ be such a representation without a trivial factor, i.e., for each $i = 1, \ldots, l$, $\mathbf{E}_i$ is either $\mathbf{E}_{12}(f)$ or $\mathbf{E}_{21}(f)$ for some $f \neq 0 \in R[\boldsymbol{x}]$. Now we will do an induction on l.

If $l = 1$, the statement of the theorem is trivial since we have either $\mathbf{A} = \left(\begin{smallmatrix} 1 & q \\ 0 & 1 \end{smallmatrix}\right)$ or $\mathbf{A} = \left(\begin{smallmatrix} 1 & 0 \\ r & 1 \end{smallmatrix}\right)$ for some $q, r \in R[\boldsymbol{x}]$, and in both cases, $\mathbf{A}$ has a zero entry.

Now let

$$\begin{aligned} \mathbf{A}' &= \mathbf{E}_1 \cdots \mathbf{E}_{l-1} \\ &= \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(R[\boldsymbol{x}]). \end{aligned}$$

Suppose that $\mathbf{A}'$ does not have a zero entry. Then by the induction hypothesis, we may assume without loss of generality that

$$(\mathrm{lt}(a), \mathrm{lt}(b)) = h \cdot (\mathrm{lt}(c), \mathrm{lt}(d))$$

for some nonzero monomial $h \in R[\boldsymbol{x}]$.

We have two cases:

$$\mathbf{A} = \begin{pmatrix} p & q \\ r & s \end{pmatrix} = \mathbf{A}'\mathbf{E}_{12}(f) = \begin{pmatrix} a & b + fa \\ c & d + fc \end{pmatrix}$$

or

$$\mathbf{A} = \begin{pmatrix} p & q \\ r & s \end{pmatrix} = \mathbf{A}'\mathbf{E}_{21}(f) = \begin{pmatrix} a + fb & b \\ c + fd & d \end{pmatrix}.$$

In the first case, since $\mathbf{A}$ is a nonconstant matrix with $\det(\mathbf{A}) = 1$, we have $\det(\mathrm{lt}(\mathbf{A})) = 0$. Hence,

$$\mathrm{lt}(c) \cdot \mathrm{lt}(b + fa) = \mathrm{lt}(a) \cdot \mathrm{lt}(d + fc) = h \cdot \mathrm{lt}(c) \cdot \mathrm{lt}(d + fc).$$

Therefore, $\mathrm{lt}(b + fa) = h \cdot \mathrm{lt}(d + fc)$.

Now

$$(\mathrm{lt}(r), \mathrm{lt}(s)) = (\mathrm{lt}(c), \mathrm{lt}(d + fc))$$

and

$$\begin{aligned} (\mathrm{lt}(p), \mathrm{lt}(q)) &= (\mathrm{lt}(a), \mathrm{lt}(b + fa)) \\ &= (h \cdot \mathrm{lt}(c), h \cdot \mathrm{lt}(d + fc)) \\ &= h \cdot (\mathrm{lt}(c), \mathrm{lt}(d + fc)) \\ &= h \cdot (\mathrm{lt}(r), \mathrm{lt}(s)) \end{aligned}$$

as desired.

The same method gives the same conclusion in the second case.

To complete the proof, we have to consider the case when $\mathbf{A}'$ has a zero entry. Let us assume $c = 0$. The case when another entry of $\mathbf{A}$ is zero can be treated in the same way.

Again there are two cases:

$$\mathbf{A}'\mathbf{E}_{12}(f) = \begin{pmatrix} a & b + fa \\ 0 & d \end{pmatrix}$$

and

$$\mathbf{A}'\mathbf{E}_{21}(f) = \begin{pmatrix} a + fb & b \\ fd & d \end{pmatrix}.$$

In the first case, $\mathbf{A} = \mathbf{A}'\mathbf{E}_{12}(f)$ has a zero entry; therefore, we are done.

In the second case, $\det(\mathbf{A}') = ad = 1$ implies that a and d are invertible elements of $R[\boldsymbol{x}]$, i.e., a and d are invertible elements of R. Since

$$\mathbf{A} = \begin{pmatrix} a + fb & b \\ fd & d \end{pmatrix}$$

is not a constant matrix, fb is not a constant. Hence we have $a < fb$ with respect to an arbitrary fixed term order $<$, and

$$\mathrm{lt}(\mathbf{A}) = \begin{pmatrix} \mathrm{lt}(f)\mathrm{lt}(b) & \mathrm{lt}(b) \\ \mathrm{lt}(f)d & d \end{pmatrix}.$$

Since d is an invertible element of R, we have

$$\begin{aligned} \text{1st row of } \mathrm{lt}(\mathbf{A}) &= (\mathrm{lt}(f)\mathrm{lt}(b), \mathrm{lt}(b)) \\ &= d^{-1}\mathrm{lt}(b) \cdot (\mathrm{lt}(f)d, d) \\ &= \text{a monomial multiple of the 2nd row of } \mathrm{lt}(\mathbf{A}). \end{aligned}$$ □

3. Realization algorithm for $\mathbf{SL_2(k[x_1, \ldots, x_m])}$. Let us now see how to obtain a realization algorithm for $\mathrm{E}_2(R[\boldsymbol{x}])$ from Theorem 2.1.

If $\mathbf{A} = \mathrm{SL}_2(R[\boldsymbol{x}])$ is a constant matrix, then the standard Gaussian elimination (based on Euclidean division algorithm on R) produces an explicit factorization of $\mathbf{A}$ into elementary matrices over R. Hence we have a trivial realization algorithm in this case.

Suppose that one of the entries of

$$\mathbf{A} := \begin{pmatrix} p & q \\ r & s \end{pmatrix} \in \mathrm{SL}_2(k[\boldsymbol{x}]),$$

say, r, is zero. Then $\det(\mathbf{A}) = ps = 1$ implies that p and s are invertible elements of R, and thus

$$\mathbf{E}_{12}(-s^{-1}q)\mathbf{A} = \begin{pmatrix} p & 0 \\ 0 & s \end{pmatrix}$$

is a constant matrix for which we already have a realization algorithm, say, $\mathbf{E}_{12}(-s^{-1}q)\mathbf{A} = \mathbf{E}_1 \cdots \mathbf{E}_l$. Then

$$\mathbf{A} = \mathbf{E}_{12}(s^{-1}q)\mathbf{E}_1 \cdots \mathbf{E}_l$$

is a desired realization for $\mathbf{A}$.

Definition 3.1.

1. *For a polynomial $f \in R[x_1, \ldots, x_m]$ with $\mathrm{lt}(f) = \boldsymbol{x}^{\mathbf{i}}$, we define its* multidegree, $\mathrm{multideg}(f)$, *to be $\mathbf{i} \in (\mathbb{Z}_{\geq 0})^m$.*
2. *Suppose $\mathbf{B} = (b_{ij}) \in \mathrm{SL}_n(R[\boldsymbol{x}])$. Then we define* $\mathrm{multideg}(\mathbf{B})$ *by*

$$\mathrm{multideg}(\mathbf{B}) = \max_{1 \leq i,j \leq n} \{\mathrm{multideg}(b_{ij})\} \in (\mathbb{Z}_{\geq 0})^m.$$

Remark 3.1. If $\mathrm{multideg}(\mathbf{B}) = (0, \ldots, 0) \in (\mathbb{Z}_{\geq 0})^m$, then clearly all the entries of $\mathbf{B}$ are constants.

Now out of a given $\mathbf{A} \in \mathrm{SL}_2(R[\boldsymbol{x}])$, we try to make a constant matrix or matrix with a zero entry by applying elementary operations. If $\mathbf{A} = \left(\begin{smallmatrix} p & q \\ r & s \end{smallmatrix}\right) \in \mathrm{SL}_2(R[\boldsymbol{x}])$ doesn't have a zero entry, then compare the two vectors $(\mathrm{lt}(p), \mathrm{lt}(q))$ and $(\mathrm{lt}(r), \mathrm{lt}(s))$.

- Step 1. If neither of them is a monomial multiple of the other, then by Theorem 2.1, $\mathbf{A}$ is not realizable. Otherwise, go to step 2.
- Step 2. Assume without loss of generality that

$$(\mathrm{lt}(p), \mathrm{lt}(q)) = h \cdot (\mathrm{lt}(r), \mathrm{lt}(s))$$

for some monomial $h \in R[\boldsymbol{x}]$. There are two cases to consider. If h is a constant, replace $\mathbf{A}$ by $\mathbf{E}_{12}(-h)\mathbf{A}$ and go to step 1. Otherwise, go to step 3.

- Step 3. Note that the matrix

$$\mathbf{E}_{12}(-h)\mathbf{A} = \begin{pmatrix} p - hr & q - hs \\ r & s \end{pmatrix}$$

has a strictly smaller multidegree than $\mathbf{A}$. Now replace $\mathbf{A}$ by $\mathbf{E}_{12}(-h)\mathbf{A}$. If $\mathbf{A}$ is a constant matrix or has a zero entry, then terminate the process. Otherwise go to step 1 with the new $\mathbf{A}$.

If any intermediate matrix in the above process is not realizable by step 1, then $\mathbf{A}$ itself is not realizable. Otherwise, since the above procedure strictly reduces the multidegree of $\mathbf{A}$ each time, by Lemma 2.2, we get either a constant matrix or a matrix with a zero entry in $\mathrm{SL}_2(k[\boldsymbol{x}])$ after a **finite** number of steps. By keeping track of all the intermediate matrices, we find elementary matrices $\mathbf{E}_1, \ldots, \mathbf{E}_l$ over $k[\boldsymbol{x}]$ such that $\mathbf{A} = \mathbf{E}_1 \cdots \mathbf{E}_l$.

4. A conjecture of Tolhuizen–Hollmann–Kalker. In [8, p. 647], the difficulty of extending a realization algorithm to matrices of FIR filters was mentioned. Mathematically, this is about realization of matrices over a Laurent polynomial ring $R[x_1^{\pm}, \ldots, x_m^{\pm}]$. While discussing this problem, the authors of [8] conjectured that the following particular matrix in $\mathrm{SL}_2(k[x^{\pm}, y^{\pm}])$ would be nonrealizable:

$$\mathbf{A} := \begin{pmatrix} 1 + (x+y)(x-y) & (x+y)^2 \\ -(x-y)^2 & 1 - (x+y)(x-y) \end{pmatrix}. \tag{4.1}$$

We prove that this conjecture is false by providing an explicit factorization of the matrix $\mathbf{A}$ into elementary matrices over $k[x^{\pm}, y^{\pm}]$.

The definitions of lex order or degree lex order for the polynomial ring $R[x_1, \ldots, x_m]$ can be extended to the Laurent polynomial ring $R[x_1^{\pm}, \ldots, x_m^{\pm}]$. For example, in Definition 2.3 or Definition 2.4, one can replace $\mathbb{Z}_{\geq 0}$ by $\mathbb{Z}$. However, it has to be noted that the resulting ordering is not a well-ordering anymore. This may cause the realization algorithm of the previous section, when applied to matrices in $\mathrm{SL}_2(R[x_1^{\pm}, \ldots, x_m^{\pm}])$, not to terminate in a finite number of steps. Although we do not have a complete solution to this problem, the realization algorithm with respect to degree lex order terminates in a finite number of steps for the above matrix.

Let "$\longrightarrow$" indicate that we are applying elementary operations, and apply the realization algorithm of the previous section with respect to degree lex order with $x < y$:

$$\begin{aligned}
\mathbf{A} &= \begin{pmatrix} 1+x^2-y^2 & x^2+2xy+y^2 \\ -x^2+2xy-y^2 & 1-x^2+y^2 \end{pmatrix} \xrightarrow{\uparrow -1} \begin{pmatrix} 1+2x^2-2xy & -1+2x^2+2xy \\ -x^2+2xy-y^2 & 1-x^2+y^2 \end{pmatrix} \\
&\xrightarrow{\downarrow -\frac{y}{2x}} \begin{pmatrix} 1+2x^2-2xy & -1+2x^2+2xy \\ -\frac{y}{2x}-x^2+xy & 1+\frac{y}{2x}-xy-x^2 \end{pmatrix} \xrightarrow{\uparrow 2} \begin{pmatrix} 1-\frac{y}{x} & 1+\frac{y}{x} \\ -\frac{y}{2x}-x^2+xy & 1+\frac{y}{2x}-xy-x^2 \end{pmatrix} \\
&\xrightarrow{\downarrow x^2} \begin{pmatrix} 1-\frac{y}{x} & 1+\frac{y}{x} \\ -\frac{y}{2x} & 1+\frac{y}{2x} \end{pmatrix} \xrightarrow{\uparrow -2} \begin{pmatrix} 1 & -1 \\ -\frac{y}{2x} & 1+\frac{y}{2x} \end{pmatrix} \xrightarrow{\downarrow \frac{y}{2x}} \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \xrightarrow{\uparrow 1} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.
\end{aligned}$$

In terms of elementary matrices, this reduction process can be written as:

$$\mathbf{E}_{12}(1)\mathbf{E}_{21}(\tfrac{y}{2x})\mathbf{E}_{12}(-2)\mathbf{E}_{21}(x^2)\mathbf{E}_{12}(2)\mathbf{E}_{21}(\tfrac{-y}{2x})\mathbf{E}_{12}(-1) \cdot \mathbf{A} = \mathbf{I},$$

from which, one gets a desired factorization of $\mathbf{A}$ over $k[x^{\pm}, y^{\pm}]$:

$$\mathbf{A} = \mathbf{E}_{12}(1)\mathbf{E}_{21}(\tfrac{y}{2x})\mathbf{E}_{12}(-2)\mathbf{E}_{21}(-x^2)\mathbf{E}_{12}(2)\mathbf{E}_{21}(-\tfrac{y}{2x})\mathbf{E}_{12}(-1). \tag{4.2}$$

Acknowledgment. The author wishes to thank Ton Kalker for showing him the interaction between polynomial ring theory and signal processing.

REFERENCES

[1] P. M. Cohn, *On the structure of the GL_2 of a ring*, Inst. Hautes Études Sci. Publ. Math., 30 (1966), pp. 365–413.
[2] D. Cox, J. Little, and D. O'Shea, *Ideals, Varieties, and Algorithms*, Undergrad. Texts Math., Springer-Verlag, 1992.
[3] T. Kalker, H. Park, and M. Vetterli, *Groebner bases techniques in multidimensional multirate systems*, in Proceedings of ICASSP 95, 1995, pp. 2121–2124.
[4] T. Y. Lam, *Serre's Conjecture*, Lecture Notes in Math., 635, Springer-Verlag, New York, 1978.
[5] H. Park, T. Kalker, and M. Vetterli, *Gröbner bases and multidimensional FIR multirate systems*, Multidimens. Systems Signal Process., 8 (1997), pp. 11–30.
[6] H. Park and C. Woodburn, *An algorithmic proof of Suslin's stability theorem for polynomial rings*, J. Algebra, 178 (1995), pp. 277–298.
[7] A. A. Suslin, *On the structure of the special linear group over polynomial rings*, Math. USSR Izv., 11 (1977), pp. 221–238.
[8] L. Tolhuizen, H. Hollmann, and A. Kalker, *On the realizability of bi-orthogonal m-dimensional 2-band filter banks*, IEEE Trans. Signal Process., 43 (1995), pp. 640–648.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 1, pp. 185–194
© 1999 Society for Industrial and Applied Mathematics

TIKHONOV REGULARIZATION AND TOTAL LEAST SQUARES*

GENE H. GOLUB†, PER CHRISTIAN HANSEN‡, AND DIANNE P. O'LEARY§

Abstract. Discretizations of inverse problems lead to systems of linear equations with a highly ill-conditioned coefficient matrix, and in order to compute stable solutions to these systems it is necessary to apply regularization methods. We show how Tikhonov's regularization method, which in its original formulation involves a least squares problem, can be recast in a total least squares formulation suited for problems in which both the coefficient matrix and the right-hand side are known only approximately. We analyze the regularizing properties of this method and demonstrate by a numerical example that, in certain cases with large perturbations, the new method is superior to standard regularization methods.

Key words. total least squares, discrete ill-posed problems, regularization, bidiagonalization

AMS subject classifications. 65F20, 65F30

PII. S0895479897326432

1. Introduction.

In this paper we study a class of methods for producing an approximate solution to a linear system of equations $A\,x \approx b$, where A is $m \times n$ with $m \geq n$. We assume that the elements of A and those of b are contaminated by some noise. An appropriate statement of the problem in the case of noisy data is the *total least squares* (TLS) formulation [5], [6, section 12.3], [17]:

$$\min_{\tilde{A},\tilde{b},x} \|(A\,,\,b) - (\tilde{A}\,,\,\tilde{b})\|_F \qquad \text{subject to} \qquad \tilde{b} = \tilde{A}\,x. \tag{1}$$

In contrast, if the elements of A are exact and only b contains noise, then the corresponding formulation is the *least squares* (LS) problem

$$\min_{x} \|Ax - b\|_2. \tag{2}$$

Our focus in this work is on very ill-conditioned problems where the singular values of A decay gradually to zero. Such problems arise, for example, from the discretization of ill-posed problems such as integral equations of the first kind. See, e.g., [2], [3], [7], [8], and [10] for examples and details.

In these problems, the solutions to the two formulations in (1) and (2) can be hopelessly contaminated by the noise in directions corresponding to small singular values of A or $(A\,,\,b)$. Because of this, it is necessary to compute a *regularized solution* in which the effect of such noise is filtered out. For LS problems, the filtering is often done by Tikhonov's regularization method, where a damping is added to each *singular value decomposition* (SVD) component of the solution, thus effectively filtering out

*Received by the editors August 25, 1997; accepted for publication (in revised form) by J. Varah October 30, 1998; published electronically August 26, 1999.

http://www.siam.org/journals/simax/21-1/32643.html

†Department of Computer Science, Stanford University, Stanford, CA 94305 (golub@sccm.stanford.edu). The work of this author was supported by National Science Foundation grant CCR 95-05393.

‡Department of Mathematical Modelling, Technical University of Denmark, Building 321, DK-2800 Lyngby, Denmark (pch@imm.dtu.dk).

§Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 (oleary@cs.umd.edu). The work of this author was supported by National Science Foundation grant CCR 95-03126 and by a General Research Board grant from the Office of Graduate Studies and Research of the University of Maryland.

the components corresponding to the small singular values. Alternatively, one can truncate the SVD expansion of the solution, leaving out all the SVD components corresponding to the small singular values.

For the TLS problem, the truncation approach has already been studied by Fierro et al. [4]. In the present work we focus on the Tikhonov regularization approach for TLS. We thus arrive at a new regularization method in which stabilization enters the formulation in a natural way and which is able to produce regularized solutions with superior properties for certain problems in which the perturbations are large.

Our paper is organized as follows. In section 2 we introduce the regularized TLS method and we study its regularizing properties. Computational aspects are described in section 3, and we conclude the paper with a numerical example in section 4. A preliminary report on this work appeared as [12].

2. The regularized TLS method. Our regularization of the TLS problem is based on Tikhonov regularization. For the linear LS problem (2), a general version of Tikhonov's method [16] takes the form

$$\min \left\{ \|A\,x - b\|_2^2 + \lambda\,\|L\,x\|_2^2 \right\}, \tag{3}$$

where λ is a positive constant chosen to control the size of the solution vector and L is a matrix that defines a (semi)norm on the solution through which the "size" is measured [11]. Often, L represents the first or second derivative operator. If L is the identity matrix, then the Tikhonov problem is said to be in standard form. The solution x_λ to (3) solves the problem

$$(A^T A + \lambda\,L^T L)x = A^T b. \tag{4}$$

As λ increases, the (semi)norm $\|L\,x\|_2$ of the solution vector decreases monotonically while the residual $\|A\,x - b\|_2$ increases monotonically.

Tikhonov regularization has an important equivalent formulation as

$$\min \|A\,x - b\|_2 \qquad \text{subject to} \qquad \|L\,x\|_2 \le \delta, \tag{5}$$

where δ is a positive constant. Problem (5) is an LS problem with a quadratic constraint and, using the Lagrange multiplier formulation,

$$\mathcal{L}(x,\lambda) = \|A\,x - b\|_2^2 + \lambda\,\left(\|L\,x\|_2^2 - \delta^2\right), \tag{6}$$

it can be shown that if $\delta \le \|x_{\mathrm{LS}}\|_2$, where x_{LS} is the LS solution to (2), then the solution x_δ to (5) is identical to the solution x_λ to (3) for an appropriately chosen λ, and there is a monotonic relation between the parameters δ and λ.

To carry this idea over to the TLS setting, we add an upper bound on the (semi)norm $\|L\,x\|_2$ of the solution vector x in the TLS problem (1). The formulation of the regularized TLS (R-TLS) problem thus becomes

$$\min \|(A\,,\,b) - (\tilde{A}\,,\,\tilde{b})\|_F \qquad \text{subject to} \qquad \tilde{b} = \tilde{A}\,x, \quad \|L\,x\|_2 \le \delta, \tag{7}$$

and the corresponding Lagrange multiplier formulation is

$$\hat{\mathcal{L}}(\tilde{A},x,\mu) = \|(A\,,\,b) - (\tilde{A}\,,\,\tilde{A}\,x)\|_F^2 + \mu\,(\|L\,x\|_2^2 - \delta^2), \tag{8}$$

where μ is the Lagrange multiplier, zero if the inequality constraint is inactive. The solution $\bar{x}_\delta$ to this problem is different from the solution x_{TLS} to (1) whenever δ is

less than $\|L\,x_{\mathrm{TLS}}\|_2$. The two solutions $\bar{x}_\delta$ and x_δ to the two regularized problems in (5) and (7) have a surprising relationship, explained by the following theorem.

THEOREM 2.1. *The R-TLS solution $\bar{x}_\delta$ to* (7), *with the inequality constraint replaced by equality, is a solution to the problem*

$$(A^TA + \lambda_I I_n + \lambda_L L^TL)\,x = A^Tb, \tag{9}$$

where the parameters λ_I and λ_L are given by

$$\lambda_I = -\,\frac{\|b - A\,x\|_2^2}{1 + \|x\|_2^2}, \tag{10}$$

$$\lambda_L = \mu\,(1 + \|x\|_2^2) \tag{11}$$

and where μ is the Lagrange multiplier in (8). *The two parameters are related by*

$$\lambda_L\,\delta^2 = b^T(b - A\,x) + \lambda_I. \tag{12}$$

Moreover, the TLS residual satisfies

$$\|(A\,,\,b) - (\tilde{A}\,,\,\tilde{b})\|_F^2 = -\lambda_I. \tag{13}$$

Proof. Define $\tilde{r} \equiv b - \tilde{b} = b - \tilde{A}\,x$. We characterize the solution to (7) by setting the partial derivatives of the Lagrangian $\hat{\mathcal{L}}$ (8) to zero. Differentiation with respect to the entries in $\tilde{A}$ yields

$$\tilde{A} - A - \tilde{r}x^T = 0,$$

and differentiation with respect to the entries in x yields

$$-\tilde{A}^T\tilde{r} + \mu\,L^TL\,x = 0 \tag{14}$$

or

$$(\tilde{A}^T\tilde{A} + \mu\,L^TL)\,x = \tilde{A}^Tb.$$

Since $A = \tilde{A} - \tilde{r}x^T$ and $\tilde{A}^T\tilde{r} = \mu\,L^TL\,x$, we see that

$$\begin{aligned} A^TA &= (\tilde{A}^T - x\tilde{r}^T)(\tilde{A} - \tilde{r}x^T) \\ &= \tilde{A}^T\tilde{A} - \mu\,xx^TL^TL + x\tilde{r}^T\tilde{r}x^T - \mu\,L^TL\,xx^T \end{aligned}$$

and

$$\tilde{A}^Tb = A^Tb + x\tilde{r}^Tb.$$

Gathering terms, we arrive at (9), with

$$\begin{aligned} \lambda_I &= \mu\,\delta^2 - \|\tilde{r}\|_2^2\,\|x\|_2^2 - \tilde{r}^Tb, \\ \lambda_L &= \mu\,(1 + \|x\|_2^2). \end{aligned}$$

To simplify the expression for λ_I, we first rewrite $\tilde{r}$ as

$$\tilde{r} = b - \tilde{A}\,x = b - (A + \tilde{r}\,x^T)\,x = b - A\,x - \tilde{r}\,x^Tx,$$

from which we obtain the relation

$$\tilde{r}\,(1+\|x\|_2^2) = b - A\,x \tag{15}$$

and, therefore, $(1+\|x\|_2^2)\|\tilde{r}\|_2^2 = (1+\|x\|_2^2)^{-1}\|b - A\,x\|_2^2$. In (14), multiplication from the left by x^T leads to

$$\mu = \frac{x^T\tilde{A}^T\tilde{r}}{x^T L^T L x} = \frac{b^T\tilde{r} - \|\tilde{r}\|_2^2}{\delta^2}\ . \tag{16}$$

Inserting (15) and (16) into the above expression for λ_I, we obtain (10). Equation (12) is proved by multiplying λ_L by δ^2 and inserting (15) and (16). Finally, to prove the expression for the TLS residual, we use the relation

$$(A\,,\,b) - (\tilde{A}\,,\,\tilde{b}) = (A-\tilde{A}\,,\,b-\tilde{b}) = (-\tilde{r}\,x^T\,,\,\tilde{r}) = -\tilde{r}\begin{pmatrix} x \\ -1 \end{pmatrix}^T .$$

Taking the Frobenius norm and using (15), we obtain (13). □

This theorem characterizes the solution to the problem in (7) whenever δ is small enough that the constraint on Lx is binding. When the solution $\bar{x}_\delta$ to (7) satisfies $\|L\bar{x}_\delta\|_2 < \delta$(with $L = 0$ as a special case), then (7) is just the standard TLS problem.

Below, we discuss the implications of this theorem for both the standard-form case $L = I_n$ (the identity matrix) and the general-form case $L \neq I_n$, both of which are important in applications.

2.1. The standard-form case. In the standard-form case, the problems simplify considerably. The Tikhonov problem (5) becomes

$$\min \|A\,x - b\|_2 \qquad \text{subject to} \qquad \|x\|_2 \leq \delta, \tag{17}$$

with solution x_δ satisfying

$$(A^T A + \lambda\, I_n)x_\delta = A^T b \tag{18}$$

for some value of $\lambda \geq 0$. Similarly, the R-TLS problem takes the form

$$\min \|(A\,,\,b) - (\tilde{A}\,,\,\tilde{b})\|_F \qquad \text{subject to} \qquad \tilde{b} = \tilde{A}\,x, \quad \|x\|_2 \leq \delta, \tag{19}$$

and (9) says that the solution $\bar{x}_\delta$ satisfies

$$(A^T A + \lambda_{IL} I_n)\,x = A^T b \tag{20}$$

whenever $\|\bar{x}_\delta\|_2 > \delta$, with $\lambda_{IL} = \lambda_I + \lambda_L$. Clearly, these two solutions are closely related.

THEOREM 2.2. *Let $L = I_n$ and let $\bar{\sigma}_{n+1}$ denote the smallest singular value of $(A\,,\,b)$. Then, for each value of δ, the resulting solutions $\bar{x}_\delta$ and x_δ are related as in the following chart.*

δ	Solutions	λ_{IL}
$\delta < \|x_{\rm LS}\|_2$	$\bar{x}_\delta = x_\delta$	$\lambda_{IL} > 0$
$\delta = \|x_{\rm LS}\|_2$	$\bar{x}_\delta = x_\delta = x_{\rm LS}$	$\lambda_{IL} = 0$
$\|x_{\rm LS}\|_2 < \delta < \|x_{\rm TLS}\|_2$	$\bar{x}_\delta \neq x_\delta = x_{\rm LS}$	$0 > \lambda_{IL} > -\bar{\sigma}_{n+1}^2$
$\delta \geq \|x_{\rm TLS}\|_2$	$\bar{x}_\delta = x_{\rm TLS},\ x_\delta = x_{\rm LS}$	$\lambda_{IL} = -\bar{\sigma}_{n+1}^2$

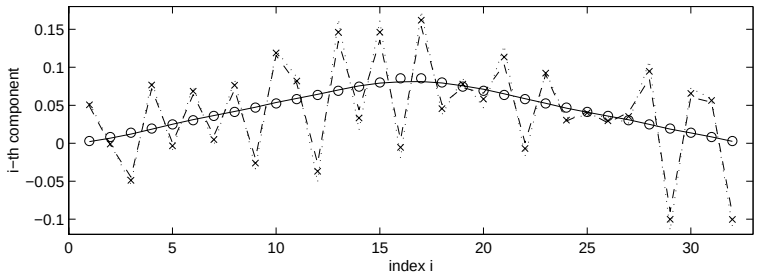


FIG. 1. *Regularized solutions to* (3) *for various values of the regularization parameter* λ; *see Table* 1 *for details.*

TABLE 1
The solutions shown in Figure 1.

Solution	λ	Solution norm	Symbol
x_λ	10^{-5}	0.289	solid line
x_{LS}	0	0.385	dashed line
x_λ	$-\frac{1}{2}\bar{\sigma}_{n+1}^2$	0.408	crosses
x_{TLS}	$-\bar{\sigma}_{n+1}^2$	0.443	dotted line
x^{exact}	–	0.288	circles

Proof. Consider the LS problem (17) first. If $\delta \geq \|x_{\mathrm{LS}}\|$, then the constraint is not binding and the solution to the problem is x_{LS}, with $\lambda = 0$. As δ is reduced below $\|x_{\mathrm{LS}}\|$, the value of λ increases monotonically, as can be seen by using the SVD $A = \Sigma_{i=1}^n u_i \sigma_i v_i^T$ in (18) to obtain

$$x_\delta = \sum_{i=1}^{n} \frac{\sigma_i u_i^T b}{\sigma_i^2 + \lambda} v_i \tag{21}$$

and then taking norms.

Similarly, as long as $\delta \geq \|x_{\mathrm{TLS}}\|$, the constraint in (19) is not binding, and in this case we have the standard TLS problem with $\lambda_{IL} = -\bar{\sigma}_{n+1}^2$ [17, Theorem 2.7]. As δ is reduced below $\|x_{\mathrm{TLS}}\|$, (21) again shows that the λ_{IL} value must increase monotonically in order to reduce the solution norm.

Noting that $\|x_{\mathrm{LS}}\| \leq \|x_{\mathrm{TLS}}\|$ [17, Corollary 6.2], the theorem's conclusion follows. □

In regularization problems, the norm $\|x_{\mathrm{LS}}\|_2$ of the LS solution is normally very large due to the errors; thus, one will choose $\delta \leq \|x_{\mathrm{LS}}\|_2$. When this is the case, we conclude from Theorem 2.2 that R-TLS produces solutions that are identical to the Tikhonov solutions. In other words, replacing the LS residual with the TLS residual in the Tikhonov formulation has no effect when $L = I_n$ and $\delta \leq \|x_{\mathrm{LS}}\|_2$.

We remark that since $\|x_{\mathrm{TLS}}\|_2 \geq \|x_{\mathrm{LS}}\|_2$, there is usually a nontrivial set of "large" values of δ for which the multiplier λ_{IL} is negative. The corresponding R-TLS solutions $\bar{x}_\delta$ are distinctly different from the Tikhonov solutions and can be expected to be even more dominated by errors than the LS solution x_{LS}.

We illustrate Theorem 2.2 with an example: discretization of a Fredholm integral equation with the second derivative operator as kernel. The implementation is deriv2 from [9], the size of the matrix A is 64×32, and both A and b are perturbed by

Gaussian noise with zero mean and standard deviation 10^{-5}. We have $\bar{\sigma}_{n+1}^2 \approx 2.38 \cdot 10^{-9}$. Figure 1 shows the solutions listed in Table 1. We see that both x_{LS} and x_{TLS} have large oscillations (the oscillations of x_{TLS} being larger than those of x_{LS}), while the regularized LS solution x_λ corresponding to $\lambda = 10^{-5}$ is much smoother. The R-TLS solution, corresponding to $\lambda = -\frac{1}{2}\bar{\sigma}_{n+1}^2 = -1.19 \cdot 10^{-9}$, has large oscillations whose amplitude is between that of x_{TLS} and x_{LS}.

We conclude that if we use Tikhonov regularization with $L = I$ in order to reduce the norm of the solution below $\|x_{\mathrm{LS}}\|_2$, then the R-TLS formulation produces the same solution as the regularized LS formulation. This is in contrast to the truncating approach, since we have shown in [4] that truncated TLS can be superior to truncated LS.

2.2. The general-form case. In many applications, it is necessary to choose a matrix L different from the identity matrix; these issues are discussed, e.g., in [10, section 4.3]. In this case, the R-TLS solution $\bar{x}_\delta$ is different from the Tikhonov solution whenever the residual $b - Ax$ is different from zero, since both λ_I and λ_L are nonzero.

From Theorem 2.1 we notice that λ_L is always positive when $\delta < \|x_{\mathrm{TLS}}\|_2$, because the Lagrange parameter μ is positive for these values of λ. On the other hand, λ_I is always negative and thus adds some *deregularization* to the solution. Statistical aspects of a negative regularization parameter in Tikhonov's method are discussed in [13].

For a given δ, there are usually several pairs of parameters λ_I and λ_L (and thus several solutions x) that satisfy relations (9)–(11), but only one of these solves the optimization problem (7). According to (13), this is the solution that corresponds to the smallest value of $|\lambda_I|$. The following relations hold.

THEOREM 2.3. *For a given value of δ, the solution $\bar{x}_\delta$ to the R-TLS problem is related to the solution to the TLS problem x_{TLS} as in the following chart.*

δ	Solution	λ_I	λ_L
$\delta < \|L\,x_{\mathrm{TLS}}\|_2$	$\bar{x}_\delta \neq x_{\mathrm{TLS}}$	$\lambda_I < 0 \quad \partial\lambda_I/\partial\delta > 0$	$\lambda_L > 0$
$\delta \geq \|L\,x_{\mathrm{TLS}}\|_2$	$\bar{x}_\delta = x_{\mathrm{TLS}}$	$\lambda_I = -\bar{\sigma}_{n+1}^2$	$\lambda_L = 0$

Proof. For $\delta < \|L\,x_{\mathrm{TLS}}\|_2$, the inequality constraint is binding and the Lagrange multiplier μ is positive, since this is a necessary condition for optimality; see, for example, [14]. Thus λ_L is positive. The residual (13) is monotonically decreasing as δ increases, since optimal solutions for smaller values of δ are candidate solutions for larger δs, so λ_I is monotonically increasing.

For $\delta = \|L\,x_{\mathrm{TLS}}\|_2$, the Lagrange multiplier is zero, and the solution becomes the unconstrained minimizer x_{TLS}. The value $-\bar{\sigma}_{n+1}^2$ follows from [5, Theorem 4.1]. The constraint is never again binding for larger δ, so the solution to the problem remains unchanged. ∎

We note that if the matrix $\lambda_I I_n + \lambda_L L^T L$ is positive definite, then the R-TLS solution corresponds to a Tikhonov solution for which the expression $\lambda\,\|L\,x\|_2$ in (3) is replaced with the norm $(\lambda_I\|x\|_2^2 + \lambda_L\|L\,x\|_2^2)^{1/2}$. If $\lambda_I I_n + \lambda_L L^T L$ is indefinite or negative definite then there is no equivalent interpretation.

3. Computational aspects. To compute the R-TLS solutions for $L \neq I_n$, we have found it most convenient to avoid explicit use of δ; instead we use λ_L as the free parameter, fixing its value and then computing the value of λ_I that satisfies (10) and is smallest in absolute value. The corresponding value of δ can then easily be computed from relation (12).

We now discuss how to solve (9) efficiently for many values of λ_I and λ_L. First, we note that the equation is equivalent to the augmented system

$$\begin{pmatrix} I_m & 0 & A \\ 0 & I_p & \lambda_L^{1/2} L \\ A^T & \lambda_L^{1/2} L^T & -\lambda_I I_n \end{pmatrix} \begin{pmatrix} r \\ s \\ x \end{pmatrix} = \begin{pmatrix} b \\ 0 \\ 0 \end{pmatrix}. \tag{22}$$

Our algorithm is based on this formulation.

We assume that the matrix L is a banded matrix, often the case when L represents a derivative operator. The key to efficiency, then, is to reduce A to $n \times n$ bidiagonal form B by means of orthogonal transformations: $H^T A K = B$. The orthogonal right transformations also should be applied to L, and simultaneously we should apply orthogonal transformations to L from the left in order to maintain its banded form. It is convenient to use sequences of Givens transformations to form J, H, and K, since this gives us the most freedom to retain the banded form of $C = J^T L K$.

Once B and C have been computed, we can recast the augmented system in (22) in the following form:

$$\begin{pmatrix} I_n & 0 & B \\ 0 & I_p & \lambda_L^{1/2} C \\ B^T & \lambda_L^{1/2} C^T & -\lambda_I I_n \end{pmatrix} \begin{pmatrix} H^T r \\ J^T s \\ K^T x \end{pmatrix} = \begin{pmatrix} H^T b \\ 0 \\ 0 \end{pmatrix}. \tag{23}$$

Since λ_I changes more frequently than λ_L in our approach, we will now use Givens rotations to annihilate $\lambda_L^{1/2} C$ using B by means of Elden's algorithm [1, section 5.3.4], which can be represented as

$$\begin{pmatrix} B \\ \lambda_L^{1/2} C \end{pmatrix} = G \begin{pmatrix} \hat{B} \\ 0 \end{pmatrix} = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix} \begin{pmatrix} \hat{B} \\ 0 \end{pmatrix}.$$

When we insert this G into the augmented system (23), it becomes

$$\begin{pmatrix} I_n & 0 & \hat{B} \\ 0 & I_p & 0 \\ \hat{B}^T & 0 & -\lambda_I I_n \end{pmatrix} \begin{pmatrix} \hat{r} \\ \hat{s} \\ K^T x \end{pmatrix} = \begin{pmatrix} G_{11}^T H^T b \\ G_{12}^T H^T b \\ 0 \end{pmatrix}.$$

The middle block row is now decoupled, and we obtain

$$\begin{pmatrix} I_n & \hat{B} \\ \hat{B}^T & -\lambda_I I_n \end{pmatrix} \begin{pmatrix} \hat{r} \\ K^T x \end{pmatrix} = \begin{pmatrix} G_{11}^T H^T b \\ 0 \end{pmatrix}.$$

Finally, we apply a symmetric perfect shuffle reordering

$$n+1, 1, n+2, 2, n+3, 3, \ldots, n, 2n$$

to the rows and columns of the above matrix, to obtain a symmetric, tridiagonal, indefinite matrix of size $2n \times 2n$:

$$\begin{pmatrix} -\lambda_I & \hat{b}_{11} & & & \\ \hat{b}_{11} & 1 & \hat{b}_{12} & & \\ & \hat{b}_{12} & -\lambda_I & \hat{b}_{22} & \\ & & \hat{b}_{22} & 1 & \ddots \\ & & & \ddots & \ddots \end{pmatrix}.$$

Up until now, we have used only orthogonal transformations in order to preserve numerical stability. We can solve this permuted system in a stable way using a general tridiagonal solver (e.g., Gauss elimination with partial pivoting).

4. Numerical results. In this section we present a numerical example that demonstrates the usefulness of the R-TLS method. Our computations are carried out in MATLAB using the *Regularization Tools* package [9].

It is a generally accepted fact that, for small noise levels, we should not expect the ordinary TLS solution to differ much from the ordinary LS solution; see [15]. The same observation is made in [4] for the T-TLS solution, and the numerical results presented below also support this observation for the R-TLS method. We emphasize that the precise meaning of "small" depends on the particular problem.

In several test problems from [9], the R-TLS algorithm was able to compute better results than Tikhonov's method. The test problem we have chosen to illustrate the R-TLS algorithm is the one for which the improvement is most dramatic. This test problem is a discretization by means of Gauss–Laguerre quadrature of the inverse Laplace transform

$$\int_0^\infty \exp(-s\,t)\,f(t)\,dt = \frac{1}{2} - \frac{1}{s+4/25}, \quad 0 \le s,$$

$$f(t) = 1 - \exp(-t\,4/25),$$

originating from [18] and implemented in the function ilaplace(n,2) in [9], except that the constant $4/25$ was chosen instead of $1/2$. The matrix L approximates the first-derivative operator. The dimensions are $m = n = 16$, the matrix A and the exact solution x^* are scaled such that $\|A\|_F = \|A\,x^*\|_2 = 1$, and the perturbed right-hand side is generated as

$$b = (A + \sigma\|E\|_F^{-1}E)x^* + \sigma\|e\|_2^{-1}e,$$

where the elements of the perturbations E and e are from a normal distribution with zero mean and unit standard deviation.

Figure 2 shows the relative errors $\|x^* - x_\delta\|_2/\|x^*\|_2$ and $\|x^* - \bar{x}_\delta\|_2/\|x^*\|_2$ in the Tikhonov and R-TLS solutions, respectively, for four values of the noise level:

$$\sigma = 0.001,\ 0.01,\ 0.1,\ 1.$$

We see that, for small values of σ, the two methods lead to almost the same minimum relative error, for almost the same value of λ_L. However, for larger values of σ, the minimum relative error for the R-TLS method is clearly smaller than that for Tikhonov's method, and it occurs for a larger value of λ_L. This shows the potential advantage of the R-TLS method, provided, of course, that a good estimate of the optimal regularization parameter can be found. This topic is outside the scope of the current paper.

In Figure 3 we have plotted the "optimal" Tikhonov and R-TLS solutions, defined as the solutions that correspond to the minima of the curves in Figure 2. In addition, we have plotted the exact solution x^*. Clearly, the addition of the term $\lambda_I I_n$ in (9) introduces a nonconstant component in the right part of the plot of the regularized solution; it is precisely this component that improves the R-TLS error compared to the Tikhonov error.

Conclusion. We have proved that the R-TLS solution to the regularized form (7) of the TLS problem is closely related to the Tikhonov solution to (3). In the standard-form case, $L = I_n$, the two solutions are identical, showing that Tikhonov

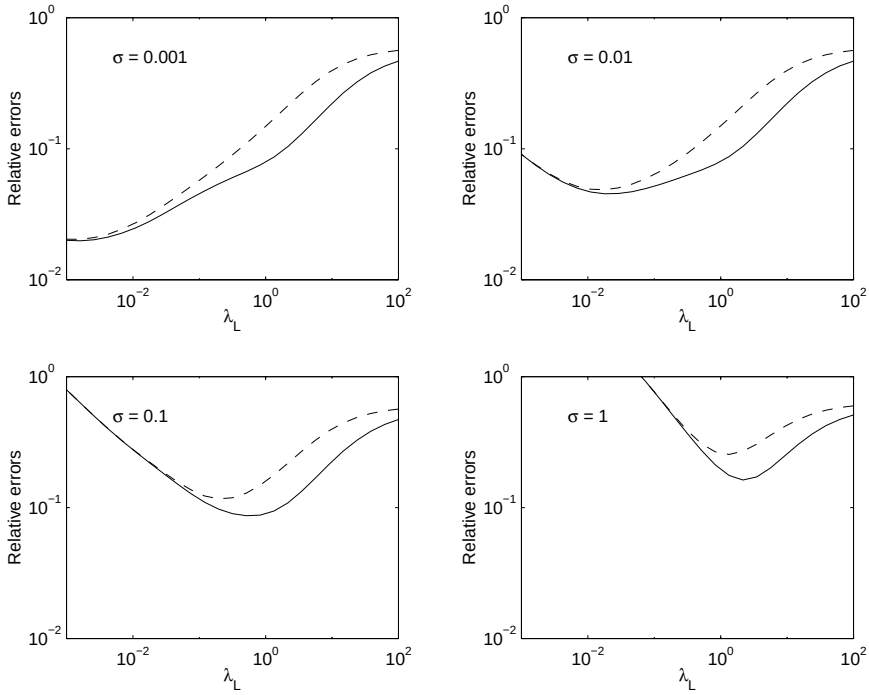


FIG. 2. *Plots of the relative errors in the Tikhonov solutions (dashed lines) and R-TLS solutions (solid lines) versus λ_L for four values of the noise level.*

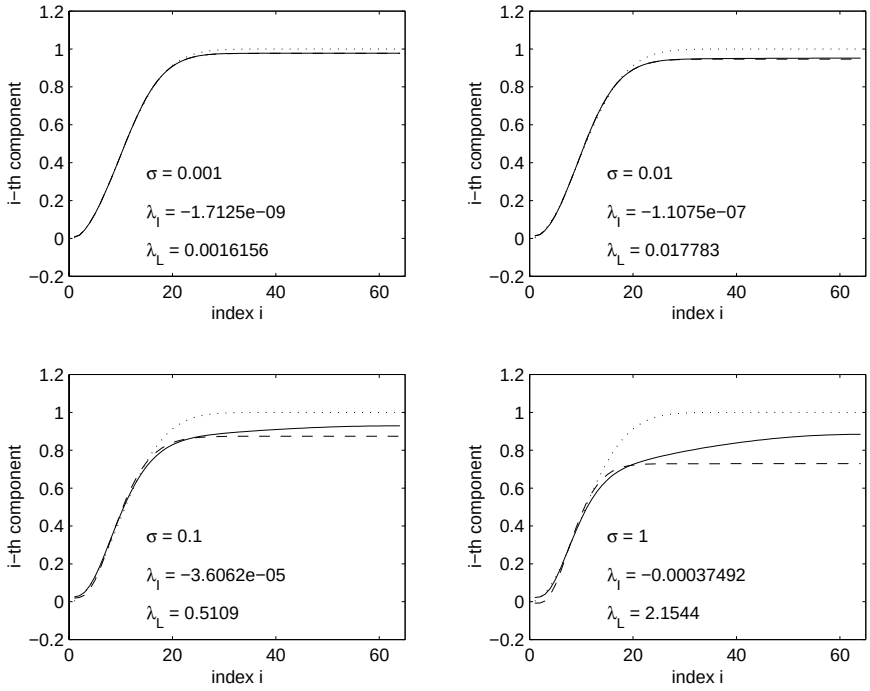


FIG. 3. *Plots of the "optimal" Tikhonov solutions (dashed lines) and R-TLS solutions (solid lines) for four values of the noise level. Also shown is the exact solution (dotted lines).*

regularization in this case is also suited for problems with a perturbed coefficient matrix. For general problems with $L \neq I_n$, the R-TLS solution is different from the corresponding Tikhonov solution in that a deregularizing term $\lambda_I I_n$ with a negative parameter λ_I is added to the formulation in (4).

REFERENCES

[1] A. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, 1996.

[2] H. W. Engl, *Regularization methods for the stable solution of inverse problems*, Surveys Math. Indust., 3 (1993), pp. 71–143.

[3] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer, Dordrecht, The Netherlands, 1996.

[4] R. D. Fierro, G. H. Golub, P. C. Hansen, and D. P. O'Leary, *Regularization by truncated total least squares*, SIAM J. Sci. Comput., 18 (1997), pp. 1223–1241.

[5] G. H. Golub and C. F. Van Loan, *An analysis of the total least squares problem*, SIAM J. Numer. Anal., 17 (1980), pp. 883–893.

[6] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, 1996.

[7] C. W. Groetsch, *Inverse Problems in the Mathematical Sciences*, Vieweg, Wiesbaden, Germany, 1993.

[8] M. Hanke and P. C. Hansen, *Regularization methods for large-scale problems*, Surveys Math. Indust., 3 (1993), pp. 253–315.

[9] P. C. Hansen, *Regularization tools, a MATLAB package for analysis of discrete regularization problems*, Numer. Algorithms, 6 (1994), pp. 1–35.

[10] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, SIAM, Philadelphia, 1998.

[11] P. C. Hansen and D. P. O'Leary, *The use of the L-curve in the regularization of discrete ill-posed problems*, SIAM J. Sci. Comput., 14 (1993), pp. 1487–1503.

[12] P. C. Hansen and D. P. O'Leary, *Regularization algorithms based on total least squares*, in *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling*, S. Van Huffel, ed., SIAM, Philadelphia, 1997, pp. 127–137.

[13] T. A. Hua and R. F. Gunst, *Generalized ridge regression: A note on negative ridge parameters*, Comm. Stat. Theory Methods, 12 (1983), pp. 37–45.

[14] S. G. Nash and A. Sofer, *Linear and Nonlinear Programming*, McGraw–Hill, New York, 1996.

[15] G. W. Stewart, *On the invariance of perturbed null vectors under column scaling*, Numer. Math., 44 (1984), pp. 61–65.

[16] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*, Wiley, New York, 1977.

[17] S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*, SIAM, Philadelphia, 1991.

[18] J. M. Varah, *On the numerical solution of ill-conditioned linear systems with applications to ill-posed problems*, SIAM J. Numer. Anal., 10 (1973), pp. 257–267.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 1, pp. 195–208
© 1999 Society for Industrial and Applied Mathematics

ON THE OPTIMAL VALUE OF THE SPECTRAL ABSCISSA FOR A SYSTEM OF LINEAR OSCILLATORS*

PEDRO FREITAS† AND PETER LANCASTER‡

Abstract. We consider the linear system of differential equations $Au''(t) + Bu'(t) + Cu(t) = 0$, where A, B, and C are $n \times n$ real symmetric matrices, and study the problem of, given A and C, finding B which optimizes the asymptotic rate of decay of solutions of this equation. The optimal value for this rate of decay is found and we show that it is attained within this class of matrices. We then study some of the properties of optimal matrices. In particular, we show that the solution is not unique in general and that optimal matrices do not have to be positive definite. A sufficient condition for all optimal matrices to be positive definite is given and some examples are discussed.

Key words. symmetric quadratic pencils, spectral abscissa

AMS subject classifications. 15A18, 15A22, 19J52

PII. S0895479897331850

1. Introduction.

By a system of linear oscillators we mean here a linear time-invariant differential system

$$\begin{cases} Au''(t) + Bu'(t) + Cu(t) = 0, \\ u(0) = u_0, \quad u'(0) = u_1, \end{cases} \tag{1.1}$$

where A, B, and C are $n \times n$ real symmetric matrices. We assume A and C to be positive definite, and write $A, C > 0$. In the engineering literature, these coefficients are called the mass and stiffness of the system, and are associated with kinetic and potential energies, respectively. The energy of a solution of (1.1) at time t is then defined to be

$$E(t) = \frac{1}{2}\langle u'(t), Au'(t)\rangle + \frac{1}{2}\langle u(t), Cu(t)\rangle, \tag{1.2}$$

where $\langle \cdot\, , \cdot \rangle$ denotes the usual inner product on $\mathbb{R}^n$.

Our objective here is the solution of the following optimization problem: given A and C as above, determine a real symmetric $n \times n$ matrix B which will, in a suitable sense (see section 2), maximize the rate of decay of $E(t)$ as t goes to infinity (more precisely, maximize the minimal asymptotic rate of decay of $E(t)$ with respect to solutions $u(t)$). In physical terms, this corresponds to the determination of the damping that causes solutions to converge to zero in the (asymptotically) fastest possible way.

This question and similar questions have received some attention in the mathematics literature but, so far, the problem has been solved only under restrictive

*Received by the editors December 23, 1997; accepted for publication (in revised form) by U. Helmke November 19, 1998; published electronically September 1, 1999.

http://www.siam.org/journals/simax/21-1/33185.html

†Departamento de Matemática, Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisboa Codex, Portugal (pfreitas@math.ist.utl.pt). The work of this author was partially supported by FCT (Portugal) under projects PRAXIS/PCEX/P/MAT/36/96 and PRAXIS XXI project 2/2.1/MAT/199/94.

‡Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta T2N 1N4, Canada (lancaste@acs.ucalgary.ca). The work of this author was partially supported by the Natural Science and Engineering Research Council of Canada.

conditions. See [3], for instance. For the general problem, Cox showed that the optimal rate of decay is bounded by the square root of the largest eigenvalue of the matrix $A^{-1}C$.

From the simple case where n is 1 it is known that the function that gives the spectral abscissa in terms of the entries of B is not differentiable at all points. This turns out to be quite a general situation, in the sense that this type of optimization problem tends to cause the coalescence of points of the spectrum, giving rise to eigenvalues which have different geometric and algebraic multiplicities (see [1]).

In [4], for instance, the authors considered the problem of optimizing the spectral abscissa for the wave equation and studied the perturbation along straight lines of a double eigenvalue occurring in the case of constant damping. At the end of that paper they point out that a similar approach can be used for the case of equation (1.1). However, this method cannot, without any further calculations, distinguish between a saddle point and an optimum. In particular, the optimum produced in [3] under one of the restrictions mentioned above also gives rise to a double eigenvalue and a similar perturbation analysis will show that this is optimal for matrices $B + \epsilon\Gamma$.

In relation to this it should be pointed out that the problem of optimizing the spectral abscissa is global, in the sense that it requires full knowledge of the spectrum. Because of this, in the present work we shall use a different strategy which will allow us to improve Cox's bounds and show that in fact the best possible decay rate for the energy is given by

$$\left[\frac{\det(C)}{\det(A)}\right]^{1/n}.$$

We then prove that this value is indeed attained for some real symmetric matrix B.

In this case we shall solve the optimization problem by showing that the above bound can only be attained if the associated eigenvalue problem has a single eigenvalue of multiplicity $2n$. This allows us to reformulate the problem as an inverse spectral problem: given A and C, find B such that

$$-\left[\frac{\det(C)}{\det(A)}\right]^{1/2n}$$

is an eigenvalue of the quadratic matrix polynomial $L(\lambda) = \lambda^2 A + \lambda B + C$ with multiplicity $2n$.

Another issue that is important when considering problems of this type is that of positive definiteness of the matrices involved. For example, if $A > 0$ and C is indefinite, then there will exist unbounded solutions irrespectively of B. Regarding the damping term B it is well known that for positive definite A and C the decay rate is necessarily positive if B is a positive definite matrix. Here we will show that optimal rates of decay can be achieved (for the same pair (A, C)) with more than one matrix B, and, under some conditions, even with matrices B which are not positive definite. This is, at least at first, quite a surprising result, as the physical interpretation is that in some sense we have to supply energy to the system for the solutions to decay faster. We give a sufficient condition for all optimal matrices to be positive definite and then, through examples, show that in some cases there is coexistence of indefinite and positive definite optimal matrices and that it may also happen that all optimal matrices are indefinite.

In systems theory language the situation can be described in a different way. Given $A > 0$, $C > 0$, and $B_0 = B_0^T$, consider the system

$$Au''(t) + B_0u'(t) + Cu(t) = B_1v(t)$$

with a symmetric feedback matrix B_1. We show that with *velocity feedback*, $v(t) = u'(t)$ there are matrices B_1 which yield the optimal decay rate described and we examine the cases for which $B := B_0 - B_1 > 0$. As indicated above, it turns out that we solve a *spectral assignment problem*, and the solution by standard system theory methods is far from clear, as we assume C to be fixed [2].

The organization of the paper is as follows. In section 2 we formulate the problem and state our main results in a precise way. In section 3 we establish the optimal value for the rate of decay and show that it can be attained. Section 4 contains a complete analysis of the case where n is two. Properties of optimal matrices are presented in section 5 and a discussion of the elementary divisor structure of optimal $L(\lambda)$ is made in section 6. Section 7 contains a sufficient condition for all optimal matrices to be positive definite and we conclude in section 8 with a discussion of the results presented and another example.

2. Statement of the problem and main results. As the matrix A is assumed to be positive definite, it is possible to make the change of variables $u = A^{-1/2}v$ in (1.1) to obtain

$$v''(t) + A^{-1/2}BA^{-1/2}v'(t) + A^{-1/2}CA^{-1/2}v(t) = 0.$$

From this we see that we can assume, without loss of generality, that A is the identity matrix and then study the system

$$u''(t) + Bu'(t) + Cu(t) = 0, \tag{2.1}$$

where B and C belong to S_n, the space of $n \times n$ real symmetric matrices. We shall assume that C is positive definite with eigenvalues $0 < c_1 \leq \cdots \leq c_n$. In order to study the decay of solutions of this equation, we are then led to the quadratic pencil

$$L(\lambda) = I\lambda^2 + B\lambda + C. \tag{2.2}$$

Denote by Σ_L the spectrum of $L(\lambda)$, that is, the set of complex numbers λ for which there exists a nontrivial vector v such that $L(\lambda)v = 0$, and for each fixed matrix C of the form above, define the function $\Omega : S_n \to \mathbb{R}$ by

$$\Omega(B) = \min_{\lambda \in \Sigma_L} \left[-\mathrm{Re}(\lambda)\right].$$

For each pair of matrices (B, C), this function gives the corresponding spectral abscissa, that is, minus the largest of the real parts of the eigenvalues of $L(\lambda)$. Thus there exists a constant κ such that the energy of solutions of system (2.1) as defined by (1.2) satisfies

$$E(t) \leq \kappa E(0) e^{(-2\omega+\delta)t}$$

with $\omega = \Omega(B)$ and for all positive δ and t. It is clear that each pair (B, C) at which ω is equal to $\Omega(B)$ gives the best possible value that can be considered in this energy inequality.

If we want to maximize the asymptotic rate at which the total energy of the system decays to zero, we must thus maximize the function Ω over the space S_n and determine if this value is attained. Note that, because Ω depends continuously on the entries of B, Ω will attain its maximum on compact subsets of S_n. Thus, since it is known that the spectral abscissa is bounded, it would in principle be possible to find a suitable compact subset of S_n where the maximum over all symmetric matrices would be attained. However, this argument cannot ensure that a larger value could not be obtained over the set of all $n \times n$ matrices. Moreover, it is not even clear at the start that the value of ω_o given below in Theorem 1 can be attained within the set of all $n \times n$ matrices.

Thus our problem reduces to determining the supremum of Ω over the space S_n, that is,

$$\omega_o = \sup_{B \in S_n} \left[\Omega(B)\right]. \tag{2.3}$$

This number will be called the optimal spectral abscissa of (2.2).

In [3] it was shown that the function Ω is bounded from above over the set of all positive definite matrices. More precisely, and in our notation, it was shown that ω_o is smaller than or equal to $c_n^{1/2}$, the square root of the largest eigenvalue of C. However, this bound will not be sharp in general—only in the trivial case when C is scalar—and in Theorem 3.1 we give an improvement of this estimate. We then go on to show that this new bound is actually the best possible, and we prove in Theorem 3.2 that it is attainable. The main results of our paper (in the context of the system (1.1)) can be summarized in the following two theorems, the first concerning existence and nonuniqueness, the second regarding positivity of optimal matrices:

THEOREM 1 (existence). *The function*

$$\Omega(B) = \min_{\lambda \in \Sigma_L} \left[-\mathrm{Re}(\lambda)\right]$$

attains its maximum

$$\omega_o = \left[\frac{\det(C)}{\det(A)}\right]^{1/2n}$$

within the set S_n. The optimal matrix is unique if and only if there exists a positive constant c such that $C = cA$, in which case $B_o = 2c^{1/2}A$.

THEOREM 2 (positivity). *Let $\gamma_1, \ldots, \gamma_n$ denote the eigenvalues of the problem*

$$Cv = \gamma Av$$

in increasing order. If

$$\gamma_1 > \frac{\left(\gamma_2^{1/2} + \cdots + \gamma_n^{1/2}\right)^{2n}}{n^{2n}\gamma_2, \ldots, \gamma_n},$$

then all optimal matrices are positive definite.

Although not optimal (see section 7), this result is important as it says when the optimum can be achieved by means of a positive definite matrix. This may be of interest in applications where there are restrictions that force the damping term to be positive. We also remark that in fact we prove a stronger result which gives a bound on the number of positive eigenvalues of an optimal matrix in terms of the separation of the spectrum (see Theorem 7.1).

3. The optimal rate of decay and attainability.

THEOREM 3.1. *The function Ω is bounded from above and the optimal spectral abscissa satisfies*

$$\omega_o \leq [\det(C)]^{1/(2n)}.$$

Furthermore, equality holds if and only if $L(\lambda)$ has one single (real) eigenvalue with algebraic multiplicity $2n$ and, in this case, $L(\lambda) \geq 0$ for all real λ.

Proof. Define the companion matrix M by

$$M = \begin{bmatrix} 0 & I \\ -C & -B \end{bmatrix}.$$

Then the eigenvalues of M and of the pencil $L(\lambda)$ coincide. Let $\{\lambda_j\}_{j=1}^{2n}$ denote these eigenvalues written in increasing order of their real parts so that $\Omega(B) = -\mathrm{Re}(\lambda_{2n})$ (note that it can thus be assumed that all eigenvalues have negative real parts, for otherwise the zero matrix would give zero as an upper bound for Ω). Now let $p(\lambda)$ denote the characteristic polynomial of M (and thus also of $L(\lambda)$), that is, $p(\lambda) = \det(\lambda I - M) = \det(L(\lambda))$. Then

$$p(0) = \det(M) = \det(C) = \prod_{j=1}^{2n} \lambda_j,$$

and thus this product is independent of the matrix B. This in turn implies that the function Ω is bounded from above.

On the other hand,

$$\det(M) = \prod_{j=1}^{2n} \lambda_j \geq \prod_{j=1}^{2n} \mathrm{Re}(\lambda_j) \geq [\mathrm{Re}(\lambda_{2n})]^{2n} = [\Omega(B)]^{2n}, \tag{3.1}$$

and thus

$$\omega_o \leq [\det(C)]^{1/(2n)}.$$

To prove the second part of the theorem, note first that if $L(\lambda)$ has a single eigenvalue with multiplicity $2n$, then clearly equality will hold. On the other hand, if there are two distinct eigenvalues, then the second inequality used in (3.1) will be strict and ω_o will be strictly smaller than $[\det(C)]^{1/(2n)}$.

To see that $L(\lambda) \geq 0$ for all λ in $\mathbb{R}$, observe that $L(\lambda) > 0$ for all λ sufficiently large, and then use the continuity of the (λ-dependent) eigenvalues of $L(\lambda)$. ∎

REMARK 3.1. This bound actually extends to the space of all real $n \times n$ matrices, as it is determined by C independently of B.

We shall now prove that this bound is actually attained, meaning that there will exist a matrix B in S_n for which $\Omega(B) = [\det(C)]^{1/(2n)}$, and thus the quadratic pencil $L(\lambda)$ will have a multiple eigenvalue with multiplicity $2n$. In this way, the original optimization problem has been transformed into an inverse spectral problem.

THEOREM 3.2. *There exists a matrix B_o in S_n such that*

$$\Omega(B_o) = [\det(C)]^{1/(2n)},$$

that is, the upper bound for ω_o given in Theorem 3.1 *is attained.*

The idea behind the proof of this theorem is as follows. Let W be a real $n \times n$ matrix (not generally symmetric) with one single eigenvalue, say $-\alpha$, with algebraic multiplicity n. Then the quadratic pencil

$$(\lambda I - W^t)(\lambda I - W) = \lambda^2 I - \lambda(W^t + W) + W^t W$$

is symmetric and will have an eigenvalue at $\lambda = -\alpha$ with algebraic multiplicity $2n$. If it is possible to find a matrix W which also satisfies $W^t W = C$, it follows that $[\det(W)]^2 = \det(C)$ and so α must satisfy

$$\alpha = [\det(C)]^{1/(2n)}.$$

If such a matrix W exists, then the matrix $B = -(W^t + W)$ is a solution to the inverse problem and thus also to the optimization problem. We remark that this construction originates in the theory of matrix polynomials which are nonnegative whenever λ is real (see [6]).

We begin our construction of the matrix W by building an upper triangular matrix with the desired properties.

PROPOSITION 3.3. *Let $\omega_o = [\det(C)]^{1/(2n)}$. Then there exists a real upper triangular $n \times n$ matrix J having one single eigenvalue equal to $-\omega_o$ with multiplicity n, and such that $J^t J$ is similar to C.*

As C is symmetric, this result is equivalent to saying that there exists a real upper triangular $n \times n$ matrix J with an eigenvalue at $-\alpha$ with multiplicity n and whose singular values are the square roots of the eigenvalues of C. In order to show that such a matrix exists, we shall make use of the following result of A. Horn [7]. See also the discussion of these inequalities in [8].

THEOREM 3.4 (Horn's theorem). *Let $(\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^n$ and $(\alpha_1, \ldots, \alpha_n) \in \mathbb{C}^n$ be a pair of n-tuples satisfying*

$$|\alpha_1| \geq \cdots \geq |\alpha_n|, \quad \sigma_1 \geq \cdots \geq \sigma_n \geq 0$$

and

$$|\alpha_1| \cdots |\alpha_k| \leq \sigma_1 \cdots \sigma_k, \tag{3.2}$$

for $k = 1, \ldots, n$, with equality for $k = n$.

Then there exists an upper triangular matrix Δ which has the n-tuples $(\alpha_1, \ldots, \alpha_n)$ and $(\sigma_1, \ldots, \sigma_n)$ as its eigenvalues and singular values respectively. Furthermore, if $\alpha_1, \ldots, \alpha_n$ are real, then the matrix Δ can be taken to be real.

The upper triangular property of Δ is not essential in the proof of Proposition 3.3, but is retained here for historical reasons. Relations (3.2) are known as Weyl inequalities, and they establish a relation between the eigenvalues and the singular values of a matrix. Basically, Horn's result is the converse of Weyl's Theorem, showing that these inequalities are also sufficient.

In our case, we have $\alpha_i = -\omega_o$ for $i = 1, \ldots, n$, and we want to build a matrix which will have the square roots of the eigenvalues of C as its singular eigenvalues. We first prove that the corresponding n-tuples built in this way satisfy Weyl's inequalities.

LEMMA 3.5. *If $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$ and $\alpha = (\sigma_1 \cdots \sigma_n)^{1/n}$, then*

$$\alpha^k \leq \sigma_1 \cdots \sigma_k, \quad k = 1, \ldots, n,$$

with equality for $k = n$.

Proof. Equality when k equals n is immediate from the definition of α. For $k < n$ notice first that $\sigma_1 \geq \cdots \geq \sigma_k$ implies

$$(\sigma_1 \cdots \sigma_k)^{n-k} \geq \sigma_k^{k(n-k)}.$$

Similarly, $\sigma_{k+1} \geq \cdots \geq \sigma_n$ gives

$$\sigma_{k+1}^{k(n-k)} \geq (\sigma_{k+1} \cdots \sigma_n)^k$$

and thus $\sigma_k \geq \sigma_{k+1}$ yields

$$(\sigma_1 \cdots \sigma_k)^{n-k} \geq (\sigma_{k+1} \cdots \sigma_n)^k.$$

Hence

$$\begin{aligned} \alpha^k &= (\sigma_1 \cdots \sigma_k)^{k/n} (\sigma_{k+1} \cdots \sigma_n)^{k/n} \\ &\leq (\sigma_1 \cdots \sigma_k)^{k/n} (\sigma_1 \cdots \sigma_k)^{(n-k)/n} \\ &= \sigma_1 \cdots \sigma_k. \qquad \square \end{aligned}$$

REMARK 3.2. This inequality can also be obtained from a well-known property of the arithmetic mean by putting $a_j = \log(\sigma_j)$.

Proof of Theorem 3.2. By Horn's theorem and Lemma 3.5 there will exist a real $n \times n$ matrix J, such that J has $-\alpha = -[\det(C)]^{1/(2n)}$ as its single eigenvalue (with multiplicity n) and the numbers $c_1^{1/2}, \ldots, c_n^{1/2}$ as its singular values. The last statement is equivalent to saying that the matrix $P = J^t J$ has eigenvalues $c_1, \ldots, c_n$, and thus there exists a real orthogonal matrix R such that $C = R^t P R$.

We now claim that the matrix $W = R^t J R$ satisfies the properties mentioned previously, that is, W has one single eigenvalue $-\alpha$ with multiplicity n, and $W^t W = C$. The first property is immediate from the definition of W. For the second, just notice that $W^t W = R^t J^t R R^t J R = R^t J^t J R = R^t P R = C$. From this it follows that the matrix $B_o = -(W^t + W)$ satisfies the conditions of the theorem.

4. The case $n = 2$. At this stage, the only situation for which it is feasible to compute explicitly all possible solutions of the optimization problem (2.3) by hand is when n equals 2. As this case already displays some of the features of the general problem, we shall present a full description of the solution.

We consider the quadratic polynomial (2.2) when n equals 2,

$$B = \begin{bmatrix} b_{11} & b_{12} \\ b_{12} & b_{22} \end{bmatrix} \quad \text{and} \quad C = \begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix}.$$

In this case, the characteristic polynomial associated with $L(\lambda)$ is given by

$$p(\lambda) = \lambda^4 + \operatorname{trace}(B)\lambda^3 + [\det(B) + \operatorname{trace}(C)]\,\lambda^2 + (b_{11}c_2 + b_{22}c_1)\lambda + \det(C),$$

and in order for it to have $\lambda = -\alpha$ as a root with multiplicity 4 it has to be of the form $\lambda^4 + 4\alpha\lambda^3 + 6\alpha^2\lambda^2 + 4\alpha^3\lambda + \alpha^4$. Comparing the two polynomials gives a set of three equations in the entries of the matrix B, namely

$$\text{(4.1)} \qquad \begin{cases} \operatorname{trace}(B) = 4\alpha, \\ \det(B) + \operatorname{trace}(C) = 6\alpha^2, \\ b_{11}c_2 + b_{22}c_1 = 4\alpha^3, \end{cases}$$

and a fourth equation which is just the statement that $\alpha = [\det(C)]^{1/4}$. Solving these equations with respect to the entries of B gives

$$\begin{cases} b_{11} = \dfrac{4c_1^{3/4}c_2^{1/4}}{c_1^{1/2}+c_2^{1/2}}, \\ b_{22} = \dfrac{4c_1^{1/4}c_2^{3/4}}{c_1^{1/2}+c_2^{1/2}}, \\ b_{12}^2 = \dfrac{16\det(C)}{\left(c_1^{1/2}+c_2^{1/2}\right)^2} + \operatorname{trace}(C) - 6\left[\det(C)\right]^{1/2}. \end{cases}$$

The expression for b_{12}^2 gives

$$b_{12} = \pm\frac{\left(c_1^{1/2}-c_2^{1/2}\right)^2}{c_1^{1/2}+c_2^{1/2}},$$

and we thus have the following

PROPOSITION 4.1. *In the case where n equals 2, there always exist two and only two optimal matrices B, except in the case where C is scalar when the solution is unique. These solutions are given by*

$$B = \frac{1}{c_1^{1/2}+c_2^{1/2}}\begin{bmatrix} 4c_1^{3/4}c_2^{1/4} & \pm\left(c_1^{1/2}-c_2^{1/2}\right)^2 \\ \pm\left(c_1^{1/2}-c_2^{1/2}\right)^2 & 4c_1^{1/4}c_2^{3/4} \end{bmatrix}.$$

Furthermore, these optimal matrices are positive definite if and only if

$$c_2 < (17+12\sqrt{2})c_1.$$

Proof. It only remains to prove the positivity result. As B must have at least one positive eigenvalue, it will be positive definite if and only if its determinant is positive. From (4.1),

$$\det(B) = 6\alpha^2 - \operatorname{trace}(C) = 6\left[\det(C)\right]^{1/2} - \operatorname{trace}(C)$$

and the result follows (note that it is assumed that $c_1 \le c_2$). ∎

5. Properties of optimal matrices. We begin by giving two simple properties of optimal matrices. The first is the fact that all optimal matrices have the same trace. This will be useful in obtaining a sufficient condition for positivity of optimal matrices.

PROPOSITION 5.1. *Let B be an optimal matrix. Then* $\operatorname{trace}(B) = 2n\omega_o$.

Proof. Consider the companion matrix M defined in the proof of Theorem 3.1. If B is optimal, then M has an eigenvalue at $-\omega_o$ with multiplicity $2n$. This means that $\operatorname{trace}(M) = -2n\omega_o$. On the other hand, $\operatorname{trace}(M) = -\operatorname{trace}(B)$ and the result follows. ∎

A second property has to do with bounds for the eigenvalues of an optimal matrix.

PROPOSITION 5.2. *Let B be an optimal matrix with eigenvalues $b_1 \le \cdots \le b_n$. Then $-2c_{n-j}^{1/2} \le b_j \le 2c_j^{1/2}$.*

Proof. If B is optimal, then, by Theorem 3.1, $L(\lambda)$ is nonnegative for all real λ. We thus have that $\lambda^2 I + B\lambda + C \geq 0$ and, for negative λ,

$$B \leq -\lambda I - \frac{1}{\lambda} C.$$

This inequality in turn implies that the (ordered) eigenvalues of B and C satisfy

$$b_j \leq -\lambda - \frac{1}{\lambda} c_j.$$

The expression on the right-hand side takes its (local) minimum value $2c_j^{1/2}$ when λ equals $-c_j^{1/2}$, and thus $b_j \leq 2c_j^{1/2}$. A similar argument for positive λ gives $b_j \geq -2c_{n-j}^{1/2}$. ∎

In the 2×2 case we saw that there existed two different optimal matrices, except in the case where C was a scalar matrix. We shall now show that a similar nonuniqueness result holds for any n.

PROPOSITION 5.3. *The solution of the optimization problem is unique if and only if the matrix C is scalar.*

Proof. If C is a scalar matrix, say $C = cI$, then it is easily seen that the optimal matrix B_o is unique and is given by $2c^{1/2}I$.

Assume now that C is not scalar but that, without loss of generality, it is in diagonal form. Let B_o be a corresponding optimal matrix. Then B_o cannot be diagonal, for this would not yield a single multiple eigenvalue. Let b_{ij} $(i \neq j)$ be an entry of B_o which is nonzero, and let δ be an n-dimensional vector whose components δ_k are either $+1$ or -1, such that $\delta_i \neq \delta_j$. Then, if $D = \operatorname{diag}\{\delta\}$, the polynomials $L(\lambda) = I\lambda^2 + B_o\lambda + C$ and $\tilde{L}(\lambda) = \lambda^2 I + DB_oD\lambda + C$ have the same (optimal) spectrum. Because of the way in which the vector δ was constructed, the entries ij of the matrices B_o and DB_oD have opposite signs, and thus DB_oD is an optimal matrix different from B_o. ∎

By using the same argument used to prove nonuniqueness it is possible to obtain the following.

PROPOSITION 5.4. *Let B_o be an optimal matrix and let m_r and m_c denote, respectively, the number of rows and the number of columns of B_o which have at least one nonzero entry above the main diagonal. If $m = \max\{m_r, m_c\}$, then there will exist at least 2^m distinct optimal matrices (including B_0). These optimal matrices have the same spectrum.*

Proof. Assume $m = m_r$: a similar argument holds if $m = m_c$ instead. Let $b_{i_k j_k}$ be m entries above the main diagonal with $i_k \neq i_l$ for $k \neq l$, and build a vector $\delta = (\delta_1, \ldots, \delta_n)$ with $\delta_{i_k} = +1$ or -1 and $\delta_j = 1$ if $j \neq i_k$ for all k. This gives 2^m possible vectors δ and if $T_p = \operatorname{diag}\{\delta^p\}$, $p = 1, \ldots, 2^m$, then $T_pCT_p = C$ and $T_pB_oT_p \neq T_qB_oT_q$ for $p \neq q$. ∎

REMARK 5.1. Note that if any of the entries δ_j which are fixed to be $+1$ are changed to -1 this will not affect the corresponding matrix $T_pB_oT_p$, so that in fact this gives all the matrices that can be obtained in this way from a single matrix.

From Example 7.1 below, we know that for n greater than 2 not all optimal matrices can be obtained in this way from a single optimal matrix.

6. The elementary divisor structure of an optimal $L(\lambda)$. As the optimal decay occurs at an eigenvalue of maximum possible multiplicity, it is natural to inquire

about the elementary divisors of $L(\lambda)$ associated with the eigenvalue $\lambda_o = \omega_o$. Recall that $L(\lambda)$ is equivalent to a diagonal matrix of monic invariant polynomials,

$$\mathrm{diag}[a_1(\lambda), \ldots, a_n(\lambda)]$$

of nonincreasing degree (the Smith normal form). Here $a_j(\lambda)$ divides $a_{j-1}(\lambda)$, $j = 2, \ldots, n$ and satisfies

$$a_j = (\lambda + \omega_o)^{k_j} q_j(\lambda),\ j = 1, \ldots, n$$

where the q_j are polynomials for which $q_j(-\omega_o) \neq 0$, and $k_1 \geq \cdots \geq k_n \geq 0$. Since $-\omega_o$ is the *only* eigenvalue of $L(\lambda)$, we have

$$\sum_{j=1}^{n} k_j = 2n$$

and $q_j(\lambda) \equiv 1$ for all j. Since $L(\lambda) \geq 0$ for all real λ (Theorem 3.1) it also follows that all the degrees k_j are *even* (see [6, section 12.5], for example). From the Smith form we see that the geometric multiplicity of the eigenvalue $-\omega_o$ (the nullity of $L(-\omega_o)$) is just the number of nonzero integers k_j and, a priori, can range from 1 (when $k_1 = 2n$) to n (when $k_1 = k_2 = \cdots = k_n = 2$). We first show that the latter case can arise only when $L(\lambda)$ has a degenerate form.

PROPOSITION 6.1. *The matrix $L(-\omega_o)$ has a zero eigenvalue with geometric multiplicity n if and only if C is a scalar matrix.*

Proof. We can consider, without loss of generality, that C is in diagonal form.

Clearly if C is a scalar matrix, say $C = cI$, then the optimal matrix B_o would also be scalar and be given, uniquely, by $B_o = 2c^{1/2}I$.

Assume now that C is not a scalar matrix. Then B_o cannot be diagonal, for this would not yield a single multiple eigenvalue. On the other hand, if the real symmetric matrix $L(-\omega_o)$ has a zero eigenvalue with multiplicity n, it must be the zero matrix. But this implies that

$$B_o = -\frac{1}{\omega_o} C - \omega_o I,$$

which is diagonal, a contradiction. ∎

Expressed in terms of the elementary divisor structure of $L(\lambda)$, this means that all n elementary divisors $(\lambda + \omega_o)^k$ of $L(\lambda)$ have $k = 2$.

When $n = 3$, for instance, the *interesting* elementary divisor cases are now $(k_1, k_2, k_3) = (6, 0, 0)$ and $(k_1, k_2, k_3) = (4, 2, 0)$. The following examples show that both can occur.

EXAMPLE 6.1. Let $C = \mathrm{diag}\{2, 4, 8\}$. Then $\omega_o = [\det(C)]^{1/6} = 2$ and the matrix

$$B_o = \frac{1}{3} \begin{bmatrix} 8 & 0 & \sqrt{2} \\ 0 & 12 & 0 \\ \sqrt{2} & 0 & 16 \end{bmatrix}$$

is an optimal matrix for this problem. In this instance,

$$L(-2) = \frac{2}{3} \begin{bmatrix} 1 & 0 & -\sqrt{2} \\ 0 & 0 & 0 \\ -\sqrt{2} & 0 & 2 \end{bmatrix},$$

which has a double zero eigenvalue. The polynomial $L(\lambda)$ has elementary divisors $(\lambda+2)^4, (\lambda+2)^2$, and $(\lambda+2)^0$.

On the other hand, the matrix

$$B_o' = \begin{bmatrix} \sqrt{33}-3 & \dfrac{\sqrt{2}(7-\sqrt{33})\sqrt{3\sqrt{33}-17}}{1+\sqrt{33}} & \dfrac{3\sqrt{33}-17}{\sqrt{2}} \\ \dfrac{\sqrt{2}(7-\sqrt{33})\sqrt{3\sqrt{33}-17}}{1+\sqrt{33}} & \dfrac{(\sqrt{33}-3)(15-\sqrt{33})}{1+\sqrt{33}} & \dfrac{(\sqrt{33}-5)\sqrt{3\sqrt{33}-17}}{2} \\ \dfrac{3\sqrt{33}-17}{\sqrt{2}} & \dfrac{(\sqrt{33}-5)\sqrt{3\sqrt{33}-17}}{2} & \dfrac{4(15-\sqrt{33})}{1+\sqrt{33}} \end{bmatrix}$$

is also an optimal matrix and now

$$L(-2) = \begin{bmatrix} 12-2\sqrt{33} & \dfrac{2\sqrt{2}(\sqrt{33}-7)\sqrt{3\sqrt{33}-17}}{1+\sqrt{33}} & \sqrt{2}(17-3\sqrt{33}) \\ \dfrac{2\sqrt{2}(\sqrt{33}-7)\sqrt{3\sqrt{33}-17}}{1+\sqrt{33}} & \dfrac{4(41-7\sqrt{33})}{1+\sqrt{33}} & (5-\sqrt{33})\sqrt{3\sqrt{33}-17} \\ \sqrt{2}(17-3\sqrt{33}) & (5-\sqrt{33})\sqrt{3\sqrt{33}-17} & \dfrac{4(5\sqrt{33}-27)}{1+\sqrt{33}} \end{bmatrix}$$

has zero as a simple eigenvalue and $L(\lambda)$ has elementary divisors $(\lambda+2)^6, (\lambda+2)^0$ and $(\lambda+2)^0$.

7. Positivity of optimal matrices. As can be seen from the 2×2 case, optimal matrices might not be positive definite when the eigenvalues of C are too far apart. For n larger than 2 the situation becomes more complex and, as will be seen from the examples below, it is possible to have situations where there are optimal matrices which are positive definite and others which are indefinite for the same matrix C. This of course poses some problems from a practical point of view in systems where structural restrictions imply that the damping has to be positive. Our main result in this section is a sufficient condition for all optimal matrices to be positive definite.

THEOREM 7.1. *If*

$$\omega_o > \frac{c_{k+1}^{1/2} + \cdots + c_n^{1/2}}{n}$$

for some integer k in $\{1, \ldots, n-1\}$, then an optimal matrix will have at most $k-1$ nonpositive eigenvalues. In particular, if

$$c_1 > \frac{\left(c_2^{1/2} + \cdots + c_n^{1/2}\right)^{2n}}{n^{2n} c_2 \cdots c_n},$$

then all optimal matrices will be positive definite.

Proof. Assume that B has k (and only k) nonpositive eigenvalues; that is,

$$b_1 \le \cdots \le b_k \le 0 < b_{k+1} \le \cdots \le b_n.$$

Then

$$\text{trace}(B) = b_1 + \cdots + b_n \leq b_{k+1} + \cdots + b_n \leq 2(c_{k+1}^{1/2} + \cdots + c_n^{1/2}), \tag{7.1}$$

where the second inequality follows from Proposition 5.2. On the other hand, we have from Proposition 5.1 that $\text{trace}(B) = 2\omega_o n$, where $\omega_o = [det(C)]^{1/(2n)}$, and thus

$$2\omega_o n \leq 2(c_{k+1}^{1/2} + \cdots + c_n^{1/2}).$$

In order that B has k nonpositive eigenvalues it is then necessary that

$$\omega_o \leq \frac{c_{k+1}^{1/2} + \cdots + c_n^{1/2}}{n}$$

and the result follows. □

Clearly this result is not optimal—just consider the case where n equals 2, and compare with the necessary and sufficient condition given in Proposition 4.1. However, the following example shows that for n larger than 2 it is no longer possible to obtain a simple necessary and sufficient condition for positive definiteness of the optimal matrices.

EXAMPLE 7.1. Let $n = 3$ and consider the matrix $C = \text{diag}\{1/4, 4, 16\}$ for which $\omega_o = 2^{2/3}$. The condition in Theorem 7.1 does not apply when k is one; in fact, it is possible to obtain an optimal matrix which is not positive definite,

$$B_o \approx \begin{bmatrix} 0.6603 & -1.0226 & -1.4899 \\ -1.0226 & 3.1176 & -1.7190 \\ -1.4899 & -1.7190 & 5.7465 \end{bmatrix},$$

with eigenvalues $-0.3876, 3.1748$ and 6.7372.

On the other hand, it is also possible to obtain a positive definite optimal matrix

$$B_o' \approx \begin{bmatrix} 0.6998 & 0.8102 & 1.6470 \\ 0.8102 & 2.2874 & 0.4608 \\ 1.6470 & 0.4608 & 6.5371 \end{bmatrix},$$

with eigenvalues $0.0619, 2.4026$, and 7.0599.

In this case Theorem 7.1 can be used to show that optimal matrices may have at most one nonpositive eigenvalue.

REMARK 7.1. The value of ω_o obtained for these matrices is not precisely $2^{2/3}$. Apart from the obvious fact that we are dealing with approximations, it is also important to note that the optimal is attained at a point where Ω is not smooth. In particular, some of the directional derivatives may become unbounded and, as a consequence, small changes in the entries of B_o may cause large deviations in the value of the spectral abscissa. This aspect of the problem will be discussed in more detail in section 8. We also note that the eigenvalues of the matrices B_o and B_o' are not as sensitive to small changes.

8. Discussion. By reformulating the problem of optimizing the decay of solutions of system (1.1) as an inverse problem, we have been able to show attainment and, in general, nonuniqueness, within the space of real symmetric matrices. However, and as we have seen in the 2×2 case and in Example 7.1, it turns out that it might be possible to obtain this optimal by means of an indefinite matrix or, in some cases, that

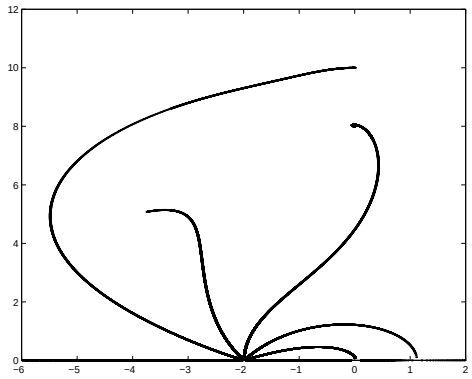


FIG. 8.1. *Eigenvalue loci for Example* 8.1.

all optimal matrices are indefinite. If in the formulation of the optimization problem we further impose the restriction that it should be possible to attain the optimal with a positive definite matrix, then clearly in this last situation the best possible value for the spectral abscissa is no longer $[\det(C)/\det(A)]^{1/2n}$.

A related aspect that should be taken into account at this point is that, due to the unbounded directional derivatives referred to in the remark following Example 7.1, it might be difficult to obtain good numerical approximations of optimal matrices. To see how this might affect the problem, we consider a parametrized version of the quadratic pencil $L(\lambda)$ given by

$$L_\epsilon(\lambda) = I\lambda^2 + \epsilon B\lambda + C,$$

where A, B, and C are as before and ϵ is a nonnegative parameter. One way of measuring the effect of small changes in the optimal matrix is to vary ϵ and see the effect that that will have on the spectral abscissa.

For indefinite damping terms, it is known from [5] that if ϵ is made large enough, then there will exist eigenvalues with positive real parts, and the system becomes unstable. Another phenomenon associated with this is that of stability switching, where the stability of the trivial solution changes as the parameter ϵ is increased. In [5] an example was given with $n = 5$ where solutions became unstable, then stable, and finally unstable again. Here we build an example of an optimal matrix for which a sequence s–u–s–u is achieved and for which two of the stability switching points are obtained for ϵ close to one.

EXAMPLE 8.1. Let $C = \mathrm{diag}\{1/100, 64, 100\}$. Then $\omega_o = 2$ and

$$B_o = \begin{bmatrix} \dfrac{4(a-493)}{17775} & \dfrac{(851201-32a)b}{35550(9887+16a)} & \dfrac{4bc}{395\ 3^{3/2}} \\ \dfrac{(851201-32a)b}{35550(9887+16a)} & \dfrac{64(493-a)(a-53818)}{17775(9887+16a)} & \dfrac{16(a-1204)c}{395\ 3^{3/2}} \\ \dfrac{4bc}{395\ 3^{3/2}} & \dfrac{64(493-a)(a-53818)}{17775(9887+16a)} & \dfrac{4(53818-a)}{9887+16a} \end{bmatrix},$$

where

$$a = \sqrt{385249}, \quad b = \sqrt{2585257 + 15776a} \quad \text{and } c = \sqrt{103a - 63929}$$

is an optimal matrix with eigenvalues $-8.6804, 4.7353$ and 15.9451. In this case, the stability switching points are $\epsilon_1 \approx 0.228$, $\epsilon_2 \approx 0.9279$, and $\epsilon_3 \approx 1.0092$. We thus see that a very small variation in the parameter ϵ might take it past the point ϵ_3 and that this will make the system unstable. In Figure 8.1 we have plotted the changes of the eigenvalues with ϵ as this parameter is increased from 0 to 1.5. Apart from the stability switching, we also see that the eigenvalues come together to give a single eigenvalue of multiplicity 6 at $\epsilon = 1$, and that when they split for ϵ larger than 1, one pair will remain real, while the other two pairs become complex again.

In the notation of section 6, the elementary divisors of $L(\lambda)$ have degrees $(6, 0, 0)$. This also accounts for the form of the splitting of the eigenvalues at $\epsilon = 1$, determining the angles that the eigencurves make with the $x-$axis: as the geometric multiplicity is 1, the tangents to the incoming and outcoming eigencurves will divide the plane in 12 equal angles.

REFERENCES

[1] J. V. Burke and M. L. Overton, *Differential properties of the spectral abscissa and the spectral radius for analytic matrix-valued mappings*, Nonlinear Anal., 23 (1994), pp. 467–488.
[2] M. T. Chu, *Inverse eigenvalue problemas*, SIAM Rev., 40 (1998), pp. 1–39.
[3] S. J. Cox, *Designing for optimal energy absorption,* I: *Lumped parameter systems*, J. Vibration Acoustics, 120 (1998), pp. 339–345.
[4] S. J. Cox and M. L. Overton, *Perturbing the critically damped wave equation*, SIAM J. Appl. Math., 56 (1996), pp. 1353–1362.
[5] P. Freitas, M. Grinfeld and P. A. Knight, *Stability for finite-dimensional systems with indefinite damping*, Adv. Math. Sci. Appl., 7 (1997), pp. 435–446.
[6] I. Gohberg, P. Lancaster, and L. Rodman, *Matrix Polynomials*, Academic Press, New York, 1982.
[7] A. Horn, *On the eigenvalues of a matrix with prescribed singular values*, Proc. Amer. Math. Soc., 5 (1954), pp. 4–7.
[8] R. Horn and C. A. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, 1995.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 1, pp. 209–229
© 1999 Society for Industrial and Applied Mathematics

ITERATIVE SOLUTION OF THE HELMHOLTZ EQUATION BY A SECOND-ORDER METHOD*

KURT OTTO† AND ELISABETH LARSSON†

Abstract. The numerical solution of the Helmholtz equation subject to nonlocal radiation boundary conditions is studied. The specific problem is the propagation of hydroacoustic waves in a two-dimensional curvilinear duct. The problem is discretized with a second-order accurate finite difference method, resulting in a linear system of equations. To solve the system of equations, a preconditioned Krylov subspace method is employed. We construct a preconditioner that is based on fast transforms and yields a direct fast Helmholtz solver for rectangular domains. Numerical experiments for curved ducts demonstrate that the rate of convergence is high. The fast transform preconditioner is compared with a symmetric successive over-relaxation (SSOR) preconditioner, and also with band Gaussian elimination. For the preconditioned iterative methods, the gains in storage requirement are large compared with band Gaussian elimination. Regarding the arithmetic complexity, the fast transform preconditioner yields a significant gain, whereas the SSOR preconditioner performs worse than band Gaussian elimination.

Key words. Helmholtz equation, waveguide problems, finite difference discretization, preconditioners, fast trigonometric transforms

AMS subject classifications. 65F10, 65N22

PII. S0895479897316588

1. Introduction.

The Helmholtz equation arises in many physical applications, e.g., scattering problems in electromagnetics and acoustics [Ernst96], [AbKr94]. In realistic applications, a wide range of wavenumbers is often of interest. For a finite element (or finite difference) discretization of the two-dimensional Helmholtz equation, it is necessary that the number of grid points grows faster than quadratically in the wavenumber in order to maintain a given accuracy [BaGoTu85a], [IhlBa97]. Thus, for high wavenumbers, the discretized Helmholtz equation "leads to a huge linear system of equations" [AbKr94, p. 2]. Due to the large bandwidth, the storage requirement renders Gaussian elimination prohibitive. To handle high wavenumbers and large domains for the Helmholtz equation in duct acoustics, Abrahamsson and Kreiss [AbKr94] devised a special iteration technique related to separation of variables. However, the effectiveness of the method relies on the degree of separability of the problem. Another way to address the computational difficulties for the discretized Helmholtz equation is to design iterative methods. Bayliss, Goldstein, and Turkel [BaGoTu83] used a preconditioned conjugate gradient method applied to the normal equations for a finite element discretization [BaGuTu82]. Due to the ill-conditioning of the normal equations, the unpreconditioned algorithm suffered from extremely slow convergence. The convergence rate was substantially improved through preconditioners based on SSOR [BaGoTu83] or a multigrid V-cycle [BaGoTu85b], [Gold82] only for

*Received by the editors February 10, 1997; accepted for publication (in revised form) by S. Vavasis October 8, 1998; published electronically September 22, 1999. This research was supported by the U. S. National Science Foundation under grant ASC-8958544 and by the Swedish National Board for Industrial and Technical Development (NUTEK). Part of the work of the first author was performed during a postdoctoral visit at the Department of Computer Science, University of Maryland, College Park, MD.

http://www.siam.org/journals/simax/21-1/31658.html

†Department of Scientific Computing, Uppsala University, Box 120, SE-751 04 Uppsala, Sweden (kurt@tdb.uu.se, bette@tdb.uu.se).

the Laplacian part of the Helmholtz operator. Recently, a block version of the quasi-minimal residual algorithm has been exploited for problems with multiple right-hand sides in structural acoustics. The method was tested in conjunction with different preconditioners, and an SSOR preconditioner gave the best performance [MaFrPi97]. The quasi-minimal residual algorithm has been applied also to capacitance matrix methods for exterior Helmholtz problems [Ernst96]. A somewhat different approach is to employ Schwarz-type domain decomposition methods. If the subdomains are reasonably small, direct solvers can be used locally. These methods are appealing because they are easy to parallelize, but in their unembellished form they converge very slowly. Kim [Kim98] introduces an automatic procedure for choosing Robin interface conditions that improve convergence significantly. However, the iteration count still depends on the number of subdomains. This may lead to poor performance for problems with many or large subdomains, in the latter case due to the excessive memory requirements and the high arithmetic costs for the direct solvers. In [CCEW98], it is demonstrated that an overlapping Schwarz method with GMRES acceleration and coarse grid corrections converges rather rapidly.

The objective of this paper is to develop a technique for solving the Helmholtz equation with an iterative method. In order to be a viable method, it should exploit the sparsity of the discretization matrix in an efficient way, converge rapidly, and be competitive with Gaussian elimination in regard to the total arithmetic complexity. Our approach is to apply a preconditioned Krylov subspace method [FrGoNa92] directly to the discretized equations. Typically, for high wavenumbers the discretization matrix is large, complex, indefinite, and ill-conditioned. As a result, standard preconditioning techniques like diagonal scaling and incomplete LU decomposition are likely to do poorly. Instead we construct preconditioners based on fast transforms; see [Otto96] and the survey in [ChanNg96]. In order to get a highly structured matrix, facilitating the design of the preconditioner, a finite difference method is used for the discretization. For the same reason, special attention is given to the choice of radiation boundary conditions. A finite element method would be more flexible for complicated geometries, but also less amenable to fast transform-based preconditioners. This is particularly noticeable for higher orders of approximation, where some of the degrees of freedom typically are *not* node values.

The paper is organized as follows. In section 2 the governing equations, the boundary conditions, and the finite difference discretization of a Helmholtz problem are derived. The specific problem is the propagation of hydroacoustic waves in a curvilinear duct. Issues concerning the preconditioner are treated in section 3. Section 4 is devoted to computational aspects with an emphasis on resolution criteria, i.e., relations between the wavenumber, the grid size, and the desired accuracy. Finally, numerical experiments are presented in section 5 followed by conclusions.

2. The model problem. In this section the theory needed to determine the system of equations for the model problem is discussed.

2.1. Notation. The quantity I_m denotes the identity matrix of order m. The square matrices $\mathrm{diag}_{j,m}(\beta_j)$ and $\mathrm{trid}_{j,m}(\alpha_j, \beta_j, \gamma_j)$ are defined in the following way:

$$\mathrm{diag}_{j,m}(\beta_j) = \begin{pmatrix} \beta_1 & & \\ & \ddots & \\ & & \beta_m \end{pmatrix},$$

$$\mathrm{trid}_{j,m}(\alpha_j,\beta_j,\gamma_j) = \begin{pmatrix} \beta_1 & \gamma_1 & & & \\ \alpha_2 & \beta_2 & \gamma_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \alpha_{m-1} & \beta_{m-1} & \gamma_{m-1} \\ & & & \alpha_m & \beta_m \end{pmatrix}.$$

2.2. Governing equations. We study the propagation of time-harmonic sound waves under water. Neglecting sound absorption and assuming that the fluid is homogeneous, the waves are governed by the Helmholtz equation

$$-\frac{\partial^2 u}{\partial x_1^2} - \frac{\partial^2 u}{\partial x_2^2} - \kappa^2 u = 0, \tag{1}$$

where $u(x_1, x_2)$ is the phasor of the acoustic pressure $\mathbb{R}\mathrm{e}(u(x_1,x_2)e^{-i2\pi ft})$ [JKPS94, p. 73]. The wavenumber is given by $\kappa = 2\pi f/c$, where f is the frequency and $c = 1500$ m/s is the sound speed. For heterogeneous media, the sound speed and consequently the wavenumber would depend on the space coordinates.

We consider a physical domain

$$\begin{cases} x_1 = x_1(\xi_1,\xi_2) \\ x_2 = x_2(\xi_1,\xi_2) \end{cases}$$

that can be mapped onto the unit square

$$\begin{cases} 0 \le \xi_1 \le 1 \\ 0 \le \xi_2 \le 1 \end{cases}$$

via an orthogonal transformation [KnSt93, p. 82]. Equation (1) is then transformed into

$$-\frac{\partial}{\partial \xi_1}\left(a\frac{\partial u}{\partial \xi_1}\right) - \frac{\partial}{\partial \xi_2}\left(a^{-1}\frac{\partial u}{\partial \xi_2}\right) - eu = 0, \tag{2}$$

where the metric coefficients a and e are given by

$$a = \sqrt{\frac{\left(\frac{\partial x_1}{\partial \xi_2}\right)^2 + \left(\frac{\partial x_2}{\partial \xi_2}\right)^2}{\left(\frac{\partial x_1}{\partial \xi_1}\right)^2 + \left(\frac{\partial x_2}{\partial \xi_1}\right)^2}},$$

$$e = \kappa^2 \sqrt{\left(\left(\frac{\partial x_1}{\partial \xi_1}\right)^2 + \left(\frac{\partial x_2}{\partial \xi_1}\right)^2\right)\left(\left(\frac{\partial x_1}{\partial \xi_2}\right)^2 + \left(\frac{\partial x_2}{\partial \xi_2}\right)^2\right)}.$$

Because of the orthogonality of the transformation, no mixed terms appear in (2).

2.3. Boundary conditions. We now choose the physical domain to be a two-dimensional duct; see the shaded area in Figure 1.

For the problem to be well posed, conditions are needed on all four boundaries. Our model problem is partly fixed by letting the physical boundary Γ_1 be a soft wall (air), whereas Γ_2 is a rigid wall (rock). The sound field is generated by a source along

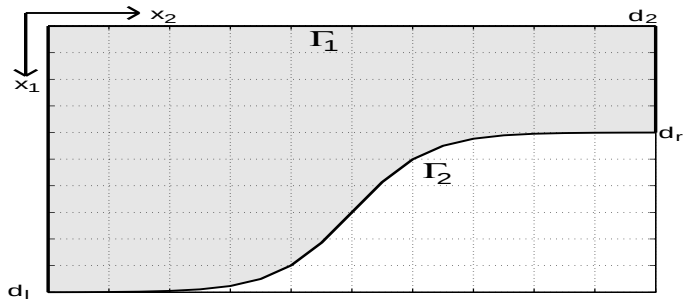


FIG. 1. *The physical domain.*

$x_2 = 0$, specified by a source term $g(\xi_1) \equiv g(x_1(\xi_1, 0))$. At $x_2 = d_2$ an artificial far-zone boundary has been introduced. Originally, the domain is semi-infinite ($d_2 \to \infty$), but for computational reasons it is truncated by assigning d_2 some finite value. For the soft wall Γ_1, the boundary condition is $u = 0$ (pressure release). This leads to

$$u(0, \xi_2) = 0, \quad 0 \le \xi_2 \le 1, \tag{3}$$

in the computational domain. Since the bottom Γ_2 is rigid, a condition on the normal derivative is imposed:

$$\frac{\partial u}{\partial n} = 0, \quad (x_1, x_2) \in \Gamma_2.$$

Due to the orthogonal transformation this becomes

$$\frac{\partial u}{\partial \xi_1}(1, \xi_2) = 0, \quad 0 \le \xi_2 \le 1. \tag{4}$$

For the radiation conditions (11) and (14) at the near- and far-zone boundaries, Dirichlet–to–Neumann (DtN) maps [KeGi89] are employed. The main reason for choosing nonlocal DtN maps, instead of the local radiation conditions described in [BaGuTu82], is that discretized DtN maps are more apt to preconditioning by fast transforms. Another desirable feature of DtN maps is that they easily achieve highly accurate radiation conditions. Our design of radiation conditions follows the principles outlined in [FixMa78], where a variational formulation of DtN conditions was derived for an axially symmetric duct parametrized by cylindrical coordinates. Boundary conditions based on DtN maps require the boundary in question to be a separable coordinate surface. Moreover, for the radiation condition in [FixMa78], it is implicitly assumed that the duct could be extended beyond the artificial boundary by parallel straight walls. This is a so-called anechoic termination [AbKr94]. Since the wave-number κ is independent of ξ_2, the above prerequisites are fulfilled by requiring the duct to be flat *only* in an infinitesimal neighborhood of $x_2 = d_2$ and $x_2 = 0$. In the present application, the wavenumber is actually a constant. Thus, without any significant loss of accuracy we can use the slightly more restrictive assumption that, in the vicinity of $x_2 = 0$, there is a local transformation

$$\begin{cases} x_1 = \xi_1 d_\ell, & 0 \le \xi_1 \le 1, \\ x_2 = \xi_2 d_2. \end{cases}$$

Substituting this into (2), together with (3) and (4), yields

$$\left\{\begin{array}{l} -\frac{d_2}{d_\ell}\frac{\partial^2 u}{\partial \xi_1^2} - \frac{d_\ell}{d_2}\frac{\partial^2 u}{\partial \xi_2^2} - \kappa^2 d_\ell d_2 u = 0, \\ u(0,\xi_2) = 0, \\ \frac{\partial u}{\partial \xi_1}(1,\xi_2) = 0. \end{array}\right. \tag{5}$$

The condition at the near-zone boundary is based on the fact that the solution for

$$0 \le \xi_1 \le 1, \quad \xi_2 \le 0,$$

can be obtained through separation of variables [MorFe53, p. 498], i.e., $u(\xi_1,\xi_2) = \psi(\xi_1)\chi(\xi_2)$. Solving (5) with this ansatz gives

$$\begin{aligned} \psi_m(\xi_1) &= \sqrt{2}\sin((m-\tfrac{1}{2})\pi\xi_1), \\ \chi_m(\xi_2) &= A_m \exp(i\sqrt{-\lambda_m}\xi_2) + B_m \exp(-i\sqrt{-\lambda_m}\xi_2), \quad m = 1,2,\ldots, \end{aligned}$$

where

$$\lambda_m = \big((m-\tfrac{1}{2})\pi d_2/d_\ell\big)^2 - (\kappa d_2)^2.$$

The eigenfunctions $\{\psi_m(\xi_1)\}_{m=1}^{\infty}$ are orthonormal with respect to the scalar product

$$\langle f, g\rangle \equiv \int_0^1 \bar{f}(\xi_1) g(\xi_1) d\xi_1.$$

The general solution to the eigenproblem (5) becomes

$$u(\xi_1,\xi_2) = \sum_{m=1}^{\infty} \psi_m(\xi_1)\big(A_m \exp(i\sqrt{-\lambda_m}\xi_2) + B_m \exp(-i\sqrt{-\lambda_m}\xi_2)\big). \tag{6}$$

For mode indices below the cutoff limit, i.e.,

$$m \le \mu_\ell = \left\lfloor \frac{\kappa d_\ell}{\pi} + \frac{1}{2} \right\rfloor, \tag{7}$$

the eigenvalues λ_m become negative, yielding propagating modes [JKPS94, p. 279]. If λ_m were positive, we would get evanescent modes [JKPS94, p. 279]. Analogously to the motivation in [FixMa78], the influence of the evanescent modes is negligible, especially on the far field. Thus, an appropriate way to truncate the series in (6) is to retain only the terms with mode indices $m \le \mu_\ell$. The situation is somewhat different for a purely exterior Helmholtz problem, where an appropriate truncation of DtN maps is a more delicate matter [GroKe95].

The A_m-terms in (6) correspond to right-going waves, and the B_m-terms correspond to left-going waves. In our model we have a source at the left boundary. We will treat the right-going waves as originating from a "truncated" point source positioned at depth $\xi_1 = \delta_s$. This is done by specifying the Fourier coefficients of the source term g:

$$A_m = \langle \psi_m, g\rangle = \psi_m(\delta_s), \quad m = 1,\ldots,\mu_\ell. \tag{8}$$

Note that left-going waves are feasible in order to handle possible reflections from the curved bottom. Inserting (7) and (8) into (6) yields

$$u(\xi_1,\xi_2) = \sum_{m=1}^{\mu_\ell} \psi_m(\xi_1)\big(\psi_m(\delta_s)\exp(i\sqrt{-\lambda_m}\xi_2) + B_m \exp(-i\sqrt{-\lambda_m}\xi_2)\big). \tag{9}$$

The coefficients B_m are determined from the solution by exploiting the orthonormality of the functions $\psi_m(\xi_1)$. From (9) we get

$$B_m = \langle \psi_m, u(\cdot, 0) \rangle - \psi_m(\delta_s). \tag{10}$$

The nonlocal boundary condition at $\xi_2 = 0$ is obtained by differentiating (9) with respect to ξ_2 and using (10). Thus,

$$-\frac{\partial u}{\partial \xi_2}(\xi_1, 0) - i \sum_{m=1}^{\mu_\ell} \sqrt{-\lambda_m} \langle \psi_m, u(\cdot, 0) \rangle \psi_m(\xi_1) = -i \sum_{m=1}^{\mu_\ell} 2\sqrt{-\lambda_m} \psi_m(\delta_s) \psi_m(\xi_1). \tag{11}$$

The boundary condition at $\xi_2 = 1$ is derived in a similar way. Due to the anechoic termination of the duct, there are no reflections, i.e., there are only right-going waves:

$$u(\xi_1, \xi_2) = \sum_{m=1}^{\mu_r} A_m \psi_m(\xi_1) \exp\bigl(i\sqrt{-\lambda_m}(\xi_2 - 1)\bigr), \tag{12}$$

where

$$\lambda_m = \bigl((m - \tfrac{1}{2})\pi d_2/d_r\bigr)^2 - (\kappa d_2)^2,$$
$$\mu_r = \left\lfloor \frac{\kappa d_r}{\pi} + \frac{1}{2} \right\rfloor.$$

The coefficients A_m are determined by

$$A_m = \langle \psi_m, u(\cdot, 1) \rangle. \tag{13}$$

Differentiation of (12) and insertion of (13) give the condition for the far-zone boundary:

$$\frac{\partial u}{\partial \xi_2}(\xi_1, 1) - i \sum_{m=1}^{\mu_r} \sqrt{-\lambda_m} \langle \psi_m, u(\cdot, 1) \rangle \psi_m(\xi_1) = 0. \tag{14}$$

2.4. Discretization. Now when the analytical problem is defined, we design the numerical method. Introduce a uniform grid as

$$\begin{cases} \xi_{1,j} = jh_1, & j = 0, \ldots, m_1 + 1, \\ \xi_{2,k} = \bigl(k - \frac{3}{2}\bigr) h_2, & k = 1, \ldots, m_2, \end{cases}$$

where

$$h_1 = \frac{1}{m_1 + \frac{1}{2}}, \qquad h_2 = \frac{1}{m_2 - 2}.$$

The grid is centered around those boundaries where the boundary conditions (4), (11), (14) involve derivatives. The main reason for centering the grid [RiMor67, p. 134] is to facilitate the construction of discretized boundary conditions that have the same order of accuracy as the rest of the scheme. Let $u_{j,k}$ denote the approximate solution at the point $(\xi_{1,j}, \xi_{2,k})$. We use centered difference operators to obtain second-order accuracy. Equation (2) is approximated with

$$\begin{aligned} &- h_1^{-2}\bigl(a_{j+\frac{1}{2},k}(u_{j+1,k} - u_{j,k}) - a_{j-\frac{1}{2},k}(u_{j,k} - u_{j-1,k})\bigr) \\ &- h_2^{-2}\bigl(a^{-1}_{j,k+\frac{1}{2}}(u_{j,k+1} - u_{j,k}) - a^{-1}_{j,k-\frac{1}{2}}(u_{j,k} - u_{j,k-1})\bigr) - e_{j,k}u_{j,k} = 0 \end{aligned} \tag{15}$$

for inner points $k = 2, \ldots, m_2 - 1$ and $j = 1, \ldots, m_1$. The boundary conditions (3) and (4) become

$$u_{0,k} = 0, \quad k = 1, \ldots, m_2, \tag{16}$$

$$u_{m_1+1,k} = u_{m_1,k}, \quad k = 1, \ldots, m_2. \tag{17}$$

For the other two boundaries matters are more complicated. We discretize the modal expansions involving the eigenfunctions $\psi_m(\xi_1)$ by evaluating them on the ξ_1-grid, i.e.,

$$\psi_m \equiv \{\psi_m(\xi_{1,j})\}_{j=1}^{m_1} = \left\{\sqrt{2}\sin\left((m - \tfrac{1}{2})\pi j h_1\right)\right\}_{j=1}^{m_1}, \quad m = 1, \ldots, m_1, \tag{18}$$

and by approximating the integrals with a second-order accurate combination of the composite trapezoid rule and the rectangle rule. Our specific choice of ξ_1-grid and quadrature rule makes the column vectors (18) orthonormal with respect to the discrete scalar product

$$\langle \psi_m, \psi_n \rangle \equiv \psi_m^* h_1 \psi_n \approx \int_0^1 \bar{\psi}_m(\xi_1)\psi_n(\xi_1)d\xi_1.$$

Moreover, a second-order accurate finite difference discretization of the eigenproblem yields *exactly* the same eigenvectors as (18). The resulting discretization of conditions (11) and (14) is

$$h_2^{-1}(u_1 - u_2) - i\sum_{m=1}^{\mu_\ell} \sqrt{-\lambda_m}\psi_m\psi_m^* h_1 \tfrac{1}{2}(u_1 + u_2) = -i\sum_{m=1}^{\mu_\ell} 2\sqrt{-\lambda_m}\psi_m(\delta_s)\psi_m, \tag{19}$$

$$h_2^{-1}(u_{m_2} - u_{m_2-1}) - i\sum_{m=1}^{\mu_r} \sqrt{-\lambda_m}\psi_m\psi_m^* h_1 \tfrac{1}{2}(u_{m_2-1} + u_{m_2}) = 0, \tag{20}$$

where

$$u_k \equiv (u_{1,k} \cdots u_{m_1,k})^T.$$

This can be written

$$\begin{aligned}(I_{m_1} - C_\ell)u_1 + (-I_{m_1} - C_\ell)u_2 &= g_1,\\ (-I_{m_1} - C_r)u_{m_2-1} + (I_{m_1} - C_r)u_{m_2} &= 0,\end{aligned}$$

where

$$C_\ell = ih_2\sum_{m=1}^{\mu_\ell} \sqrt{-\lambda_m}\psi_m\psi_m^*\frac{h_1}{2}, \qquad C_r = ih_2\sum_{m=1}^{\mu_r} \sqrt{-\lambda_m}\psi_m\psi_m^*\frac{h_1}{2}, \tag{21}$$

$$g_1 = -ih_2\sum_{m=1}^{\mu_\ell} 2\sqrt{-\lambda_m}\psi_m(\delta_s)\psi_m.$$

Notice that λ_m depends on the depth and is different for the left and right boundaries.

Applying (15) to inner grid points, using (19) and (20), and eliminating the boundary values defined by (16) and (17) gives the complete system of equations

$$Bu = g,$$

$$u \equiv \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{m_2} \end{pmatrix}, \qquad g = \begin{pmatrix} g_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad B = \mathrm{trid}_{k,m_2}(B_{k,-1}, B_{k,0}, B_{k,1}),$$

where

$$\begin{aligned} B_{1,0} &= I_{m_1} - C_\ell, \quad B_{1,1} = -I_{m_1} - C_\ell, \\ B_{m_2,-1} &= -I_{m_1} - C_r, \quad B_{m_2,0} = I_{m_1} - C_r, \end{aligned} \tag{22}$$

and for $k = 2, \ldots, m_2 - 1$,

$$\begin{aligned} B_{k,-1} &= \mathrm{diag}_{j,m_1}(-\gamma a^{-1}_{j,k-\frac{1}{2}}), \\ B_{k,0} &= \mathrm{trid}_{j,m_1}(-a_{j-\frac{1}{2},k}, \alpha_{j,k}, -a_{j+\frac{1}{2},k}), \\ B_{k,1} &= \mathrm{diag}_{j,m_1}(-\gamma a^{-1}_{j,k+\frac{1}{2}}), \end{aligned}$$

where

$$\begin{aligned} \alpha_{1,k} &= a_{\frac{1}{2},k} + a_{\frac{3}{2},k} + \gamma(a^{-1}_{1,k-\frac{1}{2}} + a^{-1}_{1,k+\frac{1}{2}}) - h_1^2 e_{1,k}, \\ \alpha_{j,k} &= a_{j-\frac{1}{2},k} + a_{j+\frac{1}{2},k} + \gamma(a^{-1}_{j,k-\frac{1}{2}} + a^{-1}_{j,k+\frac{1}{2}}) - h_1^2 e_{j,k}, \quad j = 2, \ldots, m_1 - 1, \\ \alpha_{m_1,k} &= a_{m_1-\frac{1}{2},k} + \gamma(a^{-1}_{m_1,k-\frac{1}{2}} + a^{-1}_{m_1,k+\frac{1}{2}}) - h_1^2 e_{m_1,k}, \\ \gamma &= \frac{h_1^2}{h_2^2}. \end{aligned}$$

By some minor modifications, the discretization could accommodate other combinations of boundary conditions on the boundaries Γ_1 and Γ_2. For Neumann conditions at Γ_1 and Γ_2, a suitable grid for ξ_1 would be

$$\xi_{1,j} = (j - \tfrac{1}{2})h_1, \quad j = 0, \ldots, m_1 + 1, \quad h_1 = \frac{1}{m_1}.$$

The resulting alteration of the matrix B would mainly be

$$\alpha_{1,k} = a_{\frac{3}{2},k} + \gamma(a^{-1}_{1,k-\frac{1}{2}} + a^{-1}_{1,k+\frac{1}{2}}) - h_1^2 e_{1,k}.$$

For Dirichlet conditions at Γ_1 and Γ_2, a convenient choice of ξ_1-grid would be

$$\xi_{1,j} = jh_1, \quad j = 0, \ldots, m_1 + 1, \quad h_1 = \frac{1}{m_1 + 1}.$$

This would cause $\alpha_{m_1,k}$ to change into

$$\alpha_{m_1,k} = a_{m_1-\frac{1}{2},k} + a_{m_1+\frac{1}{2},k} + \gamma(a^{-1}_{m_1,k-\frac{1}{2}} + a^{-1}_{m_1,k+\frac{1}{2}}) - h_1^2 e_{m_1,k},$$

but leave the rest of B essentially intact. Nota bene that the mth eigenvector ψ_m appearing in (21) and (18) changes with the boundary conditions. For the Neumann–Neumann case, it becomes $\psi_m \propto \{\cos((m-1)\pi\xi_{1,j})\}_{j=1}^{m_1}$, whereas the Dirichlet–Dirichlet case leads to $\psi_m \propto \{\sin(m\pi\xi_{1,j})\}_{j=1}^{m_1}$.

3. Preconditioning. We employ a Krylov subspace method to solve the system of equations. For simplicity and robustness, we choose the restarted generalized minimal residual (GMRES(ℓ)) algorithm [SaadSch86], where ℓ is the restarting length. For the iterative method to be competitive, an effective preconditioner is needed. Otherwise the cost of computing the solution would be too high. After left preconditioning, the original system $Bu = g$ is transformed into $M^{-1}Bu = M^{-1}g$.

3.1. Normal block preconditioners. We construct a preconditioner that preserves the block structure of B, thus exploiting sparsity. Moreover, it should be possible to form and apply the preconditioner at low arithmetic costs. To meet these demands, we use a normal block preconditioner [Otto96] based on fast trigonometric transforms [VLoan92, section 4.4], [BaiSw91]. The main idea in the design is to approximate the matrix B with a preconditioner having the same block structure, and where all the blocks have the same prescribed eigenvectors. These eigenvectors depend on the boundary conditions, but are chosen so that the corresponding similarity transformation is associated with a fast transform.

For the discretization matrix B in section 2.4, a Dirichlet condition was imposed on Γ_1 and a Neumann condition on Γ_2. Hence, a suitable choice for the unitary eigenvector matrix is

$$Q \equiv [q_1, \ldots, q_{m_1}], \qquad q_m = \sqrt{h_1}\psi_m,$$

which is connected to a slightly modified sine transform-II [VLoan92, p. 229]. Form a preconditioner

$$M = \mathrm{trid}_{k,m_2}(M_{k,-1}, M_{k,0}, M_{k,1}),$$

the blocks of which are diagonalized by Q, i.e.,

$$M_{k,r} \equiv Q\Lambda_{k,r}Q^*, \tag{23}$$

where $\Lambda_{k,r}$, $r = -1, 0, 1$, are diagonal matrices. There are several possible choices for $\Lambda_{k,r}$. The specific choice

$$\Lambda_{k,r} = \mathrm{diag}(Q^* B_{k,r} Q)$$

minimizes $\|B_{k,r} - M_{k,r}\|_F$ for matrices of type (23), and it also minimizes $\|B - M\|_F$. Observe that the blocks defined by (21) can be rewritten as linear combinations of outer products $q_m q_m^*$. This means that the matrix blocks (22) corresponding to the left and right boundaries will be diagonalized by Q. In fact, for a duct with a flat bottom, all the blocks in B would be diagonalized by Q, yielding $M = B$ [Otto96]. Hence, the operator M^{-1} is a direct fast Helmholtz solver for rectangular domains. For a duct with a curved bottom, blocks corresponding to inner grid lines will not be completely diagonalized. However, when the domain is moderately curved, the preconditioner presumably acts like a viable convergence accelerator.

For the Neumann–Neumann and Dirichlet–Dirichlet boundary conditions discussed in section 2.4, the eigenvector matrices Q should be chosen to match the boundary conditions, i.e., as those associated with the cosine and sine transforms, respectively. The preconditioners thus arising also yield direct fast solvers for rectangular domains, and indeed perform well [Otto96] when the domains are moderately curved.

For each iteration, the computation $x = M^{-1}y$ has to be performed. Due to the structure of the blocks of M, it holds that

$$\Lambda \equiv (I_{m_2} \otimes Q^*)M(I_{m_2} \otimes Q) = \mathrm{trid}_{k,m_2}(\Lambda_{k,-1}, \Lambda_{k,0}, \Lambda_{k,1}),$$

leading to

$$M^{-1} = (I_{m_2} \otimes Q)\Lambda^{-1}(I_{m_2} \otimes Q^*).$$

The computation $x = M^{-1}y$ can now be done in three steps:

1. $v = (I_{m_2} \otimes Q^*)y$,
2. solve $\Lambda z = v$,
3. $x = (I_{m_2} \otimes Q)z$.

Step 2 consists of solving a block tridiagonal system, where each block is diagonal. By permuting the unknowns, we get m_1 independent tridiagonal systems of order m_2. Steps 1 and 3 consist of m_2 sine transforms-II and inverse sine transforms-II of length m_1. We can utilize fast Fourier transform methods [BaiSw91] for computing these transforms.

3.2. Generalizations. The key principle for the normal block preconditioner is that a discretized two-dimensional problem is reduced to a set of one-dimensional problems by applying fast transforms in one coordinate direction. The same technique is applicable to three-dimensional waveguide problems and scattering problems. As an example, we consider a three-dimensional waveguide parametrized by cylindrical coordinates (r, ϕ, z). The waveguide is enclosed by two physical boundaries

$$\begin{aligned} \Gamma_1 : &\quad z = z_1(r,\phi), \quad d_0 \le r \le d_\infty, \quad 0 \le \phi \le 2\pi, \\ \Gamma_2 : &\quad z = z_2(r,\phi), \quad d_0 \le r \le d_\infty, \quad 0 \le \phi \le 2\pi, \end{aligned}$$

a near-zone boundary, and an artificial far-zone boundary,

$$\begin{aligned} \Gamma_0 : &\quad z_1(d_0,\phi) \le z \le z_2(d_0,\phi), \quad 0 \le \phi \le 2\pi, \\ \Gamma_\infty : &\quad z_1(d_\infty,\phi) \le z \le z_2(d_\infty,\phi), \quad 0 \le \phi \le 2\pi. \end{aligned}$$

The parameters d_0 and $d_\infty \gg d_0$ are the distances to the near-zone and far-zone boundaries, whereas $z_1(r,\phi)$ and $z_2(r,\phi)$ are smooth functions. The special case $z_1 \equiv 0$ could be viewed as the natural three-dimensional generalization of the domain depicted in Figure 1 or the geometry considered in [FixMa78]. For the near-zone boundary Γ_0, a Dirichlet condition $u(d_0, \phi, z) = g(\phi, z)$ is given, where the boundary data g acts like a source term. Depending on the physical nature of the walls of the waveguide, the appropriate combinations of boundary conditions for Γ_1 and Γ_2 involve Dirichlet–Neumann, Neumann–Neumann, or Dirichlet–Dirichlet. The computational domain

$$\begin{cases} 0 \le \rho(x,y,z) \le 1, \\ 0 \le \varphi(x,y,z) \le 2\pi, \\ 0 \le \zeta(x,y,z) \le 1, \end{cases}$$

is mapped onto the hollow, deformed cylinder in the physical domain via a boundary conforming transformation [KnSt93, p. 5]. At the far-zone boundary, an outgoing radiation condition is imposed by employing a DtN map. By making the separability assumption that the surfaces z_1 and z_2 are horizontal *only* in an infinitesimal

neighborhood of Γ_∞, the outgoing modes are built up by Hankel functions H_n of the first kind [MorFe53, p. 623], complex exponentials, and the eigenfunctions ψ_m and eigenvalues λ_m of a Sturm–Liouville problem. Thus, in the vicinity of Γ_∞, a solution containing only propagating modes has an approximate representation

$$u(\rho,\varphi,\zeta) \approx \sum_{n=0}^{\nu} \sum_{m=1}^{\mu} H_n(\sqrt{-\lambda_m}\rho) A_{\pm n,m} \exp(\pm in\varphi)\psi_m(\zeta),$$

where μ is the number of axial modes that propagate (cf. (12)), and $2\nu+1$ is the number of azimuthal modes that are selected. An adequate choice of ν depends mainly on the energy contents of the azimuthal modes for the source term, i.e., the magnitude of the Fourier coefficients of g with respect to the functions $\{\exp(\pm in\varphi)\}_{n=0}^{\infty}$. By differentiating u with respect to ρ, and determining the coefficients $A_{\pm n,m}$ by exploiting the orthogonality of the functions $\exp(\pm in\varphi)\psi_m(\zeta)$, a DtN condition like (14) is obtained. If we define a uniform grid for the coordinates (ρ,φ,ζ) of the computational domain, and construct a finite difference discretization of the same type as in section 2.4 for the transformed Helmholtz problem, then a linear system of equations $Bu = g$ arises. By ordering the components of u properly, the matrix B will get a two-level block structure, where the block sizes are given by the numbers of grid points m_φ and m_ζ for the φ- and ζ-directions.

As a convergence accelerator that matches the block structure of B, we employ a normal block preconditioner [Otto96] with two block levels. For such a preconditioner, it is optional to apply transforms either for only one of the block levels or for both. In the current application, we obtain at least two possible types of preconditioners, which will be referred to as being monotransform and ditransform. For the monotransform preconditioner, the Fourier transform is applied for the block level corresponding to the φ-direction, thus matching the condition of azimuthal periodicity. In the ditransform case, the Fourier transform is used for the φ-coordinate, whereas the sine transform-II, cosine transform, or sine transform is applied in the ζ-direction when the matching boundary conditions are Dirichlet–Neumann, Neumann–Neumann, or Dirichlet–Dirichlet, respectively. The effect of applying the transforms in the ditransform preconditioner is to decouple the three-dimensional problem with $m_\varphi m_\zeta m_\rho$ degrees of freedom into $m_\varphi m_\zeta$ independent one-dimensional problems of size m_ρ. The discretization matrices for those one-dimensional problems will be narrow banded. Analogously, the monotransform preconditioner reduces the three-dimensional problem into m_φ two-dimensional problems with $m_\zeta m_\rho$ degrees of freedom. The two-dimensional discretization matrices will not be narrow banded. This will cause fill-in, but to a lesser degree than for the original three-dimensional problem. Moreover, the rate of convergence for the monotransform preconditioner is presumably better than for its ditransform counterpart, since the former retains more information from the original discretization matrix B. If z_1, z_2, and the wavenumber κ are axially symmetric, i.e., independent of φ, the boundary conforming transformation can be simplified in such a way[1] that the monotransform preconditioner becomes a direct solver. Note that this holds even for source terms without axial symmetry and wavenumbers depending on ζ and ρ, which is of importance for applications in ocean acoustics where the sound speed might vary with depth and range [JKPS94, p. 3]. If z_1 and z_2 also are radially independent, and $\kappa = \kappa(\rho)$, both the monotransform and ditransform preconditioners become direct solvers. Based on these observations, we expect the

[1] The φ-coordinate enters the transformation simply as a rotation around the $z(\rho,\zeta)$-axis.

monotransform or the ditransform preconditioner to be an effective convergence accelerator when z_1, z_2, and κ vary moderately with φ or in the latter case with ρ, φ, and ζ.

Fast transform-based preconditioners could also be applied to discretizations of the Helmholtz equation arising in scattering problems. Grote and Keller [GroKe95] have designed DtN conditions for Helmholtz problems where the scatterer aligns with the coordinate surfaces of, e.g., spherical coordinates or prolate spheroidal coordinates [MorFe53, p. 661]. The latter coordinate system accommodates elongated scattering obstacles such as hulls of submarines. For both the spherical and prolate spheroidal cases described in [GroKe95], a monotransform preconditioner using the Fourier transform for the azimuthal coordinate would yield a direct solution method, provided that the wavenumber is azimuthally independent. For the more general case where the obstacle does not align with a coordinate surface, Ernst [Ernst96] has constructed and analyzed iterative capacitance matrix methods. A key observation is that the irregular shape of the obstacle gives rise to a low rank perturbation of a regular discretization matrix, which is invertible by means of fast transform methods. The inverse of the original matrix is then updated by employing the Sherman–Morrison–Woodbury formula.

The techniques described in this section would easily carry over to, e.g., electromagnetic waveguides and scatterers. Even though the phase velocity and the range of frequencies would be different from those in hydroacoustics, the relevant parameters determining the size of the discretized problem would still be the size of the domain in terms of wavelengths and the resolution requirement.

4. Computational issues.

4.1. Resolution. Since the solutions to the Helmholtz equation are waves, it is evident that the grid size h must follow the wavenumber κ in order to achieve a given accuracy. A naïve approach would be to use a fixed number of grid points per wavelength, i.e., keeping κh constant. Bayliss, Goldstein, and Turkel [BaGoTu85a] established that such a resolution criterion is insufficient. Instead they presented estimates predicting that the L_2 norm of the error behaves like $\mathcal{O}(\kappa^{p+1}h^p)$ for a pth-order finite element discretization. Similar estimates have been rigorously proved [IhlBa97] for a one-dimensional model problem with Dirichlet–Robin boundary conditions. The estimates are in accordance with results conjected from numerical experiments [ThoPin94]. The objective of this section is to specify convenient resolution criteria for the finite difference discretization in section 2.4, resembling those in [BaGoTu85a].

The analysis is based on a one-dimensional counterpart of (2), i.e.,

$$-\frac{d^2v}{d\xi_2^2} - (\kappa d_2)^2 v = 0, \quad 0 < \xi_2 < 1, \tag{24}$$

with Robin boundary conditions

$$\begin{aligned} -\frac{dv}{d\xi_2}(0) - i\kappa d_2 v(0) &= -2i\kappa d_2 A, \\ \frac{dv}{d\xi_2}(1) - i\kappa d_2 v(1) &= 0, \end{aligned} \tag{25}$$

replacing (11) and (14). Note that for a one-dimensional problem, the Sommerfeld condition (25) is exact inasmuch as it allows only right-going waves. Applying the

same discretization as in section 2.4 to (24) results in

$$-(v_{k+1} - 2v_k + v_{k-1}) - (\kappa d_2)^2 h_2^2 v_k = 0, \qquad k = 2, \ldots, m_2 - 1, \tag{26}$$

$$-(v_2 - v_1) - i\kappa d_2 \frac{h_2}{2}(v_1 + v_2) = -2i\kappa d_2 h_2 A, \tag{27}$$

$$(v_{m_2} - v_{m_2-1}) - i\kappa d_2 \frac{h_2}{2}(v_{m_2-1} + v_{m_2}) = 0 \tag{28}$$

for the finite difference approximation $v_k \approx v(\xi_{2,k})$. The difference equation (26) has the following characteristic equation

$$r^2 - (2 - (\kappa d_2)^2 h_2^2) r + 1 = 0$$

with roots denoted by r_1 and r_2. The root

$$r_1 = 1 - \frac{(\kappa d_2)^2 h_2^2}{2} + \sqrt{-(\kappa d_2)^2 h_2^2 + \frac{(\kappa d_2)^4 h_2^4}{4}}$$

corresponds to the right-going mode; whereas the remaining root

$$r_2 = 1 - \frac{(\kappa d_2)^2 h_2^2}{2} - \sqrt{-(\kappa d_2)^2 h_2^2 + \frac{(\kappa d_2)^4 h_2^4}{4}}$$

is associated with the left-going mode. Thus, the solution to (26) is

$$v_k = C_1 r_1^{k-\frac{3}{2}} + C_2 r_2^{k-\frac{3}{2}},$$

where the coefficients C_1 and C_2 are determined from (27) and (28), yielding

$$C_1 = \left(1 + \mathcal{O}((\kappa d_2)^2 h_2^2)\right) A, \qquad C_2 = \mathcal{O}\left((\kappa d_2)^2 h_2^2 A\right).$$

Combining this with a Taylor expansion of $r_1^{k-\frac{3}{2}}$ leads to

$$v_k = A \exp\left(i\kappa d_2 \xi_{2,k} + \frac{i(\kappa d_2)^3 h_2^2}{24} \xi_{2,k} + \mathcal{O}((\kappa d_2)^4 h_2^3) \right) + \mathcal{O}((\kappa d_2)^2 h_2^2 A).$$

Comparing this with the true solution to (24), i.e.,

$$v(\xi_{2,k}) = A \exp(i\kappa d_2 \xi_{2,k}),$$

we conclude that the leading phase error of magnitude $\frac{(\kappa d_2)^3 h_2^2}{24} \xi_{2,k}$ grows linearly in ξ_2. Furthermore, a reasonable resolution criterion is

$$\frac{(\kappa d_2)^3 h_2^2}{24} = \tau,$$

where τ is a given tolerance. Notice that for this resolution we obtain

$$v_k = A \exp\left(i\kappa d_2 \xi_{2,k} + i\tau \xi_{2,k} + \mathcal{O}(\tau^{\frac{3}{2}} (\kappa d_2)^{-\frac{1}{2}})\right) + \mathcal{O}(\tau (\kappa d_2)^{-1} A).$$

When κd_2 is sufficiently large and τ is less than one, the terms $\mathcal{O}(\tau(\kappa d_2)^{-1} A)$ and $\mathcal{O}(\tau^{\frac{3}{2}} (\kappa d_2)^{-\frac{1}{2}})$, representing artificial reflections and amplitude errors, are negligible compared with the phase error $\tau \xi_{2,k}$. Under these circumstances, the phase error is a

measure of the pointwise relative error. Extensive numerical experiments, comparing the numerical solution v_k with the true solution $v(\xi_{2,k})$, corroborate that the phase error prediction above is sharp.

Thus, for the two-dimensional problem in section 2.4, we are led to the following resolution in the ξ_2-direction:

$$h_2 \leftarrow \frac{(24\tau)^{\frac{1}{2}}}{(\kappa d_2)^{\frac{3}{2}}}, \qquad m_2 \leftarrow \left\lceil \frac{1}{h_2} + 2 \right\rceil . \tag{29}$$

For the ξ_1-direction, the choice of resolution is more subtle. Lacking a more sophisticated analysis, a rescaling of condition (29) is advocated:

$$\begin{aligned} d_1 &\leftarrow \max(d_\ell, d_r), \\ h_1 &\leftarrow \frac{(24\tau d_1/d_2)^{\frac{1}{2}}}{(\kappa d_1)^{\frac{3}{2}}}, \qquad m_1 \leftarrow \left\lceil \frac{1}{h_1} - \frac{1}{2} \right\rceil . \end{aligned} \tag{30}$$

4.2. Complexity. In this section we discuss the efficiency of the normal block preconditioned iterative method regarding memory requirement and arithmetic complexity. Note that only the highest order terms will be considered and that the number of arithmetic operations will be normalized by the number of unknowns $m_1 m_2$. A complex addition will be counted as two arithmetic operations, a complex multiplication as six arithmetic operations, and a complex division as nine arithmetic operations.

In order to determine the arithmetic complexity, we must specify how the initial approximation and the stopping criterion are computed. As an initial approximation we use the preconditioned right-hand side $M^{-1}g$, which is advantageous if $M^{-1}B$ is close to the identity matrix. We have imposed the following stopping criterion

$$\frac{\|M^{-1}(g - Bu^{(i)})\|_2}{\|M^{-1}g\|_2} < \epsilon$$

with tolerance $\epsilon = 10^{-4}$.

The arithmetic work can be divided into initialization and iteration. The initial part consists of forming [Otto96] the preconditioner and factorizing the tridiagonal systems at a cost of $a_{pf} = 20\frac{\hat{m}}{m_1}\log_2 \hat{m} + 125$, where $\hat{m} = 2^{\lceil \log_2(2m_1+1)\rceil+1}$. The computation of the initial approximation is done with a preconditioner solve that requires $a_{ps} = 20\frac{\hat{m}}{m_1}\log_2 \hat{m} + 115$ arithmetic operations per unknown. The iterative method also goes through some initial steps. The cost for these is $a_{in} = 2a_{ps} + a_m + 10$, where $a_m = 40$ is the work required for a matrix–vector product $y = Bx$. Accordingly, the total arithmetic cost for the initialization becomes $a_{init} = a_{pf} + a_{ps} + a_{in}$.

The work for one iteration of GMRES(ℓ) is taken as the average over a complete cycle of ℓ iterations and is given by $a_{it} = a_m + a_{ps} + 8\ell + 38$.

If we let n_{it} be the number of iterations required for convergence, then the total work for solving $M^{-1}Bu = M^{-1}g$ with the GMRES(ℓ) method is $a_{init} + n_{it}a_{it}$.

The memory requirement for our method is $m_m + m_p + m_{it}$; where $m_m = 7m_1m_2$ is the number of memory positions needed for the coefficient matrix, the right-hand side, and the solution; $m_p = 8m_1m_2 + 4\hat{m}m_2$ is the number of memory positions used by the preconditioner; and $m_{it} = 2(\ell+1)m_1m_2$ denotes the storage requirement for the iterative method. Note that a complex value is considered to take up two memory positions.

TABLE 1
Comparison of preconditioned GMRES(ℓ) *and band Gaussian elimination.*

	Arithmetic complexity	Memory requirement
Band GE	$8m_1^2 + 25m_1$	$4m_1 + 4$
GMRES(ℓ,`ssor`)	$203 + n_{it}(8\ell + 128)$	$2\ell + 12$
GMRES(ℓ,`nb`)	$80\frac{\hat{m}}{m_1}\log_2 \hat{m} + 520$ $+ n_{it}(20\frac{\hat{m}}{m_1}\log_2 \hat{m} + 8\ell + 193)$	$2\ell + 17 + 4\frac{\hat{m}}{m_1}$

In Table 1, our method, denoted by GMRES(ℓ,`nb`), is compared with band Gaussian elimination, which is the standard direct method. We also consider SSOR preconditioning, which is a technique commonly used for Helmholtz problems [BaGoTu83], [MaFrPi97]. For the implementation of GMRES(ℓ) combined with the SSOR preconditioner, the so-called Eisenstat trick [MaFrPi97, and the references therein] was used. The storage requirements have been normalized by the number of unknowns.

5. Numerical experiments. In this section the results from some numerical experiments are presented. In all experiments, the systems of equations have been solved using the GMRES(ℓ) method combined with the preconditioner defined in section 3.1 and in some cases also combined with the SSOR preconditioner. The orthogonal grid is generated by a code based on the method described in [Abra91]. The implementations are made in Fortran 90, utilizing 64-bit precision for the grid generation, and 32-bit precision for the iterative method and the preconditioner. The numerical experiments were performed on a DEC AlphaServer 8200 EV5/300.[2] The geometry of the duct, i.e., the bottom profile, is defined by the following functions:

$$\begin{cases} x_1(\theta) = d_\ell + (d_r - d_\ell)\frac{\tanh(s(\theta-\delta_c))-\tanh(-s\delta_c)}{\tanh(s(1-\delta_c))-\tanh(-s\delta_c)}, \\ x_2(\theta) = \theta d_2, \end{cases} \qquad 0 \le \theta \le 1,$$

where

$$\delta_c = 0.5, \qquad s = \frac{4}{\min(\delta_c, 1-\delta_c)}.$$

By this choice the depth varies smoothly from d_ℓ at the left boundary to d_r at the right boundary. The parameter δ_c determines the center of the slope, whereas s controls the steepness. By increasing s, the slope steepens and the bottom flattens out at the ends. The relative source depth δ_s is set to 0.5 in all experiments. We use resolution criteria (29) and (30) with a phase error tolerance $\tau = 8\%$.

It would be interesting to investigate the arithmetic speedup for GMRES(ℓ,`nb`) compared with restarted GMRES using a very simple preconditioner such as symmetric diagonal scaling, but the latter method with restarting length $\ell \le 17$ does not converge in 8361 iterations for any of the problem sizes considered. However, the effectiveness of the normal block preconditioner M is indicated when comparing the corresponding preconditioned spectrum with the one arising after symmetric diagonal scaling. The spectra for a small (2730 unknowns) problem are shown in Figures 2 and 3. The preconditioned spectrum exhibits a high degree of clustering around one, which is favorable for Krylov subspace methods [Green97, p. 55], [Axel88].

[2]The actual computer is part of the Yggdrasil computing facilities at the Department of Scientific Computing, Uppsala University.

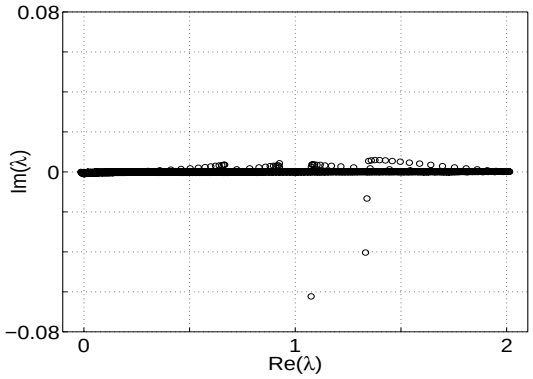


FIG. 2. *The spectrum of B (after symmetric diagonal scaling) for $d_2 = 300$, $d_\ell = 50$, $d_r = 20$, and $f = 25$.*

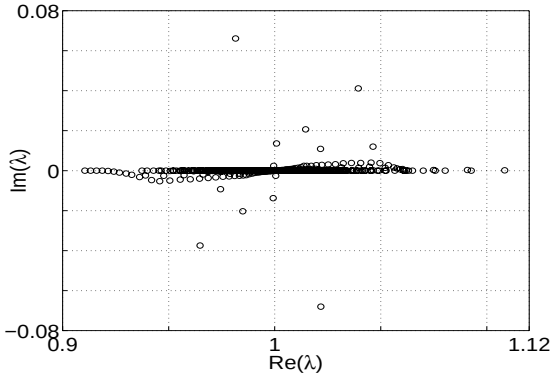


FIG. 3. *The spectrum of $M^{-1}B$ for $d_2 = 300$, $d_\ell = 50$, $d_r = 20$, and $f = 25$.*

Since the preconditioner coincides with the discretization matrix for the model problem in a duct with a flat bottom, it is to be expected that the rate of convergence will be affected by the geometry. When the bottom of the duct gets more curved, the preconditioner is not as good an approximation of B. Figure 4 shows how the geometry influences the number of iterations for GMRES(6,nb). Notice that here the number of iterations decreases when the problem size increases (and the duct gets less curved).

Another interesting issue is how the frequency affects the number of iterations. This is demonstrated in Figure 5 for a duct of medium steepness and frequencies in the low–to–intermediate range. For the same duct and frequencies, we investigated the effect of mismatching the transform and the Dirichlet–Neumann boundary conditions. To this end, preconditioners based on the cosine and the sine transforms were used. The result was that preconditioned GMRES(6) did not converge in 695 iterations. For Helmholtz problems with Neumann–Neumann or Dirichlet–Dirichlet boundary conditions, a similar severe degradation of the convergence rate is triggered by choosing inapt transforms [Otto96].

In Figures 6 and 7, the results from comparative experiments are shown. The number of unknowns depends cubicly on the frequency and ranges from 7452 to 2563902. We define the arithmetic gain for one method over another as the total

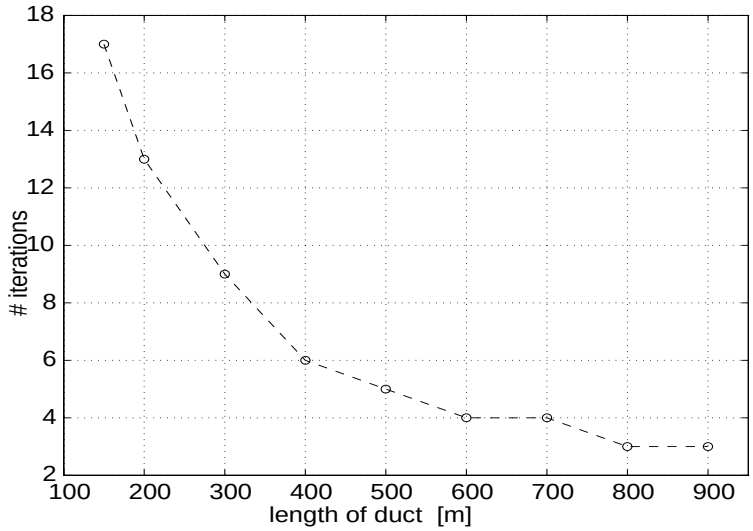


FIG. 4. *Number of iterations for ducts where the length d_2 is varied. All the other parameters are held constant, $d_\ell = 50$, $d_r = 20$, and $f = 100$.*

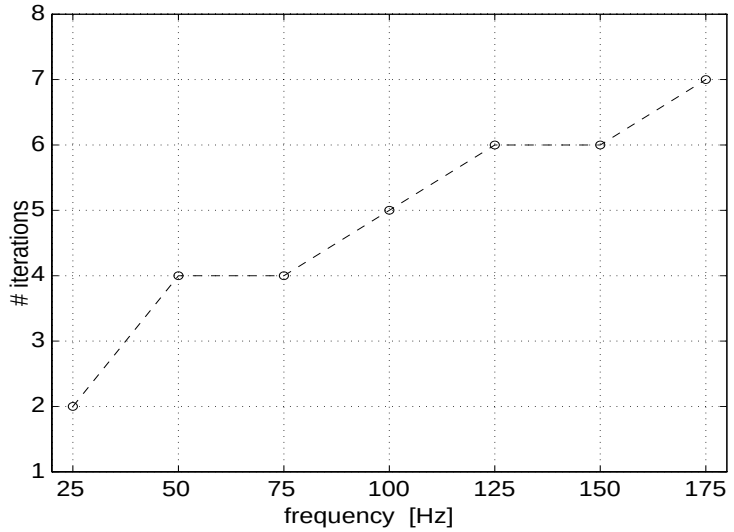


FIG. 5. *Number of iterations for different frequencies, $d_2 = 500$, $d_\ell = 50$ and $d_r = 20$.*

number of arithmetic operations for the latter divided by the total arithmetic complexity of the former. Analogously, the memory gain is defined in terms of memory requirements. To make a fair comparison between the two preconditioned methods, the restarting length ℓ_s for the GMRES(ℓ_s,`ssor`) method is chosen such that the memory requirement equals that of GMRES(ℓ,`nb`). In addition, the relaxation parameter for SSOR is set to its optimal value, which is determined within three decimal digits by a bisecting procedure in each test case. It is clear that GMRES(6,`nb`) is more efficient than band Gaussian elimination both regarding arithmetic complexity and memory requirement for all problem sizes considered. Furthermore, the gains increase as the

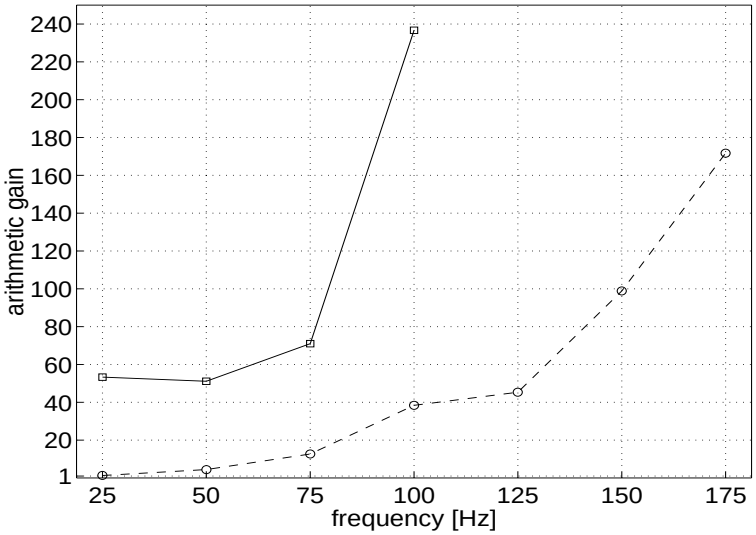


FIG. 6. *Arithmetic gain for* GMRES(6,`nb`) *compared with* GMRES(17,`ssor`) *(solid line) and band Gaussian elimination (dashed line).*

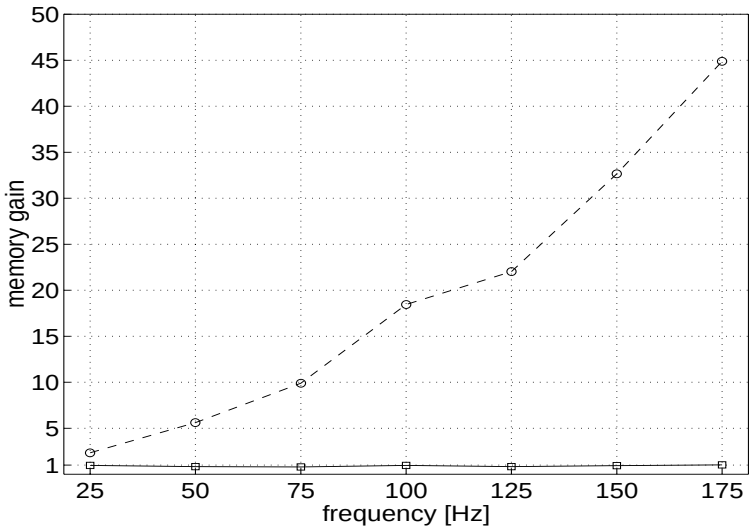


FIG. 7. *Memory gain for* GMRES(6,`nb`) *compared with* GMRES(17,`ssor`) *(solid line) and band Gaussian elimination (dashed line).*

frequency increases. Notice that, with respect to arithmetic complexity, the SSOR preconditioned method is even more outperformed than band Gaussian elimination. The actual iteration counts for GMRES(17,`ssor`) are 1006, 2141, 3504, and 9092 for the smaller problems, whereas the method fails to converge in 10000 iterations for the three largest problems.

Finally, Figures 8 and 9 display the solutions for two different frequencies. We have chosen rather low frequencies because those solutions are easier to visualize. Note that, for the lower frequency, only one wave mode is transmitted in the narrow part

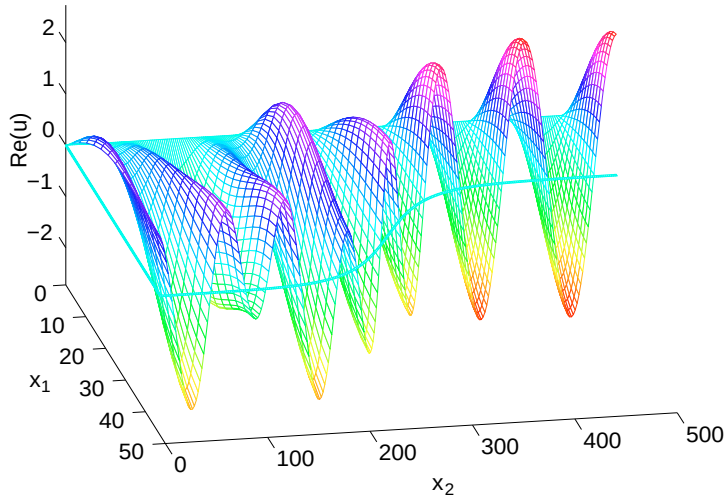


FIG. 8. *The solution for* $f = 25$, $d_2 = 500$, $d_\ell = 50$ *and* $d_r = 20$. *The contour of the duct is also depicted.*

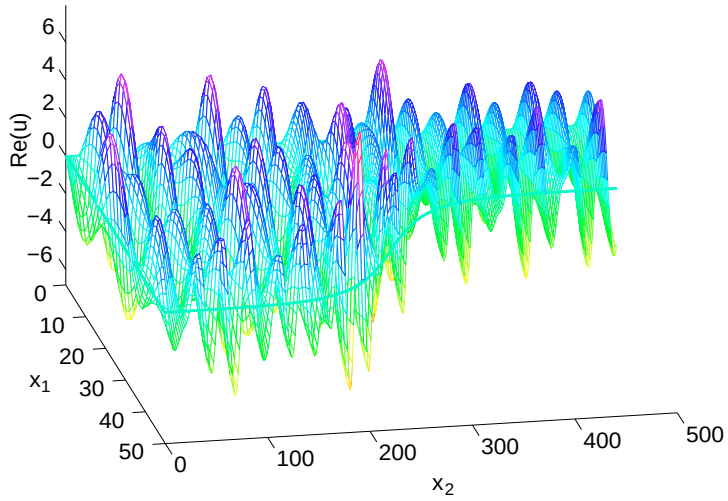


FIG. 9. *The solution for* $f = 75$, $d_2 = 500$, $d_\ell = 50$ *and* $d_r = 20$. *The contour of the duct is also depicted.*

of the duct. For the higher frequency, several modes are transmitted and interfere.

6. Conclusions. We have applied a preconditioned GMRES(ℓ) algorithm to a second-order finite difference discretization of the Helmholtz equation subject to Dirichlet, Neumann, and DtN boundary conditions for a two-dimensional waveguide problem. We utilize a normal block preconditioner, which results in a direct fast Helmholtz solver for rectangular domains, provided that the transform matches the boundary conditions. The memory requirement for this preconditioned method is linear in the number of unknowns. Thus, the sparsity of the original discretization

matrix is efficiently exploited. Numerical experiments for a hydroacoustic wave propagation problem show that the iterative method accelerated by the normal block preconditioner yields a significant gain both in storage requirement and arithmetic complexity when it is compared with band Gaussian elimination. In comparison with an SSOR preconditioned method having the same memory requirement, the gain in arithmetic complexity is even larger. For the normal block method versus band Gaussian elimination, the gain increases when the wavenumber is raised. Moreover, the number of iterations required for convergence grows moderately (or even decreases) as the number of unknowns increases.

In order to suppress the phase error, the number of unknowns has to grow cubicly in the wavenumber due to the second-order accurate discretization. Thus, for high wavenumbers, the *discretization* is less tractable from a computational point of view. The memory requirement might be a bottle-neck. To mitigate this adverse effect, high-order discretizations will be investigated in a forthcoming paper. Another pertinent concern is to perform a more rigorous phase error analysis. Further directions of research will also entail applications to heterogeneous media, e.g., cases where the sound speed depends on the depth due to temperature gradients, changes in hydrostatic pressure, and variable salinity.

Acknowledgments. The authors would like to thank Dr. Leif Abrahamsson for supplying the grid generation code. The first author also expresses his gratitude to Prof. Howard Elman, who invited the author to a postdoctoral visit at the Department of Computer Science, University of Maryland, where this research was completed.

REFERENCES

[Abra91] L. Abrahamsson, *Orthogonal grid generation for two-dimensional ducts*, J. Comput. Appl. Math., 34 (1991), pp. 305–314.

[AbKr94] L. Abrahamsson and H.-O. Kreiss, *Numerical solution of the coupled mode equations in duct acoustics*, J. Comput. Phys., 111 (1994), pp. 1–14.

[Axel88] O. Axelsson, *A restarted version of a generalized preconditioned conjugate gradient method*, Comm. Appl. Numer. Methods, 4 (1988), pp. 521–530.

[BaiSw91] D. H. Bailey and P. N. Swarztrauber, *The fractional Fourier transform and applications*, SIAM Rev., 33 (1991), pp. 389–404.

[BaGoTu83] A. Bayliss, C. I. Goldstein, and E. Turkel, *An iterative method for the Helmholtz equation*, J. Comput. Phys., 49 (1983), pp. 443–457.

[BaGoTu85a] A. Bayliss, C. I. Goldstein, and E. Turkel, *On accuracy conditions for the numerical computation of waves*, J. Comput. Phys., 59 (1985), pp. 396–404.

[BaGoTu85b] A. Bayliss, C. I. Goldstein, and E. Turkel, *The numerical solution of the Helmholtz equation for wave propagation problems in underwater acoustics*, Comput. Math. Appl., 11 (1985), pp. 655–665.

[BaGuTu82] A. Bayliss, M. Gunzburger, and E. Turkel, *Boundary conditions for the numerical solution of elliptic equations in exterior regions*, SIAM J. Appl. Math., 42 (1982), pp. 430–451.

[CCEW98] X.-C. Cai, M. A. Casarin, F. W. Elliott, and O. B. Widlund, *Overlapping Schwarz algorithms for solving Helmholtz's equation*, in Domain Decomposition Methods 10, J. Mandel, C. Farhat, and X.-C. Cai, eds., Contemp. Math. 218, AMS, Providence, RI, 1998, pp. 391–399.

[ChanNg96] R. H. Chan and M. K. Ng, *Conjugate gradient methods for Toeplitz systems*, SIAM Rev., 38 (1996), pp. 427–482.

[Ernst96] O. G. Ernst, *A finite-element capacitance matrix method for exterior Helmholtz problems*, Numer. Math., 75 (1996), pp. 175–204.

[FixMa78] G. J. Fix and S. P. Marin, *Variational methods for underwater acoustic problems*, J. Comput. Phys., 28 (1978), pp. 253–270.

[FrGoNa92] R. W. Freund, G. H. Golub, and N. M. Nachtigal, *Iterative solution of linear systems*, Acta Numerica, 1 (1992), pp. 57–100.

[Gold82] C. I. Goldstein, *A finite element method for solving Helmholtz type equations in waveguides and other unbounded domains*, Math. Comp., 39 (1982), pp. 309–324.

[Green97] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, PA, 1997.

[GroKe95] M. J. Grote and J. B. Keller, *On nonreflecting boundary conditions*, J. Comput. Phys., 122 (1995), pp. 231–243.

[IhlBa97] F. Ihlenburg and I. Babuška, *Finite element solution of the Helmholtz equation with high wave number Part* II: *The h-p version of the FEM*, SIAM J. Numer. Anal., 34 (1997), pp. 315–358.

[JKPS94] F. B. Jensen, W. A. Kuperman, M. B. Porter, and H. Schmidt, *Computational Ocean Acoustics*, AIP Press, New York, 1994.

[KeGi89] J. B. Keller and D. Givoli, *Exact non-reflecting boundary conditions*, J. Comput. Phys., 82 (1989), pp. 172–192.

[Kim98] S. Kim, *Domain decomposition iterative procedures for solving scalar waves in the frequency domain*, Numer. Math., 79 (1998), pp. 231–259.

[KnSt93] P. Knupp and S. Steinberg, *Fundamentals of Grid Generation*, CRC Press, Boca Raton, FL, 1993.

[MaFrPi97] M. Malhotra, R. W. Freund, and P. M. Pinsky, *Iterative solution of multiple radiation and scattering problems in structural acoustics using a block quasi-minimal residual algorithm*, Comput. Methods Appl. Mech. Engrg., 146 (1997), pp. 173–196.

[MorFe53] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, McGraw-Hill, New York, 1953.

[Otto96] K. Otto, *A unifying framework for preconditioners based on fast transforms*, Report 187, Department of Scientific Computing, Uppsala University, Uppsala, Sweden, 1996.

[RiMor67] R. D. Richtmyer and K. W. Morton, *Difference Methods for Initial-Value Problems*, 2nd ed., John Wiley & Sons, New York, 1967.

[SaadSch86] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[ThoPin94] L. L. Thompson and P. M. Pinsky, *Complex wavenumber Fourier analysis of the p-version finite element method*, Comput. Mech., 13 (1994), pp. 255–275.

[VLoan92] C. F. Van Loan, *Computational Frameworks for the Fast Fourier Transform*, SIAM, Philadelphia, PA, 1992.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 1, pp. 230–244
© 1999 Society for Industrial and Applied Mathematics

COMPRESSIONS AND DILATIONS OF NUMERICAL RANGES*

J. MAROULAS† AND M. ADAM†

Abstract. Inner and outer approximation of numerical ranges of $n \times n$ complex matrices and matrix polynomials is investigated in this paper, which is based on the numerical ranges of matrices of smaller or double dimensions.

Key words. numerical range, matrix polynomials

AMS subject classifications. 15A60

PII. S0895479898340524

1. Introduction. Let M_n be the algebra of all $n \times n$ complex matrices. For a matrix $A \in M_n$, the *numerical range* $NR[A]$, also known as the *field of values*, is the set of complex numbers

$$NR[A] = \{\langle Ax, x\rangle : \quad x \in C^n, \ \ \|x\| = 1\}. \tag{1.1}$$

The radius

$$r(A) = \ \max \{|z| : \quad z \in NR[A]\}$$

of the smallest circle $|z| = r(A)$ that encloses $NR[A]$ is called the *numerical radius*. The usefulness of the numerical range and numerical radii is well known. As an extensive background for this active research topic we refer to [HJ] and [AL]. Replacing in (1.1) the Euclidean inner product with the *indefinite scalar product* on $\mathbf{C}^n$, it is known [GLR] that there exists an invertible indefinite hermitian matrix S such that $\langle x, y\rangle_S = \langle Sx, y\rangle$. Hence, we obtain the *S-numerical range* of A:

$$W_S(A) = \left\{ \frac{\langle Ax, x\rangle_S}{\langle x, x\rangle_S} : \ \ x \in \mathbf{C}^n, \quad \langle x, x\rangle_S \neq 0 \right\} = W_S^+(A) \bigcup W_{-S}^+(A), \tag{1.2}$$

where

$$W_S^+ = \{\langle Ax, x\rangle_S : \ \ x \in \mathbf{C}^n, \ \ \langle x, x\rangle_S = 1\}. \tag{1.3}$$

In particular, the set $W_S^+(A)$ is called the *positive S-numerical range*. The S-numerical ranges generalize the classical numerical range, and some properties of the $NR[A]$ can be extended to $W_S(A)$. In [B] it is proved that $W_S^+(A)$ is a convex set and also that the closure of set $W_S(A)$ contains all eigenvalues of A if A is positive definite. Moreover, we can readily verify the following:

(i) $W_S^+(A + kI) = W_S^+(A) + k, \quad k \in \mathbf{C}$.
(ii) $W_S^+(kA) = kW_S^+(A), \quad k \in \mathbf{C}$.
(iii) $W_S^+(A + B) \subseteq W_S^+(A) + W_S^+(B)$.
(iv) $W_S^+(A_1) \subseteq W_S^+(A), A_1,$ a submatrix of A.

* Received by the editors June 4, 1998; accepted for publication (in revised form) by R. Bhatia November 6, 1998; published electronically September 22, 1999.
http://www.siam.org/journals/simax/21-1/34052.html
†Department of Mathematics, National Technical University, Zografou Campus, Athens 15780, Greece (maroulas@math.ntua.gr, maria@math.ntua.gr).

(v) $z \in W_S^+(A)$, S real symmetric $\Rightarrow \overline{z} \in W_S^+(\bar{A})$.

A further study of both $W_S(A)$ and $W_S^+(A)$ has been presented by [LTU], where it is shown that $W_S(A)$ is always p-convex, i.e., for any distinct pair of points $z_1, z_2 \in W_S(A)$ either the closed line segment $\overline{z_1 z_2} \subset W_S(A)$ or the line $\{az_1 + (1-a)z_2 : a \leq 0 \text{ or } 1 \leq a\} \subset W_S(A)$. Clearly, if S is positive definite, then

$$W_S(A) = W_S^+(A) = NR[S^{1/2}AS^{-1/2}].$$

The idea of the numerical range has also been extended to matrix polynomials [LR], [M]:

$$L(\lambda) = I\lambda^q + A_1\lambda^{q-1} + \cdots + A_q; \quad A_i \in M_n,$$

where

$$NR[L(\lambda)] = \{\lambda : \ \langle L(\lambda)x, x\rangle = 0 \ \text{ for some } \ x \in \mathbf{C}^n\}. \tag{1.4}$$

$NR[L(\lambda)]$ has peculiar geometric properties with respect to the boundedness and connectedness of this set, which are important to the factorization of matrix polynomials [M], [MP1, MP2, MP3], [MMP], and [LMZ].

In this paper we continue our effort for further development of the subject and the study of related problems. In section 2 we express $NR[A]$ as the union of the numerical ranges of matrices of dimensions $k \times k$ for $2 \leq k < n$. In this way, each set in the union can be considered as an *inner approximation* or *dilation* of $NR[A]$. This result generalizes the approach of Markus and Pesce [MP], where $NR[A]$ is the union of the numerical ranges of 2×2 matrices, i.e., of ellipsoid regions using only real orthogonal vectors. Taking advantage of the fact that $NR[A]$ and $NR[e^{2i\theta}\bar{A}]$ are symmetric with respect to the straight line $y = (\tan\theta)x$, we have set the convex hull of $NR[A] \bigcup NR[e^{2i\theta}\bar{A}]$ equal to the numerical range of a suitable matrix. This result for $\theta = 0$ leads to the convenient equality

$$\text{Conv.hull}\left(NR[A] \cup NR[\bar{A}]\right) = NR\left(\begin{bmatrix} M & N \\ -N & M \end{bmatrix}\right),$$

where M, N are real matrices defined by $A = M + iN$. Therefore, $NR[A]$ is presented as the intersection of numerical ranges as the line $y = (\tan\theta)x$ rotates around the origin for $0 \leq \theta \leq \pi$. The last equality is generalized further for the matrix

$$C = \begin{bmatrix} A & B \\ -\bar{B} & \bar{A} \end{bmatrix},$$

which, as is known, corresponds to a quartenionic matrix. Clearly, the outer set Conv.hull$(NR[A] \cup NR[\bar{A}])$ can be considered as a *compression* of $NR[A]$. Some additional comments on projections on the axes, on the joint numerical range [HJ], and on the numerical radius have been presented. The MATLAB procedure is exposed and examples are given.

In the third section we refer to $W_S^+(A)$. We show that for any indefinite hermitian matrix S,

$$NR[A] \cap W_S^+(A) \neq \emptyset,$$

and also that

$$W^+_{I_2\otimes S}(A\oplus B) = \text{Conv.hull}(W^+_S(A)\cup W^+_S(B)).$$

This equality clarifies Conv.hull$(W^+_S(A)\cup W^+_S(e^{2i\theta}\bar{A}))$ as the $(I_2\otimes S)$-numerical range of a matrix depending on A.

Finally, in the fourth section, we present two properties of $NR[L(\lambda)]$. In particular, the first one refers to the reduction of this set to $NR[M(\lambda)]$, where the $k\times k$ $(k\geq 2)$ matrix polynomial $M(\lambda)$ depends on $L(\lambda)$.

2. The numerical range of a matrix. An approximation or a *dilation* of $NR[A]$ in (1.1) is presented as a first statement.

THEOREM 2.1. *For a prefix number $k<n$,*

$$NR[A] = \bigcup_{\xi_1,\ldots,\xi_k} NR\left(\begin{bmatrix} \xi_1^*A\xi_1 & \ldots & \xi_1^*A\xi_k \\ \vdots & & \vdots \\ \xi_k^*A\xi_1 & \ldots & \xi_k^*A\xi_k \end{bmatrix}\right), \tag{2.1}$$

where $\xi_1,\ \ldots,\xi_k$ run over all sets by k orthonormal vectors of $\mathbf{C}^n$.

Proof. Any vector $x\in\mathbf{C}^n$ belongs to a k-dimensional subspace $\mathbf{E}\subset\mathbf{C}^n$. Let $\xi_1,\ \ldots,\xi_k$ be an orthonormal basis of $\mathbf{E}$ such that $\xi_i\in\mathbf{C}^n$. Then

$$x = [\xi_1\ \ldots\ \xi_k]\,\omega,$$

where $\omega\in\mathbf{C}^k$. Clearly, ω is a unit vector if and only if x is a unit, and due to

$$x^*Ax = \omega^*\begin{bmatrix}\xi_1^*\\ \vdots\\ \xi_k^*\end{bmatrix} A\,[\xi_1\ldots\xi_k]\,\omega = \omega^*\begin{bmatrix} \xi_1^*A\xi_1 & \ldots & \xi_1^*A\xi_k \\ \vdots & & \vdots \\ \xi_k^*A\xi_1 & \ldots & \xi_k^*A\xi_k \end{bmatrix}\omega,$$

we verify the relationship (2.1). □

For $k=2$, the interior of $NR[A]$ is covered by the ellipses

$$NR\left(\begin{bmatrix} \xi_1^*A\xi_1 & \xi_1^*A\xi_2 \\ \xi_2^*A\xi_1 & \xi_2^*A\xi_2 \end{bmatrix}\right),$$

where $\{\xi_1,\ \xi_2\}$ is any pair of orthonormal vectors. The estimation of $NR[A]$ clearly is realized easily if $k>2$, since the fields of values in the left part of (2.1) are convex sets and even they occupy a more extensive area of $NR[A]$ as k increases. This is illustrated in example 1. In particular, for $k=2$ and for $\xi_1,\ \xi_2$ real orthogonal vectors, (2.1) has been presented in [MP] using a different approach.

A similar approximation for the q-numerical range of A:

$$NR_q[A] = \{\ y^*Ax\quad :\quad \|x\| = \|y\| = 1,\ \ y^*x = q\},\qquad q\in[0,1],$$

is given by the formula:

$$NR_q[A] = \bigcup_{\xi_1,\xi_2} NR_q\left(\begin{bmatrix} \xi_1^*A\xi_1 & \xi_1^*A\xi_2 \\ \xi_2^*A\xi_1 & \xi_2^*A\xi_2 \end{bmatrix}\right),$$

since for $x \in \mathbf{\Delta} = \mathrm{span}\{\xi_1, \xi_2\}$, there exists $y \in \mathbf{\Delta}$ such that $y^*x = q$.

For the *compression* of $NR[A]$ we state the following theorem.

THEOREM 2.2. *For any matrix* A,

$$\text{(2.2)}\quad \text{Conv.hull}(NR[A] \cup NR[e^{2i\theta}\bar{A}]) = NR\left(\frac{1}{2}\begin{bmatrix} A + e^{2i\theta}\bar{A} & -i(A - e^{2i\theta}\bar{A}) \\ i(A - e^{2i\theta}\bar{A}) & A + e^{2i\theta}\bar{A} \end{bmatrix})\right),$$

where $0 \le \theta \le \pi$.

Proof. Since

$$\text{Conv.hull}(NR[A] \cup NR[e^{2i\theta}\bar{A}]) = NR\left(\begin{bmatrix} A & O \\ O & e^{2i\theta}\bar{A} \end{bmatrix}\right),$$

we have to show that

$$NR\left(\frac{1}{2}\begin{bmatrix} A + e^{2i\theta}\bar{A} & -i(A - e^{2i\theta}\bar{A}) \\ i(A - e^{2i\theta}\bar{A}) & A + e^{2i\theta}\bar{A} \end{bmatrix}\right) = NR\left(\begin{bmatrix} A & O \\ O & e^{2i\theta}\bar{A} \end{bmatrix}\right).$$

In fact, for a unit vector $x \in \mathbf{C}^{2n}$ and for $A = M + iN$, where $M, N \in \mathbf{R}_{n\times n}$, we have

$$\begin{aligned} x^*\begin{bmatrix} A & O \\ O & e^{2i\theta}\bar{A} \end{bmatrix}x &= \frac{1}{2}x^*\begin{bmatrix} I & O \\ O & e^{2i\theta}I \end{bmatrix}\begin{bmatrix} I & -iI \\ -iI & I \end{bmatrix}\begin{bmatrix} M & N \\ -N & M \end{bmatrix}\begin{bmatrix} I & iI \\ iI & I \end{bmatrix}x \\ &= \frac{1}{4}x^*\begin{bmatrix} I & -iI \\ -iI & I \end{bmatrix}\begin{bmatrix} (1+e^{2i\theta})I & -i(1-e^{2i\theta})I \\ i(1-e^{2i\theta})I & (1+e^{2i\theta})I \end{bmatrix} \\ &\quad \times \begin{bmatrix} M & N \\ -N & M \end{bmatrix}\begin{bmatrix} I & iI \\ iI & I \end{bmatrix}x \\ &= \frac{1}{2}x^*\left(\begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix} \otimes I\right)\begin{bmatrix} \frac{A+e^{2i\theta}\bar{A}}{2} & \frac{-i(A-e^{2i\theta}\bar{A})}{2} \\ \frac{i(A-e^{2i\theta}\bar{A})}{2} & \frac{A+e^{2i\theta}\bar{A}}{2} \end{bmatrix} \\ &\quad \times \left(\begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix} \otimes I\right)x \\ &= y^*\begin{bmatrix} \frac{A+e^{2i\theta}\bar{A}}{2} & \frac{-i(A-e^{2i\theta}\bar{A})}{2} \\ \frac{i(A-e^{2i\theta}\bar{A})}{2} & \frac{A+e^{2i\theta}\bar{A}}{2} \end{bmatrix}y, \end{aligned}$$

where $y = \frac{1}{\sqrt{2}}x^*\left(\begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix} \otimes I\right)x$. Since

$$y^*y = \frac{1}{2}x^*\left(\begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix} \otimes I\right)\left(\begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix} \otimes I\right)x = x^*x = 1,$$

the relationship (2.2) is verified readily. ∎

COROLLARY 2.3. *For any matrix* $A = M + iN$, *where* $M, N \in \mathbf{R}_{n\times n}$,

$$\text{(2.3)}\qquad \text{Conv.hull } (NR[A] \cup NR[\bar{A}]) = NR\left(\begin{bmatrix} M & N \\ -N & M \end{bmatrix}\right).$$

Proof. Equation (2.3) comes from (2.2) with $\theta = 0$. □

From Corollary 2.3, we have the following comments.

(I) For any matrix $A = M + iN$, the numerical range of a $2n \times 2n$ matrix $\begin{bmatrix} M & N \\ -N & M \end{bmatrix}$ is symmetric to the real axis.

(II)

$$NR\left(\begin{bmatrix} M & -N \\ N & M \end{bmatrix}\right) = NR\left(\begin{bmatrix} M & N \\ -N & M \end{bmatrix}\right). \tag{2.4}$$

This is evident, due to

$$\begin{bmatrix} M & -N \\ N & M \end{bmatrix} = \begin{bmatrix} O & I \\ I & O \end{bmatrix}\begin{bmatrix} M & N \\ -N & M \end{bmatrix}\begin{bmatrix} O & I \\ I & O \end{bmatrix}.$$

(III)

$$\mathrm{proj}_{ox} NR[A] = \mathrm{proj}_{ox} NR\left(\begin{bmatrix} M & N \\ -N & M \end{bmatrix}\right). \tag{2.5}$$

In fact, by the symmetry of $NR[A]$ and $NR[\bar{A}]$ to the real axis, it is implied that the field Conv.hull$(NR[A] \cup NR[\bar{A}])$ is also symmetric to the same axis. Hence, by (2.3) we obtain (2.5).

(IV)

$$\mathrm{proj}_{oy} NR[A] = \mathrm{proj}_{ox} NR\left(\begin{bmatrix} N & M \\ -M & N \end{bmatrix}\right). \tag{2.6}$$

By the equalities (2.4) and $\mathrm{Im}(x^*Ax) = \mathrm{Re}(x^*(-iA)x)$, (2.6) is obvious.

(V) The real field of values [HJ, p. 85]

$$JNR\left(\begin{bmatrix} M & -N \\ N & M \end{bmatrix}, \begin{bmatrix} N & M \\ -M & N \end{bmatrix}\right) \equiv NR[A]. \tag{2.7}$$

In fact, if we denote the unit vector $x = u + iv$, then by the relationship

$$x^*Ax = w^T \begin{bmatrix} M & -N \\ N & M \end{bmatrix} w + i\left(w^T \begin{bmatrix} N & M \\ -M & N \end{bmatrix} w\right),$$

where $w = \begin{bmatrix} u \\ v \end{bmatrix} \in \mathbf{R}^{2n}$, we confirm the truth of the contention.

THEOREM 2.4. *For any matrix* $A \in \mathbf{C}_{n\times n}$,

$$NR[A] = \bigcap_{0\le\theta\le\pi} NR\left(\frac{1}{2}\begin{bmatrix} A + e^{2i\theta}\bar{A} & -i(A - e^{2i\theta}\bar{A}) \\ i(A - e^{2i\theta}\bar{A}) & A + e^{2i\theta}\bar{A} \end{bmatrix}\right). \tag{2.8}$$

Proof. Let $z \in NR[A]$. Then the complex number

$$w = e^{2i\theta}\overline{z} = \frac{1 + i\lambda}{1 - i\lambda}\overline{z},$$

where $\lambda = \tan\theta$, is symmetric in z with respect to the straight line $y = \lambda x$. Since $w \in NR[e^{2i\theta}\bar{A}]$, the fields $NR[A]$ and $NR[e^{2i\theta}\bar{A}]$ are symmetric with respect to the line $y = \lambda x$, and hence by (2.2),

$$\bigcap_{0\le\theta\le\pi} NR\left(\frac{1}{2}\begin{bmatrix} A+e^{2i\theta}\bar{A} & -i(A-e^{2i\theta}\bar{A}) \\ i(A-e^{2i\theta}\bar{A}) & A+e^{2i\theta}\bar{A} \end{bmatrix}\right)$$
$$= \bigcap_{0\le\theta\le\pi} \text{Conv.hull}\left(NR[A]\cup NR\left[\frac{1+i\lambda}{1-i\lambda}\bar{A}\right]\right) = NR[A]. \qquad \square$$

Setting $e^{-i\theta}A = S_\theta + iT_\theta$, where S_θ, T_θ are real matrices, by (2.8) we have directly the following corollary.

COROLLARY 2.5.

$$NR[A] = \bigcap_{0\le\theta\le\pi} NR\left(e^{i\theta}\begin{bmatrix} S_\theta & T_\theta \\ -T_\theta & S_\theta \end{bmatrix}\right). \tag{2.9}$$

Hence, by the theory of the numerical range of matrices [HJ], we can say that Theorems 2.1 and 2.4 give a better approximation of $NR[A]$ than the polygonal approximation. This happens because the boundary of the numerical ranges of matrices in (2.1) and also of those in (2.8) contain parts of $\partial NR[A]$ and not points, i.e., the vertices of polygons. Hence, the verification of $NR[A]$ by the intersection in (2.8), due to the convexity of $NR[A]$, needs a small number for θ.

A further generalization of (2.3) is the investigation of a relationship between the numerical ranges of matrices:

$$C = \begin{bmatrix} A & B \\ -\bar{B} & \bar{A} \end{bmatrix} \quad K = \begin{bmatrix} M & N & \vdots & U & V \\ -N & M & \vdots & -V & U \\ \dots & \dots & \dots & \dots & \dots \\ -U & V & \vdots & M & -N \\ -V & -U & \vdots & N & M \end{bmatrix}, \tag{2.10}$$

where $A = M + iN$ and $B = U + iV$.

THEOREM 2.6. *For the matrices C and K in* (2.10) *it is true that* $NR[C] = NR[K]$.

Proof. Denote by R the unitary matrix

$$R = \frac{1}{2}\begin{bmatrix} 1 & i & 1 & -i \\ i & 1 & -i & 1 \\ -i & 1 & i & 1 \\ -1 & i & -1 & -i \end{bmatrix} \otimes I.$$

Then

$$R^* K R = \text{diag}(C, \bar{C}),$$

and thus we obtain

$$NR[K] = NR[\text{diag}(C,\bar{C})] = \text{Conv.hull}\left(NR[C]\cup NR[\bar{C}]\right).$$

Moreover, by the equation

$$\begin{bmatrix} O & -I \\ I & O \end{bmatrix} C \begin{bmatrix} O & I \\ -I & O \end{bmatrix} = \bar{C},$$

we have $NR[C] = NR[\bar{C}]$ and consequently the equality of $NR[C]$ and $NR[K]$ is evident. □

In the special case where

$$C = \begin{bmatrix} p+iq & u+iv \\ -u+iv & p-iq \end{bmatrix}, \qquad p,q,u,v \in \mathbf{R},$$

we can check that $NR[C]$ is the interval $[p - i\sqrt{q^2+u^2+v^2},\ p + i\sqrt{q^2+u^2+v^2}]$.

For the *numerical radius* $r(A)$, defined by $r(A) = \max\{|z| : z \in NR[A]\}$, we verify the following theorem.

THEOREM 2.7.

$$r\left(\begin{bmatrix} \xi_1^* A\xi_1 & \dots & \xi_1^* A\xi_k \\ \vdots & & \vdots \\ \xi_k^* A\xi_1 & \dots & \xi_k^* A\xi_k \end{bmatrix}\right) \le r(A) = r\left(\begin{bmatrix} A & O \\ O & e^{2i\theta}\bar{A} \end{bmatrix}\right). \tag{2.11}$$

Proof. By the relationship (2.1), clearly the NR of the matrix on the right is a subset of $NR[A]$, which implies the inequality in (2.11). For the equality, note that

$$r\left(\begin{bmatrix} A & O \\ O & e^{2i\theta}\bar{A} \end{bmatrix}\right) = \max\left(r(A),\ r(e^{2i\theta}\bar{A})\right) = r(A),$$

since $r(e^{2i\theta}\bar{A}) = |e^{2i\theta}| r(\bar{A}) = r(A)$. □

MATLAB procedure. The inner approximation of $NR[A]$, as it is expressed in Theorem 2.1, can be illustrated as follows:

Step 1. Introduce the matrix.

Step 2. For a prefix number $k < n$ introduce k arbitrary linearly independent vectors of $\mathbf{C}^n$.

Step 3. Orthonormalize this set to $\xi_1, \dots, \xi_k$.

Step 4. Calculate the matrix $B = [\xi_i^* A \xi_j], \quad i,j = 1,2,\dots,k$.

Step 5. Determine the numerical range of B.

Step 6. Illustrate $NR[B]$.

Step 7. Repeat this procedure for some other set of vectors.

Example 1. Let

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 2 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}.$$

In Figure 2.1, the $NR[A]$ are defined as the union of the numerical ranges of 2×2 and 3×3 matrices, an application of (2.1) for $k = 2$ and $k = 3$.

As an implication of Theorem 2.4 we give the next example.

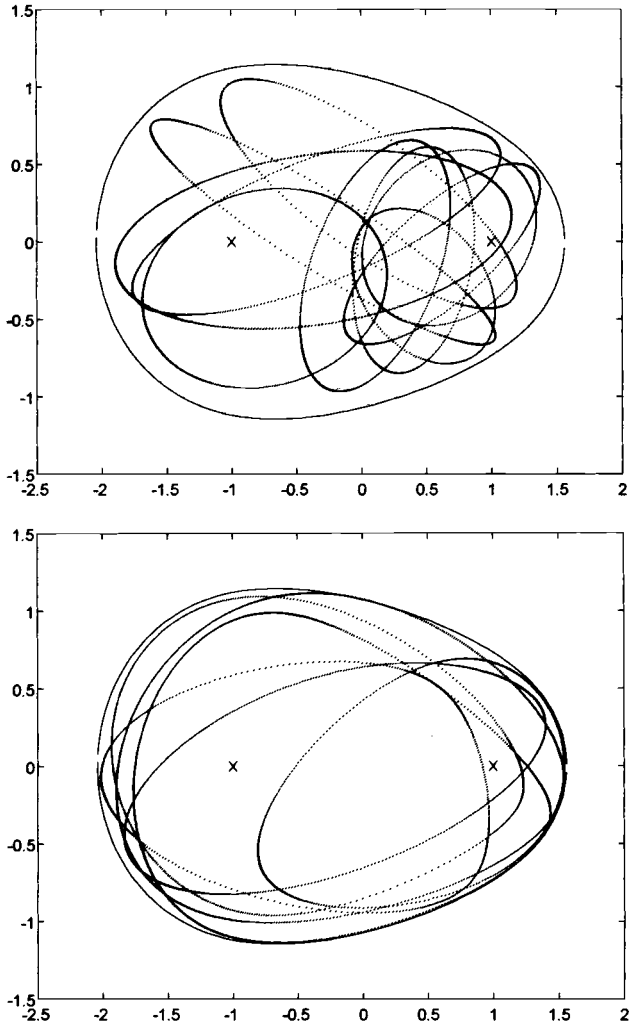


FIG. 2.1

Example 2. Let

$$A = \begin{bmatrix} 2-i & 1 \\ 0 & 1-i \end{bmatrix}.$$

Then for $\theta = 0,\ \frac{\pi}{6},\ \frac{\pi}{3},\ \frac{\pi}{2},\ \frac{2\pi}{3}$ we illustrate the NR of the matrix

$$\frac{1}{2}\begin{bmatrix} A + e^{2i\theta}\bar{A} & -i(A - e^{2i\theta}\bar{A}) \\ i(A - e^{2i\theta}\bar{A}) & A + e^{2i\theta}\bar{A} \end{bmatrix}$$

(see Figure 2.2).

From this, (2.8) is evident.

3. Numerical range on an indefinite inner product. Properties of the numerical range of a matrix on an indefinite inner product space have been presented

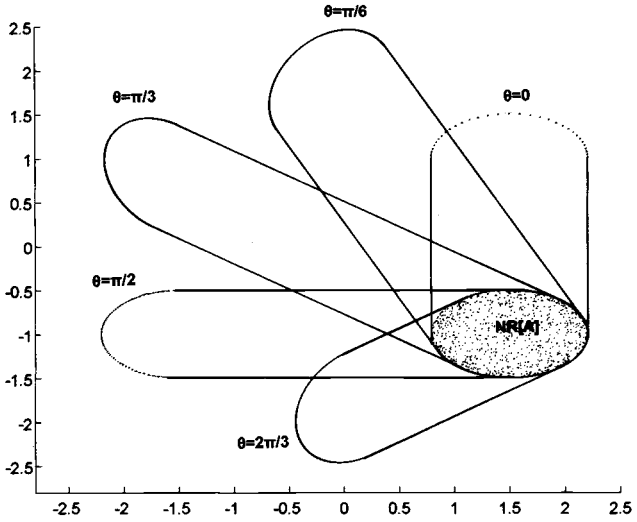


FIG. 2.2

in a recent publication [LTU]. The authors have extended some properties of the classical numerical range and have given a detailed description of $W_S^+(A)$ and

$$V_S^+(A) = \{\langle Ax, x\rangle : \quad x \in \mathbf{C}^n, \quad \langle x, x\rangle_S = 1\}.$$

Obviously, by (1.3), $W_S^+(A) = V_S^+(SA)$ and it is proved that $W_S^+(A)$ and $V_S^+(A)$ are always convex in a different approach from that in [B]. We next state a connection of $NR[A]$ and $W_S^+(A)$.

THEOREM 3.1. *For any hermitian matrix S, the set $NR[A] \cap W_S^+(A)$ is nonempty.*

Proof. For the hermitian S, let P be an orthonormal matrix such that $S = P^*DP$, where D is a real diagonal matrix. Then for any S-unit vector $x \in \mathbf{C}^n$ (i.e., $x^*Sx = 1$), we have

$$\langle Ax, x\rangle_S = x^*SAx = x^*P^*DPAx = y^*DBy = \langle By, y\rangle_D,$$

where $y = Px \in \mathbf{C}^n$ and $B = PAP^*$. Since y is a D-unit vector, $W_S^+(A) = W_D^+(B)$. Moreover, for any unit vector $w \in \mathbf{C}^n$, the equations

$$w^*Aw = w^*P^*BPw = z^*Bz, \qquad z = Pw, \ \|z\| = 1$$

imply $NR[A] = NR[B]$.

Therefore, it is enough to prove that $NR[B] \cap W_D^+(B) \neq \emptyset$, where D is a real diagonal matrix. Next we denote $D = \operatorname{diag}(D_1, -D_2)$, where $D_1 > 0$ and $D_2 \geq 0$ are the subdiagonal matrices. Let $\mathbf{E}$ be a subspace of $\mathbf{C}^n$ and $\xi_1, \ldots, \xi_k$ be a D-orthonormal basis of $\mathbf{E}$, i.e., $\langle \xi_j, \xi_i\rangle_D = \xi_i^* D\xi_j = \pm\delta_{ij}$, where δ_{ij} is Kronecker's symbol. In particular, we can consider that

$$1 = \langle \xi_i, \xi_i\rangle_D = \langle \xi_{io}, \xi_{io}\rangle_{D_1} \qquad \text{for } \xi_i = \begin{bmatrix} \xi_{io} \\ 0 \end{bmatrix}, \qquad i = 1, 2, \ldots, p$$

and

$$-1 = \langle \xi_j, \xi_j\rangle_D = \langle \xi_{jo}, \xi_{jo}\rangle_{D_2} \quad \text{for } \xi_j = \begin{bmatrix} 0 \\ \xi_{jo} \end{bmatrix}, \quad j = p+1, \ldots, k.$$

Then, for $\lambda \in W_D^+(B)$, there exists a vector $x_0 = [\xi_1, \ldots, \xi_k]\,\eta$ of $\mathbf{E}$ such that

$$\lambda = \langle Bx_0, x_0\rangle_D = x_0^* DBx_0 = \eta^* \begin{bmatrix} \langle B\xi_1, \xi_1\rangle_D & \ldots & \langle B\xi_k, \xi_1\rangle_D \\ \vdots & & \vdots \\ \langle B\xi_1, \xi_k\rangle_D & \ldots & \langle B\xi_k, \xi_k\rangle_D \end{bmatrix} \eta$$

and

$$1 = \langle x_0, x_0\rangle_D = \eta^* \begin{bmatrix} \xi_1^* \\ \vdots \\ \xi_k^* \end{bmatrix} D\,[\xi_1 \ldots \xi_k]\,\eta = \eta^* \mathrm{diag}\,(I_p, -I_{k-p})\,\eta.$$

For $\eta = \begin{bmatrix} \eta_1 \\ 0 \end{bmatrix}$, with $\eta_1 \in \mathbf{C}^p$, clearly $\|\eta_1\| = 1$ and

$$\lambda = \eta_1^* \begin{bmatrix} \langle B\xi_1, \xi_1\rangle_D & \ldots & \langle B\xi_p, \xi_1\rangle_D \\ \vdots & & \vdots \\ \langle B\xi_1, \xi_p\rangle_D & \ldots & \langle B\xi_p, \xi_p\rangle_D \end{bmatrix} \eta_1$$

$$= \eta_1^* \begin{bmatrix} \langle B_1\xi_{10}, \xi_{10}\rangle_{D_1} & \ldots & \langle B_1\xi_{p0}, \xi_{10}\rangle_{D_1} \\ \vdots & & \vdots \\ \langle B_1\xi_{10}, \xi_{p0}\rangle_{D_1} & \ldots & \langle B_1\xi_{p0}, \xi_{p0}\rangle_{D_1} \end{bmatrix} \eta_1,$$

where B_1 is the $p \times p$ principal submatrix of B. Since $\xi_{10}, \ldots, \xi_{p0}$ are D_1-orthonormal vectors and $D_1 > 0$, the vectors $D_1^{1/2}\xi_{10}, \ldots, D_1^{1/2}\xi_{p0}$ are orthonormal, and by Theorem 2.1,

$$\lambda \in NR\left[D_1^{1/2} B_1 D_1^{-1/2}\right] = NR\,[B_1] \subset NR\,[B]\,. \qquad \square$$

Due to the fact that $W_{-S}^+(A) = W_{-D}^+(B)$ and $W_S(A) = W_S^+(A) \cup W_{-S}^+(A) = W_D^+(B) \cup W_{-D}^+(B)$, clearly we have the following.

COROLLARY 3.2. *For any hermitian matrix S, the sets $NR[A] \cap W_{-S}^+(A)$ and $NR[A] \cap W_S(A)$ are nonempty.*

COROLLARY 3.3. *For any hermitian matrix S, $NR[A] \cap V_S^+(A) \neq \emptyset$.*

Proof. Let $S = P^*DP$, where $D = \mathrm{diag}(D_1, -D_2)$ is a real diagonal matrix with $D_1 > 0$ and $D_2 \geq 0$. If x is an S-unit vector, then the vector $y = Px$ is D-unit and $V_S^+(A) = V_D^+(B)$, where $A = P^*BP$. For $y = \begin{bmatrix} \theta \\ 0 \end{bmatrix}$, we have

$$1 = \langle Dy, y\rangle = \theta^* D_1 \theta = (\theta^* D_1^{1/2})(D_1^{1/2}\theta)$$

and

$$\langle By, y\rangle = \langle B_1\theta, \theta\rangle = \langle \tilde{B}_1\theta_1, \theta_1\rangle,$$

where B_1 is a submatrix of B and $\theta_1 = D_1^{1/2}\theta$, $B_1 = D^{1/2}\tilde{B}_1 D^{1/2}$.
Therefore,

$$NR[\tilde{B}_1] = NR[B_1] \subset NR[B] = NR[A]$$

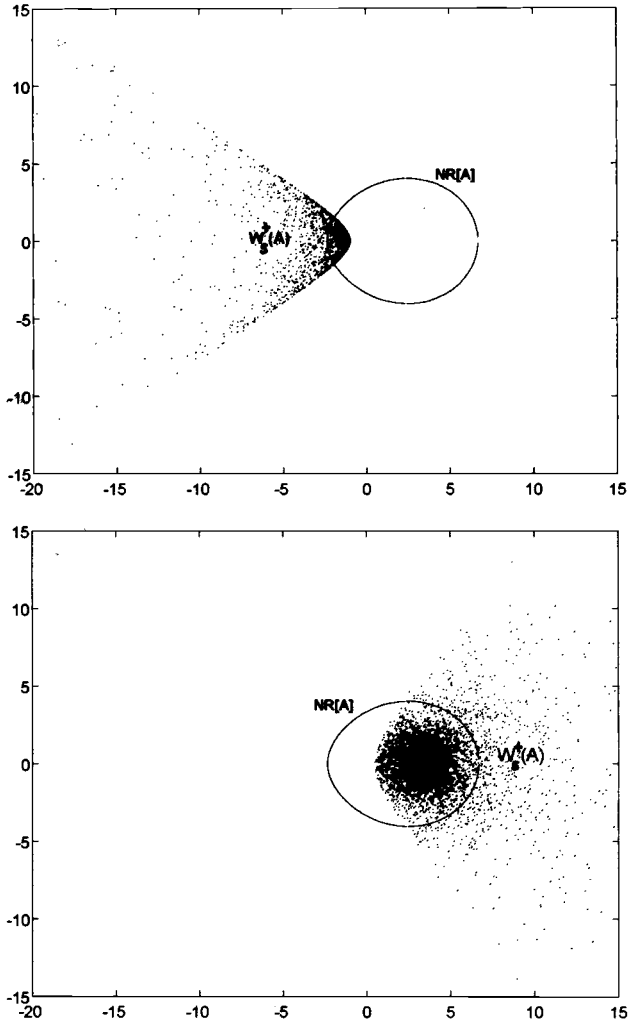


FIG. 3.1

and

$$NR[\tilde{B}_1] \,=\, V_{D_1}^+(B_1) \subset V_D^+(B) = V_S^+(A). \qquad \square$$

Example 3. The statement of Theorem 3.1 is illustrated in Figure 3.1 for

$$A = \begin{bmatrix} -1 & 0 & 2 \\ 0 & 2 & 8 \\ 1 & 0 & 3 \end{bmatrix},$$

first for $S = \mathrm{diag}(2, 0, -4)$, and then for $S = \mathrm{diag}(-2, 1, 4)$.

Another property of the numerical range of a matrix on an indefinite inner product, an analogue of Theorem 2.2, is as follows.

THEOREM 3.4. *For any matrix A,*

$$W^+_{I_2\otimes S}\left(\begin{bmatrix} A & O \\ O & e^{2i\theta}\bar{A} \end{bmatrix}\right) = W^+_{I_2\otimes S}\left(\frac{1}{2}\begin{bmatrix} A+e^{2i\theta}\bar{A} & -i(A-e^{2i\theta}\bar{A}) \\ i(A-e^{2i\theta}\bar{A}) & A+e^{2i\theta}\bar{A} \end{bmatrix}\right), \tag{3.1}$$

where $0 \le \theta \le \pi$.

Proof. Let a vector $x \in \mathbf{C}^{2n}$ such that $x^*(I_2 \otimes S)x = 1$. Then

$$x^*(I_2\otimes S)\begin{bmatrix} A & O \\ O & e^{2i\theta}\bar{A} \end{bmatrix}x = \frac{1}{2}y^*(I_2\otimes S)\begin{bmatrix} A+e^{2iy}\bar{A} & -i(A-e^{2i\theta}\bar{A}) \\ i(A-e^{2i\theta}\bar{A}) & A+e^{2i\theta}\bar{A} \end{bmatrix}y,$$

where $y = \frac{1}{\sqrt{2}}\left(\begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}\otimes I\right)x$. Since

$$y^*\,(I_2\otimes S)\,y \;=\; x^*\,(I_2\otimes S)\,x \;=\; 1,$$

(3.1) is obtained. □

On the other hand we have the following theorem.

THEOREM 3.5. *Let the hermitian matrix S have at least one positive eigenvalue. Then*

$$W^+_{I_2\otimes S}\left(\begin{bmatrix} A & O \\ O & B \end{bmatrix}\right) = \text{Conv.hull}(W^+_S(A)\cup W^+_S(B)). \tag{3.2}$$

Proof. Let

$$\lambda \in W^+_{I_2\otimes S}\left(\begin{bmatrix} A & O \\ O & B \end{bmatrix}\right)$$

correspond to the $I_2 \otimes S$-unit vector $\upsilon = \begin{bmatrix} x \\ y \end{bmatrix}$ (i.e., $x^*Sx + y^*Sy = 1$). According to Lemma 2.2 in [LTU], we can consider that $x^*Sx > 0$ and $y^*Sy > 0$ and then

$$\lambda = \upsilon^*(I_2\otimes S)\begin{bmatrix} A & O \\ O & B \end{bmatrix}\upsilon = x^*SAx + y^*SBy = x^*Sx\frac{x^*SAx}{x^*Sx} + y^*Sy\frac{y^*SBy}{y^*Sy}.$$

Hence, λ is a convex combination of

$$\frac{x^*SAx}{x^*Sx} \in W^+_S(A) \qquad \text{and} \qquad \frac{y^*SBy}{y^*Sy} \in W^+_S(B)$$

and consequently

$$W^+_{I_2\otimes S}\left(\begin{bmatrix} A & O \\ O & B \end{bmatrix}\right) \subset \text{Conv.hull}(W^+_S(A)\cup W^+_S(B)).$$

For the reverse relationship let $x^*Sx = 1$. Then the vector $z = \begin{bmatrix} x \\ 0 \end{bmatrix}$ is $I_2 \otimes S$-unit and due to

$$z^*(I_2\otimes S)\begin{bmatrix} A & O \\ O & B \end{bmatrix}z = x^*SAx,$$

it is implied that

$$W_S^+(A) \subset W_{I_2\otimes S}^+\left(\begin{bmatrix} A & O \\ O & B \end{bmatrix}\right).$$

Similarly, for $y^*Sy = 1$,

$$W_S^+(B) \subset W_{I_2\otimes S}^+\left(\begin{bmatrix} A & O \\ O & B \end{bmatrix}\right).$$

Therefore,

$$W_S^+(A) \cup W_S^+(B) \subset W_{I_2\otimes S}^+\left(\begin{bmatrix} A & O \\ O & B \end{bmatrix}\right),$$

and due to the convexity of

$$W_{I_2\otimes S}^+\left(\begin{bmatrix} A & O \\ O & B \end{bmatrix}\right)$$

(3.2) follows. □

These two statements lead to an analogue of (2.2) in Theorem 2.2:

$$\begin{aligned}&\text{Conv.hull}(W_S^+(A) \cup W_S^+(e^{2i\theta}\bar{A}))\\ &\quad = W_{I_2\otimes S}^+\left(\frac{1}{2}\begin{bmatrix} A+e^{2i\theta}\bar{A} & -i(A-e^{2i\theta}\bar{A}) \\ i(A-e^{2i\theta}\bar{A}) & A+e^{2i\theta}\bar{A} \end{bmatrix}\right).\end{aligned} \tag{3.3}$$

Denoting $A = M + iN$ for $\theta = 0$, by (3.3) we obtain

$$\text{Conv.hull}(W_S^+(A) \cup W_S^+(\bar{A})) = W_{I_2\otimes S}^+\left(\begin{bmatrix} M & N \\ -N & M \end{bmatrix}\right), \tag{3.4}$$

and by (3.1)

$$W_{I_2\otimes S}^+\left(\begin{bmatrix} A & O \\ O & \bar{A} \end{bmatrix}\right) = W_{I_2\otimes S}^+\left(\begin{bmatrix} M & N \\ -N & M \end{bmatrix}\right).$$

Similar results of (3.3) and (3.4) are obtained for $V_S^+(A)$.

4. Reduction of NR of matrix polynomials. Transferring the idea of Theorem 2.1 for the numerical range of matrix polynomials $L(\lambda)$ we state the next proposition.

THEOREM 4.1. *For $k \le n$,*

$$NR[L(\lambda)] = \bigcup_{\xi_1,\ldots,\xi_k} NR[M_\xi(\lambda)], \tag{4.1}$$

where the $k \times k$ matrix polynomial

$$M_\xi(\lambda) = G^*(E \otimes L(\lambda))G,$$

with

$$E = \begin{bmatrix} 1 & \dots & 1 \\ \vdots & & \vdots \\ 1 & \dots & 1 \end{bmatrix}_{k\times k}, \qquad G = \operatorname{diag}(\xi_1, \dots, \xi_k)_{nk\times k},$$

and $\xi_1, \dots, \xi_k$ *are orthonormal vectors of* $\mathbf{C}^n$. *In* (4.1) *the union runs over all sets by* k *orthonormal vectors of* $\mathbf{C}^n$.

Proof. Let the vectors $\xi_1, \dots, \xi_k \in \mathbf{C}^n$ be an orthonormal basis of a subspace $\mathbf{E} \subset \mathbf{C}^n$ and let $x \in \mathbf{E}$. Then

$$x = [\xi_1 \dots \xi_k]y, \quad y \in \mathbf{C}^k$$

and

$$\begin{aligned} x^*L(\lambda)x &= y^*[\xi_1 \dots \xi_k]^*L(\lambda)[\xi_1 \dots \xi_k]y \\ &= y^* \begin{bmatrix} \xi_1^*L(\lambda)\xi_1 & \dots & \xi_1^*L(\lambda)\xi_k \\ \vdots & & \vdots \\ \xi_k^*L(\lambda)\xi_1 & \dots & \xi_k^*L(\lambda)\xi_k \end{bmatrix} y = y^*M_\xi(\lambda)y. \end{aligned}$$

Hence, running over all sets by k orthonormal vectors, (4.1) results. ∎

Obviously (4.1) for $k = 2$ is simple. Denoting $L(\lambda)$ as

$$\overline{L}(\lambda) = I\lambda^q + \bar{A}_1\lambda^{q-1} + \dots + \bar{A}_q,$$

a nearly similar relation to (2.2) arises.

THEOREM 4.2.

$$\begin{aligned} &NR\left(\begin{bmatrix} L(\lambda) & O \\ O & e^{2i\theta}\overline{L}(\lambda) \end{bmatrix}\right) \\ &= NR\left(\frac{1}{2}\begin{bmatrix} L(\lambda) + e^{2i\theta}\overline{L}(\lambda) & -i(L(\lambda) - e^{2i\theta}\overline{L}(\lambda)) \\ i(L(\lambda) - e^{2i\theta}\overline{L}(\lambda)) & L(\lambda) + e^{2i\theta}\overline{L}(\lambda) \end{bmatrix}\right). \end{aligned} \tag{4.2}$$

Proof. Let the vector $z \in \mathbf{C}^{2n}$. Then, as in Theorem 2.2,

$$z^*\begin{bmatrix} L(\lambda) & O \\ O & e^{2i\theta}\overline{L}(\lambda) \end{bmatrix} z = \frac{1}{2}w^*\begin{bmatrix} L(\lambda) + e^{2i\theta}\overline{L}(\lambda) & -i(L(\lambda) - e^{2i\theta}\overline{L}(\lambda)) \\ i(L(\lambda) - e^{2i\theta}\overline{L}(\lambda)) & L(\lambda) + e^{2i\theta}\overline{L}(\lambda) \end{bmatrix} w,$$

where

$$w = \frac{1}{\sqrt{2}}\begin{bmatrix} I & iI \\ iI & I \end{bmatrix} z.$$

Note that $\|w\| = \|z\|$. ∎

Writing $L(\lambda) = L_1(\lambda) + iL_2(\lambda)$, where the coefficients of $L_1(\lambda)$ and $L_2(\lambda)$ are real matrices, by (4.2) and for $\theta = 0$, we have the following corollary.

COROLLARY 4.3.

$$NR\left(\begin{bmatrix} L(\lambda) & O \\ O & \overline{L}(\lambda) \end{bmatrix}\right) = NR\left(\begin{bmatrix} L_1(\lambda) & L_2(\lambda) \\ -L_2(\lambda) & L_1(\lambda) \end{bmatrix}\right).$$

REFERENCES

[AL] T. Ando and C. K. Li, *The numerical range and numerical radii*, Linear and Multilinear Algebra, 37 (1994), pp. 1–3.

[B] T. Bayasgalan, *The numerical range of linear operators in spaces with an indefinite metric*, Acta Math. Hungar., 57 (1991), pp. 7–9.

[GLR] I. Gohberg, P. Lancaster, and L. Rodman, *Matrices and Indefinite Scalar Products*, Oper. Theory Adv. Appl. 8, Birkhauser, Basel, 1983.

[HJ] R. Horn and C. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, London. 1991.

[LTU] C.-K. Li, N.-K. Tsing, and F. Uhlig, *Numerical ranges of an indefinite inner product space*, Electron. J. Linear Algebra, 1 (1996), pp. 1–17.

[LMZ] P. Lancaster, J. Maroulas, and P. Zizler, *The Numerical Range of Selfadjoint Matrix Polynomials*, Oper. Theory Adv. Appl. 106, Birkhauser, Basel, 1998, pp. 291–304.

[LR] C.-K. Li and L. Rodman, *Numerical range of matrix polynomials*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1256–1265.

[M] A. Markus, *Introduction to Spectral Theory of Polynomial Operator Pencils*, American Mathematical Society, Providence, RI, 1988.

[MMP] A. Markus, J. Maroulas, and P. Psarrakos, *Spectral Properties of a Matrix Polynomial Connected with a Component of its Numerical Range*, Oper. Theory Adv. Appl. 106, Birkhauser, Basel, 1998, pp. 305–308.

[MP] M. Markus and C. Pesce, *Computer generated numerical ranges and some resulting theorems*, Linear and Multilinear Algebra, 20 (1987), pp. 121–157.

[MP1] J. Maroulas and P. Psarrakos, *The boundary of numerical range of matrix polynomials*, Linear Algebra Appl., 267 (1997), pp. 101–111.

[MP2] J. Maroulas and P. Psarrakos, *On the connectedness of numerical range of matrix polynomials*, Linear Algebra Appl., 280 (1998), pp. 97–108.

[MP3] J. Maroulas and P. Psarrakos, *A connection between numerical ranges of selfadjoint matrix polynomials*, Linear and Multilinear Algebra, 44 (1998), pp. 327–340.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 1, pp. 245–278
© 1999 Society for Industrial and Applied Mathematics

MINIMAL DEGREE COPRIME FACTORIZATION OF RATIONAL MATRICES*

CRISTIAN OARĂ† AND ANDRAS VARGA‡

Abstract. Given a rational matrix G with complex coefficients and a domain Γ in the closed complex plane, both arbitrary, we develop a complete theory of coprime factorizations of G over Γ, with denominators of McMillan degree as small as possible. The main tool is a general pole displacement theorem which gives conditions for an invertible rational matrix to dislocate by multiplication a part of the poles of G. We apply this result to obtain the parametrized class of all coprime factorizations over Γ with denominators of minimal McMillan degree n_b—the number of poles of G outside Γ. Specific choices of the parameters and of Γ allow us to determine coprime factorizations, as for instance, with polynomial, proper, or stable factors. Further, we consider the case in which the denominator has a certain symmetry, namely it is J all-pass with respect either to the imaginary axis or to the unit circle. We give necessary and sufficient solvability conditions for the problem of coprime factorization with J all-pass denominator of McMillan degree n_b and, when a solution exists, we give a construction of the class of coprime factors. When no such solution exists, we discuss the existence of, and give solutions to, coprime factorizations with J all-pass denominators of minimal McMillan degree ($> n_b$). All the developments are carried out in terms of descriptor realizations associated with rational matrices, leading to explicit and computationally efficient formulas.

Key words. rational matrices, coprime factorizations, J all-pass, descriptor realizations, numerical algorithms

AMS subject classifications. 26C15, 65F30, 93B40, 93C35, 93C45, 93C55, 93C60

PII. S0895479898339979

1. Introduction.

Let G be an arbitrary rational matrix (possible improper) and Γ be a given domain of the closed complex plane. A *left coprime factorization* (LCF) over Γ of G is a fractional representation of the form $G = M^{-1}N$, with N and M rational matrices having poles only in Γ and satisfying $MU + NV = I$ for certain rational matrices U and V with all poles in Γ. Analogously, a factorization in the form $G = NM^{-1}$, with N and M having poles only in Γ and satisfying $UM + VN = I$ for certain rational matrices U and V with poles in Γ, is called a *right coprime factorization* over Γ.

In this paper we develop a complete theory of LCFs over Γ with denominators M of smallest possible McMillan degree (see the precise definition below). For brevity, we call such factorizations of *minimal degree*, and define the degree of the factorization to be the McMillan degree of the denominator M. We address only LCFs since all results for right coprime factorizations follow by duality.

Apparently, the theory of minimal degree coprime factorizations has not been previously considered as such in the literature, although it is crucial to solving various problems encountered in the theory of linear systems and networks [29, 32], to canonical and noncanonical spectral factorizations of unstable rational matrices [6, 11, 4, 5], to several conjugation-based approaches to nonstandard H_∞ control problems [30, 31],

*Received by the editors June 3, 1998; accepted for publication (in revised form) November 18, 1998; published electronically September 22, 1999.

http://www.siam.org/journals/simax/21-1/33997.html

†Faculty of Automatic Control and Computers, University Polytechnica Bucharest, str. Austrului 34, RO-73115, Bucharest, Romania (oara@popov.riccati.pub.ro). The research of this author was supported by a Fellowship of the Alexander von Humboldt Foundation, Bonn, Germany.

‡German Aerospace Center, DLR-Oberpfaffenhofen, Institute of Robotics and System Dynamics, D-82234 Wessling, Germany (Andras.Varga@dlr.de).

and it brings important numerical advantages as for instance in computing coprime factorizations with denominators which are polynomial, proper, stable, all-pass, or J lossless. The only noticeable exception is [2], where coprime factorizations are addressed with another requirement of minimality, namely, the sum of the McMillan degrees of the denominator and nominator to be as small as possible. However, even in the case treated in [2] in which the matrix to be factored is square and of full rank over rationals this sum minimality leads to combinatorial problems over an infinite set which have little interest from an algorithmic viewpoint as well as for the type of applications that we have mentioned above.

Our main tool in the derivations, which is also of independent interest, is a general pole displacement theorem which gives a characterization in terms of realizations for an invertible rational matrix M to cancel in the product MG all poles of G in Γ. An important feature of our result is the possibility of coping with domains containing infinity, without using conformal mapping techniques. In particular, we show that the minimal degree of an LCF over Γ is n_b, where n_b is the number of poles of G outside Γ, and give a description in terms of realizations of all factors solving the minimal degree LCF over Γ. It turns out that the basic ingredient in computing LCFs of minimal degree lies in the solution of a generalized eigenvalue assignment problem of order n_b.

We develop also a theory of minimal degree LCFs with the additional requirement that the denominator has a certain symmetry. We consider here two cases, in which the denominator is J all-pass, either with respect to the imaginary axis or to the unit circle, and the domain Γ is adequately chosen to reflect the respective symmetry. We study first LCFs with J all-pass denominator of degree n_b, which we call the *canonical case*, and give necessary and sufficient solvability conditions. When solutions exist we give a parameterized description of the factors in terms of associated realizations. When such canonical solutions do not exist, we discuss the solution of *noncanonical* LCFs with J all-pass denominators of McMillan degree as small as possible ($> n_b$). The computation of factors in the canonical case relies on solving a generalized Lyapunov equation of order n_b. In the noncanonical case, the rank r of the solution of the same Lyapunov equation allows us to determine the minimal degree of the factorization as $2n_b - r$. The noncanonical factors can then be constructed by embedding the solution of this Lyapunov equation of order n_b into a $(2n_b - r)$-order Lyapunov equation with nonsingular solution. The noncanonical case treated here has important connections with the noncanonical Wiener–Hopf and J spectral factorizations as discussed in [6] for proper and invertible rational matrices. Another connected work is [31], which in contrast to our results is restricted to the canonical case with J lossless denominator with respect to the imaginary axis, and involves the solution of a Riccati equation of order equal to the McMillan degree of G.

For an extensive treatment of related topics in factorization theory of rational matrices we refer to the excellent collections of papers [4, 6, 11] for the regular case, and to [19, 20, 3] for the nonregular and nonsquare cases.

The paper is organized as follows. The main definitions and notation used throughout the paper are presented in a preliminary section together with a brief discussion on various descriptor realizations of rational matrices. In section 3 we obtain formulas for an invertible factor that dislocates by left multiplication certain prescribed poles of a given rational matrix. In section 4 we give a parameterization of the class of factors that solve the minimal degree LCF over Γ. Section 5 is devoted to the theory of minimal degree LCFs with J all-pass denominators with respect to the

imaginary axis. In section 6 we develop the analogous results in the case in which the denominator is J all-pass with respect to the unit circle. Several numerical examples are presented in section 7 together with a discussion of appropriate computational methods. Some conclusions summarize the main aspects of the proposed theory and provide hints to possible applications and extensions. We defer to an appendix the characterization in terms of realizations of an arbitrary J all-pass rational matrix.

2. Preliminaries.

2.1. Basic notation. We start with some notation and definitions. For a matrix A we denote by A^T and A^* its transpose and its conjugate transpose, respectively, and, if A is invertible, by A^{-T} and A^{-*} its transposed inverse and its conjugated transposed inverse, respectively.

By $\mathbb{C}$, $\mathbb{C}^-$, $\mathbb{C}^+$, $\mathbb{C}^0$, and $\mathbb{R}$ we denote the complex plane, the open left half plane, the open right half plane, the imaginary axis, and the real axis, respectively, and let $\overline{\mathbb{C}} := \mathbb{C} \cup \{\infty\}$ be the closed complex plane, and $\overline{\mathbb{C}}^- := \mathbb{C}^- \cup \mathbb{C}^0 \cup \{\infty\}$, $\overline{\mathbb{C}}^+ := \mathbb{C}^+ \cup \mathbb{C}^0 \cup \{\infty\}$, $\overline{\mathbb{C}}^0 := \mathbb{C}^0 \cup \{\infty\}$. Here the overbar denotes closure. By $\mathbb{D}$ we denote the open unit disk, and $\mathbb{D}^c = \overline{\mathbb{C}} \setminus \overline{\mathbb{D}}$ stands for the exterior of the closed unit disk containing the infinity.

Consider the disjoint partition of the closed complex plane into a "good" region Γ_g and a "bad" region Γ_b as

$$\overline{\mathbb{C}} = \Gamma_g \cup \Gamma_b. \tag{2.1}$$

A frequent interpretation of Γ_g in system theory is related to the standard stability concept, that is, for linear continuous-time systems $\Gamma_g = \mathbb{C}^-$, while for linear discrete-time systems $\Gamma_g = \mathbb{D}$ (or, sometimes, their closures). However, we use also other interpretations, as for instance $\Gamma_g = \{\infty\}$ to obtain coprime factorizations with polynomial factors. When dealing with LCFs of rational matrices with real coefficients, we assume that Γ_g is symmetric with respect to the imaginary axis, that is, if $\lambda \in \Gamma_g$ then $\overline{\lambda} \in \Gamma_g$. This is in particular true for all the aforementioned examples of Γ_g and guarantees that the resulting factors of the LCF have real coefficients as well.

2.2. Rational matrices. Throughout the paper we consider matrices with coefficients in the field $\mathbb{F}$, where $\mathbb{F}$ denotes either $\mathbb{R}$ or $\mathbb{C}$. We view a $p \times m$ rational matrix $G(\lambda)$ as a matrix whose entries are rational functions with coefficients in $\mathbb{F}$ and of the complex variable λ. Alternatively, we shall assimilate the rational matrix $G(\lambda)$ with the *transfer-function matrix* (TFM) of a linear time-invariant continuous-time or discrete-time descriptor system, and λ is either s or z, the complex variables appearing in the Laplace- or Z-transform, respectively, according to the type of system. We denote by $G^\sim$ the *adjoint* of G, where $G^\sim(s) = G^*(-\overline{s})$ in continuous time and $G^\sim(z) = G^*(1/\overline{z})$ in discrete time. In particular, if G has real coefficients we have $G^\sim(s) = G^T(-s)$ and $G^\sim(z) = G^T(1/z)$, respectively. Let J be a *signature* matrix, i.e., a matrix satisfying $J = J^* = J^{-1}$. We say that the rational matrix G is J *all-pass with respect to the imaginary axis* if $G^\sim JG = J$, where $G^\sim$ denotes the adjoint in continuous time. Accordingly, we say that G is J *all-pass with respect to the unit circle* if $G^\sim JG = J$, where $G^\sim$ denotes now the adjoint in discrete time.

2.3. Structural elements of rational matrices. For an arbitrary rational matrix we recall now several structural elements that play an important role in the sequel: normal rank, finite and infinite poles and zeros, and their orders. It is easy to realize that $G(\lambda)$ has constant rank for all but a finite number of points $\lambda \in \mathbb{C}$ for

which it has smaller rank. The *normal rank* of G is by definition the rank of $G(\lambda)$ for almost all $\lambda \in \mathbb{C}$. A rational matrix which is square (say $p \times p$) and has full normal rank p is called nonsingular; otherwise it is called singular. Clearly, a nonsingular rational matrix has an inverse which is a rational matrix as well.

In this paper, a particular focus is on the *poles* and *zeros* of a rational matrix, defined now. For an arbitrary $p \times m$ rational matrix G of normal rank r there always exists a $p \times p$ polynomial matrix $U(\lambda)$ with $\det U(\lambda) = 1$ and an $m \times m$ polynomial matrix $V(\lambda)$ with $\det V(\lambda) = 1$ that bring G to the so-called Smith–McMillan form

$$D(\lambda) = U(\lambda)G(\lambda)V(\lambda),$$

where

$$D(\lambda) = \operatorname{diag}\left\{\frac{\epsilon_1(\lambda)}{\eta_1(\lambda)}, \frac{\epsilon_2(\lambda)}{\eta_2(\lambda)}, \ldots, \frac{\epsilon_r(\lambda)}{\eta_r(\lambda)}, 0, \ldots, 0\right\},$$

and the polynomials $\epsilon_i(\lambda)$, $\eta_i(\lambda)$ are pairwise coprime for $i = 1, \ldots, r$ and satisfy the divisibility properties

$$\epsilon_i(\lambda) | \epsilon_{i+1}(\lambda), \qquad \eta_{i+1}(\lambda) | \eta_i(\lambda), \qquad i = 1, \ldots, r-1.$$

The polynomials $\epsilon_i(\lambda)$ and $\eta_i(\lambda)$ are uniquely defined by $G(\lambda)$ and are called *elementary divisors* of G. Define

$$z(\lambda) = \epsilon_1(\lambda) \cdots \epsilon_r(\lambda), \qquad p(\lambda) = \eta_1(\lambda) \cdots \eta_r(\lambda).$$

Then the n_z (n_p) roots of $z(\lambda)$ ($p(\lambda)$) are by definition *the finite zeros (poles)* of G and their *multiplicities* are by definition the multiplicities of the roots of $z(\lambda)$ ($p(\lambda)$). The *orders* of the zero (pole) λ_0 of G are by definition the nonzero multiplicities of λ_0 as a root of $\epsilon_i(\lambda)$ ($\eta_i(\lambda)$) for $i = 1, \ldots, r$. We say that ∞ is a zero (pole) of G if 0 is a zero (pole) of the rational matrix $\widehat{G}(\lambda) = G(\frac{1}{\lambda})$, and we define the orders of the (zero) pole $\lambda = \infty$ of $G(\lambda)$ as the orders of the pole $\lambda = 0$ of $\widehat{G}(\lambda)$.

By definition, the *McMillan degree* of G—denoted $\delta(G)$—is the sum of the orders of all its poles (finite and infinite). Once a partition (2.1) is fixed, we have $\delta(G) = n_b + n_g$, where n_g denotes the number of "good" poles in Γ_g and n_b denotes the number of "bad" poles in Γ_b (counting multiplicities).

A rational matrix G which is J all-pass with respect to the imaginary axis features a pole-zero symmetry with respect to the imaginary axis. Precisely, if $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ is the union of poles of G, then $\{-\overline{\lambda}_1, -\overline{\lambda}_2, \ldots, -\overline{\lambda}_n\}$ is the union of zeros of G. A similar property holds for a rational matrix G which is J all-pass with respect to the unit circle where now the pole–zero symmetry is with respect to the unit circle.

2.4. Descriptor realizations of rational matrices.

Standard descriptor realizations of rational matrices. It is well known (see, for example, [28, 27]) that any $p \times m$ rational matrix $G(\lambda)$ with coefficients in $\mathbb{F}$ (even improper or polynomial) has a descriptor realization of the form

$$G(\lambda) = D + C(\lambda E - A)^{-1}B =: \left[\begin{array}{c|c} A - \lambda E & B \\ \hline C & D \end{array}\right], \tag{2.2}$$

where $A, E \in \mathbb{F}^{n \times n}$, $B \in \mathbb{F}^{n \times m}$, $C \in \mathbb{F}^{p \times n}$, $D \in \mathbb{F}^{p \times m}$, and the so-called pole pencil $A - \lambda E$ is *regular*, i.e., it is square and $\det(A - \lambda E) \not\equiv 0$. The dimension n of the

square matrices A and E is called the *order of the realization* (2.2). We use $\Lambda(A-\lambda E)$ to denote the union of generalized eigenvalues of the regular pencil $A-\lambda E$ (finite and infinite, multiplicities counting). Occasionally, we shall also use the more compact notation $G=(A-\lambda E,B,C,D)$ to denote (2.2). The descriptor realization (2.2) of G is called *minimal* if its order is as small as possible among all realizations of this kind. Well-known criteria for minimality of a descriptor realization (2.2) are (see [27])

$$\operatorname{rank}\begin{bmatrix} A-\lambda E & B \end{bmatrix} = n \quad \forall \lambda \in \mathbb{C}, \tag{2.3a}$$

$$\operatorname{rank}\begin{bmatrix} E & B \end{bmatrix} = n, \tag{2.3b}$$

$$\operatorname{rank}\begin{bmatrix} A-\lambda E \\ C \end{bmatrix} = n \quad \forall \lambda \in \mathbb{C}, \tag{2.3c}$$

$$\operatorname{rank}\begin{bmatrix} E \\ C \end{bmatrix} = n, \tag{2.3d}$$

$$A\ker(E) \subseteq \operatorname{Im}(E). \tag{2.3e}$$

The conditions of minimality (2.3) are usually known as finite and infinite controllability, finite and infinite observability, and absence of nondynamic modes, respectively. For a minimal descriptor realization (2.2) of order n we have $\delta(G)=\operatorname{rank} E \le n$.

The principal inconvenience of realizations of the form (2.2) is that their minimal possible order is greater than the McMillan degree of G unless G is proper, and this brings important technical difficulties in factorization problems in which the McMillan degree plays a paramount role. A remedy to this is to use a generalization of (2.2) in which either the B or the C matrix is replaced by a matrix pencil, as explained in the next section.

Alternative descriptor realizations. Any rational $p\times m$ matrix G has a realization

$$G(\lambda) = D + C(\lambda E - A)^{-1}(B-\lambda F) =: \left[\begin{array}{c|c} A-\lambda E & B-\lambda F \\ \hline C & D \end{array}\right], \tag{2.4}$$

and for any fixed $\alpha,\beta\in\mathbb{F}$, not both zero, there exists a realization

$$G(\lambda) = D + C(\lambda E - A)^{-1}B(\alpha-\lambda\beta) =: \left[\begin{array}{c|c} A-\lambda E & B(\alpha-\lambda\beta) \\ \hline C & D \end{array}\right], \tag{2.5}$$

where $A,E\in\mathbb{F}^{n\times n}$, $B,F\in\mathbb{F}^{n\times m}$, $C\in\mathbb{F}^{p\times n}$, $D\in\mathbb{F}^{p\times m}$, and the pole pencil $A-\lambda E$ is regular. A realization (2.5) will be called centered at $\frac{\alpha}{\beta}$. (If $\beta=0$ we interpret $\frac{\alpha}{\beta}$ as ∞.) Occasionally, we shall use also the more compact notation $G=(A-\lambda E,B-\lambda F,C,D)$ and $G=(A-\lambda E,B(\alpha-\lambda\beta),C,D)$ to denote (2.4) and (2.5), respectively. Realizations of type (2.5) have been used for example in [12, 13, 14] for various problems related to factorizations of rational matrix functions.

All realizations (2.2), (2.4), and (2.5) are formally contained in the following general realization of a rational matrix due to Rosenbrock [21]:

$$G(\lambda) = W(\lambda) - V(\lambda)T^{-1}(\lambda)U(\lambda) =: \left[\begin{array}{c|c} T(\lambda) & U(\lambda) \\ \hline V(\lambda) & W(\lambda) \end{array}\right],$$

where $T(\lambda)$, $U(\lambda)$, $V(\lambda)$, and $W(\lambda)$ are matrix polynomials, with $T(\lambda)$ square and invertible. It is easy to see that

$$\begin{bmatrix} \widetilde{T}(\lambda) & \widetilde{U}(\lambda) \\ \widetilde{V}(\lambda) & \widetilde{W}(\lambda) \end{bmatrix} = \begin{bmatrix} Q & O \\ X & I \end{bmatrix} \begin{bmatrix} T(\lambda) & U(\lambda) \\ V(\lambda) & W(\lambda) \end{bmatrix} \begin{bmatrix} Z & Y \\ O & I \end{bmatrix}, \tag{2.6}$$

where Q and Z are constant invertible matrices and X and Y are arbitrary constant matrices, defines an *equivalent realization* in the form

$$G(\lambda) = \left[\begin{array}{c|c} \widetilde{T}(\lambda) & \widetilde{U}(\lambda) \\ \hline \widetilde{V}(\lambda) & \widetilde{W}(\lambda) \end{array}\right].$$

(See also [21] for a more general case.) This type of transformation performed on a realization is called *equivalence*, and if in addition $X = 0$ and $Y = 0$ it is called *state-space equivalence*.

We call realizations of the type (2.4) or (2.5) *minimal* if the dimension of the square matrices A and E (also called the order of the realization) is as small as possible among all realizations of the respective kind. It can be easily shown that any rational matrix $G(\lambda)$ has a minimal realization of type (2.4) of order equal to $\delta(G)$. For any fixed α and β, not both zero, and such that $\frac{\alpha}{\beta}$ is not a pole of G there also exists a minimal realization of type (2.5) of order equal to $\delta(G)$. The condition imposed on $\frac{\alpha}{\beta}$ is needed only for writing down minimal realizations (2.5) which have order equal to $\delta(G)$. More precisely, even if $\frac{\alpha}{\beta}$ is a pole of G we can still write a realization (2.5) but the minimal order will necessarily be greater than $\delta(G)$. This is exactly what is happening for realizations (2.2) which are obtained from (2.5) for $\alpha = 1$ and $\beta = 0$ and for which the minimal order is necessarily greater than $\delta(G)$, provided $\frac{\alpha}{\beta} = \infty$ is a pole of G. Notice that for (2.5) we can always freely choose α and β such as to ensure that $\frac{\alpha}{\beta}$ is not a pole of G. For the rest of the paper, if not otherwise stated, we assume this choice implicitly. The nice feature of (2.4) and (2.5) that their minimal order equals the McMillan degree of G recommends them for the kind of problems treated in this paper.

It is relatively straightforward to show that the criteria of minimality for a descriptor realization of type (2.4) are

$$\text{rank}\left[\begin{array}{cc} A-\lambda E & B-\lambda F \end{array}\right] = n \quad \forall \lambda \in \mathbb{C}, \tag{2.7a}$$

$$\text{rank}\left[\begin{array}{cc} E & F \end{array}\right] = n, \tag{2.7b}$$

$$\text{rank}\left[\begin{array}{c} A-\lambda E \\ C \end{array}\right] = n \quad \forall \lambda \in \mathbb{C}, \tag{2.7c}$$

$$\text{rank}\left[\begin{array}{c} E \\ C \end{array}\right] = n, \tag{2.7d}$$

while for realizations of type (2.5) similar conditions result by simply replacing (a) and (b) in (2.7) with

$$\text{rank}\left[\begin{array}{cc} A-\lambda E & B(\alpha-\lambda\beta) \end{array}\right] = n, \quad \forall \lambda \in \mathbb{C}, \tag{2.8a}$$

$$\text{rank}\left[\begin{array}{cc} E & B \end{array}\right] = n. \tag{2.8b}$$

Of particular importance in the case of (2.5) is that any two minimal realizations $G = (A-\lambda E, B(\alpha-\lambda\beta), C, D)$ and $G = (\widetilde{A}-\lambda\widetilde{E}, \widetilde{B}(\alpha-\lambda\beta), \widetilde{C}, \widetilde{D})$ are always related by an equivalence transformation as

$$\widetilde{E} = QEZ, \quad \widetilde{A} = QAZ, \quad \widetilde{B} = QB, \quad \widetilde{C} = CZ, \quad \widetilde{D} = D,$$

where Q and Z are unique invertible matrices. Moreover, for an invertible rational matrix $G(\lambda)$, it is possible to choose a realization (2.5) with D invertible, provided α

and β are such that $\frac{\alpha}{\beta}$ is neither a pole nor a zero of $G(\lambda)$. In this case we can use explicit formulas for the inverse G^{-1} in the form

$$G^{-1}(\lambda) = \left[\begin{array}{c|c} A - \lambda E - BD^{-1}C(\alpha - \lambda\beta) & BD^{-1}(\alpha - \lambda\beta) \\ \hline -D^{-1}C & D^{-1} \end{array}\right]. \tag{2.9}$$

An additional nice feature of realizations (2.5) is revealed in the appendix, where we show that an appropriate choice of α and β leads to nice formulas in case of rational matrices featuring certain symmetries.

Conversions between the representations (2.2), (2.4), and (2.5) are straightforward. Here we only show how to obtain from an nth-order realization (2.2) a realization of the form (2.4) of order equal to $\operatorname{rank} E$. We can always determine two invertible (even unitary) matrices Q and Z such that

$$G(\lambda) = \left[\begin{array}{c|c} Q(A - \lambda E)Z & QB \\ \hline CZ & D \end{array}\right] = \left[\begin{array}{cc|c} A_d - \lambda E_d & A_{dn} - \lambda E_{dn} & B_d \\ O & A_n & B_n \\ \hline C_d & C_n & D \end{array}\right], \tag{2.10}$$

where the invertible matrix $A_n \in \mathbb{R}^{\kappa\times\kappa}$ contains the so-called infinite *nondynamic* modes, and $\operatorname{rank}\left[\begin{array}{cc} E_d & E_{dn} \end{array}\right] = \operatorname{rank} E$. By using a general transformation as in (2.6) we get an equivalent realization of the type (2.4) in the form

$$G(\lambda) = \left[\begin{array}{c|c} A_d - \lambda E_d & B_d - A_{dn}A_n^{-1}B_n + \lambda E_{dn}A_n^{-1}B_n \\ \hline C_d & D - C_nA_n^{-1}B_n \end{array}\right], \tag{2.11}$$

which has order equal to $\operatorname{rank} E$. Provided (2.2) is minimal, the realization (2.11) is minimal as well.

All the above stated features of realizations (2.4) and (2.5), although not well known, follow by standard arguments from the realization theory [21, 27], and therefore we do not give detailed proofs here.

Structural elements in terms of realizations. The following result adapted from [21] and [28] will be instrumental in subsequent discussion.

THEOREM 2.1. *Let $G(\lambda)$ be an arbitrary rational matrix with a realization* (2.4) (*or* (2.5)).

Poles. *The poles of $G(\lambda)$ are among the generalized eigenvalues of the pole pencil $A - \lambda E$. If the realization* (2.4) *is minimal (or if $\frac{\alpha}{\beta}$ is not a pole of $G(\lambda)$, and* (2.5) *is minimal), then the poles of $G(\lambda)$ are exactly the generalized eigenvalues of $A - \lambda E$, and the orders of the poles are pairwise equal to the multiplicities of the generalized eigenvalues of $A - \lambda E$.*

Zeros. *The zeros of $G(\lambda)$ are among the zeros of the* system pencil *associated with* (2.4),

$$\left[\begin{array}{cc} A - \lambda E & B - \lambda F \\ C & D \end{array}\right], \tag{2.12}$$

or of the system pencil associated with (2.5),

$$\left[\begin{array}{cc} A - \lambda E & B(\alpha - \lambda\beta) \\ C & D \end{array}\right]. \tag{2.13}$$

For a careful discussion of the delicate notions of pole and zero at infinity of a rational matrix and connection with the realization theory we refer to [28].

Separated realizations. As a methodological question, throughout the paper we assume that the rational matrix G is given by a minimal descriptor realization (2.2), as this type of realization is most frequently used in the literature to represent arbitrary (possibly improper) rational matrices. The solutions M and N to the LCF $G = M^{-1}N$ will be given directly by minimal realizations of the type (2.4) or (2.5) of order equal to their respective McMillan degree. Furthermore, once a partition (2.1) is fixed, we assume that G is given by a *separated realization* with respect to (2.1), namely,

$$G(\lambda) = \left[\begin{array}{cc|c} A_b - \lambda E_b & A_{bg} - \lambda E_{bg} & B_b \\ O & A_g - \lambda E_g & B_g \\ \hline C_b & C_g & D \end{array}\right], \tag{2.14}$$

where the $n_b \times n_b$ pencil $A_b - \lambda E_b$ contains the n_b poles of $G(\lambda)$ in Γ_b and

$$\operatorname{rank} \left[\begin{array}{cc} E_b & E_{bg} \end{array}\right] = n_b; \tag{2.15}$$

that is, all infinite nondynamic modes are included in $A_g - \lambda E_g$. Starting with an arbitrary minimal realization (2.2) it is always possible to arrive to a separated realization by first determining a realization as in (2.10) and further making a spectral decomposition of the pencil $A_d - \lambda E_d$ with respect to the partition (2.1). Furthermore, if the realization to start with has real coefficients, and Γ_g is symmetric, we can always determine a separated realization with real coefficients as well.

3. Basic pole displacement result. In this section we solve the general *pole displacement problem* (PDP) formulated below.

PDP: Given a $p \times m$ rational matrix G and a disjoint partition $\overline{\mathbb{C}} = \Gamma_g \cup \Gamma_b$, both arbitrary, determine an invertible $p \times p$ rational matrix M with all poles in Γ_g which cancels by left multiplication all poles of G located in Γ_b, i.e., such that $N(\lambda) := M(\lambda)G(\lambda)$ has all poles in Γ_g.

The following theorem gives an answer to the PDP in terms of descriptor realizations of G and M. It is a rich extension of Theorem 3.1 and Lemma 4.1 in [24] and can be viewed also as an extension to the nonsquare case of the results in [10].

THEOREM 3.1. *Given a rational matrix $G(\lambda)$ and a disjoint partition $\overline{\mathbb{C}} = \Gamma_g \cup \Gamma_b$, both arbitrary, let n_b be the number of poles of G in Γ_b. Assume that* (2.14) *is a minimal descriptor realization of G separated with respect to the given partition and satisfying the condition* (2.15). *Then the class of solutions to the PDP is given by*

$$M(\lambda) = \left[\begin{array}{c|c} A_x - \lambda E_x & B_x p(\lambda) \\ \hline C_x & D_x \end{array}\right], \tag{3.1}$$

where the realization (3.1) *is minimal and satisfies the following: $A_x - \lambda E_x$ has all eigenvalues in Γ_g; D_x is invertible; $p(\lambda) := \alpha - \lambda\beta$, where $\frac{\alpha}{\beta} \notin \Lambda(A_b - \lambda E_b)$; and*

$$\left[\begin{array}{cc} A_x - \lambda E_x & B_x p(\lambda) \\ C_x & D_x \end{array}\right] \left[\begin{array}{c} X \\ C_b \end{array}\right] = \left[\begin{array}{c} Y \\ O \end{array}\right] (A_b - \lambda E_b) \tag{3.2}$$

holds for certain injective matrices X and Y.

The minimal McMillan degree of a solution to the PDP is n_b.

Before proceeding with the proof, we make the following remark.

REMARK 3.2. *Using Theorem* 3.1, *it is easy to see that by an appropriate equivalence transformation, each solution* $M(\lambda)$ *of the PDP has a minimal realization* (3.1) *such that*

$$\left[\begin{array}{cc|c} A_{x,11}-\lambda E_{x,11} & A_{x,12}-\lambda E_{x,12} & B_{x,1}(\alpha-\beta\lambda) \\ A_{x,21}-\lambda E_{x,21} & A_{x,22}-\lambda E_{x,22} & B_{x,2}(\alpha-\beta\lambda) \\ \hline C_{x,1} & C_{x,2} & D_x \end{array}\right]\left[\begin{array}{c} I \\ O \\ C_b \end{array}\right]=\left[\begin{array}{c} I \\ O \\ O \end{array}\right](A_b-\lambda E_b), \tag{3.3}$$

where (3.3) *follows directly from* (3.2). *Then it can be rapidly checked that* $M^{-1}(\lambda)$ *has a realization of the form*

$$M^{-1}(\lambda)=\left[\begin{array}{cc|c} A_b-\lambda E_b & * & * \\ O & * & * \\ \hline C_b & * & * \end{array}\right]; \tag{3.4}$$

that is, $M^{-1}(\lambda)$ *and* $G(\lambda)$ *have a "common" pair* $(C_b, A_b-\lambda E_b)$. *This fact shows the connections of our result with the "spectral triples" discussed in* [10] *for an invertible and proper* $G(\lambda)$ *and will be instrumental in obtaining solutions to the noncanonical problems in sections* 5 *and* 6.

Proof. Before proceeding with the proof we make several preparations. Using conversion formulas as for getting from (2.10) to (2.11), we can put (2.14) in the form

$$G(\lambda)=\left[\begin{array}{cc|c} A_b-\lambda E_b & \widetilde{A}_{bg}-\lambda\widetilde{E}_{bg} & \widetilde{B}_b-\lambda\widetilde{F}_b \\ O & \widetilde{A}_g-\lambda\widetilde{E}_g & \widetilde{B}_g-\lambda\widetilde{F}_g \\ \hline C_b & \widetilde{C}_g & \widetilde{D} \end{array}\right], \tag{3.5}$$

where

$$\operatorname{rank}\left[\begin{array}{ccc} E_b & \widetilde{E}_{bg} & \widetilde{F}_b \end{array}\right]=n_b, \tag{3.6}$$

$$\Lambda(A_b-\lambda E_b)\cap\Lambda(\widetilde{A}_g-\lambda\widetilde{E}_g)=\emptyset. \tag{3.7}$$

This conversion is necessary so as to enforce (3.7) and can be skipped if $\infty\in\Gamma_g$. Using (3.5) and (3.1) we can write a realization of $N(\lambda)=M(\lambda)G(\lambda)$ as

$$N(\lambda)=\left[\begin{array}{ccc|c} A_x-\lambda E_x & B_xC_bp(\lambda) & B_x\widetilde{C}_gp(\lambda) & B_x\widetilde{D}p(\lambda) \\ O & A_b-\lambda Ethat_b & \widetilde{A}_{bg}-\lambda\widetilde{E}_{bg} & \widetilde{B}_b-\lambda\widetilde{F}_b \\ O & O & \widetilde{A}_g-\lambda\widetilde{E}_g & \widetilde{B}_g-\lambda\widetilde{F}_g \\ \hline C_x & D_xC_b & D_x\widetilde{C}_g & D_x\widetilde{D} \end{array}\right]. \tag{3.8}$$

Assume that X and Y are two matrices that satisfy the generalized Sylvester equation

$$(A_x-\lambda E_x)X-Y(A_b-\lambda E_b)+B_xC_b(\alpha-\beta\lambda)=0, \tag{3.9}$$

and let $\widetilde{X}$ and $\widetilde{Y}$ be the unique solutions that satisfy the generalized Sylvester equation

$$(A_b-\lambda E_b)\widetilde{X}-\widetilde{Y}(\widetilde{A}_g-\lambda\widetilde{E}_g)+\widetilde{A}_{bg}-\lambda\widetilde{E}_{bg}=0. \tag{3.10}$$

Note that $\widetilde{X}$ and $\widetilde{Y}$ always exist due to (3.7). Define

$$Q:=\left[\begin{array}{ccc} I & -Y & O \\ O & I & -\widetilde{Y} \\ O & O & I \end{array}\right],\quad Z:=\left[\begin{array}{ccc} I & X & O \\ O & I & \widetilde{X} \\ O & O & I \end{array}\right]. \tag{3.11}$$

By using the transformation matrices (3.11) we obtain an equivalent realization in the form

$$N(\lambda) = \left[\begin{array}{ccc|c} A_x - \lambda E_x & O & * & * \\ O & A_b - \lambda E_b & O & \widehat{B}_b - \lambda \widehat{F}_b \\ O & O & \widetilde{A}_g - \lambda \widetilde{E}_g & \widetilde{B}_g - \lambda \widetilde{F}_g \\ \hline C_x & D_x C_b + C_x X & * & D_x \widetilde{D} \end{array}\right], \tag{3.12}$$

where from the minimality of $G(\lambda)$ and from (3.6) we have

$$\begin{aligned} \operatorname{rank} \left[\begin{array}{cc} A_b - \lambda E_b & \widehat{B}_b - \lambda \widehat{F}_b \end{array}\right] &= n_b \quad \forall \lambda \in \mathbb{C}, \\ \operatorname{rank} \left[\begin{array}{cc} E_b & \widehat{F}_b \end{array}\right] &= n_b. \end{aligned} \tag{3.13}$$

We show first that if $M(\lambda)$ is given by the minimal realization (3.1) satisfying the respective conditions in the statement, then it is a solution to the PDP. Notice first that $M(\lambda)$ is invertible. Since $\Lambda(A_x - \lambda E_x) \subset \Gamma_g$ we get that all poles of $M(\lambda)$ are in Γ_g. From (3.2) it will follow that $N(\lambda) = M(\lambda)G(\lambda)$ has all its poles in Γ_g. Indeed, (3.2) shows that (3.9) is satisfied and let $\widetilde{X}$ and $\widetilde{Y}$ be the unique solutions to (3.10). We perform an equivalence transformation with Q and Z defined by (3.11) to get to (3.12). From (3.2) we also get that the entry $D_x C_b + C_x X$ in (3.12) is zero. Thus, after removing the unobservable part we get

$$N(\lambda) := \left[\begin{array}{cc|c} A_x - \lambda E_x & * & * \\ O & \widetilde{A}_g - \lambda \widetilde{E}_g & \widetilde{B}_g - \lambda \widetilde{F}_g \\ \hline C_x & * & D_x \widetilde{D} \end{array}\right], \tag{3.14}$$

which clearly has *all* its poles in Γ_g. Thus $M(\lambda)$ has canceled by left multiplication all poles of $G(\lambda)$ located in Γ_b and thus it is a solution to the PDP.

Conversely, now let $M(\lambda)$ be a solution to the PDP and let (3.1) be a minimal realization of it, with D_x invertible. From minimality it follows that $\Lambda(A_x - \lambda E_x) \subset \Gamma_g$. Hence $\Lambda(A_x - \lambda E_x) \cap \Lambda(A_b - \lambda E_b) = \emptyset$ and thus there exist unique matrices X and Y to satisfy the generalized Sylvester equation (3.9). For similar reasons, the equation (3.10) has unique solutions $\widetilde{X}$ and $\widetilde{Y}$. Define Q and Z as in (3.11) and perform an equivalence transformation to get (3.12). Since $M(\lambda)$ is a solution to the PDP it follows that $N(\lambda)$ has no poles in Γ_b. However, (3.12) shows that this is possible only if

$$(D_x C_b + C_x X)(A_b - \lambda E_b)^{-1}(\widehat{B}_b - \lambda \widehat{F}_b) \tag{3.15}$$

has no poles in Γ_b. Since (3.13) holds, it follows by standard minimality arguments that (3.15) has no poles in Γ_b only if $D_x C_b + C_x X = 0$. Combining this with (3.9) we obtain

$$\left[\begin{array}{cc} A_x - \lambda E_x & B_x(\alpha - \beta\lambda) \\ C_x & D_x \end{array}\right] \left[\begin{array}{c} X \\ C_b \end{array}\right] = \left[\begin{array}{c} Y \\ O \end{array}\right] (A_b - \lambda E_b). \tag{3.16}$$

It only remains to show that X and Y are injective. Since $M(\lambda)$ is invertible, the leftmost pencil in (3.16) is regular. Thus, there exists $\lambda = \lambda_0$ for which the rightmost and leftmost pencils in (3.16) are invertible, from which it follows that

$$\operatorname{rank} \left[\begin{array}{c} X \\ C_b \end{array}\right] = \operatorname{rank} \left[\begin{array}{c} Y \\ O \end{array}\right]. \tag{3.17}$$

We show that X is injective and the injectiveness of Y follows then from (3.17). Let $\mathcal{V} = \mathrm{Ker}X$. Since D_x is invertible, we get from (3.16) that

$$C_b\mathcal{V} = 0. \tag{3.18}$$

From (3.16) we further get

$$\begin{bmatrix} Y \\ O \end{bmatrix} (A_b - \lambda E_b)\mathcal{V} = 0,$$

from which

$$A_b\mathcal{V} + E_b\mathcal{V} \subset \mathrm{Ker}Y. \tag{3.19}$$

But (3.17) and (3.18) show that $\dim \mathcal{V} = \dim \mathrm{Ker}\left[\begin{smallmatrix} X \\ C_b \end{smallmatrix}\right] = \dim \mathrm{Ker}\left[\begin{smallmatrix} Y \\ O \end{smallmatrix}\right] = \dim \mathrm{Ker}Y$. Thus

$$\dim(A_b\mathcal{V} + E_b\mathcal{V}) \leq \dim \mathrm{Ker}Y = \dim \mathcal{V}, \tag{3.20}$$

where only equality can take place, since, in general, for a regular pencil $A_b - \lambda E_b$ and an arbitrary space $\mathcal{V}$ we have $\dim(A_b\mathcal{V} + E_b\mathcal{V}) \geq \dim \mathcal{V}$. It follows that

$$\dim(A_b\mathcal{V} + E_b\mathcal{V}) = \dim \mathcal{V} \tag{3.21}$$

and thus $\mathcal{V}$ is a deflating subspace (see [23]) of the pencil $A_b - \lambda E_b$ contained in $\mathrm{Ker}\, C_b$. However, this is possible only if $\mathcal{V} = 0$ [22] because (3.5) is a minimal realization of $G(\lambda)$ and thus

$$\mathrm{rank}\begin{bmatrix} A_b - \lambda E_b \\ C_b \end{bmatrix} = n_b \quad \forall \lambda \in \mathbb{C}, \quad \mathrm{rank}\begin{bmatrix} E_b \\ C_b \end{bmatrix} = n_b.$$

It follows that X and Y are both injective, and thus any solution to the PDP satisfies the requirements in the statement.

Finally, from the injectiveness of X it follows that $\delta(M) \geq n_b$. ∎

4. Minimal degree coprime factorization. For a given rational matrix $G(\lambda)$ and a domain Γ_g, both arbitrary, we characterize now the class of LCFs over Γ_g having minimal degree which we show to be n_b. Recall that n_b is the number of poles of $G(\lambda)$ in Γ_b. For our proofs, we need the following result on the solution in a particular form to the generalized eigenvalue assignment problem.

LEMMA 4.1. *Let $A - \lambda E$ be a regular pencil, with A, $E \in \mathbb{F}^{n\times n}$, $B \in \mathbb{F}^{n\times m}$; let $\Gamma \subset \overline{\mathbb{C}}$ be a set of n elements (not necessarily distinct and assumed to be symmetric if $\mathbb{F} = \mathbb{R}$); and let $\alpha, \beta \in \mathbb{F}$, not both zero, such that $\frac{\alpha}{\beta} \notin \Lambda(A - \lambda E)$ and $\frac{\alpha}{\beta} \notin \Gamma$. Assume that*

$$\mathrm{rank}\begin{bmatrix} A - \lambda E & B \end{bmatrix} = n \quad \forall \lambda \in \mathbb{C}, \tag{4.1a}$$

$$\mathrm{rank}\begin{bmatrix} E & B \end{bmatrix} = n. \tag{4.1b}$$

Then there exists a matrix $F \in \mathbb{F}^{m\times n}$ such that

$$\Lambda(A - \lambda E + BF(\alpha - \beta\lambda)) = \Gamma. \tag{4.2}$$

Proof. To reduce (4.2) to a standard eigenvalue assignment problem we use the conformal mapping $\lambda = (\alpha z + \beta)/(\beta z - \alpha)$. Under this mapping Γ changes to Γ_z,

which will contain only finite points. Define $A_z - zE_z := \alpha A + \beta E - z(\beta A - \alpha E)$ and $B_z := (\alpha^2 + \beta^2)B$. With the transformed data we have reduced (4.2) to

$$\Lambda(A_z + B_z F - zE_z) = \Gamma_z, \tag{4.3}$$

with E_z invertible. From assumption (4.1a), for $z \neq \frac{\alpha}{\beta}$, and (4.1b), for $z = \frac{\alpha}{\beta}$, we get

$$\operatorname{rank}[\, A_z - zE_z \;\; B_z \,] = n, \quad \forall z \in \mathbb{C}.$$

Thus (4.3) is a standard eigenvalue assignment problem for the controllable pair $(E_z^{-1}A_z, E_z^{-1}B_z)$ and the set Γ_z, and has always a solution F which is the solution of the original problem (4.2) as well. If $\mathbb{F} = \mathbb{R}$, and Γ is symmetric, then Γ_z will be symmetric as well, and F can be chosen with elements in $\mathbb{R}$. ∎

THEOREM 4.2. *Given a rational matrix $G(\lambda)$ and a disjoint partition $\overline{\mathbb{C}} = \Gamma_g \cup \Gamma_b$, both arbitrary, let n_b be the number of poles of G in Γ_b. Assume that* (2.14) *is a minimal descriptor realization of G separated with respect to the given partition and satisfying condition* (2.15). *Then we have the following:*

1. *The minimal degree of a LCF of G over Γ_g is n_b.*

2. *The class of solutions to the minimal degree LCF problem $G(\lambda) = M^{-1}(\lambda)N(\lambda)$ is given by*

$$\begin{aligned} &N(\lambda) \\ &= \left[\begin{array}{cc|c} A_b - \lambda E_b + KC_b p(\lambda) & A_{bg} - \lambda E_{bg} + KC_g p(\lambda) & B_b + KDp(\lambda) \\ O & A_g - \lambda E_g & B_g \\ \hline WC_b & WC_g & WD \end{array}\right], \end{aligned} \tag{4.4}$$

$$M(\lambda) = \left[\begin{array}{c|c} A_b - \lambda E_b + KC_b p(\lambda) & Kp(\lambda) \\ \hline WC_b & W \end{array}\right], \tag{4.5}$$

where W is an arbitrary invertible matrix, K is any matrix that solves the generalized eigenvalue assignment problem $\Lambda(A_b - \lambda E_b + KC_b p(\lambda)) \subset \Gamma_g$, $p(\lambda) := \alpha - \lambda\beta$, and α and β are chosen to satisfy the requirements of Lemma 4.1.

If G has real coefficients and Γ_g is a symmetric set then the class of solutions with real coefficients to the minimal degree LCF problem is also given by (4.4) *and* (4.5), *where now all the intervening elements can be chosen real.*

Proof. We show first that $N(\lambda)$ and $M(\lambda)$ of the form (4.4)–(4.5) solve the minimal LCF problem with degree n_b. To show that $G(\lambda) = M^{-1}(\lambda)N(\lambda)$ is a matter of straightforward computations involving realizations and equivalence transformations. Further, we see from (4.5) that $\delta(M(\lambda)) \leq n_b$. But we have also from Theorem 3.1 that $\delta(M(\lambda)) \geq n_b$ since $M(\lambda)$ solves in particular the corresponding PDP formulated for $G(\lambda)$. Thus $\delta(M(\lambda)) = n_b$ which proves part 1 of the theorem. We show further that $M(\lambda)$ and $N(\lambda)$ are left coprime. As both $N(\lambda)$ and $M(\lambda)$ have poles only in Γ_g, it is enough to show that the compound rational matrix $[\, N(\lambda) \;\; M(\lambda) \,]$ has no zeros outside Γ_g. (See [21] and [29] for equivalent characterizations of left coprimeness.) Indeed, using the descriptor realization of $G = (A - \lambda E, B, C, D)$ separated as in (2.14), a realization of $[\, N(\lambda) \;\; M(\lambda) \,]$ is

$$[\, N(\lambda) \;\; M(\lambda) \,] = \left[\begin{array}{c|cc} A - \lambda E + \overline{K}Cp(\lambda) & B + \overline{K}Dp(\lambda) & \overline{K}p(\lambda) \\ \hline WC & WD & W \end{array}\right], \tag{4.6}$$

where $\overline{K} = \left[\begin{smallmatrix} K \\ O \end{smallmatrix}\right]$. Incidentally, it is easy to see that (4.6) is a minimal realization. From Theorem 2.1 we have that the zeros of $[\, N(\lambda)\ M(\lambda)\,]$ are among the zeros (including infinity) of the system pencil associated with the above realization and given by

$$\left[\begin{array}{ccc} A - \lambda E + \overline{K}Cp(\lambda) & B + \overline{K}Dp(\lambda) & \overline{K}p(\lambda) \\ WC & WD & W \end{array}\right], \tag{4.7}$$

which is strictly equivalent to (see [8] for the definition of strict equivalence for matrix pencils)

$$\left[\begin{array}{ccc} A - \lambda E & B & \overline{K}p(\lambda) \\ O & O & W \end{array}\right].$$

Thus the zeros of $[\, N(\lambda)\ M(\lambda)\,]$ are among those of $[\, A - \lambda E\ B\,]$. However, this last pencil has no zeros since the representation of $G(\lambda)$ is minimal, and thus $\operatorname{rank}[\, A - \lambda E\ B\,] = n\ \forall \lambda \in \mathbb{C}$, λ finite, and $\operatorname{rank}[\, E\ B\,] = n$. We conclude that $[\, N(\lambda)\ M(\lambda)\,]$ has no zeros as well and therefore $M(\lambda)$ and $N(\lambda)$ result in left coprime over Γ_g. Actually, $M(\lambda)$ and $N(\lambda)$ result in left coprime (over the whole $\overline{\mathbb{C}}$).

Conversely, we show now that if $M(\lambda)$ and $N(\lambda)$ solve the LCF over Γ_g with minimal degree n_b then they are of the form (4.4)–(4.5). Since $M(\lambda)$ and $N(\lambda)$ have all poles in Γ_g, it follows from Theorem 3.1 that there exists a minimal realization (3.1) of $M(\lambda)$ which satisfies (3.2). Since the McMillan degree of $M(\lambda)$ is precisely n_b, the injective matrices X and Y in (3.2) are actually square and thus nonsingular. Therefore, using Remark 3.2, we can simply take $X = Y = -I_{n_b}$. Then, the expression for $M(\lambda)$ follows immediately from (3.2), where $B_x := K$ solves the eigenvalue assignment problem $\Lambda(A_b - \lambda E_b + KC_b(\alpha - \lambda\beta)) \subset \Gamma_g$. Lemma 4.1 shows that this last problem always has a solution since from the minimality of the realization (2.14) we get that the pair $(A_b^* - \lambda E_b^*, C_b^*)$ satisfies (4.1a) written for actual data, and α and β can always be chosen a priori to satisfy the respective requirements. The expression (4.4) results immediately after a simple equivalence transformation and removing the unobservable part from the descriptor representation of the product $N(\lambda) = M(\lambda)G(\lambda)$.

For the case of real coefficients, we can start with a separated realization for G having real elements. Then we choose α and β real, and since Γ_g is symmetric it follows from Lemma 4.1 that the feedback matrix K can be chosen with real elements as well. □

Notice from the above theorem that the class of all solutions depends essentially only on the pair $(C_b, A_b - \lambda E_b)$ that is unique (up to an equivalence transformation) for a given $G(\lambda)$.

5. LCF with J all-pass denominator with respect to the imaginary axis. In this section we solve the minimal degree LCF with the additional requirement on the denominator to have a certain symmetry, namely we consider here the case in which the denominator is J all-pass with respect to the imaginary axis. To reflect this symmetry accordingly, we take throughout this section the disjoint partition $\overline{\mathbb{C}} = \Gamma_g \cup \Gamma_b$ defined by

$$\Gamma_b := \mathbb{C}^+ \text{ (or } \Gamma_b := \mathbb{C}^-) \qquad \Gamma_g := \overline{\mathbb{C}} \setminus \Gamma_b. \tag{5.1}$$

However, due to the additional requirement on the denominator to be J all-pass it is not always possible to solve the LCF over Γ_g with minimal degree n_b. When this is

possible, we call the factorization *canonical*; otherwise we call it *noncanonical*. Notice that for both of the choices in (5.1), Γ_b is an open set while Γ_g is a closed one. This will be instrumental for the solution in the noncanonical case.

The following result that will prove useful later gives a characterization of *proper* J all-pass rational matrices in terms of minimal realizations. It is a particular case (with $\alpha = 1, \beta = 0$) of the general theorem Theorem A.1 given in the appendix.

COROLLARY 5.1. *Let $M(s)$ be a proper square invertible rational matrix, having a minimal realization*

$$M(s) := \left[\begin{array}{c|c} A - sE & B \\ \hline C & D \end{array}\right].$$

Then M is J all-pass with respect to the imaginary axis if and only if $D^*JD = J$ and there exists an invertible Hermitian matrix X such that

$$\begin{aligned} A^*XE + E^*XA - C^*JC &= 0, \\ C - DJB^*XE &= 0. \end{aligned}$$

5.1. The canonical case. The following theorem gives a complete solution to the LCF problem with J all-pass denominator of McMillan degree n_b.

THEOREM 5.2. *Given an arbitrary rational matrix $G(s)$ and a disjoint partition $\overline{\mathbb{C}} = \Gamma_g \cup \Gamma_b$ defined by* (5.1)*, let n_b be the number of poles of G in Γ_b. Assume* (2.14) *is a minimal descriptor realization of G separated with respect to the given partition and satisfying condition* (2.15)*. Then the LCF with J all-pass denominator with respect to the imaginary axis has a solution $G(s) = M^{-1}(s)N(s)$ of minimal degree n_b if and only if the equation*

$$A_b^*XE_b + E_b^*XA_b - C_b^*JC_b = 0 \tag{5.2}$$

has an invertible Hermitian solution X. In this case, the class of all solutions is given by

$$N(s) = \left[\begin{array}{cc|c} A_b + KC_b - sE_b & A_{bg} + KC_g - sE_{bg} & B_b + KD \\ O & A_g - sE_g & B_g \\ \hline WC_b & WC_g & WD \end{array}\right], \tag{5.3}$$

$$M(s) = \left[\begin{array}{c|c} A_b + KC_b - sE_b & K \\ \hline WC_b & W \end{array}\right], \tag{5.4}$$

where

$$K = -X^{-1}E_b^{-*}C_b^*J, \tag{5.5}$$

and W is any J unitary matrix satisfying

$$W^*JW = J. \tag{5.6}$$

If G has real coefficients then the class of solutions with real coefficients is also given by (5.3) *and* (5.4)*, where now all the intervening coefficients can be chosen real.*

Proof. Notice first that E_b in a minimal realization (2.14) is invertible due to the choice of Γ_b in (5.1).

We show first that any solution $G = M^{-1}N$ to the LCF with J all-pass denominator of McMillan degree n_b is given by (5.3) and (5.4). In particular, M and N solve the minimal degree LCF problem (without any requirement on M) and must have minimal realizations of the form (4.4) and (4.5), respectively, where we can take $\alpha = 1$ and $\beta = 0$. Thus N and M are of the form (5.3) and (5.4), respectively. Since the realization (5.4) is minimal, and $M(s)$ is J all-pass, it follows from Corollary 5.1 that (5.2) has an invertible Hermitian solution and (5.5) and (5.6) are satisfied. Moreover, it follows that K given by (5.5) is such that $\Lambda(A_b + KC_b - sE_b) \subset \Gamma_g$.

Conversely, we have to show that N and M given by (5.3) and (5.4) are a solution to the LCF with J all-pass denominator of McMillan degree n_b. But this follows immediately from Theorem 4.2 and Corollary 5.1.

For the case of real coefficients, we can start with a separated realization for G having real elements, and it will follow from (5.2) that X is real and symmetric. □

If in the statement of the above theorem we add the condition $X > 0$ or $X < 0$ we obtain the solution to the LCF with J lossless or J expansive denominator, respectively (see [16]). Notice that for the chosen partition (5.1) equation (5.2) always has a unique solution, and thus the above theorem is an effective tool for checking the existence of, and computing the solutions to, the LCF with J all-pass denominator or, as a particular case, with J lossless denominator. We do not elaborate further in this direction.

For $J = I$, the above theorem provides the solution of an important particular case: LCF with all-pass denominator $M(s)$, (i.e., $M^*(-\overline{s})M(s) = I$). In this case, (5.2) has always a sign definite (positive or negative) solution. If we take $\Gamma_b = \mathbb{C}^+$, then we obtain the well known results for the LCF with inner denominator.

5.2. The noncanonical case. We study now LCFs with J all-pass denominator in the noncanonical case, i.e., when there is no solution with a J all-pass denominator of minimal McMillan degree n_b. This case is considerably more intricate than the canonical case. The idea is to introduce additional poles and zeros in $M(s)$ such that it could simultaneously be J all-pass and solve the LCF problem for $G(s)$. At the same time, we want to keep the McMillan degree of M as small as possible. It turns out that the additional poles/zeros can be taken only on the imaginary axis (including infinity) since the pole-zero symmetry featured by a J all-pass factor implies with necessity that all additional poles will be reflected into symmetric additional zeros that will be also zeros of the compound matrix $[N(s)\ M(s)]$.

Before stating our main result we construct a particular separated realization of G which facilitates the subsequent developments. Let $G(s)$ be an arbitrary rational matrix given by a minimal realization (2.14) separated with respect to $\Gamma_g \cup \Gamma_b$ and satisfying (2.15), and let n_b be the number of poles of G in Γ_b. Then the generalized Lyapunov equation

$$A_b^* X E_b + E_b^* X A_b - C_b^* J C_b = 0 \tag{5.7}$$

has a unique Hermitian solution X, and let $r := \operatorname{rank}(X)$. Let

$$U^* X U = \begin{bmatrix} \Sigma_r & O \\ O & O \end{bmatrix} \tag{5.8}$$

be a spectral decomposition of X where Σ_r is diagonal and nonsingular and U is unitary. Further, let

$$U^* E_b Q^* = \begin{bmatrix} E_{11} & O \\ E_{21} & E_{22} \end{bmatrix} \tag{5.9}$$

be an RQ decomposition of U^*E_b, where Q is unitary; E_{11} and E_{22} are lower triangular, square, and invertible; and the partition in (5.9) corresponds to the partition in (5.8). With (5.8) and (5.9) it is easy to see that

$$G(s) = \left[\begin{array}{cc|c} U^*(A_b-\lambda E_b)Q^* & U^*(A_{bg}-\lambda E_{bg}) & U^*B_b \\ O & A_g-\lambda E_g & B_g \\ \hline C_bQ^* & C_g & D \end{array}\right] \tag{5.10}$$

is another realization of G separated with respect to $\Gamma_g \cup \Gamma_b$ which we call *balanced with respect to the Lyapunov equation* (5.7). Thus we may assume from the beginning that for the separated realization (2.14) we have E_b lower triangular and the corresponding Lyapunov equation (5.7) has a diagonal solution. We are now ready for our main result.

THEOREM 5.3. *Given an arbitrary rational matrix $G(s)$ and a disjoint partition $\overline{\mathbb{C}} = \Gamma_g \cup \Gamma_b$ defined by* (5.1)*, let n_b be the number of poles of G in Γ_b. Assume that* (2.14) *is a minimal descriptor realization of G, separated with respect to the given partition and satisfying condition* (2.15)*. Let $r := \mathrm{rank}(X)$, where X is the unique Hermitian solution to the Lyapunov equation*

$$A_b^*XE_b + E_b^*XA_b - C_b^*JC_b = 0. \tag{5.11}$$

Then the minimal degree of the LCF with J all-pass denominator is $2n_b - r$.

Proof. Without restricting the generality, we may assume from the beginning that the realization (2.14) is balanced with respect to the Lyapunov equation (5.11). Further, as E_b is invertible we could, as a first simplifying step in the whole proof, reduce (5.11) to a standard Lyapunov equation with $E_b = I$. However, we prefer not to invert E_b as far as possible as this will bring benefits in terms of the reliability of the associated numerical algorithms.

The proof is quite lengthy and we divide it in several steps: we show first that any solution $G = M^{-1}N$ to the LCF with J all-pass denominator satisfies

$$\delta(M) \geq 2n_b - r \tag{5.12}$$

and further prove that M^{-1} has exactly n_b poles in Γ_b (which are the poles of G in Γ_b) and the rest (additional) $\delta(M) - n_b$ poles are with necessity on $\overline{\mathbb{C}}^0$. Finally, we construct a solution of minimal degree $2n_b - r$ with the additional $n_b - r$ poles placed arbitrary on $\overline{\mathbb{C}}^0$.

Proof of $\delta(M) \geq 2n_b - r$. Let $G(s) = M^{-1}(s)N(s)$ be a solution to the LCF with J all-pass denominator. As M cancels in the product $N = MG$ all poles of G in Γ_b and M^{-1} is J all-pass, we get with Remark 3.2 and Theorem A.1 in the appendix that there exists a minimal realization of M^{-1} of the form

$$M^{-1}(s) = \left[\begin{array}{c|c} A_M - sE_M & B_Mp(s) \\ \hline C_M & D_M \end{array}\right] = \left[\begin{array}{cc|c} A_b - sE_b & A_{bx} - sE_{bx} & B_{x1}p(s) \\ O & A_x - sE_x & B_{x2}p(s) \\ \hline C_b & C_x & D_M \end{array}\right], \tag{5.13}$$

where $p(s) := \alpha - s\beta$, and there exists a Hermitian invertible matrix X_M such that

$$A_M^*X_ME_M + E_M^*X_MA_M - C_M^*JC_M = 0. \tag{5.14}$$

According to the partition of the right–hand side of (5.13), X can be written in the partitioned form

$$X_M = \left[\begin{array}{cc} X & X_{12} \\ X_{12}^* & X_{22} \end{array}\right], \tag{5.15}$$

where it follows that X is the unique solution to (5.11). Since X_M is invertible and the rank of X is r, it follows from (5.15) that $\operatorname{rank}(X_M) \geq 2n_b - r$ and thus (5.12) holds.

Location of the poles of M^{-1}. Let $G = M^{-1}N$ be a solution to the LCF with J all-pass denominator. From (5.13) we see that the union of poles of $M^{-1}(s)$ contains $\Lambda(A_b - sE_b) \subset \mathbb{C}^+$ which are also the poles of G in Γ_b. From the pole-zero symmetry of a J all-pass rational matrix it follows that the poles of $M(s)$ are the conjugated of the poles of $M^{-1}(s)$. Since M has all poles in Γ_g we infer that the poles of M^{-1} are in $\overline{\Gamma}_b := \Gamma_b \cup \overline{\mathbb{C}}^0$, from which we have

$$\Lambda(A_x - sE_x) \subset \overline{\Gamma}_b. \tag{5.16}$$

We show now that

$$\Lambda(A_x - sE_x) \subset \Gamma_g, \tag{5.17}$$

from where we shall conclude with (5.16) that $\Lambda(A_x - sE_x) \subset \overline{\mathbb{C}}^0$ and thus M^{-1} has exactly n_b poles in Γ_b and the additional $\delta(M) - n_b$ poles are on $\overline{\mathbb{C}}^0$. We prove in fact that the coprimeness of N and M over Γ_g implies (5.17). To this end we compute realizations for M and N. Using (5.13) we write a realization of M in the form

$$\begin{aligned} M(s) &= \left[\begin{array}{c|c} A_M - sE_M - B_M W C_M(\alpha - s\beta) & -B_M W(\alpha - s\beta) \\ \hline WC_M & W \end{array}\right] \\ &= \left[\begin{array}{cc|c} A_b - sE_b - B_{x1}WC_b p(s) & A_{bx} - sE_{bx} - B_{x1}WC_x p(s) & -B_{x1}Wp(s) \\ -B_{x2}WC_b p(s) & A_x - sE_x - B_{x2}WC_x p(s) & -B_{x2}Wp(s) \\ \hline WC_b & WC_x & W \end{array}\right], \end{aligned} \tag{5.18}$$

where $W := D_M^{-1}$. With (5.18) and (2.14) we get successively a realization of

$$\begin{aligned} N(s) &= M(s)G(s) \\ &= \left[\begin{array}{ccc|c} A_b - sE_b - B_{x1}WC_b p(s) & A_{bx} - sE_{bx} - B_{x1}WC_x p(s) & -B_{x1}WC_b p(s) & -B_{x1}WC_g p(s) & -B_{x1}WDp(s) \\ -B_{x2}WC_b p(s) & A_x - sE_x - B_{x2}WC_x p(s) & -B_{x2}WC_b p(s) & -B_{x2}WC_g p(s) & -B_{x2}WDp(s) \\ O & O & A_b - sE_b & A_{bg} - sE_{bg} & B_b \\ O & O & O & A_g - sE_g & B_g \\ \hline WC_b & WC_x & WC_b & WC_g & WD \end{array}\right] \\ &= \left[\begin{array}{cc|ccc} A_b - sE_b - B_{x1}WC_b p(s) & A_{bx} - sE_{bx} - B_{x1}WC_x p(s) & A_{bg} - B_{x1}WC_g p(s) - sE_{bg} & B_b - B_{x1}WDp(s) \\ -B_{x2}WC_b p(s) & A_x - sE_x - B_{x2}C_x p(s) & -B_{x2}WC_g p(s) & -B_{x2}WDp(s) \\ O & O & A_g - sE_g & B_g \\ \hline WC_b & WC_x & WC_g & WD \end{array}\right], \end{aligned} \tag{5.19}$$

where we have removed the unobservable part with the help of a state-space equivalence transformation. From (5.19) we see that the poles of $N(s)$ are among $\Lambda(A_M - sE_M - B_M W C_M(\alpha - s\beta)) \cup \Lambda(A_g - sE_g)$ and thus they are indeed included in Γ_g. We use now (5.18) and (5.19) to write the system pencil associated with the compound matrix $[N(s)\ M(s)]$ and we get after some strict equivalence operations that the zeros of $[N(s)\ M(s)]$ are among the zeros of the pencil

$$\left[\begin{array}{ccccc} A_b - sE_b & A_{bg} - sE_{bg} & B_b & A_{bx} - sE_{bx} & -B_{x1}Wp(s) \\ O & A_g - sE_g & B_g & O & O \\ O & O & O & A_x - sE_x & -B_{x2}Wp(s) \\ O & O & O & O & W \end{array}\right]. \tag{5.20}$$

The realization (2.14) of G is minimal and thus controllable, from which it follows that the only zeros of $[N(s)\ M(s)]$ are given by $\Lambda(A_x - sE_x)$ which should be included in Γ_g since N and M are coprime over Γ_g. Thus we proved that the coprimeness of N and M over Γ_g implies that (5.17) holds.

Construction of a minimal solution. We show now that the LCF with J all-pass denominator has a solution $G = M^{-1}N$ such that $\delta(M) \leq 2n_b - r$, from which it will follow with (5.12) that the minimal degree $2n_b - r$ can be achieved. To this end, we shall construct $M(s)$, define $N(s) = M(s)G(s)$, and prove that they both satisfy the required properties. Of course, M should simultaneously be J all-pass, and M^{-1} should have the form (5.13), with the additional poles given by $\Lambda(A_x - sE_x) \subset \overline{\mathbb{C}}^0$.

For the sake of clarity we assume first that the additional poles/zeros $\Lambda(A_x - sE_x)$ are placed in finite locations on $\mathbb{C}^0$. The case with poles at infinity follows analogously, although the formulas become more intricate.

Placing additional poles/zeros in finite locations on $\mathbb{C}^0$. In this case we can take for M^{-1} a realization (5.13) with $\alpha = 1$, $\beta = 0$, i.e.,

$$M^{-1}(s) = \left[\begin{array}{c|c} A_M - sE_M & B_M \\ \hline C_M & D_M \end{array}\right] = \left[\begin{array}{cc|c} A_b - sE_b & A_{bx} - sE_{bx} & B_{x1} \\ O & A_x - sE_x & B_{x2} \\ \hline C_b & C_x & D_M \end{array}\right]. \tag{5.21}$$

The key idea of the construction is to embed the Lyapunov equation (5.11) into a larger Lyapunov equation (5.14) with an invertible solution X_M as in (5.15) and where the matrices A_x, E_x, A_{bx}, E_{bx}, B_{x1}, B_{x2}, C_x have to be determined. We can simply take $E_{bx} = 0$, $\quad E_x = I_{n_b - r}$ in case of placing additional poles/zeros at finite locations. We choose further $X_{12} = \left[\begin{smallmatrix} O \\ I_{n_b - r} \end{smallmatrix}\right]$, $X_{22} = O$ and get

$$\begin{aligned} X_M &= \left[\begin{array}{ccc} \Sigma_r & O & O \\ O & O & I_{n_b - r} \\ O & I_{n_b - r} & O \end{array}\right], \quad E_M = \left[\begin{array}{ccc} E_{11} & O & O \\ E_{21} & E_{22} & O \\ O & O & I_{n_b - r} \end{array}\right], \\ A_M &= \left[\begin{array}{ccc} A_{11} & A_{12} & A_{x1} \\ A_{21} & A_{22} & A_{x2} \\ O & O & A_x \end{array}\right], \quad C_M = \left[\begin{array}{ccc} C_1 & C_2 & C_x \end{array}\right], \end{aligned} \tag{5.22}$$

where we have taken into account that the realization (2.14) is balanced with respect to (5.11) and we have partitioned

$$A_b = \left[\begin{array}{cc} A_{11} & A_{12} \\ A_{21} & A_{22} \end{array}\right], \quad A_{bx} = \left[\begin{array}{c} A_{x1} \\ A_{x2} \end{array}\right], \quad C_b = \left[\begin{array}{cc} C_1 & C_2 \end{array}\right]$$

conformably with (5.8). The matrices A_{x1}, A_{x2}, A_x, C_x remain to be determined as to satisfy (5.14). We use (5.22) to write (5.14) componentwise as

$$A_{11}^*\Sigma_r E_{11} + E_{11}^*\Sigma_r A_{11} - C_1^* J C_1 = 0, \tag{5.23a}$$

$$E_{11}^*\Sigma_r A_{12} - C_1^* J C_2 = 0, \tag{5.23b}$$

$$C_2^* J C_2 = 0, \tag{5.23c}$$

$$A_{21}^* + E_{11}^*\Sigma_r A_{x1} + E_{21}^* A_x - C_1^* J C_x = 0, \tag{5.23d}$$

$$A_{22}^* + E_{22}^* A_x - C_2^* J C_x = 0, \tag{5.23e}$$

$$A_{x2}^* + A_{x2} - C_x^* J C_x = 0, \tag{5.23f}$$

where the rest of equations are transpose conjugated versions of the ones above. Equations (5.23a), (5.23b) and (5.23c) are automatically fulfilled due to (5.11). Before showing how the remaining equations can be satisfied, we prove that the pair $(C_2, A_{22} - sE_{22})$ is observable. Indeed, since the pair $(C_b, A_b - sE_b)$ is observable, the standard PHB observability tests shows that

$$\operatorname{rank}\begin{bmatrix} A_b - sE_b \\ C_b \end{bmatrix} = \operatorname{rank}\begin{bmatrix} A_{11} - sE_{11} & \Sigma_r^{-1}E_{11}^{-*}C_1^* J C_2 \\ A_{21} - sE_{22} & A_{22} - sE_{22} \\ C_1 & C_2 \end{bmatrix} = n_b \ \forall s \in \mathbb{C} \tag{5.24}$$

where we have replaced $A_{12} = \Sigma_r^{-1}E_{11}^{-*}C_1^* J C_2$ from (5.23b). From (5.24) we get

$$\operatorname{rank}\begin{bmatrix} A_{22} - sE_{22} \\ C_2 \end{bmatrix} = n_b - r, \qquad \forall s \in \mathbb{C}.$$

Thus $(C_2, A_{22} - sE_{22})$ is observable. It follows from (5.23e) that we can take the matrix C_x such that the poles of $A_x = E_{22}^{-*}(-A_{22}^* + C_2^* J C_x)$ are on the imaginary axis by simply solving an eigenvalue assignment problem for the controllable pair

$$(-E_{22}^{-*}A_{22}^*, E_{22}^{-*}C_2^* J). \tag{5.25}$$

Alternatively, we may compute C_x as to solve a generalized eigenvalue assignment problem for the controllable pair $(-A_{22}^* + sE_{22}^*, -C_2^* J)$ with the benefit of not inverting E_{22}^*. Finally, we choose A_{x1} as the unique solution of (5.23d) and A_{x2} as any solution to (5.23f). With all these choices it results that (5.14) is satisfied.

We take D_M to be any J unitary matrix and define

$$B_M = \begin{bmatrix} B_{x1} \\ B_{x2} \end{bmatrix} := E_M^{-*} X_M^{-1} C_M^* D_M^* J. \tag{5.26}$$

We get from Corollary 5.1 that M^{-1} given in (5.21) is J all-pass. It follows that M has all its poles in Γ_g, is J all-pass, and has McMillan degree less than or equal to $2n_b - r$, and a realization is given by

$$\begin{aligned} M(s) &= \left[\begin{array}{c|c} A_M - sE_M - B_M W C_M & -B_M W \\ \hline W C_M & W \end{array}\right] \\ &= \left[\begin{array}{cc|c} A_b - sE_b - B_{x1} W C_b p(s) & A_{bx} - B_{x1} W C_x p & -B_{x1} W \\ -B_{x2} W C_b & A_x - sI - B_{x2} W C_x & -B_{x2} W \\ \hline W C_b & W C_x & W \end{array}\right], \end{aligned} \tag{5.27}$$

where $W = D_M^{-1}$. We define $N(s) := M(s)G(s)$ and get

$$N(s) = \left[\begin{array}{ccc|c} A_b - sE_b - B_{x1}WC_b & A_{bx} - B_{x1}WC_x & A_{bg} - B_{x1}WC_g - sE_{bg} & B_b - B_{x1}WD \\ -B_{x2}WC_b & A_x - sI - B_{x2}C_x & -B_{x2}WC_g & -B_{x2}WD \\ O & O & A_g - sE_g & B_g \\ \hline WC_b & WC_x & WC_g & WD \end{array}\right]. \tag{5.28}$$

We can check successively that N has all poles in Γ_g, and that N and M are coprime over Γ_g. The details are skipped since the argument is identical as in the first part of the proof for showing that the additional poles of M^{-1} should be placed on $\overline{\mathbb{C}}^0$. Thus N and M define a solution to the LCF with J all-pass denominator, having degree less than or equal to $2n_b - r$. This together with (5.12) shows that the minimal degree $2n_b - r$ is attained and concludes the proof.

Placing additional poles/zeros arbitrary on $\overline{\mathbb{C}^0}$. We show that this case can be reduced to the case of finite poles/zeros. Indeed, we chose a realization for $M(s)$ of the form (5.13), where we take $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{C}^0$ such that $\frac{\alpha}{\beta}$ is neither an eigenvalue of $A_b - sE_b$ nor a desired location of additional poles/zeros. A characterization of J all-pass rational matrices with a realization centered at $\frac{\alpha}{\beta}$ is given in Theorem A.1 in the appendix. Thus we have again to embed a Lyapunov equation (5.11) into a larger Lyapunov equation (5.14) but where now E_M will be singular if we would like to place additional poles/zeros at infinity. To this end we transform the Lyapunov equation (5.14) into the equivalent Lyapunov equation

$$\widehat{A}_M^* X_M \widehat{E}_M + \widehat{E}_M^* X_M \widehat{A}_M - \widehat{C}_M^* J \widehat{C}_M = 0, \tag{5.29}$$

where the coefficients to be determined are

$$\widehat{A}_M := \alpha A_M + \beta E_M, \quad \widehat{E}_M := \alpha E_M - \beta A_M, \quad \widehat{C}_M = \sqrt{\alpha^2 + \beta^2} C_M, \tag{5.30}$$

with $\widehat{E}_M$ invertible. Thus we apply exactly the same procedure as for finite poles/zeros but now we determine the coefficients of the equation (5.29) instead of (5.14). Finally, the coefficients of (5.14) follow immediately from (5.30). The only precaution is with respect to the locations in which the additional poles/zeros have to be placed which are mapped by a bilinear transformations moving ∞ into the finite point $\frac{\alpha}{\beta} \in \mathbb{C}^0$. The rest of the argument is identical and is therefore not repeated. ∎

The proof of Theorem 5.3 also gives a procedure to compute solutions of minimal McMillan degree. Moreover, the result below shows that from (5.21) one can get the class of all solutions.

THEOREM 5.4. *Assume the same notation and hypotheses as for Theorem* 5.3. *The class of all solutions to the minimal degree LCF with J all-pass denominator $G = M^{-1}N$, with M proper, is given by* (5.27) *and* (5.28), *where C_x and A_x follow by solving the eigenvalue assignment problem for the pair* (5.25) *with $A_x = E_{22}^{-*}(-A_{22}^* + C_2^* J C_x) \subset \mathbb{C}^0$, A_{x2} is any solution to* (5.23f), *A_{x1} is the unique solution to* (5.23d), *W is any J unitary matrix, and B_{x1} and B_{x2} are given by* (5.26).

Proof. Most of the proof is already contained in the proof of Theorem 5.3. We have only to prove that any solution may be obtained by the procedure given in the statement. Let $G = M^{-1}N$ be a solution to the LCF with J all-pass denominator M of minimal degree $2n_b - r$, and proper. Let G be given by a separated realization

(2.14) balanced with respect to the Lyapunov equation (5.11). Then M^{-1} is given by a realization (5.21), and (5.14) is satisfied for an invertible Hermitian matrix X_M as in (5.15), where X has the form in the right-hand side of (5.8). We show that by equivalence state-space transformations on (5.21) we obtain a realization for which the corresponding Lyapunov equation (5.14) has all the intervening matrices of the form depicted in (5.22), and this will end the proof.

Writing X_M in partitioned form we get

$$X_M = \begin{bmatrix} \Sigma_r & O & X_1 \\ O & O & X_2 \\ X_1^* & X_2^* & X_{22} \end{bmatrix}$$

where we have denoted $X_{12} = \left[\begin{smallmatrix} X_1 \\ X_2 \end{smallmatrix}\right]$. Since the matrix X_M is square and invertible, and has dimension $2n_b - r$ we get that X_2 is invertible, and thus we can write

$$U_M^* X_M U_M = \begin{bmatrix} \Sigma_r & O & O \\ O & O & I_{n_b-r} \\ O & I_{n_b-r} & O \end{bmatrix}, U_M := \begin{bmatrix} I & O & -\Sigma_r X_1 X_2^{-1} \\ O & I & -\frac{1}{2} X_2^{-*}(X_{22} - X_1^* \Sigma_r^{-1} X_1) X_2^{-1} \\ O & O & X_2^{-1} \end{bmatrix}. \tag{5.31}$$

Further, let

$$\widetilde{E}_M := U_M^{-1} E_M = \begin{bmatrix} \widetilde{E}_b & \widetilde{E}_{bx} \\ O & \widetilde{E}_x \end{bmatrix},$$

where the block upper triangular structure in the right-hand side follows from the block upper triangular structure of both E_M and U_M^{-1}. With $\widehat{E}_M := I$, $\widehat{A}_M := U_M^{-1} A_M \widetilde{E}_M^{-1}$, $\widehat{C}_M := C_M \widetilde{E}_M^{-1}$ and $\widehat{X}_M := U_M^* X_M U_M$ we get that (5.14) can be written equivalently as

$$\widehat{A}_M^* \widehat{X}_M \widehat{E}_M + \widehat{E}_M^* \widehat{X}_M \widehat{A}_M - \widehat{C}_M^* J \widehat{C}_M = 0,$$

where the intervening matrices with "hat" have the same block structure as their counterparts in (5.22). It follows that the corresponding equations (5.23a)–(5.23f) are satisfied from where the conclusion is immediate. ∎

The class of all solutions, including the case of improper M, is obtained from (5.18) and (5.19), where the unknown coefficients are obtained by first using the transformation (5.30) to reduce the problem to the case of a proper M. We do not elaborate further in this direction.

REMARK 5.5. *From the proof of Theorem* 5.3 *we can give an answer to the general problem of existence of a LCF with J lossless denominator as follows. If the solution X to* (5.11) *is positive definite, then the problem is solvable; otherwise there is no solution of any degree. Indeed, assume by contrary that X is not positive definite and there is a solution to the LCF with J lossless denominator of a certain degree. Then* (5.14) *should be satisfied with* $X_M > 0$ *given in* (5.15)*, from which* $X > 0$ *which contradicts our assumption.*

6. LCF with *J* all-pass denominators with respect to the unit circle. In this section we give the discrete-time version of the results presented in section 6. More precisely, we solve the minimal degree LCF problem with the additional requirement that the denominator have another type of symmetry, namely, that it be

J all-pass with respect to the unit circle. Throughout this section we take the disjoint partition $\overline{\mathbb{C}} = \Gamma_g \cup \Gamma_b$ defined by

$$\Gamma_b := \mathbb{D}^c \text{ (or } \Gamma_b := \mathbb{D}), \qquad \Gamma_g := \overline{\mathbb{C}} \setminus \Gamma_b. \tag{6.1}$$

Similarly as for the symmetry discussed in the previous section, it is not always possible to solve the LCF over Γ_g with minimal degree n_b. Again, when this is possible, we call the factorization *canonical*, otherwise we call it *noncanonical*.

Since our theory encompasses the cases in which $M(z)$ has to cancel a pole of G at ∞ or at 0, it follows that $M(z)$ or $M^{-1}(z)$ could, in general, be improper. Therefore, to achieve the full generality we have to use for $M(z)$ a descriptor representation of type (2.5), with singular E. The following result gives a characterization of J all-pass rational matrices (possible improper) without poles at 1 in terms of associated realizations. It is a particular case for $\alpha = \beta = 1$ of the general theorem Theorem A.2 given in the appendix.

COROLLARY 6.1. *Let $M(z)$ be a square invertible rational matrix, without poles at 1, having a minimal realization*

$$M(z) := \left[\begin{array}{c|c} A - zE & (1-z)B \\ \hline C & D \end{array}\right].$$

Then M is J all-pass with respect to the unit circle if and only if $D^*JD = J$ and there exists an invertible Hermitian matrix X such that

$$\begin{aligned} E^*XE - A^*XA - C^*JC &= 0, \\ C - DJB^*X(E-A) &= 0. \end{aligned}$$

6.1. The canonical case. The following theorem gives a complete solution to the LCF with J all-pass denominator of McMillan degree n_b.

THEOREM 6.2. *Given an arbitrary rational matrix $G(z)$ and a disjoint partition $\overline{\mathbb{C}} = \Gamma_g \cup \Gamma_b$ defined by* (6.1)*, let n_b be the number of poles of G in Γ_b. Assume* (2.14) *is a minimal descriptor realization of G separated with respect to the given partition and satisfying the condition* (2.15)*. Then the LCF with J all-pass denominator with respect to the unit circle has a solution $G(z) = M^{-1}(z)N(z)$ of minimal degree n_b if and only if the Stein equation*

$$E_b^*XE_b - A_b^*XA_b - C_b^*JC_b = 0 \tag{6.2}$$

has an invertible Hermitian solution X. In this case the class of all solutions is given by

$$N(z) = \left[\begin{array}{cc|c} A_b - zE_b + KC_b(1-z) & A_{bg} - zE_{bg} + KC_g(1-z) & B_b + KD(1-z) \\ O & A_g - zE_g & B_g \\ \hline WC_b & WC_g & WD \end{array}\right], \tag{6.3}$$

$$M(z) = \left[\begin{array}{c|c} A_b - zE_b + KC_b(1-z) & K(1-z) \\ \hline WC_b & W \end{array}\right] \tag{6.4}$$

where $K = -X^{-1}(E_b - A_b)^{-*}C_b^*J$ and W is a J unitary matrix, $W^*JW = J$.

If G has real coefficients, then the class of solutions with real coefficients is also given by (6.3) *and* (6.4)*, where now all the intervening coefficients can be chosen real.*

Proof. The proof is identical as for Theorem 5.2, where we replace Corollary 5.1 with Corollary 6.1. □

If in the statement of the above theorem we add the condition $X > 0$ or $X < 0$, we obtain the solution to the LCF with J lossless or J expansive denominator, respectively. Notice that for the chosen partition (6.1) equation (6.2) has always a unique solution, and thus the above theorem is an effective tool for checking the existence of, and for computing solutions to, the LCF with J all-pass denominator or, as a particular case, with J lossless denominator.

For $J = I$, the above theorem provides the solution of an important particular case: LCF with all-pass denominator $M(z)$, (i.e., $M^*(1/\overline{z})M(z) = I$). In this case, (6.2) has always a sign definite (positive or negative) solution. If we take $\Gamma_b = \mathbb{D}^c$, then we obtain the LCF with inner denominator.

6.2. The noncanonical case. We give now solvability conditions in the noncanonical case.

THEOREM 6.3. *Given an arbitrary rational matrix $G(z)$ and a disjoint partition $\overline{\mathbb{C}} = \Gamma_g \cup \Gamma_b$ defined by* (6.1)*, let n_b be the number of poles of G in Γ_b. Assume that* (2.14) *is a minimal descriptor realization of G, separated with respect to the given partition, and satisfying condition* (2.15)*. Let $r := \mathrm{rank}(X)$ where X is the unique Hermitian solution to the Stein equation* (6.2)*. Then the minimal degree of the LCF with J all-pass denominator is $2n_b - r$.*

Proof. The proof is almost identical to the proof of Theorem 5.3. The key is to first transform (6.2) into the equivalent equation

$$\widehat{A}_b^* X \widehat{E}_b + \widehat{E}_b^* X \widehat{A}_b - \widehat{C}_b^* J \widehat{C}_b = 0, \tag{6.5}$$

where $\widehat{A}_b := E_b + A_b$, $\widehat{E}_b := E_b - A_b$, $\widehat{C}_b := \sqrt{2}C_b$. Since $\widehat{E}_b$ is invertible and (6.5) is of the form (5.11) we may use further a technique similar to that in the proof of Theorem 5.3. We note that the role of ∞ in placing additional poles/zeros in the proof of Theorem 5.3 is played in this case by 1. □

A remark similar to Remark 5.5 holds in this case as well.

7. Numerical aspects and examples. In this section we comment briefly on the numerical algorithms that are recommended for the methods proposed so far, and give several numerical examples.

7.1. Numerical aspects. Separated realizations as used in this paper can always be obtained by starting with an arbitrary realization (2.2) and performing unitary state-space equivalence transformations. If $\infty \in \Gamma_b$ then the preliminary transformation depicted in (2.10) is needed as to enforce further condition (2.15). The unitary matrices Q and Z in (2.10) can be obtained by a row compression of E followed by a column compression of A. The row/column compressions can be achieved by using any rank revealing procedure [7] based on unitary transformations. Further, we can arrive at (2.14) satisfying (2.15) by performing a spectral decomposition of the regular pencil $A_d - \lambda E_d$ (or $A - \lambda E$ if the preliminary step is skipped). The spectral decomposition of a regular pencil can be achieved by unitary transformations using the QZ algorithm [17] followed by a reordering of the generalized eigenvalues [23].

For solving the generalized eigenvalue assignment problem in Lemma 4.1 we can adapt almost all existing algorithms for pole assignment. (See [18] for a collection of such algorithms.)

For solving generalized Lyapunov equations (5.11) or Stein equations (6.2) one can use the algorithms proposed in [9]. Once the solution is computed, the balanced

form can be obtained by solving a symmetric eigenvalue problem for X [15]. Again, we can rely solely on unitary transformations.

7.2. Numerical examples.

Example 1. Minimal degree factorization with proper factors. Consider the improper rational matrix

$$G(\lambda) = \begin{bmatrix} \lambda^2 & \dfrac{\lambda}{\lambda - 1} \\ 0 & \dfrac{1}{\lambda} \end{bmatrix}$$

having finite poles at $\lambda_1 = 1$ and $\lambda_2 = 0$ and two poles at infinity, and let $\Gamma_g = \mathbb{C}^-$. A minimal order descriptor realization for G in the form (2.14) is given by

$$\left[\begin{array}{c|c} A_b - \lambda E_b & A_{bg} - \lambda E_{bg} \\ \hline O & A_g - \lambda E_g \end{array}\right] = \left[\begin{array}{cccc|c} -\lambda & 1 & 0 & 0 & 0 \\ 0 & 1-\lambda & 0 & 0 & 0 \\ 0 & 0 & 1 & -\lambda & 0 \\ 0 & 0 & 0 & 1 & -\lambda \\ \hline 0 & 0 & 0 & 0 & 1 \end{array}\right], \left[\begin{array}{c} B_b \\ \hline B_g \end{array}\right] = \left[\begin{array}{cc} 0 & 0 \\ 0 & -1 \\ 0 & 0 \\ 0 & 0 \\ \hline 1 & 0 \end{array}\right],$$

$$\left[\begin{array}{c|c} C_b & C_g \end{array}\right] = \left[\begin{array}{cccc|c} 0 & -1 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 \end{array}\right], \quad D = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

To compute a minimal degree LCF over Γ_g we use Theorem 4.2 and first determine a matrix K such that

$$\Lambda(A_b - \lambda E_b + K C_b(\alpha - \beta\lambda)) = \{-1, -1, -1, -1\}, \tag{7.1}$$

where we chose for convenience $\alpha = -2$ and $\beta = 1$. By using Lemma 4.1 for the dual pair $(A_b^T - \lambda E_b^T, C_b^T)$ we can assign the eigenvalues for the modified pair

$$(A_z - \lambda E_z, C_z) := (\alpha A_b^T - \beta E_b^T - \lambda(\beta A_b^T - \alpha E_b^T), (\alpha^2 + \beta^2) C_b^T)$$

to $\Gamma_z := \{3, 3, 3, 3\}$, where E_z is nonsingular. By using a pole assignment algorithm for generalized eigenvalues, we determine

$$K = \begin{bmatrix} -8 & 1 \\ -8 & 0 \\ 6 & 0 \\ 3 & 0 \end{bmatrix},$$

which fulfills (7.1). Using (4.4) and (4.5), we get the following solution to the minimal degree LCF over Γ_g:

$$N(\lambda) = \begin{bmatrix} \dfrac{(\lambda-1)\lambda^2}{3(\lambda+1)^3} & \dfrac{\lambda}{3(\lambda+1)^3} \\ 0 & \dfrac{1}{2(\lambda+1)} \end{bmatrix}, \quad M(\lambda) = \begin{bmatrix} \dfrac{\lambda-1}{3(\lambda+1)^3} & 0 \\ 0 & \dfrac{\lambda}{2(\lambda+1)} \end{bmatrix}$$

where the minimal degree of the LCF is clearly 4.

Example 2. Minimal degree coprime factorization with J all-pass denominator with respect to the imaginary axis. Consider the TFM of a continuous-time system

$$G(s) = \begin{bmatrix} \dfrac{1}{s-2} & \dfrac{1}{s-1} \\ \dfrac{\gamma_1}{s-2} & \dfrac{\gamma_2}{s-1} \end{bmatrix},$$

where $\gamma_1, \gamma_2 \in \mathbb{R}$ are two parameters, and let $\Gamma_b = \mathbb{C}^+$. It is easy to check that G has two unstable finite poles at $s_1 = 1$ and $s_2 = 2$. A minimal order realization for G in the form (2.14) is given by

$$A_b - sE_b = \begin{bmatrix} 1-s & 0 \\ 0 & 2-s \end{bmatrix}, B_b = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, C_b = \begin{bmatrix} 1 & 1 \\ \gamma_1 & \gamma_2 \end{bmatrix}, D = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

With $J = \left[\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right]$, the solution to the Lyapunov equation (5.2) in Theorem 5.2 is

$$X = \begin{bmatrix} \dfrac{1-\gamma_1^2}{2} & \dfrac{1-\gamma_1\gamma_2}{3} \\ \dfrac{1-\gamma_1\gamma_2}{3} & \dfrac{1-\gamma_2^2}{4} \end{bmatrix},$$

which for $\gamma_1(\gamma_2+3) - 1 - 3\gamma_2 \neq 0$ is nonsingular, for $\gamma_1(\gamma_2+3) - 1 - 3\gamma_2 = 0$ and $\gamma_2 \neq 1$ has rank equal to 1 and for $\gamma_1 = \gamma_2 = 1$ has rank equal to 0.

The canonical case. Consider $\gamma_1 = 2$ and $\gamma_2 = 1$ for which the solution of (5.2) is

$$X = \begin{bmatrix} -\dfrac{3}{2} & -\dfrac{1}{3} \\ -\dfrac{1}{3} & 0 \end{bmatrix}$$

and the corresponding K in Theorem 5.2 is

$$K = -X^{-1}C_b^T J = \begin{bmatrix} 3 & -3 \\ -\dfrac{21}{2} & \dfrac{15}{2} \end{bmatrix}.$$

The resulting factors $N(s)$ and $M(s)$ are

$$N(s) = \frac{1}{2}\begin{bmatrix} \dfrac{2\,s+11}{(s+1)\,(s+2)} & \dfrac{2}{(s+1)} \\ \dfrac{4\,s+13}{(s+1)\,(s+2)} & \dfrac{2}{(s+1)} \end{bmatrix},$$

$$M(s) = \frac{1}{2}\begin{bmatrix} \dfrac{2\,s^2-9\,s-5}{(s+1)\,(s+2)} & \dfrac{9\,s-3}{(s+1)\,(s+2)} \\ \dfrac{-9\,s+3}{(s+1)\,(s+2)} & \dfrac{2\,s^2+9\,s-5}{(s+1)\,(s+2)} \end{bmatrix}.$$

The noncanonical case: rank$(X) = 1$. Consider $\gamma_1 = 2$ and $\gamma_2 = 5$ for which the solution of (5.2) is

$$X = \begin{bmatrix} -\frac{3}{2} & -3 \\ -3 & -6 \end{bmatrix}$$

and has only a nonzero eigenvalue at $-\frac{15}{2}$. With

$$U = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix}$$

we obtain

$$U^T X U = \begin{bmatrix} -\frac{15}{2} & 0 \\ 0 & 0 \end{bmatrix},$$

and the known matrices in (5.22) result with $Q = U^T$ as

$$U^T A U := \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right] = \frac{1}{5} \begin{bmatrix} 9 & 2 \\ 2 & 6 \end{bmatrix}, \quad CU := \left[\begin{array}{c|c} C_1 & C_2 \end{array}\right] = \frac{1}{\sqrt{5}} \begin{bmatrix} 3 & -1 \\ 12 & 1 \end{bmatrix},$$
$$U^T E U = \begin{bmatrix} E_{11} & O \\ E_{21} & E_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The matrix

$$X_M = \begin{bmatrix} -\frac{15}{2} & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

fulfills (5.14) for

$$A_{x1} = \frac{8}{15}, \quad A_{x2} = \frac{18}{5}, \quad A_x = 0, \quad C_x = \begin{bmatrix} -\frac{6}{\sqrt{5}} \\ 0 \end{bmatrix},$$

where we have chosen C_x such that A_x has a zero eigenvalue. The corresponding B_M is

$$B_M = X_M^{-1} C_M^T J = \frac{1}{\sqrt{5}} \begin{bmatrix} -\frac{2}{5} & \frac{8}{5} \\ -6 & 0 \\ -1 & -1 \end{bmatrix}.$$

With M^{-1} defined by (5.21) we finally obtain

$$M(s) = \frac{1}{75} \begin{bmatrix} \dfrac{75\,s^3 + 63\,s^2 - 336\,s - 50}{s\,(s+2)\,(s+1)} & -\dfrac{2(81\,s^2 - 168\,s + 25)}{s\,(s+2)\,(s+1)} \\ \dfrac{2(81\,s^2 + 168\,s + 25)}{s\,(s+2)\,(s+1)} & \dfrac{75\,s^3 - 63\,s^2 - 336\,s + 50}{s\,(s+2)\,(s+1)} \end{bmatrix},$$

$$N(s) := M(s)G(s) = \frac{1}{25} \begin{bmatrix} \dfrac{25\,s^2 - 62\,s + 50}{s\,(s+2)\,(s+1)} & \dfrac{25\,s^2 - 199\,s + 50}{s\,(s+2)\,(s+1)} \\ \dfrac{2(25\,s^2 + 31\,s - 25)}{s\,(s+2)\,(s+1)} & \dfrac{125\,s^2 + 199\,s - 50}{s\,(s+2)\,(s+1)} \end{bmatrix}.$$

The noncanonical case: rank(X) = 0. Consider $\gamma_1 = \gamma_2 = 1$ for which the solution of (5.2) is $X = 0$. With $U = I$, we further have $A_{22} = A$ and $C_2 = C$. The matrix X_M has the form

$$X_M = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

and fulfills (5.14) for

$$A_{x2} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad A_x = \begin{bmatrix} -3 & 5 \\ -2 & 3 \end{bmatrix}, \quad C_x = \begin{bmatrix} -1 & \frac{5}{2} \\ 1 & -\frac{5}{2} \end{bmatrix},$$

where this time we have chosen C_x such that A_x has two eigenvalues on the imaginary axis at $\pm j$. The corresponding B_M results as

$$B_M = X_M^{-1} C_M^T J = \begin{bmatrix} -1 & -1 \\ \frac{5}{2} & \frac{5}{2} \\ 1 & -1 \\ 1 & -1 \end{bmatrix}.$$

With M^{-1} defined by (5.21) we eventually obtain

$$M(s) = \frac{1}{2} \begin{bmatrix} \dfrac{2\,s^4 - 3\,s^2 + 5}{(s^2+1)\,(s+2)\,(s+1)} & \dfrac{-7\,s^2 + 3}{(s^2+1)\,(s+2)\,(s+1)} \\ \dfrac{-7\,s^2 + 3}{(s^2+1)\,(s+2)\,(s+1)} & \dfrac{2\,s^4 - 3\,s^2 + 5}{(s^2+1)\,(s+2)\,(s+1)} \end{bmatrix},$$

$$N(s) := M(s)G(s) = \begin{bmatrix} \dfrac{s-2}{s^2+1} & \dfrac{s-1}{s^2+1} \\ \dfrac{s-2}{s^2+1} & \dfrac{s-1}{s^2+1} \end{bmatrix}.$$

Example 3. Minimal degree coprime factorization with J all-pass denominator with respect to the unit circle. Consider the TFM of a discrete-time system

$$G(z) = \begin{bmatrix} z^2 & \dfrac{1}{z-2} \\ 0 & z \end{bmatrix}$$

having an unstable finite pole at $z_1 = 2$ and three poles at infinity, and let $\Gamma_b := \mathbb{D}^c$. A minimal order descriptor realization for G in the form (2.14) is given by

$$\left[\begin{array}{c|c} A_b - zE_b & A_{bg} - zE_{bg} \\ \hline O & A_g - zE_g \end{array}\right] = \left[\begin{array}{cccc|cc} 2-z & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -z & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -z & 0 \\ 0 & 0 & 0 & 1 & 0 & -z \\ \hline 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right],$$

$$\left[\begin{array}{c} B_b \\ \hline B_g \end{array}\right] = \left[\begin{array}{cc} 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ \hline 1 & 0 \\ 0 & 1 \end{array}\right],$$

$$\left[\begin{array}{c|c} C_b & C_g \end{array}\right] = \left[\begin{array}{cccc|cc} 1 & -1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 & 0 \end{array}\right], \quad D = \left[\begin{array}{cc} 0 & 1 \\ 0 & 0 \end{array}\right].$$

The solution of (6.2) is

$$X = \left[\begin{array}{cccc} -1/3 & 1/2 & 1/4 & 0 \\ 1/2 & -1 & 0 & 0 \\ 1/4 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right],$$

and the corresponding K in Theorem 6.2 is

$$K = -X^{-1}(E_b - A_b)^{-T} C_b^T J = \left[\begin{array}{cc} -12 & 0 \\ -5 & 0 \\ -2 & 0 \\ 0 & 1 \end{array}\right].$$

The factors $N(z)$ and $M(z)$ resulting from Theorem 6.2 are

$$N(z) = \left[\begin{array}{cc} -\dfrac{z-2}{2\,z-1} & -\dfrac{1}{z^2\,(2\,z-1)} \\ 0 & 1 \end{array}\right], \quad M(z) = \left[\begin{array}{cc} -\dfrac{z-2}{z^2\,(2\,z-1)} & 0 \\ 0 & \dfrac{1}{z} \end{array}\right].$$

8. Conclusions. We have presented a comprehensive theory of minimal degree coprime factorization of rational matrices over a given domain of the closed complex plane. The given formulas allow us to compute the coprime factors by a one-shot approach and, in many case, also recursively by dislocating one pole (or a pair of poles) at a time (see for example the pole dislocation technique in [24]). In general, the one-shot approach brings advantages in terms of numerical reliability, efficiency, and versatility as any algorithm for pole assignment may be used, while the recursive method is more advantageous if a certain form of the coprime factors is needed at further computational steps. Nevertheless, the recursive techniques are applicable only when the involved Lyapunov equations have sign definite solutions (positive or negative), and then these equations are solved only implicitly. The recursive techniques cannot be employed in the more general cases discussed in this paper, not even in the canonical J all-pass case. Essentially, the recursive techniques fail since a J all-pass rational matrix cannot be written in general as a product of elementary J all-pass factors. In this respect, the paper provides a systematic theoretical framework for the recursive algorithms proposed in [25, 26] and elucidates in which cases they are applicable and provide the right answer.

A particular feature of the results developed in this paper is that the methods are exactly tailored to the dimension of the problem to be solved avoiding unnecessary

redundancy. As an example, the solution of a full-order Riccati equation (order equal to the McMillan degree of the rational matrix to be factored) that is usually employed in such factorizations (see for example [16], [31]) is completely avoided. Instead, we solve a Lyapunov equation of lower dimension with the benefits of increased numerical accuracy and computational efficiency.

The approach taken in this paper also gives ground for extensions to LCFs with denominators satisfying other symmetries, as for example with respect to the real line, or with respect to a certain contour in the complex plane. The theory presented here has been already applied as a preliminary step to the computation of the most general inner-outer, spectral, and J lossless factorizations and it is a promising step toward computing the more general J spectral factorizations, either canonical or noncanonical.

Appendix. In this appendix we give a characterization in terms of associated realizations for an arbitrary J all-pass rational matrix, with respect to either the imaginary axis or the unit circle. In this context we see the importance of centered realizations of the type (2.5) for which a suitable choice of α and β provides very handy formulas. A suitable choice means that α and β are such that $\frac{\alpha}{\beta}$ is on the contour with respect to which the symmetry is defined (either the imaginary axis or the unit circle). This choice guarantees, for a given rational matrix M with a minimal realization centered at $\frac{\alpha}{\beta}$, that its adjoint $M^{\sim}$ has a minimal realization centered at $\frac{\alpha}{\beta}$ which can be simply expressed in terms of the realization of M. This is especially important when working with TFMs of discrete-time systems as it avoids the intricacies related to the feature that ∞ is the symmetrized value of 0.

We start with the case in which the symmetry occurs with respect to the imaginary axis.

THEOREM A.1. *Let $M(s)$ be a square and invertible rational matrix. Let $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{C}^0$ such that $\frac{\alpha}{\beta}$ is not a pole of M, let $p(s) := \alpha - s\beta$, and assume*

$$M(s) := \left[\begin{array}{c|c} A - sE & Bp(s) \\ \hline C & D \end{array}\right] \tag{A.1}$$

is a minimal realization. Then M is J all-pass with respect to the imaginary axis if and only if

$$D^*JD = J \tag{A.2a}$$

and there exists an invertible Hermitian matrix X such that

$$A^*XE + E^*XA - C^*JC = 0, \tag{A.2b}$$

$$C - DJB^*X(\alpha E - \beta A) = 0. \tag{A.2c}$$

Proof. Notice first that $M(s)$ having no poles at $\frac{\alpha}{\beta}$ implies that in a minimal realization (A.1) $\alpha E - \beta A$ is invertible.

Only if. Since $M(s)$ is J all-pass, we have

$$M^*(-\overline{s})JM(s) = J \tag{A.3}$$

$\forall\, s = j\omega$ which are not poles of $M(s)$. In particular, for $s = \frac{\alpha}{\beta}$ we get (A.2a). Hence D is invertible. Further, (A.3) may be rewritten as $M^*(-\overline{s})J = JM^{-1}(s)$ or,

equivalently, in terms of realizations

$$\left[\begin{array}{c|c} -A^*-sE^* & C^*Jp(s) \\ \hline -B^* & D^*J \end{array}\right] = \left[\begin{array}{c|c} A-\alpha BD^{-1}C-s(E-\beta BD^{-1}C) & BD^{-1}p(s) \\ \hline -JD^{-1}C & JD^{-1} \end{array}\right]. \tag{A.4}$$

Since both realizations in (A.4) are minimal and they realize the same rational matrix, there exist unique invertible matrices Q and Z such that

$$Q(A-\alpha BD^{-1}C)Z = -A^*, \tag{A.5a}$$

$$Q(E-\beta BD^{-1}C)Z = E^*, \tag{A.5b}$$

$$QBD^{-1} = C^*J, \tag{A.5c}$$

$$JD^{-1}CZ = B^*. \tag{A.5d}$$

It is easy to see that equations (A.5a)–(A.5d) also hold with Q and Z replaced by Z^{-*} and Q^{-*}, respectively. Due to the uniqueness of Q and Z we get that $Q = Z^{-*}$. Substituting this in (A.5a) and (A.5b) we get

$$(A-\alpha BD^{-1}C)Z = -Z^*A^*, \tag{A.6a}$$

$$(E-\beta BD^{-1}C)Z = Z^*E^*. \tag{A.6b}$$

Rewriting (A.5d) we get

$$C - DJB^*Z^{-1} = 0. \tag{A.7}$$

Comparing (A.7) with (A.2c) we define

$$X := Z^{-1}(\alpha E-\beta A)^{-1} \tag{A.8}$$

that is clearly invertible. To show that X is Hermitian, we subtract (A.6a) multiplied by β from (A.6b) multiplied by α and get

$$X^{-1} = (\alpha E-\beta A)Z = Z^*(\alpha E^*+\beta A^*) = X^{-*}. \tag{A.9}$$

It remains only to show (A.2b). We rewrite the left-hand side of (A.2b) successively as

$$\begin{aligned} &\frac{1}{\alpha^2+\beta^2}[(-\beta E^*+\alpha A^*)X(\alpha E-\beta A)+(\alpha E^*+\beta A^*)X(\alpha A+\beta E)] - C^*JC \\ &\quad \overset{(A.8)}{=} \frac{1}{\alpha^2+\beta^2}[(-\beta E^*+\alpha A^*)Z^{-1}+Z^{-*}(\alpha A+\beta E)] - C^*JC \\ &\quad \overset{(A.6a)-(A.6b)}{=} Z^{-*}BD^{-1}C - C^*JC \overset{(A.7)}{=} 0, \end{aligned}$$

which ends the first part of the proof.

If. We show that if (A.2a)–(A.2c) hold for an invertible Hermitian matrix X, then $M(s)$ is J all-pass. Starting with (A.1) we get a realization of

$$M^{\sim}(s)JM(s) = \left[\begin{array}{cc|c} -A^*-sE^* & C^*JC & C^*JDp(s) \\ O & A-sE & Bp(s) \\ \hline -B^* & D^*JC & D^*JD \end{array}\right].$$

Let

$$Q := \left[\begin{array}{cc} I & -(\beta A^*+\alpha E^*)X \\ O & I \end{array}\right], \qquad Z := \left[\begin{array}{cc} I & X(\alpha E-\beta A) \\ O & I \end{array}\right],$$

and performing an equivalence state-space transformation with Q to the left and Z to the right we get

$$M^*(-\overline{s})JM(s) = \left[\begin{array}{cc|c} -A^*-sE^* & C^*JC-E^*XA-A^*XE & [C^*JD-(\beta A^*+\alpha E^*)XB]p(s) \\ O & A-sE & Bp(s) \\ \hline -B^* & -B^*X(\alpha E-\beta A)+D^*JC & D^*JD \end{array}\right]. \tag{A.10}$$

Now using (A.2a)–(A.2c) we get that the $(1,2)$, $(1,3)$ and $(3,2)$ entries of (A.10) are identically zero. Thus removing the uncontrollable and unobservable part of (A.10) we finally get

$$M^{\sim}(s)JM(s) = D^*JD$$

from which the conclusion follows that $M(s)$ is J all-pass. □

Here follows the case of a J all-pass rational matrix with respect to the unit circle.

THEOREM A.2. *Let $M(z)$ be a square and invertible rational matrix. Let $\alpha \in \mathbb{C}$, with $\|\alpha\| = 1$ and such that α^2 is not a pole of M, let $p(z) := (\alpha - \overline{\alpha}z)$, and assume that*

$$M(z) := \left[\begin{array}{c|c} A-zE & Bp(z) \\ \hline C & D \end{array}\right] \tag{A.11}$$

is a minimal realization. Then M is J all-pass with respect to the unit circle if and only if

$$D^*JD = J \tag{A.12a}$$

and there exists an invertible Hermitian matrix X such that

$$E^*XE - A^*XA - C^*JC = 0, \tag{A.12b}$$

$$C - DJB^*X(\alpha E - \overline{\alpha}A) = 0. \tag{A.12c}$$

Proof. Notice first that since $M(z)$ has no pole at α^2 and (A.11) is a minimal realization it follows that $\overline{\alpha}A - \alpha E$ is invertible.

Only if. Since $M(z)$ is J all-pass, we have

$$M^{\sim}(z)JM(z) = J \tag{A.13}$$

for all $z = e^{j\theta}$ which are not poles of $M(z)$. In particular, for $z = \alpha$ we get (A.12a) and it follows that D is invertible. Further, (A.13) may be rewritten as

$$M^*\left(\frac{1}{\overline{z}}\right)J = JM^{-1}(z)$$

or, equivalently, in terms of realizations

$$\left[\begin{array}{c|c} E^*-zA^* & C^*Jp(z) \\ \hline B^* & D^*J \end{array}\right] = \left[\begin{array}{c|c} (A-\alpha BD^{-1}C)-z(E-\overline{\alpha}BD^{-1}C) & BD^{-1}p(z) \\ \hline -JD^{-1}C & JD^{-1} \end{array}\right]. \tag{A.14}$$

Since both realizations in (A.14) are minimal and they realize the same rational matrix, there exist two unique invertible matrices Q and Z such that

$$Q(A-\alpha BD^{-1}C)Z = E^*, \tag{A.15a}$$
$$Q(E-\overline{\alpha}BD^{-1}C)Z = A^*, \tag{A.15b}$$
$$QBD^{-1} = C^*J, \tag{A.15c}$$
$$-JD^{-1}CZ = B^*. \tag{A.15d}$$

It is easy to see that equations (A.15a)–(A.15d) hold if we replace Q and Z with $-Z^{-*}$ and $-Q^{-*}$, respectively. Due to the uniqueness of Q and Z we get

$$Q = -Z^{-*}. \tag{A.16}$$

Thus substituting (A.16) in (A.15a)–(A.15d) results in

$$(A-\alpha BD^{-1}C)Z = -Z^*E^*, \tag{A.17a}$$
$$(E-\overline{\alpha}BD^{-1}C)Z = -Z^*A^*. \tag{A.17b}$$

Rewriting (A.15d) we get

$$C + DJB^*Z^{-1} = 0. \tag{A.18}$$

Comparing (A.12c) with (A.18) we define

$$X := -Z^{-1}(\alpha E - \overline{\alpha}A)^{-1}, \tag{A.19}$$

which is clearly invertible. To show that X is Hermitian we subtract (A.17b) multiplied by α from (A.17a) multiplied by $\overline{\alpha}$ and get

$$X^{-1} = (\overline{\alpha}A - \alpha E)Z = Z^*(\alpha A^* - \overline{\alpha}E^*) = X^*.$$

It remains only to show (A.12b). We rewrite the left-hand side of (A.12b) successively as

$$\begin{aligned}
&\frac{1}{2}[(\alpha A^* + \overline{\alpha}E^*)X(\alpha E - \overline{\alpha}A) + (\overline{\alpha}E^* - \alpha A^*)X(\alpha E + \overline{\alpha}A)] - C^*JC \\
&\overset{(A.19)}{=} -\frac{1}{2}[(\alpha A^* + \overline{\alpha}E^*)Z^{-1} + Z^{-*}(\alpha E + \overline{\alpha}A)] - C^*JC \\
&\overset{(A.17a)-(A.17b)}{=} -Z^{-*}BD^{-1}C - C^*JC \overset{(A.18)}{=} 0
\end{aligned}$$

which ends this part of the proof.

If. We show that if (A.12a)–(A.12c) hold for an invertible Hermitian matrix X, then $M(z)$ is J all-pass. Starting with (A.11) we get a realization of

$$M^*\left(\frac{1}{\overline{z}}\right)JM(z) = \left[\begin{array}{cc|c} E^* - zA^* & C^*JCp(z) & C^*JDp(z) \\ O & A - zE & Bp(z) \\ \hline B^* & D^*JC & D^*JD \end{array}\right].$$

Let

$$Q := \begin{bmatrix} I & (\alpha A^* - \overline{\alpha}E^*)X \\ O & I \end{bmatrix}, \qquad Z := \begin{bmatrix} I & X(\overline{\alpha}A - \alpha E) \\ O & I \end{bmatrix}.$$

Performing an equivalence state-space transformation with Q to the left and Z to the right we obtain

$$M^*\left(\frac{1}{\overline{z}}\right)JM(z) = \left[\begin{array}{ccc|c} E^* - zA^* & O & & O \\ O & A - zE & & Bp(z) \\ \hline B^* & O & & D^*JD \end{array}\right], \tag{A.20}$$

where we have used (A.12b) and (A.12c) to get that the $(1,2)$, $(1,3)$ and $(3,2)$ entries are zero. Thus, removing the uncontrollable and unobservable part of (A.20) we finally get $M^{\sim}(z)JM(z) = D^*JD$, from which with (A.12a) we conclude that $M(z)$ is J all-pass. □

REFERENCES

[1] E. Anderson, Z. Bai, C. Bishof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen, *LAPACK Users' Guide,* 2nd ed., SIAM, Philadelphia, PA, 1995.

[2] J.A. Ball, J. Kim, L. Rodman, and M. Verma, *Minimal degree coprime factorization of rational matrix functions,* Linear Algebra Appl., 186 (1993), pp. 117–164.

[3] J.A. Ball, I. Gohberg, and M. Rakowski, *Reconstruction of rational nonsquare matrix functions from local data,* Integral Equations Operator Theory, 20 (1994), pp. 249–305.

[4] H. Bart, I. Gohberg, and M. A. Kaashoek, *Minimal Factorization of Matrix and Operator Functions*, Oper. Theory Adv. Appl. 1, Birkhäuser, Basel, 1979.

[5] H. Bart, I. Gohberg, M. Kaashoek, and P. Van Dooren, *Factorizations of transfer functions,* SIAM J. Control Optim., 18 (1980), pp. 675–696.

[6] H. Bart, I. Gohberg, and M.A. Kaashoek, *Constructive Methods of Wiener–Hopf Factorization*, Oper. Theory Adv. Appl. 21, Birkhäuser, Basel, 1986.

[7] T. Chan, *Rank revealing QR factorizations,* Linear Algebra Appl., 88/89 (1987), pp. 67–82.

[8] F.R. Gantmacher, *The Theory of Matrices*, Chelsea, New York, 1960.

[9] J.D. Gardiner, A.J. Laub, J.J. Amato, and C.B. Moler, *Solution of the Sylvester matrix equation* $AXB^T + CXD^T = E$, ACM Trans. Math. Software, 18 (1992), pp. 223–231.

[10] I. Gohberg, M.A. Kaashoek, and A.C.M. Ran, *Partial pole and zero displacement by cascade connection,* SIAM J. Matrix Anal. Appl., 10 (1989), pp. 316–325.

[11] I. Gohberg, ed., *Topics in Interpolation Theory of Rational Matrix-Valued Functions,* Oper. Theory Adv. Appl. 33, Birkhäuser, Basel, 1988.

[12] I. Gohberg and M.A. Kaashoek, *Regular rational matrix functions with prescribed pole and zero structure,* in Topics in Interpolation Theory of Rational Matrix-Valued Functions, Oper. Theory Adv. Appl. 33, Birkhäuser, Basel, 1988, pp. 109–122.

[13] I. Gohberg, M.A. Kaashoek, and A.C.M. Ran, *Inverse spectral problems for regular improper rational matrix functions,* in Topics in Interpolation Theory of Rational Matrix-Valued Functions, Oper. Theory Adv. Appl. 33, Birkhäuser, Basel, 1988, pp. 123–173.

[14] I. Gohberg, M.A. Kaashoek, and A.C.M. Ran, *Factorizations of and extensions to J–unitary rational matrix functions on the unit circle,* Integral Equations Operator Theory, 15 (1992), pp. 262–300.

[15] G.H. Golub, and C.F. van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1989.

[16] H. Kimura, *Chain-Scattering Approach to H_∞-Control*, Birkhäuser, Boston, 1996.

[17] C.B. Moler and G.W. Stewart, *An algorithm for generalized matrix eigenvalue problem,* SIAM J. Numer. Anal., 10 (1973), pp. 241–256.

[18] R.V. Patel, A. Laub, and P. Van Dooren, eds., *Numerical Linear Algebra Techniques for Systems and Control*, IEEE Press, Los Alamitos, CA, 1994.

[19] M. Rakowski, *Zero–Pole Interpolation for Nonregular Rational Matrix Functions*, Ph.D. thesis, Virginia Tech, Blacksburg, VA, 1989.

[20] M. Rakowski, *Generalized pseudoinverses of matrix valued functions,* Integral Equations Operator Theory, 14 (1991), pp. 564–585.

[21] H. H. Rosenbrock, *State-Space and Multivariable Theory*, John Wiley, New York, 1970.

[22] P. Van Dooren, *The generalized eigenstructure problem in linear systems theory,* IEEE Trans. Automat. Control, 26 (1981), pp. 111–129.

[23] P. Van Dooren, *A generalized eigenvalue approach for solving Riccati equations,* SIAM J. Sci. Stat. Comput., 2 (1981), pp. 121–135.

[24] P. Van Dooren, *Rational and polynomial matrix factorizations via recursive pole-zero cancellation,* Linear Algebra Appl., 137/138 (1990), pp. 663–697.
[25] A. Varga, *Computation of coprime factorizations of rational matrices,* Linear Algebra Appl., 271 (1998), pp. 83–115.
[26] A. Varga and T. Katayama, *Computation of J–inner–outer factorizations of rational matrices,* Internat. J. Robust Nonlinear Control, 8 (1998), pp. 245–263.
[27] G. Verghese, B. Lévy, and T. Kailath, *A generalized state-space for singular systems,* IEEE Trans. Automat. Control, AC-26 (1981), pp. 811–831.
[28] G. Verghese, P. Van Dooren, and T. Kailath, *Properties of the system matrix of a generalized state-space system,* Internat. J. Control, 30 (1979), pp. 235–243.
[29] M. Vidyasagar, *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985.
[30] X. Xin and H. Kimura, *(J, J')–lossless factorization for descriptor systems,* Linear Algebra Appl., 205/206 (1994), pp. 1289–1318.
[31] X. Xin and H. Kimura, *Singular (J, J')–lossless factorization for strictly proper functions,* Internat. J. Control, 59 (1994), pp. 1383–1400.
[32] K. Zhou, J.C. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice Hall, Upper Saddle River, NJ, 1996.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 1, pp. 279–299
© 1999 Society for Industrial and Applied Mathematics

BILUTM: A DOMAIN-BASED MULTILEVEL BLOCK ILUT PRECONDITIONER FOR GENERAL SPARSE MATRICES*

YOUSEF SAAD† AND JUN ZHANG‡

Abstract. This paper describes a domain-based multilevel block ILU preconditioner (BILUTM) for solving general sparse linear systems. This preconditioner combines a high accuracy incomplete LU factorization with an algebraic multilevel recursive reduction. Thus, in the first level the matrix is permuted into a block form using (block) independent set ordering and an ILUT factorization for the reordered matrix is performed. The reduced system is the approximate Schur complement associated with the partitioning, and it is obtained implicitly as a by-product of the partial ILUT factorization with respect to the complement of the independent set. The incomplete factorization process is repeated with the reduced systems recursively. The last reduced system is factored approximately using ILUT again. The successive reduced systems are not stored. This implementation is efficient in controlling the fill-in elements during the multilevel block ILU factorization, especially when large size blocks are used in domain decomposition-type implementations. Numerical experiments are used to show the robustness and efficiency of the proposed technique for solving some difficult problems.

Key words. incomplete LU factorization, ILUT, multilevel ILU preconditioner, Krylov subspace methods, multielimination ILU factorization

AMS subject classifications. 65F10, 65N06

PII. S0895479898341268

1. Introduction.

The preconditioning technique proposed in this paper is based on a multilevel block incomplete LU factorization. It is intended for solving general sparse linear systems of the form

$$Ax = b, \tag{1.1}$$

where A is an unstructured matrix of order n. Such linear systems are often solved by Krylov subspace methods coupled with a suitable preconditioner [52]. The research and design of preconditioners with inherent parallelism have received much attention recently, spurred by the popularity of distributed memory architectures. The main considerations when comparing preconditioners are their intrinsic efficiency, generality, parallelism, and robustness. An experimental study on robustness of a few general purpose preconditioners has been conducted in [22] and a number of ILU-type preconditioners have been tested for solving some difficult problems from computational fluid dynamics in [19, 20].

The incomplete LU factorization without fill-in (ILU(0)) is probably the best known general purpose preconditioner [40]. However, this preconditioner is not robust and is inefficient and fails for many real-life problems. Many extensions of ILU(0), which increase its accuracy and robustness, have been designed, and we refer to [52]

*Received by the editors June 29, 1998; accepted for publication (in revised form) by A. Greenbaum November 23, 1998; published electronically September 22, 1999. This work was supported in part by NSF grant CCR-9618827 and in part by the Minnesota Supercomputer Institute.
http://www.siam.org/journals/simax/21-1/34126.html

†Department of Computer Science and Engineering, University of Minnesota, 4-192 EE/CS Building, 200 Union Street SE, Minneapolis, MN 55455 (saad@cs.umn.edu, http://www.cs.umn.edu/~saad).

‡Department of Computer Science and Engineering, University of Minnesota, 4-192 EE/CS Building, 200 Union Street SE, Minneapolis, MN 55455. Current address: Department of Computer Science, University of Kentucky, 773 Anderson Hall, Lexington, KY 40506-0046 (jzhang@cs.uky.edu, http://www.cs.uky.edu/~jzhang).

for a partial account of the literature and to [25, 26, 33, 44, 60, 64] for just a few ideas described.

The moderate parallelism which can be extracted from the triangular solves in standard ILU factorization [1, 52] is limited, and becomes inadequate for the more accurate ILU factorizations. Alternatives have been considered in the past to develop preconditioners with inherently more parallelism than standard ILU; see, e.g., [48, 52, 58] for references. A standard technique used for this purpose is to exploit "multicolor orderings" or "independent sets" [49]. A well-known drawback of using a multicolor ordering prior to building an ILU factorization is that the quality of the preconditioning on the reordered system is generally worse than that on the original system [29, 30, 32]. However, numerical results in [50] show that some high accuracy ILU-type preconditioners with red-black ordering may eventually outperform their counterparts with natural ordering, if enough fill-in is allowed. Conversely, the higher amount of fill-in usually reduces the parallelism that is available in ILU(0). Therefore, a desirable goal is to achieve high accuracy while retaining parallelism achieved by using multicoloring or independent sets. Other alternatives for developing parallel preconditioners have been proposed that are based on sparse approximate inverse techniques; see, e.g., [9, 16, 21, 23, 35, 37, 62]. These preconditioners afford potential maximum parallelism both in their construction stage (except [9, 62]) and in the application stage which requires only matrix-vector operations. However, these methods tend to become inefficient for handling very large matrices because of their local nature.

The multielimination ILU preconditioner (ILUM), introduced in [51], is based on exploiting the idea of successive independent set orderings. It has a multilevel structure and offers a good degree of parallelism without sacrificing overall effectiveness. Similar preconditioners developed in [12, 55] show near grid-independent convergence for certain types of problems. In a recent report, some of these multilevel preconditioners were tested and compared favorably with other preconditioned iterative methods and direct methods, at least for the Laplace equation [10].

The idea of combining multilevel techniques with ILU is not new. Alternative multilevel approaches that require grid information have been developed. Examples of such approaches include the nested recursive two-level factorization, repeated red-black orderings, and generalized cyclic reduction [2, 4, 11, 31, 43] (see also the survey paper by Axelsson and Vassilevski [3]). Some recently developed methods require only the adjacency graph of the coefficient matrices [12, 46, 51, 55]. For the repeated red-black ordering approach, a near-optimal bound for the condition number of the preconditioned matrix was reported [42]. Other methods which bear some similarity with ILUM-type techniques are the algebraic multigrid methods [13, 18, 46, 47, 59] and certain types of multigrid methods which consider matrix entries [28, 27, 45]. Equally interesting are the multilevel preconditioning techniques based on hierarchical basis, multigraph, or ILU decomposition associated with the finite difference or finite element analysis [6, 8, 7, 15, 36].

A block version of ILUM (BILUM) was recently defined by using small dense matrices as pivots instead of scalars [55, 57]. For some hard-to-solve problems, BILUM may perform much better than ILUM. Tests with large blocks indicate that the larger the block, the more robust the resulting preconditioner. The solution with the independent blocks in BILUM uses the exact inverse or a regularized inverse based on the singular value decomposition (SVD) [55, 57]. These strategies are efficient for blocks of small size, but the cost of such inversion strategies grows cubically as the size of the blocks increases.

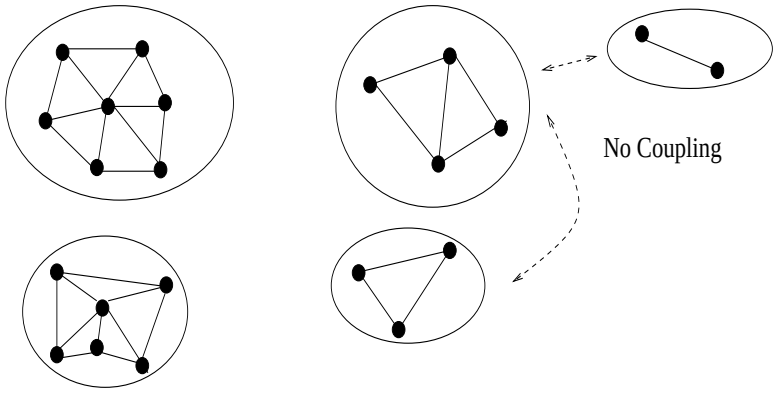


FIG. 2.1. *Independent groups or blocks.*

In this paper we focus on *domain decomposition-based* block ILUM. In this case the blocks can be very large and are associated with a subdomain, as in domain decomposition methods. For these large size blocks, the computational and memory costs of computing exact or regularized inverses used for BILUM become prohibitive. The dropping strategies that have been proposed [57] are not able to deal with both problems (computation and memory costs) simultaneously. It is therefore vital to exploit sparsity for domain decomposition-based block ILUM. This paper introduces an efficient approach to constructing BILUM based on this principle. The construction of such a preconditioner is based on a restricted ILU factorization with a dual dropping strategy (ILUT); see [50]. This multilevel block ILUT preconditioner (BILUTM) retains the efficiency and flexibility of ILUT and offers inherent parallelism that can be exploited on parallel or distributed architectures.

This paper is organized as following. Section 2 gives an overview and background on multigrid and multilevel preconditioning techniques. Section 3 provides some details on the construction of the reduced system by partial Gaussian elimination. Section 4 discusses the proposed BILUTM. Section 5 describes some numerical experiments, and section 6 gives a few concluding remarks.

2. Multilevel preconditioning techniques. Multilevel preconditioners take advantage of the fact that different parts of the error spectrum can be treated independently on different levels. In construction, multilevel preconditioners also exploit, explicitly or implicitly, the property that a set of unknowns that are not coupled to each other can be eliminated simultaneously in a Gaussian elimination–type process. Such a set is usually called an *independent set* [34]. This concept of independent set can easily be generalized to blocks. Thus a block independent set is a set of groups (blocks) of unknowns such that there is no coupling between unknowns of any two different groups (blocks) [55]. Unknowns within the same group (block) may be coupled. This is illustrated in Figure 2.1.

Thus, point (scalar) independent sets are a particular case which use blocks of uniform size of 1. Various heuristic strategies may be used to find a block independent set with different properties [51, 55]. A simple and usually efficient strategy is a greedy algorithm, which groups the nearest nodes together. Since the focus of this paper is not on finding block independent sets, we assume that this greedy algorithm is used throughout to find block independent sets.

A maximal independent set is an independent set that cannot be augmented by other nodes and still remain independent. Independent sets are often constructed with some other conditions such as guaranteeing certain diagonal dominance for the nodes of the independent set or the vertex cover, which is defined as the complement of the independent set. Thus, in practice, the maximality of an independent set is rarely guaranteed, especially when some dropping strategies are applied [57].

Algebraic and "black box" multigrid methods try to mimic geometric multigrid methods by defining a prolongation operator $\mathbf{I}_{\alpha+1}^{\alpha}$ based on some heuristic arguments; here $0 \leq \alpha < \mathcal{L}$ is an integer used to label the level. For convenience and for satisfying certain conservation laws, the restriction operator, $\mathbf{I}_{\alpha}^{\alpha+1}$, is traditionally defined as the adjoint of the prolongation operator (possibly scaled by a constant), i.e., $\mathbf{I}_{\alpha}^{\alpha+1} = {\mathbf{I}_{\alpha+1}^{\alpha}}^T$ [47]. With $A_0 = A$, the recursive coarse grid operators are then generated by using the Galerkin technique as

$$A_{\alpha+1} = \mathbf{I}_{\alpha}^{\alpha+1} A_{\alpha} \mathbf{I}_{\alpha+1}^{\alpha}. \tag{2.1}$$

Note that, in order for the grid transfer operators to be defined efficiently, a logically rectangular grid is explicitly or implicitly assumed for the black box or matrix-dependent approaches [28]. Most such multigrid methods are designed for two-dimensional problems, and their extensions to higher dimensions are not straightforward [5]. For algebraic multigrid methods, improvements have recently been reported by defining more accurate grid transfer operators [17, 18].

In independent set orderings, the unknowns may be permuted such that those associated with the independent set are listed first, followed by the other unknowns. The permutation matrix P_{α}, associated with such an ordering, transforms the original matrix into a matrix which has the following block structure:

$$A_{\alpha} \sim P_{\alpha} A_{\alpha} P_{\alpha}^T = \begin{pmatrix} D_{\alpha} & F_{\alpha} \\ E_{\alpha} & C_{\alpha} \end{pmatrix}, \tag{2.2}$$

where D_{α} is a block diagonal matrix of dimension m_{α}, and C_{α} is a square matrix of dimension $n_{\alpha} - m_{\alpha}$. In what follows, the notation is slightly abused by not distinguishing the original system from the permuted system, so both permuted and unpermuted matrices will be denoted by A_{α}.

To improve load balancing on parallel computers, it is desirable to have uniformly sized independent blocks. However, this is not a necessary requirement for the techniques described in this paper.

In algebraic multilevel preconditioning techniques, the reduced systems are recursively constructed as the Schur complement with respect to either D_{α} or C_{α}. In the case of BILUM [51, 55], such a construction amounts to performing a block LU factorization of the form

$$\begin{pmatrix} D_{\alpha} & F_{\alpha} \\ E_{\alpha} & C_{\alpha} \end{pmatrix} = \begin{pmatrix} I_{\alpha} & 0 \\ E_{\alpha} D_{\alpha}^{-1} & I_{\alpha} \end{pmatrix} \times \begin{pmatrix} D_{\alpha} & F_{\alpha} \\ 0 & A_{\alpha+1} \end{pmatrix}, \tag{2.3}$$

where $A_{\alpha+1}$ is the Schur complement with respect to C_{α} and I_{α} is the generic identity matrix on level α. Note that $n_{\alpha+1} = m_{\alpha}$. The solution process with the above factorization consists of level-by-level forward elimination, followed by an exact solution on the last reduced system $A_{\mathcal{L}}$. The solution of the original system is obtained by level-by-level backward substitution (with suitable permutation).

The procedure described above is a direct solution method and the reduced systems become denser and denser as the level number increases, a consequence of the

fill-in caused by the elimination process. In BILUM, some dropping strategies are used to control the amount of fill-in by discarding certain elements of small magnitude or by limiting the number of elements allowed in each row of the L and U factors [51, 54, 55]. The resulting incomplete multilevel block LU factorization is then used as a preconditioner in a Krylov subspace method based iterative solver.

In the implementation of BILUM in [55], the block diagonals D_α consist of small size blocks. These small blocks are usually dense and an exact inverse technique is used to compute D_α^{-1} by inverting each small block independently (in parallel). In [57], some regularized-inverse technique based on SVD is used to invert the (potentially near-singular) blocks approximately. As we noted in the introduction, such direct inversion strategies usually produce dense inverse matrices even if the original blocks are highly sparse with large sizes. Thus some heuristic approaches have been proposed to drop small elements from the exactly or approximately inverted blocks to recover sparsity. Obviously this approach cannot reduce the cost of inverting these blocks.

The link between the algebraic multigrid methods and BILUM was discussed briefly in [57]. If we define the grid transfer operators naturally based on the matrix, the reduced system based on the Schur complement technique as in (2.3) also satisfies the Galerkin condition (2.1). In other words, BILUM can be viewed as a naturally defined algebraic multigrid or multilevel technique.

3. Gaussian elimination and ILUT. ILUT is a high-order (high accuracy) preconditioner based on incomplete LU factorization. It uses a dual dropping strategy to control the storage cost (the amount of fill-in) [50]. Its implementation is based on the IKJ variant of Gaussian elimination, which we recall next.

ALGORITHM 3.1. GAUSSIAN ELIMINATION–IKJ VARIANT.

1. *For* $i = 2, n$, *Do*
2. *For* $k = 1, i-1$, *Do*
3. $a_{i,k} := a_{i,k}/a_{k,k}$
4. *For* $j = k+1, n$, *Do*
5. $a_{i,j} := a_{i,j} - a_{i,k} * a_{k,j}$
6. *End Do*
7. *End Do*
8. *End Do*

The ILUT(τ, p) preconditioner attempts to control fill-in elements by applying a dual dropping strategy in Algorithm 3.1. The accuracy of ILUT(τ, p) is controlled by two dropping parameters, τ and p. In Algorithm 3.2, w is a work array, $a_{i,\beta}$ and $u_{k,\beta}$ denote the ith and kth rows of A and U, respectively.

ALGORITHM 3.2. STANDARD ILUT(τ, p) FACTORIZATION [50, 52].

1. *For* $i = 2, n$, *Do*
2. $w := a_{i,\beta}$
3. *For* $k = 1, i-1$ *and when* $w_k \neq 0$, *Do*
4. $w_k := w_k/a_{k,k}$
5. *If* $|w_k| < \tau * \text{nzavg}(a_{i,\beta})$, *set* $w_k := 0$
6. *If* $w_k \neq 0$, *then*
7. $w := w - w_k * u_{k,\beta}$
8. *End If*
9. *End Do*
10. *Apply a dropping strategy to row* w
11. *Set* $l_{i,j} := w_j$ *for* $j = 1, \ldots, i-1$ *whenever* $w_j \neq 0$
12. *Set* $u_{i,j} := w_j$ *for* $j = i, \ldots, n$ *whenever* $w_j \neq 0$

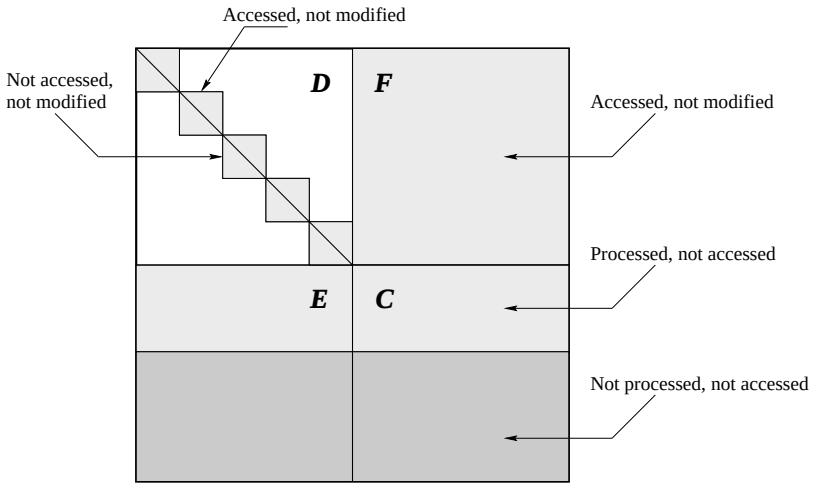


FIG. 3.1. *Illustration of restricted IKJ version of Gaussian elimination.*

13. *Set* $w := 0$
14. *End Do*

In Line 5, the function nzavg($a_{i,\beta}$) returns the average magnitude of the nonzero elements of a given sparse row. Elements with relatively small magnitude are dropped. In Line 10, a different dropping strategy is applied. First, small elements are dropped according to the relative magnitude similar to the criterion used in Line 5. Then a sorting operation is performed and only the largest p elements in absolute value of the L and U factors are kept. After the dual dropping strategy, there are at most p elements kept in each row of the L and U factors.

We now consider a slightly different elimination procedure. Assume that the first m equations are associated with the independent set as in the left-hand side of (3.2). If we perform the LU factorization (Gaussian elimination) to the upper part (the first m rows) of the matrix, i.e., to the submatrix $(D\ F)$, we have

$$(D\ \ F) = (LU\ \ L^{-1}F). \tag{3.1}$$

We then continue the Gaussian elimination to the lower part, but the elimination is only performed with respect to the submatrix E; i.e., we only eliminate those elements $a_{i,k}$ for which $m < i \le n, 1 \le k \le m$. Appropriate linear combinations are also performed with respect to the C submatrix, in connection with the eliminations in the E submatrix, as in the usual Gaussian elimination. Note that when doing these operations on the lower part, the upper part of the matrix is only accessed but not modified; see Figure 3.1. The processed rows of the lower part are never accessed again. This gives the following "restricted" version of Algorithm 3.1.

ALGORITHM 3.3. RESTRICTED IKJ VERSION OF GAUSSIAN ELIMINATION.

1. *For* $i = 2, n$, *Do*
2. *For* $k = 1, \min(i-1, m)$, *Do*
3. $a_{i,k} := a_{i,k}/a_{k,k}$
4. *For* $j = k+1, n$, *Do*
5. $a_{i,j} := a_{i,j} - a_{i,k} * a_{k,j}$
6. *End Do*
7. *End Do*
8. *End Do*

Here m is a parameter which defines the size of the matrix D. Algorithm 3.3 performs a block factorization of the form

$$\begin{pmatrix} D & F \\ E & C \end{pmatrix} = \begin{pmatrix} L & 0 \\ EU^{-1} & I \end{pmatrix} \times \begin{pmatrix} U & L^{-1}F \\ 0 & A_1 \end{pmatrix} = \mathbf{LU}. \tag{3.2}$$

In other words, the $a_{i,k}$'s (of the lower part) for $k \leq m$ are the elements in EU^{-1} and the other elements are those in A_1.

PROPOSITION 3.4. *The matrix A_1 computed by Algorithm* 3.3 *is the Schur complement of A with respect to C.*

Proof The part of the matrix after the upper part Gaussian elimination (with respect to the independent set, see (3.1)) that is accessed is $(U\ L^{-1}F)$; the L part is never accessed again. So we may write the active part of the (partially processed) matrix A as

$$\begin{pmatrix} U & \tilde{F} \\ E & C \end{pmatrix} = \begin{pmatrix} U & L^{-1}F \\ E & C \end{pmatrix}.$$

In order to eliminate an element in E, say, $e_{i,j}(= a_{i,j})$ with $m < i \leq n, 1 \leq j \leq m$, we perform a linear combination of the ith row of A and jth row of the U-part $(U\ L^{-1}F)$. Hence, the elements in C are modified according to the operations

$$\tilde{c}_{i,k} = c_{i,k} - \frac{e_{i,j}}{u_{j,j}} \tilde{f}_{j,k}.$$

After eliminating all $e_{i,j}$'s in E, the elements of the C matrix are changed to

$$\tilde{c}_{i,k} = c_{i,k} - \sum_{j=1}^{m} \frac{e_{i,j}}{u_{j,j}} \tilde{f}_{j,k}.$$

It follows that

$$A_1 = \tilde{C} = C - EU^{-1}\tilde{F} = C - EU^{-1}L^{-1}F = C - ED^{-1}F. \qquad \square$$

Note that $D = LU$ is factored. However, even in exact factorization, LU is usually sparser than D^{-1}. The submatrices DU^{-1} and $L^{-1}F$ are formed automatically, and the Schur complement is formed implicitly, during the partial Gaussian elimination with respect to the lower part of A.

Dropping strategies similar to those used in Algorithm 3.2 can be applied to Algorithm 3.3, resulting in an incomplete LU factorization with an approximate Schur complement A_1. We formally describe the restricted ILUT factorization as in Algorithm 3.5.

ALGORITHM 3.5. RESTRICTED ILUT(τ, p) FACTORIZATION.

1. *For* $i = 2, n$, *Do*
2. $\quad w := a_{i,\beta}$
3. $\quad$ *For* $k = 1, \min(i-1, m)$ *and when* $w_k \neq 0$, *Do*
4. $\quad\quad w_k := w_k / a_{k,k}$
5. $\quad\quad$ *Set* $w_k := 0$ *if* $|w_k| < \tau * \text{nzavg}(a_{i,\beta})$
6. $\quad\quad$ *If* $w_k \neq 0$, *then*
7. $\quad\quad\quad w := w - w_k * u_{k,\beta}$
8. $\quad\quad$ *End If*
9. $\quad$ *End Do*

10. *Apply a dropping strategy to row w*
11. *Set* $l_{i,j} := w_j$ *for* $j = 1, \ldots, \min(i-1, m)$ *whenever* $w_j \neq 0$
12. *Set* $u_{i,j} := w_j$ *for* $j = \min(i, m), \ldots, n$ *whenever* $w_j \neq 0$
13. *Set* $w := 0$
14. *End Do*

Algorithm 3.5 yields an ILU factorization of the form

$$A = \mathbf{LU} + R, \tag{3.3}$$

where R is the residual matrix representing the difference between A and $\mathbf{LU}$. The ILUT implementation gives an easy representation of the residual matrix.

PROPOSITION 3.6. *The elements of the residual matrix* R *as in* (3.3) *are those elements dropped in Algorithm* 3.5.

Proof. The proof can be formulated from the arguments in [52, p. 274] and [56]. □

Clearly, Algorithm 3.5 will fail when any individual ILUT fails on at least one of the blocks due to zero pivots. There are at least three strategies that can be used to deal with this situation. First, one can use pivoting as in ILUTP [52], a variant of ILUT which incorporates column pivoting. Second, we may use a diagonal threshold strategy as was done in ILUM [56]. In this technique, nodes with small absolute diagonal values are put in the vertex cover. This strategy may reduce the size of the independent set. Third, we may replace a small (absolute) diagonal value by a larger one and proceed with the normal ILUT. The third strategy is suitable and is almost free of cost since ILUT is not an exact factorization anyway. In our implementation we chose the third strategy and replaced a zero diagonal value by a value that is determined by the dropping tolerance and the absolute average nonzero elements of the current row.

We mention that in Algorithm 3.5 the diagonals of the approximate Schur complement (A_1) are not dropped regardless of their values. From Figure 3.1 the accuracy of the EU^{-1} part is related to that of the LU part; the accuracy of the A_1 part is related to that of the $L^{-1}F$ part. It may be profitable to use different dropping parameters (τ, p) for the upper and lower parts of the ILU factorizations in Algorithm 3.5. We did some numerical experiments and did not find overwhelming evidence for supporting the use of different dropping parameter set for most test problems. However, even if other problems may be tested to show certain advantages, the increased difficulty of determining more parameters for a general purpose preconditioner may offset the gain in convergence. Thus, the numerical results reported in this paper all use uniform dropping parameters during the construction phase. We even kept the parameters the same between different levels.

We point out that the inherent parallelism in the construction phase is excellent. The construction of the upper part factorization is parallelizable with respect to individual blocks. The construction of the lower part factorization is fully parallelizable relative to individual rows, as processing each row needs only information from (access to) the upper part. In addition, parallel algorithms for finding independent sets are available [38, 39].

4. Multilevel block ILUT. The BILUTM is based on the restricted ILUT Algorithm 3.5. On each level α, an incomplete LU factorization is performed and an approximate reduced system $A_{\alpha+1}$ is formed as in Algorithm 3.5. Formally, we have

$$\begin{pmatrix} D_\alpha & F_\alpha \\ E_\alpha & C_\alpha \end{pmatrix} = \begin{pmatrix} L_\alpha & 0 \\ E_\alpha U_\alpha^{-1} & I_\alpha \end{pmatrix} \times \begin{pmatrix} U_\alpha & L_\alpha^{-1} F_\alpha \\ 0 & A_{\alpha+1} \end{pmatrix} = \mathbf{L}_\alpha \mathbf{U}_\alpha. \tag{4.1}$$

The whole process of finding block independent sets, permuting the matrix, and performing the restricted ILUT factorization is recursively repeated on the matrix $A_{\alpha+1}$. The recursion is stopped when the last reduced system $A_{\mathcal{L}}$ is small enough. Then a standard ILUT factorization $\mathbf{L}_{\mathcal{L}}\mathbf{U}_{\mathcal{L}}$ is performed on $A_{\mathcal{L}}$ (Algorithm 3.2). However, we do not store any reduced systems on any level, including the last one. Instead, we store two sparse matrices on each level

$$\mathbf{L}_\alpha = \begin{pmatrix} L_\alpha & 0 \\ E_\alpha U_\alpha^{-1} & I_\alpha \end{pmatrix} \qquad \text{and} \qquad \mathbf{U}_\alpha = \begin{pmatrix} U_\alpha & L_\alpha^{-1}F_\alpha \\ 0 & 0 \end{pmatrix}, \quad \text{for} \quad 0 \le \alpha < \mathcal{L}-1,$$

along with the factors $\mathbf{L}_{\mathcal{L}}$ and $\mathbf{U}_{\mathcal{L}}$.

The approximate solution on the last level is obtained by applying one sweep of ILUT of the last reduced system using the factors $\mathbf{L}_{\mathcal{L}}\mathbf{U}_{\mathcal{L}}$. This is different from the implementation of BILUM [55], where the last reduced system is solved to certain accuracy by a Krylov subspace method preconditioned by ILUT. The advantage of BILUTM includes the added flexibility in controlling the amount of fill-in (and the computation costs during the construction), especially when large-sized blocks are used.

Suppose the right-hand side b and the solution vector x are partitioned according to the independent set ordering as in (2.2); then we would have, on each level,

$$x_\alpha = \begin{pmatrix} x_{\alpha,1} \\ x_{\alpha,2} \end{pmatrix} \qquad \text{and} \qquad b_\alpha = \begin{pmatrix} b_{\alpha,1} \\ b_{\alpha,2} \end{pmatrix}.$$

The forward elimination is performed by solving for a temporary vector y_α, i.e., for $\alpha = 0, 1, \ldots, \mathcal{L}-1$, by solving

$$\begin{pmatrix} L_\alpha & 0 \\ E_\alpha U_\alpha^{-1} & I_\alpha \end{pmatrix} \begin{pmatrix} y_{\alpha,1} \\ y_{\alpha,2} \end{pmatrix} = \begin{pmatrix} b_{\alpha,1} \\ b_{\alpha,2} \end{pmatrix}, \text{ with } \begin{array}{lrcl} \text{(F1)}: & y_{\alpha,1} & = & L_\alpha^{-1} b_{\alpha,1}, \\ \text{(F2)}: & y_{\alpha,2} & = & b_{\alpha,2} - E_\alpha U_\alpha^{-1} y_{\alpha,1}. \end{array}$$

We then solve the last reduced system as

$$\mathbf{L}_{\mathcal{L}}\mathbf{U}_{\mathcal{L}} x_{\mathcal{L}} = b_{\mathcal{L}}.$$

A backward substitution is performed to obtain the solution by solving, for $\alpha = \mathcal{L}-1, \ldots, 1, 0$,

$$\begin{pmatrix} U_\alpha & L_\alpha^{-1}F_\alpha \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_{\alpha,1} \\ x_{\alpha,2} \end{pmatrix} = \begin{pmatrix} y_{\alpha,1} \\ y_{\alpha,2} \end{pmatrix}, \text{ with } \begin{array}{lrcl} \text{(B1)}: & x_{\alpha,1} & = & y_{\alpha,1} - L_\alpha^{-1}F_\alpha x_{\alpha,2}, \\ \text{(B2)}: & x_{\alpha,1} & = & U_\alpha^{-1} x_{\alpha,1}. \end{array}$$

The backward substitution will work since $x_{\alpha,2} = x_{\alpha+1,1}$ and $x_{\mathcal{L}-1,2} = x_{\mathcal{L}}$. The preconditioned iteration process is reminiscent of a multigrid V-cycle algorithm [14], see Figure 4.1. A Krylov subspace iteration is performed on the finest level acting as a smoother; the residual is then transferred level-by-level to the coarsest level, where one sweep of ILUT is used to yield an approximate solution. In the current situation, the coarsest-level ILUT is actually a direct solver with limited accuracy comparable to the accuracy of the whole preconditioning process.

Let us rewrite (4.1) as

$$\begin{pmatrix} I_\alpha & 0 \\ E_\alpha U_\alpha^{-1} L_\alpha^{-1} & I_\alpha \end{pmatrix} \times \begin{pmatrix} L_\alpha U_\alpha & 0 \\ 0 & A_{\alpha+1} \end{pmatrix} \times \begin{pmatrix} I_\alpha & U_\alpha^{-1} L_\alpha^{-1} F_\alpha \\ 0 & I_\alpha \end{pmatrix} \tag{4.2}$$

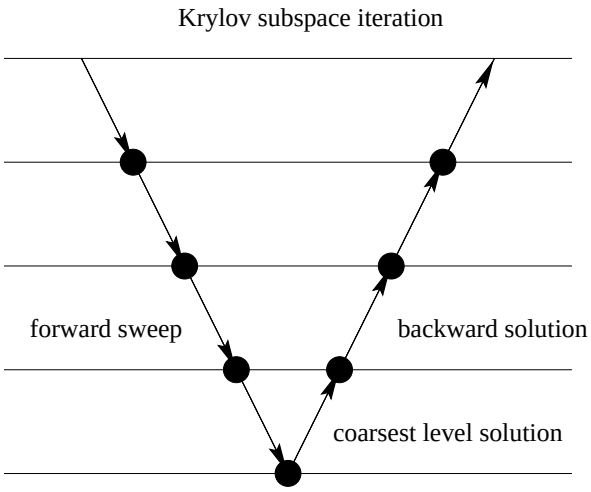


FIG. 4.1. *The multilevel structure of the BILUTM preconditioned Krylov subspace solver.*

and examine a few interesting properties. It is clear that the central part of (4.2) is an operator acting on the full vector, say, x_α ($L_\alpha U_\alpha$ on $x_{\alpha,1}$ and $A_{\alpha+1}$ on $x_{\alpha,2}$). In a two-level analysis, we may define

$$\mathbf{I}_\alpha^{\alpha+1} = \begin{pmatrix} -E_\alpha U_\alpha^{-1} L_\alpha^{-1} & I_\alpha \end{pmatrix} \qquad \text{and} \qquad \mathbf{I}_{\alpha+1}^{\alpha} = \begin{pmatrix} -U_\alpha^{-1} L_\alpha^{-1} F_\alpha \\ I_\alpha \end{pmatrix}$$

as the restriction and interpolation operators, respectively. Then the following results linking BILUTM with the algebraic multigrid methods can be verified directly; see [57].

PROPOSITION 4.1. *Suppose the factorization* (4.2) *exists and is exact. Then*

1. *the reduced system $A_{\alpha+1}$ satisfies the Galerkin condition* (2.1);
2. *if, in addition, A_α is symmetric, then $\mathbf{I}_\alpha^{\alpha+1} = \mathbf{I}_{\alpha+1}^{\alpha}{}^T$.*

The above discussion of the solution procedure omitted the permutation and inverse permutation that must be performed before and after each operation on each level. This is also the approach that we used in our current implementation (and that of BILUM [55]). On the other hand, we may permute the matrices on each level in the construction phase. In this case, only the global permutation is needed before and after the application of the preconditioner [51].

All of the steps of the procedure, including the construction of the preconditioner, are fully parallelizable, except potentially the solution of the last reduced system.[1] For example, the step (F1) of the forward solve with L_α^{-1} can be performed in parallel because only the unknowns within each block are coupled. The same is true for the backward solve with U_α^{-1} in step (B2). All other parts are just matrix-vector operations and vector updates.

Finally, the sparsity of BILUTM depends primarily on the parameter p used to control the amount of fill-in allowed and the size of the block independent sets.

PROPOSITION 4.2. *Let m_α be the size of the block independent set on level α. The number of nonzeros of BILUTM with $\mathcal{L}$ levels of reductions is bounded by $p(2n + \sum_{\alpha=1}^{\mathcal{L}} \alpha m_\alpha)$.*

[1] For this, we may employ a sparse approximate inverse technique [21] or a multicoloring strategy [52] to solve the last reduced system; see [61].

Proof. On each level $0 \le \alpha < \mathcal{L} - 1$, the L and U factors of the upper part have at most p elements in each row. The left-hand side of the lower part also has at most p elements in each row; the right-hand side (the reduced system) is not stored. Suppose the sizes of the block independent set and the vertex cover are m_α and r_α, respectively. On level α, the total number of nonzeros is bounded by $2pm_\alpha + pr_\alpha$. Since the last reduced system is factored by ILUT(τ, p), the amount of nonzeros is bounded by $2pn_{\mathcal{L}} = 2pm_{\mathcal{L}}$ and $r_{\mathcal{L}} = 0$. Summing all levels yields the bound

$$\sum_{\alpha=0}^{\mathcal{L}-1}(2pm_\alpha + pr_\alpha) + 2pm_{\mathcal{L}} = p\left(2\sum_{\alpha=0}^{\mathcal{L}} m_\alpha + \sum_{\alpha=0}^{\mathcal{L}-1} r_\alpha\right). \tag{4.3}$$

Since the nodes of all independent sets and the last reduced system constitute the order of the matrix, we have

$$\sum_{\alpha=0}^{\mathcal{L}} m_\alpha = n. \tag{4.4}$$

Note that $r_\alpha = m_{\alpha+1} + r_{\alpha+1}$ for $0 \le \alpha < \mathcal{L} - 1$, and $r_{\mathcal{L}} = 0$. It is easy to verify

$$\sum_{\alpha=0}^{\mathcal{L}-1} r_\alpha = \sum_{\alpha=1}^{\mathcal{L}} \alpha m_\alpha. \tag{4.5}$$

Substituting (4.4) and (4.5) into (4.3) gives the bound for the number of nonzeros of BILUTM as

$$2pn + p\sum_{\alpha=1}^{\mathcal{L}} \alpha m_\alpha = p\left(2n + \sum_{\alpha=1}^{\mathcal{L}} \alpha m_\alpha\right). \tag{4.6}$$

□

We remark that in (4.6), the term $2pn$ is the bound for the number of nonzeros of the standard ILUT. Due to the block structure of BILUTM, the first few rows of each block of the upper L factors have less than p elements and the overall nonzero number of BILUTM is actually smaller. The term $p\sum_{\alpha=1}^{\mathcal{L}} \alpha m_\alpha$ represents the extra nonzeros for the multilevel implementation. Note that m_0 is not in the second term and the factor α grows as the level increases. It is therefore advantageous to have large block independent sets in the first few levels.

5. Numerical experiments. Implementations of ILUM and BILUM have been described in detail in [51, 55]. One significant difference between BILUTM and BILUM and ILUM is that we do not use an inner iteration to solve the last reduced system. Instead, backward and forward solution steps are performed with the incomplete LU factors $\mathbf{L}_{\mathcal{L}}\mathbf{U}_{\mathcal{L}}$ of the last reduced system. Unless otherwise explicitly indicated, we used the following default parameters for our preconditioned iterative solver: GMRES with a restart value of 50 was used as the accelerator; the maximum number of reductions (levels) allowed was 10, i.e., $\mathcal{L} = 10$; the threshold dropping tolerance was set to be $\tau = 10^{-4}$, and the block sizes were chosen to be equal to the parameter p used to control the number of fill-in elements.

All matrices were considered general sparse matrices and any available structures were not exploited. The right-hand side was generated by assuming that the solution is a vector of all ones and the initial guess was a vector of some random numbers. The computations were terminated when the 2-norm of the residual was reduced by a factor

TABLE 5.1
Comparison of BILUTM and ILUT for solving the convection-diffusion problem with different Re.

	BILUTM					ILUT				
Re	p	iter	total	solu	spar	p	iter	total	solu	spar
1	10	56	30.8	16.1	3.53	8	58	15.0	14.2	3.21
10	10	63	32.2	17.6	3.53	8	42	13.1	11.9	3.21
100	10	39	25.4	10.6	3.55	9	21	6.83	5.30	3.60
1000	10	13	17.5	2.80	3.39	9	5	2.09	1.08	3.32
10000	20	22	24.1	6.40	5.76	17	22	10.4	7.18	5.82
100000	100	43	43.8	22.5	15.2	180	25	192.5	44.7	71.5

of 10^7. We also set an upper bound of 100 for the GMRES iteration. The numerical experiments were conducted on a Power-Challenge XL Silicon Graphics workstation equipped with 512 MB of main memory, two 190 MHZ R10000 processors, and 1 MB secondary cache.

In all tables with numerical results, "bsize" is the size of the uniform blocks (used only when bsize $\neq p$), "iter" shows the number of GMRES iterations, "total" shows the CPU time in seconds for the preprocessing and solution phases, "solu" shows the CPU time for the solution phase only, and "spar" shows the sparsity ratio which is the ratio between the number of nonzeros of the preconditioner to that of the original matrix. The symbol "–" indicates lack of convergence. We mainly compare BILUTM with (single-level) ILUT and sometimes BILUTM with different parameters.

Convection-diffusion problem. We first consider a convection-diffusion problem

$$u_{xx} + u_{yy} + \mathrm{Re}(\exp(xy-1)u_x - \exp(-xy)u_y) = 0, \tag{5.1}$$

defined on the unit square. Here Re is the so-called Reynolds number. A Dirichlet boundary condition was assumed, but the linear systems used the artificially generated right-hand side as stated above. We used the standard 5-point central difference discretization scheme with a uniform mesh $h = 1/201$. The resulting matrices with different values of Re have 40,000 unknowns and 199,200 nonzeros. The percentage of the rows with diagonal dominance becomes smaller as Re increases [63].

Table 5.1 gives some performance data of BILUTM and ILUT for solving (5.1) with different Re. Here p was varied so that BILUTM and ILUT used approximately the same storage space. There was an exception for $\mathrm{Re} = 10^5$ when ILUT did not converge for $p < 180$ while BILUTM converged for $p = 100$. It can be seen that for simple problems (small Re), ILUT was more efficient than BILUTM. They performed similarly for $\mathrm{Re} = 10^4$. For $\mathrm{Re} \geq 10^5$, ILUT failed to converge unless it used very large storage space. Conversely, BILUTM did very well for this difficult problem.

Note that BILUTM took more steps to converge for small Re problems. However, note also that the inherent parallelism in BILUTM is far superior to that in ILUT. Table 5.2 gives another set of tests with larger values for p. We see that BILUTM performed better than ILUT (solution time) with high accuracy. This improvement comes without sacrificing potential for parallelism but the cost of preprocessing increased somewhat.

The next test is to show how the performance of BILUTM is affected by the block sizes. Here we chose $p = 10$ and $\mathrm{Re} = 10^3$. The size of the uniform blocks varied from 1 to 400. Note that for the large block sizes, we actually had only three levels of reductions. The results are given in Table 5.3.

We note that for most values of the block size, the performance of BILUTM has no significant difference. This property is desirable since it implies that, for this test

TABLE 5.2
Comparison of high accuracy BILUTM and ILUT for solving the convection-diffusion problem with different Re.

	BILUTM					ILUT				
Re	p	iter	total	solu	spar	p	iter	total	solu	spar
1	50	10	14.1	3.29	9.30	24	13	9.63	5.27	9.47
10	50	11	14.3	3.64	9.29	24	14	10.1	5.71	9.46
100	50	7	12.5	2.21	8.83	22	7	6.54	2.62	8.57
1000	50	3	7.91	0.76	5.93	15	3	1.76	0.71	3.78
10000	50	6	14.2	2.04	9.95	43	6	8.47	2.38	9.81

TABLE 5.3
Performance of BILUTM$(10^{-4}, 10)$ *as a function of the block size. Convection-diffusion problem with* $\mathrm{Re} = 10^3$.

bsize	1	5	10	30	50	90	170	200	250	290	350	380	400
iter	17	14	13	12	11	11	11	11	11	12	11	12	12
total	112	28.9	17.6	9.3	7.5	6.6	6.9	7.4	8.8	8.5	10.2	14.3	11.5
solu	3.8	3.1	2.8	2.4	2.2	2.2	2.2	2.2	2.2	2.4	2.2	2.4	2.4
spar	2.5	3.2	3.4	3.2	3.2	3.3	3.4	3.4	3.4	3.4	3.4	3.5	3.5

TABLE 5.4
Description of the TOKAMAK matrices.

Name	Unknowns	Nonzeros	Condition number	Diagonal dominance
UTM300	300	3 155	1.50(+06)	no
UTM1700a	1 700	21 313	6.24(+06)	no
UTM1700b	1 700	21 509	1.16(+07)	no
UTM3060	3 060	42 211	3.94(+07)	no
UTM5940	5 940	83 842	1.91(+09)	no

TABLE 5.5
Comparison of BILUTM and ILUT for solving the first four TOKAMAK matrices.

	BILUTM					ILUT				
Matrices	p	iter	total	solu	spar	p	iter	total	solu	spar
UTM300	20	26	0.11	0.045	4.25	20	17	0.039	0.021	2.38
UTM1700a	20	36	1.09	0.63	3.98	30	30	0.82	0.42	3.64
UTM1700b	20	27	0.86	0.44	3.82	30	29	0.77	0.40	3.56
UTM3060	30	26	2.18	0.99	4.70	38	25	1.90	0.99	4.63

problem and with current test conditions, the convergence rate of BILUTM would not be very sensitive to the number of processors had our test been implemented on a parallel computer.

TOKAMAK matrices. The TOKAMAK matrices are real unsymmetric matrices which arise from nuclear fusion plasma simulations in a tokamak reactor.[2] These are part of the SPARSKIT collections and have been provided by P. Brown of Lawrence Livermore National Laboratory. Table 5.4 shows some data on these matrices.

The solution details for the first four TOKAMAK matrices are listed in Table 5.5, and those for UTM5940 are listed in Table 5.6. We note that for the first three matrices of small sizes, ILUT seemed to outperform BILUTM, given a similar memory consumption. They were almost tied for UTM3060. For the largest TOKAMAK matrix, BILUTM performed much better than ILUT by virtually all measures (Table 5.6). In fact, ILUT could not converge for $p \leq 70$, while BILUTM still converged

[2]The TOKAMAK matrices are available online from the matrix market of the National Institute of Standards Technology at http://math.nist.gov/MatrixMarket.

TABLE 5.6
Solving the UTM5940 matrix by BILUTM and ILUT with different parameters.

BILUTM						ILUT					
p	τ	iter	total	solu	spar	p	τ	iter	total	solu	spar
100	10^{-4}	19	10.4	2.29	9.93	130	10^{-4}	25	14.5	4.04	13.5
90	10^{-4}	21	9.05	2.38	9.13	120	10^{-4}	28	13.8	4.32	12.7
80	10^{-4}	23	8.67	2.56	8.78	110	10^{-4}	31	13.4	4.57	11.8
70	10^{-4}	26	8.82	2.79	8.21	100	10^{-4}	35	12.8	4.90	10.9
60	10^{-4}	26	6.64	2.52	7.13	90	10^{-4}	37	11.8	4.90	9.96
50	10^{-4}	27	6.62	2.49	6.45	80	10^{-4}	46	12.1	5.78	8.94
40	10^{-4}	36	6.58	3.14	5.64	70	10^{-5}	–	–	–	–
30	10^{-4}	75	8.65	6.05	4.72	70	10^{-4}	–	–	–	–
20	10^{-4}	96	8.64	6.82	3.44	70	10^{-3}	–	–	–	–

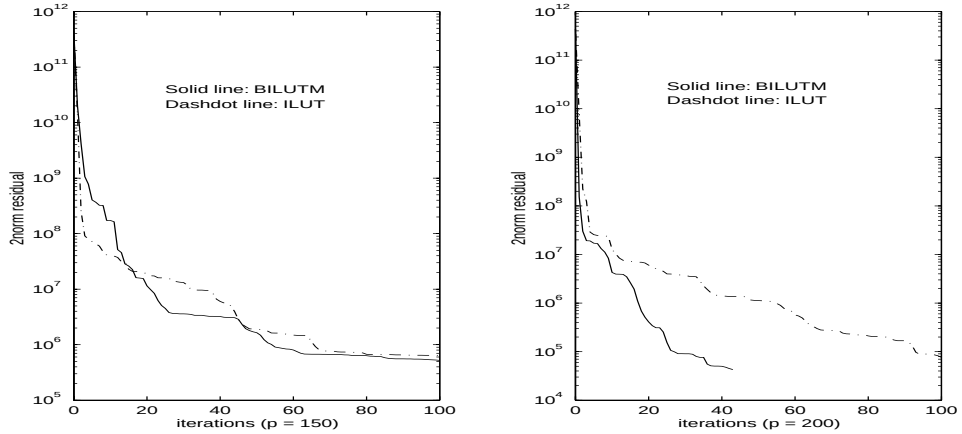


FIG. 5.1. *Convergence history of BILUTM and ILUT for solving the RAEFSKY4 matrix.*

with $p = 20$. It can be seen that BILUTM needed less than half the storage required for ILUT to converge. With more storage space made available for ILUT, BILUTM still outperformed ILUT with a faster convergence rate (and less memory consumption.)

RAEFSKY4 matrix. The RAEFSKY4 matrix[3] has 19,779 unknowns and 1,328,611 nonzeros. It is from the buckling problem for the container model and was supplied by H. Simon from Lawrence Berkeley National Laboratory (originally created by A. Raefsky from Centric Engineering). This is probably the hardest one in the total of six RAEFSKY matrices. (BILUM with diagonal threshold techniques was able to solve all but the RAEFSKY4 matrix [56].) In order for BILUTM to converge fast, we found it necessary to use a larger restart value (100) for GMRES. Figure 5.1 shows the convergence history of BILUTM and ILUT with $p = 150$ and 200, respectively. In both tests, the block size was 200 for BILUTM. We note that with $p = 150$, both preconditioners had similar lack of full convergence. However, with $p = 200$, BILUTM converged in 43 iterations while ILUT was still not fully converged in 100 iterations.

Figure 5.2 depicts performance comparisons when BILUTM was used with differ-

[3]The RAEFSKY matrices are available online from the University of Florida sparse matrix collection [24] at http://www.cise.ufl.edu/~davis/sparse.

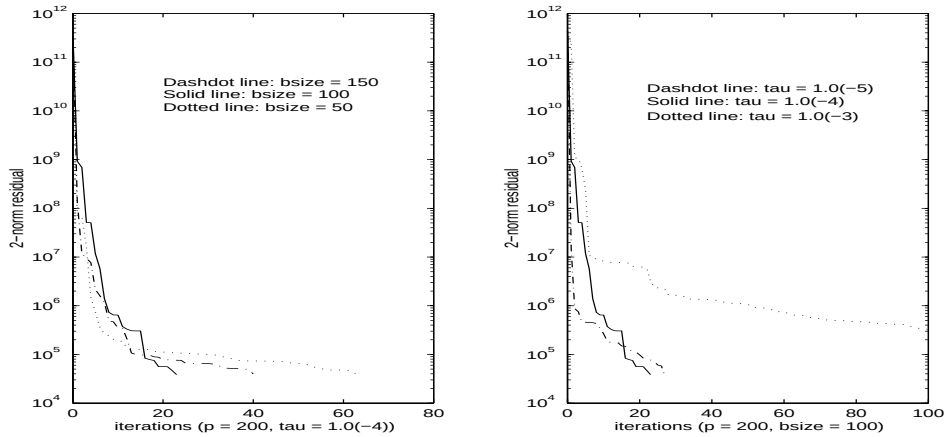


FIG. 5.2. *Convergence history of BILUTM with different parameters for solving the RAEFSKY4 matrix. Left: different block size. Right: different dropping threshold τ.*

TABLE 5.7
Solving the WIGTO966 matrix by BILUTM and ILUT with different parameters.

BILUTM							ILUT					
bsize	p	τ	iter	total	solu	spar	p	τ	iter	total	solu	spar
200	200	10^{-5}	7	29.6	1.39	6.21	400	10^{-4}	16	72.0	5.05	9.65
100	200	10^{-5}	8	27.8	1.70	6.69	400	10^{-3}	18	52.8	5.14	8.57
100	150	10^{-5}	11	21.3	2.10	5.87	360	10^{-5}	18	76.8	5.21	9.48
100	100	10^{-5}	33	18.5	5.17	4.45	360	10^{-4}	20	68.7	5.64	9.17
100	100	10^{-4}	44	19.7	6.96	4.43	360	10^{-3}	33	61.4	8.97	8.71
70	70	10^{-5}	25	11.0	3.20	3.17	340	10^{-5}	28	76.6	7.91	9.14
30	60	10^{-5}	43	11.4	5.27	2.80	340	10^{-4}	44	76.2	13.2	8.92
30	60	10^{-4}	41	11.0	4.94	2.74	340	10^{-3}	42	59.5	10.9	8.14
30	40	10^{-4}	86	12.3	8.52	2.02	320	10^{-5}	–	–	–	–
20	40	10^{-5}	93	12.7	8.96	1.89	320	10^{-4}	41	71.3	11.9	8.59
20	40	10^{-4}	86	11.8	8.16	1.86	320	10^{-3}	39	54.2	10.5	7.90
20	35	10^{-4}	89	11.3	8.04	1.70	300	10^{-4}	–	–	–	–
20	35	10^{-3}	95	11.4	8.33	1.60	300	10^{-3}	–	–	–	–

ent parameters. The left part of Figure 5.2 shows that BILUTM with block size 100 gave the best results. Larger and smaller block sizes resulted in deterioration of convergence. The right part of Figure 5.2 shows that $\tau = 10^{-4}$ was the best among the three values tested for this parameter. It is interesting to note that higher accuracy ($\tau = 10^{-5}$) did not yield faster convergence.

WIGTO966 matrix. The WIGTO966 matrix[4] has 3,864 unknowns and 238,252 nonzeros. It comes from an Euler equation model and was supplied by L. Wigton from Boeing. It is solvable by ILUT with large values of p [19]. This matrix was also used to compare BILUM with ILUT in [54] and to test point and block preconditioning techniques in [20, 22]. BILUM (with GMRES(10)) was shown to be six times faster than ILUT with only one-third of the memory required by ILUT [54]. In our current tests, we chose several values for τ and p for BILUTM and ILUT, and the size of the blocks in case of BILUTM. We tabulate the results in Table 5.7. Amazingly BILUTM converged for this problem with a sparsity ratio of 1.60. The smallest sparsity ratio that yields convergence for ILUT is 7.90. In addition, BILUTM converged almost five

[4]The WIGTO966 matrix is available from the authors.

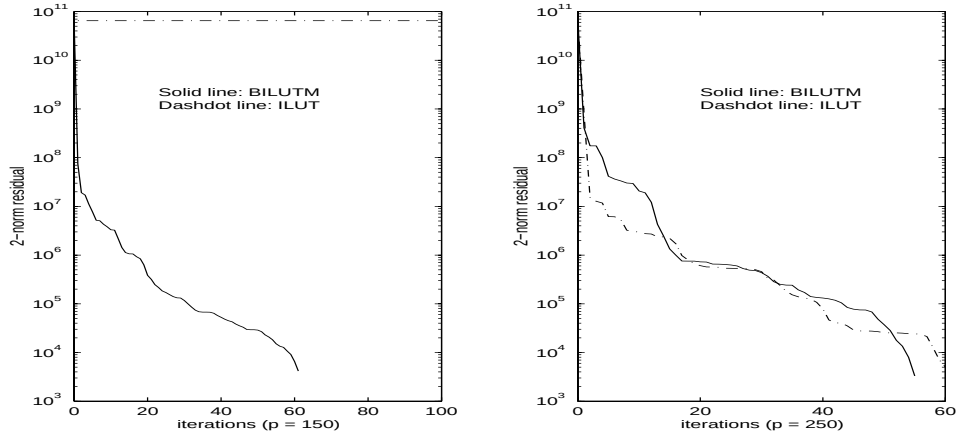


FIG. 5.3. *Convergence history of BILUTM and ILUT with different amount of fill-in (p) for solving the OLAFU matrix.*

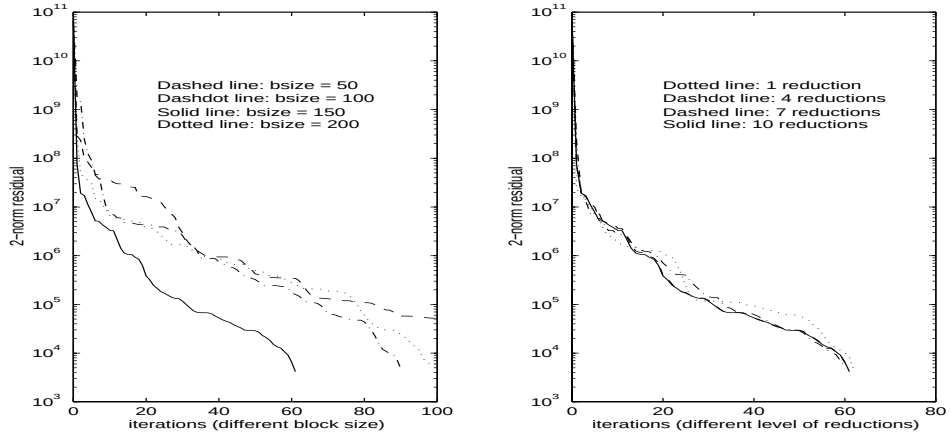


FIG. 5.4. *Convergence history of BILUTM with different parameters and* $p = 150$ *for solving the OLAFU matrix. Left: different block size. Right: different level of reductions.*

times faster (total CPU time) than ILUT and used just about one-fifth of the memory that was required by ILUT.

OLAFU matrix. The OLAFU matrix[5] has 16,146 unknowns and 1,015,156 nonzeros. It is a structural modeling problem from NASA Langley. The tests with OLAFU also used GMRES(100) as the accelerator. Figure 5.3 shows the comparison between BILUM and ILUT with two different values of p. We point out that with $p = 150$, ILUT did not show any signs of convergence (see the left part of Figure 5.3), while BILUTM converged within 61 iterations. The right part of Figure 5.3 shows that ILUT did converge with more fill-in ($p = 200$), but BILUTM was still faster. These results indicate that the OLAFU matrix cannot be solved by ILUT without sufficient accuracy. We remark that for the comparison shown in Figure 5.3, both BILUTM

[5]The OLAFU matrix is available online from the University of Florida sparse matrix collection [24] at http://www.cise.ufl.edu/~davis/sparse.

TABLE 5.8
Description of the BARTH matrices.

Name	Unknowns	Nonzeros	Descriptions
BARTHT1A	14 075	481 125	Small airfoil 2D Navier–Stokes, distance 1
BARTHS1A	15 735	539 225	Small airfoil 2D Navier–Stokes, distance 1
BARTHT2A	14 075	1 311 725	Small airfoil 2D Navier–Stokes, distance 2
BARTHS2A	15 735	1 510 325	Small airfoil 2D Navier–Stokes, distance 2

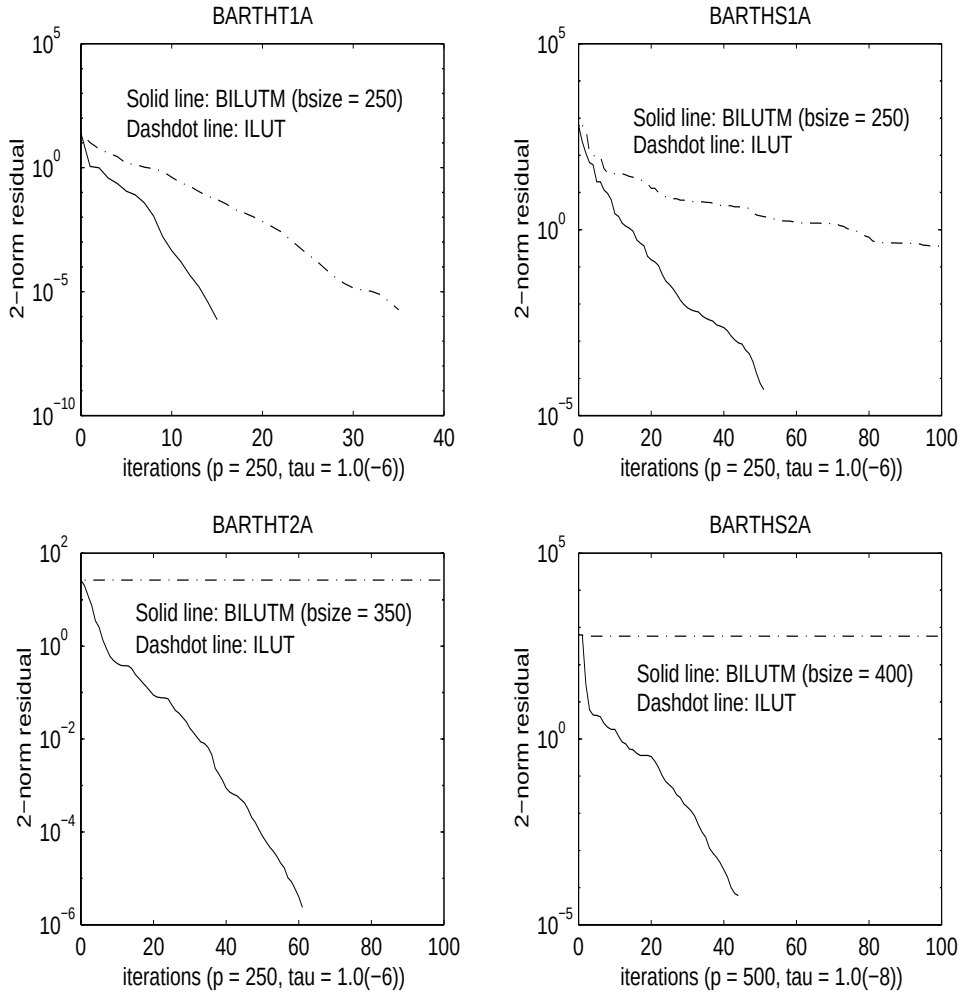


FIG. 5.5. *Convergence history of BILUTM and ILUT for solving the BARTH matrices.*

and ILUT used approximately the same memory space for the same value of p.

In Figure 5.4, we show test results of using BILUTM with different parameters to solve OLAFU. The parameter $p = 150$ was fixed. The left part of Figure 5.4 shows that the size of the blocks did affect the convergence of BILUTM for this large problem. It seems that taking the block size to be equal to the fill-in parameter p yielded the best results. The right part of Figure 5.4 indicates that the number of levels (reductions) did not have significant effect on the convergence of BILUTM. This is probably because the largest independent set had been factored out in the

first level and the use of ILUT on the coarsest level gave comparable accuracy. Recall that there is a big difference between one level of reduction and no reduction at all, since Figure 5.3 shows that BILUTM without reduction (actually equivalent to ILUT) failed to converge.

BARTH matrices. The BARTH matrices[6] were supplied by T. Barth of NASA Ames. They are for a two-dimensional high Reynolds number airfoil problem, with one equation turbulence model. The S and T matrices are results of using different grids. The grid of the T matrices has a concentration of elements unrealistically close to the airfoil. The four BARTH matrices are described in Table 5.8. Note that in order for ILUT and BILUTM to work properly, zero diagonals are added. The BARTH matrices have been used as test matrices for other ILU-type techniques in [19], but none of them has been solved by enhanced BILUM techniques [57], partly because of the prohibitive computation and memory costs associated with the use of very large-sized blocks (on the given computer).

We present in Figure 5.5 only one set of comparisons of BILUTM and ILUT by solving the BARTH matrices using large size blocks and GMRES(100). We remark that for this set of test parameters, ILUT took about three times more CPU time (BARTHT1A) and used about 20% more memory than BILUTM did. We found that BILUTM converged much faster than ILUT for these indefinite matrices with small and zero diagonals. For the two largest BARTH matrices, ILUT almost completely failed to reduce the residual norm within 100 iterations, while BILUTM converged satisfactorily.

6. Concluding remarks. We have presented a mulitlevel block ILU preconditioner with a dual dropping strategy for solving general sparse matrices. The method offers flexibility in controlling the amount of fill-in during the ILU factorization when large size blocks are used for domain decomposition based implementation of multilevel ILU preconditioning method. A particular merit of BILUTM is that both the construction and application phases of the preconditioner have a high level of inherent parallelism.

We gave an upper bound for the number of nonzeros of the preconditioner. We showed that the extra storage costs of multilevel implementation are not substantial if large block independent sets can be found in the first few reductions. It may also be beneficial not to have too many levels, especially when the size of the block independent sets becomes small.

Our numerical experiments with several matrices show that the proposed technique indeed demonstrates the anticipated flexibility and effectiveness. As a parallelizable high accuracy preconditioner, BILUTM is comparable with sequential ILUT for solving easy problems. For some difficult problems, where high accuracy preconditioning is a must, BILUTM is more robust and is more efficient than ILUT and usually requires less memory. In other words, this preconditioner does not sacrifice convergence in order to improve parallelism. This is in sharp contrast with lower order preconditioner such as ILU(0) or the high-order single-level preconditioner such as ILUT.

Although we did not directly compare BILUTM with other multilevel preconditioning techniques, we did remark that BILUTM solved several difficult matrices that might not be solved by BILUM efficiently on the given computer because of the computation and memory costs associated with the use of very large size blocks.

[6]The BARTH matrices are available from the authors.

Implementations of grid-based multilevel methods on parallel and vector computers can be found in [41] (for structured matrices) and those of domain-based (two-level) methods can be found in [53, 54]. Those and other implementations on shared-memory machines [11] demonstrate the advantage of the inherent parallelism of the multilevel preconditioning methods. Conversely, implementing multilevel preconditioning methods on distributed-memory machines requires the consideration of cost trade-off between communications and computations. It is obviously not advantageous to have too small blocks and too many levels. The parallel solution of the last reduced system may also be desirable in certain applications.

REFERENCES

[1] E. C. Anderson and Y. Saad, *Solving sparse triangular systems on parallel computers*, Internat. J. High Speed Comput., 1 (1989), pp. 73–96.

[2] O. Axelsson and M. Neytcheva, *Algebraic multilevel iteration method for Stieltjes matrices*, Numer. Linear Algebra Appl., 1 (1994), pp. 216–236.

[3] O. Axelsson and P. S. Vassilevski, *A survey of multilevel preconditioned iterative methods*, BIT, 29 (1989), pp. 769–793.

[4] O. Axelsson and P. S. Vassilevski, *Algebraic multilevel preconditioning methods.* II, SIAM J. Numer. Anal., 27 (1990), pp. 1569–1590.

[5] V. A. Bandy, *Black Box Multigrid for Convection-Diffusion Equations on Advanced Computers*, Ph.D. thesis, University of Colorado at Denver, Denver, CO, 1996.

[6] R. E. Bank and R. K. Smith, *The incomplete factorization multigraph algorithm*, SIAM J. Sci. Comput., 20 (1999), pp. 1349–1364.

[7] R. E. Bank and C. Wagner, *Multilevel ILU decomposition*, Numer. Math., to appear.

[8] R. E. Bank and J. Xu, *The hierarchical basis multigrid method and incomplete LU decomposition*, in the Seventh International Symposium on Domain Decomposition Methods for Partial Differential Equations, D. Keyes and J. Xu, eds., AMS, Providence, RI, 1994, pp. 163–173.

[9] M. Benzi and M. Tuma, *A sparse approximate inverse preconditioner for nonsymmetric linear systems*, SIAM J. Sci. Comput., 19 (1998), pp. 968–994.

[10] E. F. F. Botta, K. Dekker, Y. Notay, A. van der Ploeg, C. Vuik, F. W. Wubs, and P. M. de Zeeuw, *How fast the Laplace equation was solved in* 1995, Appl. Numer. Math., 32 (1997), pp. 439–455.

[11] E. F. F. Botta, A. van der Ploeg, and F. W. Wubs, *A fast linear system solver for large unstructured problems on a shared-memory computer*, in Proceedings of the Conference on Algebraic Multilevel Methods with Applications, O. Axelsson and B. Polman, eds., University of Nijmegen, The Netherlands, 1996, pp. 105–116.

[12] E. F. F. Botta and F. W. Wubs, *Matrix renumbering ILU: An effective algebraic multilevel ILU-preconditioner for sparse matrices*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 1007–1026.

[13] D. Braess, *Towards algebraic multigrid for elliptic problems of second order*, Computing, 55 (1995), pp. 379–393.

[14] W. L. Briggs, *A Multigrid Tutorial*, SIAM, Philadelphia, PA, 1987.

[15] T. F. Chan, S. Go, and J. Zou, *Multilevel Domain Decomposition and Multigrid Methods for Unstructured Meshes: Algorithms and Theory*, Technical report CAM 95-24, Department of Mathematics, UCLA, Los Angeles, CA, 1995.

[16] T. F. Chan, W. P. Tang, and W. L. Wan, *Wavelet sparse approximate inverse preconditioners*, BIT, 37 (1997), pp. 644–660.

[17] Q. S. Chang, Y. S. Wong, and L. Z. Feng, *New interpolation formulas of using geometric assumptions in the algebraic multigrid method*, Appl. Math. Comput., 50 (1992), pp. 223–254.

[18] Q. S. Chang, Y. S. Wong, and H. Q. Fu, *On the algebraic multigrid method*, J. Comput. Phys., 125 (1996), pp. 279–292.

[19] A. Chapman, Y. Saad, and L. Wigton, *High-Order ILU Preconditioners for CFD Problems*, Technical report UMSI 96/14, Minnesota Supercomputer Institute, University of Minnesota, Minneapolis, MN, 1996.

[20] E. Chow and M. A. Heroux, *An object-oriented framework for block preconditioning*, ACM Trans. Math. Software, 24 (1998), pp. 159–183.

[21] E. Chow and Y. Saad, *Approximate inverse techniques for block-partitioned matrices*, SIAM J. Sci. Comput., 18 (1997), pp. 1657–1675.
[22] E. Chow and Y. Saad, *Experimental study of ILU preconditioners for indefinite matrices*, J. Comput. Appl. Math., 86 (1997), pp. 387–414.
[23] E. Chow and Y. Saad, *Approximate inverse preconditioners via sparse-sparse iterations*, SIAM J. Sci. Comput., 19 (1998), pp. 995–1023.
[24] T. Davis, *University of Florida sparse matrix collection*, NA Digest, 97(23), June 7, 1997.
[25] E. F. D'Azevedo, P. A. Forsyth, and W. P. Tang, *Ordering methods for preconditioned conjugate gradient methods applied to unstructured grid problems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 944–961.
[26] E. F. D'Azevedo, P. A. Forsyth, and W. P. Tang, *Towards a cost effective ILU preconditioner with high level fill*, BIT, 31 (1992), pp. 442–463.
[27] P. M. de Zeeuw, *Matrix-dependent prolongations and restrictions in a blackbox multigrid solver*, J. Comput. Appl. Math., 33 (1990), pp. 1–25.
[28] J. E. Dendy, Jr., *Black box multigrid*, J. Comput. Phys., 48 (1982), pp. 366–386.
[29] I. S. Duff and G. A. Meurant, *The effect of reordering on preconditioned conjugate gradients*, BIT, 29 (1989), pp. 635–657.
[30] L. C. Dutto, *The effect of reordering on the preconditioned GMRES algorithm for solving the compressible Navier-Stokes equations*, Internat. J. Numer. Methods Engrg., 36 (1993), pp. 457–497.
[31] H. C. Elman, *Approximate Schur complement preconditioners on serial and parallel computers*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 581–605.
[32] H. C. Elman and E. Agron, *Ordering techniques for the preconditioned conjugate gradient method on parallel computers*, Comput. Phys. Commun., 53 (1989), pp. 253–269.
[33] K. Gallivan, A. Sameh, and Z. Zlatev, *A parallel hybrid sparse linear system solver*, Comput. Systems Engrg., 1 (1990), pp. 183–195.
[34] J. A. George and J. W. Liu, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1981.
[35] N. I. M. Gould and J. A. Scott, *Sparse approximate-inverse preconditioners using norm-minimization techniques*, SIAM J. Sci. Comput., 19 (1998), pp. 605–625.
[36] M. Griebel and G. Starke, *Multilevel preconditioning based on discrete symmetrization for convection-diffusion equations*, J. Comput. Appl. Math., 83 (1997), pp. 165–183.
[37] M. J. Grote and T. Huckle, *Parallel preconditioning with sparse approximate inverses*, SIAM J. Sci. Comput., 18 (1998), pp. 838–853.
[38] M. T. Jones and P. E. Plassman, *A parallel graph coloring heuristic*, SIAM J. Sci. Comput., 14 (1993), pp. 654–669.
[39] M. Luby, *A simple parallel algorithm for the maximal independent set problem*, SIAM J. Comput., 15 (1986), pp. 1036–1053.
[40] J. A. Meijerink and H. A. van der Vorst, *An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix*, Math. Comp., 31 (1977), pp. 148–162.
[41] M. Neytcheva, A. Padiy, M. Mellaard, K. Georgiev, and O. Axelsson, *Scalable and Optimal Iterative Solvers for Linear and Nonlinear Problems*, Technical report MRI 9613, Mathematical Research Institute, University of Nijmegen, The Netherlands, 1996.
[42] Y. Notay and Z. Ould Amar, *Incomplete factorization preconditioning may lead to multigrid like speed of convergence*, in Advanced Mathematics: Computation and Applications, A. S. Alekseev and N. S. Bakhvalov, eds., NCC Publisher, Novosibirsk, Russia, 1996, pp. 435–446.
[43] Y. Notay and Z. Ould Amar, *A nearly optimal preconditioning based on recursive red-black orderings*, Numer. Linear Algebra Appl., 4 (1997), pp. 369–391.
[44] O. Orterby and Z. Zlatev, *Direct Methods for Sparse Matrices*, Springer-Verlag, New York, 1983.
[45] A. A. Reusken, *Multigrid with matrix-dependent transfer operators for convection-diffusion problems*, in Multigrid Method IV, Proceedings of the Fourth European Multigrid Conference, P. W. Hemker and P. Wesseling, eds., International Series of Numerical Mathematics 116, Birkhäuser Verlag, Basel, 1994, pp. 269–280.
[46] A. A. Reusken, *Approximate Cyclic Reduction Preconditioning*, Technical report RANA 97-02, Department of Mathematics and Computing Science, Eindhoven University of Technology, The Netherlands, 1997.
[47] J. W. Ruge and K. Stüben, *Algebraic multigrid*, in Multigrid Methods, Frontiers Appl. Math. 3, S. McCormick, ed., SIAM, Philadelphia, PA, 1987, pp. 73–130.
[48] Y. Saad, *Krylov subspace methods on supercomputers*, SIAM J. Sci. Stat. Comput., 10 (1989),

pp. 1200–1232.
[49] Y. Saad, *Highly parallel preconditioners for general sparse matrices*, in Recent Advances in Iterative Methods, IMA Volumes in Mathematics and Its Applications 60, G. Golub, M. Luskin, and A. Greenbaum, eds., Springer Verlag, New York, NY, 1994, pp. 165–199.
[50] Y. Saad, *ILUT: A dual threshold incomplete ILU preconditioner*, Numer. Linear Algebra Appl., 1 (1994), pp. 387–402.
[51] Y. Saad, *ILUM: A multi-elimination ILU preconditioner for general sparse matrices*, SIAM J. Sci. Comput., 17 (1996), pp. 830–847.
[52] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing, Boston, MA, 1996.
[53] Y. Saad and M. Sosonkina, *Distributed Schur Complement Techniques for General Sparse Linear Systems*, Technical report UMSI 97/159, Minnesota Supercomputer Institute, University of Minnesota, Minneapolis, MN, 1997.
[54] Y. Saad, M. Sosonkina, and J. Zhang, *Domain decomposition and multi-level type techniques for general sparse linear systems*, in Domain Decomposition Methods 10, Contemporary Mathematics 218, J. Mandel, C. Farhat, and X.-C. Cai, eds., AMS, Providence, RI, 1998, pp. 174–190.
[55] Y. Saad and J. Zhang, *BILUM: Block versions of multielimination and multilevel ILU preconditioner for general sparse linear systems*, SIAM J. Sci. Comput., 20 (1999), pp. 2103–2121.
[56] Y. Saad and J. Zhang, *Diagonal Threshold Techniques in Robust Multi-Level ILU Preconditioners for General Sparse Linear Systems*, Technical report UMSI 98/7, Minnesota Supercomputer Institute, University of Minnesota, Minneapolis, MN, 1998.
[57] Y. Saad and J. Zhang, *Enhanced Multi-Level Block ILU Preconditioning Strategies for General Sparse Linear Systems*, Technical report UMSI 98/98, Minnesota Supercomputer Institute, University of Minnesota, Minneapolis, MN, 1998.
[58] H. A. van der Vorst, *A vectorizable version of some ICCG methods*, SIAM J. Sci. Stat. Comput., 3 (1982), pp. 350–356.
[59] P. Vanek, J. Mandel, and M. Brezina, *Algebraic multigrid by smoothed aggregation for second and fourth order elliptic problems*, Computing, 56 (1996), pp. 179–196.
[60] D. P. Young, R. G. Melvin, F. T. Johnson, J. E. Bussoletti, L. B. Wigton, and S. S. Samant, *Application of sparse matrix solvers as effective preconditioners*, SIAM J. Sci. Stat. Comput., 10 (1989), pp. 1186–1199.
[61] J. Zhang, *Sparse Approximate Inverse and Multi-Level Block ILU Preconditioning Techniques for General Sparse Matrices*, Technical report 279-98, Department of Computer Science, University of Kentucky, Lexington, KY, 1998.
[62] J. Zhang, *A Sparse Approximate Inverse Technique for Parallel Preconditioning of General Sparse Matrices*, Technical report 281-98, Department of Computer Science, University of Kentucky, Lexington, KY, 1998.
[63] J. Zhang, *Preconditioned iterative methods and evaluation of finite difference schemes for convection-diffusion*, Appl. Math. Comput., to appear.
[64] Z. Zlatev, *Use of iterative refinement in the solution of sparse linear systems*, SIAM J. Numer. Anal., 19 (1982), pp. 381–399.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 1, pp. 300–312
© 1999 Society for Industrial and Applied Mathematics

THE CHANGE-OF-VARIABLES FORMULA USING MATRIX VOLUME*

ADI BEN-ISRAEL†

Abstract. The matrix volume is a generalization, to rectangular matrices, of the absolute value of the determinant. In particular, the matrix volume can be used in change-of-variables formulae, instead of the determinant (if the Jacobi matrix of the underlying transformation is rectangular). This result is applicable to integration on surfaces, illustrated here by several examples.

Key words. determinants, Jacobians, matrix volume, change-of-variables in integration, surface integrals, Radon transform, Fourier transform, generalized Pythagorean theorem

AMS subject classifications. Primary, 15A15, 26B15; Secondary, 53A05, 42B10, 65R10

PII. S0895479895296896

1. Introduction.

The change-of-variables formula is

$$\int_{\mathcal{V}} f(\mathbf{v})\, d\mathbf{v} = \int_{\mathcal{U}} (f \circ \phi)(\mathbf{u})\ |\det J_\phi(\mathbf{u})|\ d\mathbf{u}, \tag{1}$$

where $\mathcal{U}, \mathcal{V}$ are sets in $\mathbb{R}^n$, ϕ is a sufficiently well-behaved function: $\mathcal{U} \to \mathcal{V}$, and f is integrable on $\mathcal{V}$. Here $d\mathbf{x}$ denotes the volume element $|dx_1 \wedge dx_2 \wedge \cdots \wedge dx_n|$, and J_ϕ is the Jacobi matrix (or Jacobian)

$$J_\phi := \left(\frac{\partial \phi_i}{\partial u_j}\right), \quad \text{also denoted} \quad \frac{\partial(v_1, v_2, \ldots, v_n)}{\partial(u_1, u_2, \ldots, u_n)},$$

representing the derivative of ϕ. An advantage of (1) is that integration on $\mathcal{V}$ is translated to (perhaps simpler) integration on $\mathcal{U}$.

This formula was given in 1841 by Jacobi [8], following Euler (the case $n = 2$) and Lagrange ($n = 3$). It gave prominence to functional (or symbolic) determinants, i.e., (nonnumerical) determinants of matrices including functions or operators as elements.

If $\mathcal{U}$ and $\mathcal{V}$ are in spaces of different dimensions, say, $\mathcal{U} \subset \mathbb{R}^n$ and $\mathcal{V} \subset \mathbb{R}^m$ with $n > m$, then the Jacobian J_ϕ is a rectangular matrix and (1) cannot be used in its present form. However, if J_ϕ is of full column rank throughout $\mathcal{U}$, we can replace $|\det J_\phi|$ in (1) by the volume $\operatorname{vol} J_\phi$ of the Jacobian to get

$$\int_{\mathcal{V}} f(\mathbf{v})\, d\mathbf{v} = \int_{\mathcal{U}} (f \circ \phi)(\mathbf{u}) \operatorname{vol} J_\phi(\mathbf{u})\, d\mathbf{u}. \tag{2}$$

Recall that the volume of an $m \times n$ matrix of rank r is

$$\operatorname{vol} A := \sqrt{\sum_{(I,J)\in\mathcal{N}} \det^2 A_{IJ}}, \tag{3}$$

*Received by the editors November 25, 1995; accepted for publication (in revised form) by G. Styan November 1, 1998; published electronically September 30, 1999. This research was done at the Indian Statistical Institute, New Delhi 110 016, India, and was supported by a Fulbright grant and the U.S. Educational Foundation in India.

http://www.siam.org/journals/simax/21-1/29689.html

†Rutgers Center for Operations Research, Rutgers University, 640 Bartholomew Road, Piscataway, NJ 08854-8003 (bisrael@rutcor.rutgers.edu).

where A_{IJ} is the submatrix of A with rows I and columns J, and $\mathcal{N}$ is the index set of $r \times r$ nonsingular submatrices of A; see, e.g., [1]. Alternatively, $\operatorname{vol} A$ is the product of the singular values of A. If A is of full column rank, its volume is simply

$$\operatorname{vol} A = \sqrt{\det A^T A}. \tag{4}$$

If $m = n$, then $\operatorname{vol} J_\phi = |\det J_\phi|$ and (2) reduces to the classical result (1).

The formula (2) is well known in differential geometry; see, e.g., [2, Proposition 6.6.1] and [6, section 3.2.3]. Although there are elementary accounts of this formula (see, e.g., [3, Volume II, Chapter IV, section 4], [7, section 8.1], and [13, section 3.4]), it is infrequently used in applications.

The objectives of this note are (i) to establish the usefulness of (2) for various surface integrals, (ii) to simplify the computation of the Radon and Fourier transforms and general integrals in $\mathbb{R}^n$ (see Examples 7–9 and Appendix A below), and (iii) to introduce the *functional matrix volume*, in analogy with the functional determinant.

We illustrate (2) for an elementary calculus example. Let $\mathcal{S}$ be a subset of a surface in $\mathbb{R}^3$ represented by

$$z = g(x, y) \tag{5}$$

and let $f(x, y, z)$ be a function integrable on $\mathcal{S}$. Let $\mathcal{A}$ be the projection of $\mathcal{S}$ on the xy-plane. Then $\mathcal{S}$ is the image of $\mathcal{A}$ under a mapping ϕ

$$\mathcal{S} = \phi(\mathcal{A}) \quad \text{or} \quad \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x \\ y \\ g(x,y) \end{pmatrix} = \phi \begin{pmatrix} x \\ y \end{pmatrix}, \quad \begin{pmatrix} x \\ y \end{pmatrix} \in \mathcal{A}. \tag{6}$$

The Jacobi matrix of ϕ is the 3×2 matrix

$$J_\phi(x, y) = \frac{\partial(x, y, z)}{\partial(x, y)} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ g_x & g_y \end{pmatrix}, \tag{7}$$

where $g_x = \frac{\partial g}{\partial x}$, $g_y = \frac{\partial g}{\partial y}$. The volume of (7) is, by (4),

$$\operatorname{vol} J_\phi(x, y) = \sqrt{1 + g_x^2 + g_y^2}. \tag{8}$$

Substituting (8) into (2) we get the well-known formula

$$\int_{\mathcal{S}} f(x, y, z)\, ds = \int_{\mathcal{A}} f(x, y, g(x, y)) \sqrt{1 + g_x^2 + g_y^2}\, dx\, dy, \tag{9}$$

giving an integral over $\mathcal{S}$ as an integral over its projection in the xy-plane.

The simplicity of this approach is not lost in higher dimensions or with different coordinate systems, as demonstrated below by 11 elementary examples from calculus and analysis.

- Example 1 concerns line integrals, in particular the *arc length* of a curve in $\mathbb{R}^n$.
- Example 2 is an application to surface integration in $\mathbb{R}^3$ using cylindrical coordinates.
- Examples 3 and 4 concern integration on an *axially symmetric surface* (or *surface of revolution*) in $\mathbb{R}^3$. In these integrals, the volume of J_ϕ contains the necessary information on the surface symmetry.

- Example 5 shows integration on an $(n-1)$-dimensional surface in $\mathbb{R}^n$.
- The area of the unit sphere in $\mathbb{R}^n$, a classical exercise, is computed in Example 6.
- Example 7 uses (2) to compute the Radon transform of a function $f : \mathbb{R}^n \to \mathbb{R}$.
- Example 8 computes an integral over $\mathbb{R}^n$ as an integral on $\mathbb{R}^{n-1}$ followed by an integral on $\mathbb{R}$.
- Example 9 applies this to the computation of the Fourier transforms of a function $f : \mathbb{R}^n \to \mathbb{R}$.
- The last two examples concern simplex faces in $\mathbb{R}^n$. Example 10 is the generalized Pythagorean theorem. Example 11 computes the largest face of the n-dimensional regular simplex.

These examples show that the full rank assumption for J_ϕ is quite natural and presents no real restriction in applications.

The solutions given here should be compared with the "classical" solutions, as taught in calculus. We see that (2) offers a unified method for a variety of curve and surface integrals and coordinate systems, without having to construct (and understand) the differential geometry in each application. The computational tractability of (2) is illustrated in Appendix A.

By a blanket assumption, throughout this paper all functions are continuously differentiable as needed, all surfaces are smooth, and all curves are rectifiable.

2. Examples and applications. If the mapping $\phi : \mathcal{U} \to \mathcal{V}$ is given by

$$y_i = \phi_i(x_1, x_2, \ldots, x_n), \quad i \in \overline{1, m},$$

we denote its Jacobi matrix J_ϕ by

$$\frac{\partial(y_1, y_2, \ldots, y_m)}{\partial(x_1, x_2, \ldots, x_n)} = \left(\frac{\partial \phi_i}{\partial x_j}\right), \quad i \in \overline{1, m}\,,\ j \in \overline{1, n}, \tag{10}$$

the customary notation for Jacobi matrices (in the square case).

Example 1. Let $\mathcal{C}$ be an arc on a curve in $\mathbb{R}^n$, represented in parametric form as

$$\mathcal{C} := \phi([0,1]) = \{(x_1, x_2, \ldots, x_n) \,:\, x_i := \phi_i(t),\ 0 \le t \le 1\}. \tag{11}$$

The Jacobi matrix $J_\phi(t) = \frac{\partial(x_1, x_2, \ldots, x_n)}{\partial t}$ is the column matrix $(\phi_i'(t))$ and its volume is

$$\operatorname{vol} J_\phi = \sqrt{\sum_{i=1}^{n} \left(\phi_i'(t)\right)^2}.$$

The *line integral* (assuming it exists) of a function f along $\mathcal{C}$, $\int_{\mathcal{C}} f$, is given in terms of the volume of J_ϕ as follows:

$$\int_{\mathcal{C}} f = \int_0^1 f(\phi_1(t), \ldots, \phi_n(t)) \sqrt{\sum_{i=1}^{n} \left(\phi_i'(t)\right)^2}\, dt. \tag{12}$$

In particular, $f \equiv 1$ gives

$$\text{arc length } \mathcal{C} = \int_0^1 \sqrt{\sum_{i=1}^{n} \left(\phi_i'(t)\right)^2}\, dt. \tag{13}$$

If one of the variables, say, $a \le x_1 \le b$, is used as parameter, (13) gives the familiar result

$$\text{arc length } \mathcal{C} = \int_a^b \sqrt{1 + \sum_{i=2}^{n} \left(\frac{dx_i}{dx_1}\right)^2}\, dx_1.$$

Example 2. Let $\mathcal{S}$ be a surface in $\mathbb{R}^3$ represented by

$$z = z(r, \theta), \tag{14}$$

where $\{r, \theta, z\}$ are *cylindrical coordinates*

$$x = r\,\cos\theta, \tag{15a}$$
$$y = r\,\sin\theta, \tag{15b}$$
$$z = z. \tag{15c}$$

The Jacobi matrix of the mapping (15a), (15b), and (14) is

$$\frac{\partial(x, y, z)}{\partial(r, \theta)} = \begin{pmatrix} \cos\theta & -r\,\sin\theta \\ \sin\theta & r\,\cos\theta \\ \frac{\partial z}{\partial r} & \frac{\partial z}{\partial \theta} \end{pmatrix}; \tag{16}$$

see also (A.4), Appendix A. The volume of (16) is

$$\text{vol } J_\phi = \sqrt{r^2 + r^2 \left(\frac{\partial z}{\partial r}\right)^2 + \left(\frac{\partial z}{\partial \theta}\right)^2} = r\sqrt{1 + \left(\frac{\partial z}{\partial r}\right)^2 + \frac{1}{r^2}\left(\frac{\partial z}{\partial \theta}\right)^2}; \tag{17}$$

see also (A.7), Appendix A. An integral over a domain $\mathcal{V} \subset \mathcal{S}$ is therefore

$$\int_{\mathcal{V}} f(x, y, z)\, dV = \int_{\mathcal{U}} f(r\,\cos\theta,\, r\,\sin\theta, z(r, \theta))\, r \sqrt{1 + \left(\frac{\partial z}{\partial r}\right)^2 + \frac{1}{r^2}\left(\frac{\partial z}{\partial \theta}\right)^2}\, dr\, d\theta. \tag{18}$$

Example 3. Let $\mathcal{S}$ be a surface in $\mathbb{R}^3$, symmetric about the z-axis. This axial symmetry is expressed in cylindrical coordinates by

$$z = z(r) \quad \text{or} \quad \frac{\partial z}{\partial \theta} = 0 \quad \text{in} \quad (16)\text{–}(18).$$

The volume (17) thus becomes

$$\text{vol } J_\phi = r\sqrt{1 + z'(r)^2} \tag{19}$$

with the axial symmetry "built in." An integral over a domain $\mathcal{V}$ in a z-symmetric surface $\mathcal{S}$ is therefore

$$\int_{\mathcal{V}} f(x, y, z)\, dV = \int_{\mathcal{U}} f(r\,\cos\theta, r\,\sin\theta, z(r))\, r\sqrt{1 + z'(r)^2}\, dr\, d\theta.$$

Example 4. Again let $\mathcal{S}$ be a z-symmetric surface in $\mathbb{R}^3$. We use *spherical coordinates*

$$x = \rho \sin\phi \cos\theta, \tag{20a}$$

$$y = \rho \sin\phi \sin\theta, \tag{20b}$$

$$z = \rho \cos\phi. \tag{20c}$$

The axial symmetry is expressed by

$$\rho \;:=\; \rho(\phi),$$

showing that $\mathcal{S}$ is given in terms of the two variables ϕ and θ. The volume of the Jacobi matrix is easily computed,

$$\operatorname{vol} \frac{\partial(x,y,z)}{\partial(\phi,\theta)} = \rho \sqrt{\rho^2 + (\rho'(\phi))^2}\, \sin\phi,$$

and the change-of-variables formula gives

$$\int_{\mathcal{V}} f(x,y,z)\, dV = \int_{\mathcal{U}} f(\rho(\phi)\, \sin\phi\, \cos\theta, \rho(\phi)\, \sin\phi\, \sin\theta, \rho(\phi)\, \cos\phi)\, \rho(\phi) \sqrt{\rho(\phi)^2 + (\rho'(\phi))^2}\, \sin\phi\, d\phi\, d\theta. \tag{21}$$

Example 5. Let a surface $\mathcal{S}$ in $\mathbb{R}^n$ be given by

$$x_n \;:=\; g(x_1, x_2, \ldots, x_{n-1}), \tag{22}$$

let $\mathcal{V}$ be a subset on $\mathcal{S}$, and let $\mathcal{U}$ be the projection of $\mathcal{V}$ on $\mathbb{R}^{n-1}$, the space of variables $(x_1, \ldots, x_{n-1})$. The surface $\mathcal{S}$ is the graph of the mapping $\phi : \mathcal{U} \to \mathcal{V}$, given by its components $\phi \;:=\; (\phi_1, \phi_2, \ldots, \phi_n)$,

$$\begin{aligned} &\phi_i(x_1, \ldots, x_{n-1}) := x_i, \quad i = 1, \ldots, n-1, \\ &\phi_n(x_1, \ldots, x_{n-1}) := g(x_1, \ldots, x_{n-1}). \end{aligned}$$

The Jacobi matrix of ϕ is

$$J_\phi = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \\ \dfrac{\partial g}{\partial x_1} & \dfrac{\partial g}{\partial x_2} & \cdots & \dfrac{\partial g}{\partial x_{n-2}} & \dfrac{\partial g}{\partial x_{n-1}} \end{pmatrix}$$

and its volume is

$$\operatorname{vol} J_\phi = \sqrt{1 + \sum_{i=1}^{n-1} \left(\frac{\partial g}{\partial x_i}\right)^2}. \tag{23}$$

For any function f integrable on $\mathcal{V}$ we therefore have

(24)
$$\int_{\mathcal{V}} f(x_1, x_2, \ldots, x_{n-1}, x_n)\, dV$$
$$= \int_{\mathcal{U}} f(x_1, x_2, \ldots, x_{n-1}, g(x_1, x_2, \ldots, x_{n-1})) \sqrt{1 + \sum_{i=1}^{n-1} \left(\frac{\partial g}{\partial x_i}\right)^2}\, dx_1 dx_2 \cdots dx_{n-1}.$$

In particular, $f \equiv 1$ gives the area of $\mathcal{V}$

$$\int_{\mathcal{V}} 1\, dV = \int_{\mathcal{U}} \sqrt{1 + \sum_{i=1}^{n-1} \left(\frac{\partial g}{\partial x_i}\right)^2}\, dx_1 dx_2 \cdots dx_{n-1}. \tag{25}$$

Example 6. Let $\mathcal{B}_n$ be the unit ball in $\mathbb{R}^n$, $\mathcal{S}_n$ the unit sphere, and a_n the area of $\mathcal{S}_n$. Integrals on $\mathcal{S}_n$, in particular the area a_n, can be computed using spherical coordinates, e.g., [12, section VII.2], or the surface element of $\mathcal{S}_n$, e.g., [11]. An alternative, simpler, approach is to use the results of Example 5, representing the "upper hemisphere" as $\phi(\mathcal{B}_{n-1})$, where $\phi = (\phi_1, \phi_2, \ldots, \phi_n)$ is

$$\phi_i(x_1, x_2, \ldots, x_{n-1}) = x_i \ , \ i \in \overline{1, n-1},$$
$$\phi_n(x_1, x_2, \ldots, x_{n-1}) = \sqrt{1 - \sum_{i=1}^{n-1} x_i^2}.$$

The Jacobi matrix is

$$J_\phi = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ -\dfrac{x_1}{x_n} & -\dfrac{x_2}{x_n} & \cdots & -\dfrac{x_{n-1}}{x_n} \end{pmatrix}$$

and its volume is easily computed:

$$\operatorname{vol} J_\phi = \sqrt{1 + \sum_{i=1}^{n-1} \left(\frac{x_i}{x_n}\right)^2} = \frac{1}{|x_n|} = \frac{1}{\sqrt{1 - \sum_{i=1}^{n-1} x_i^2}}. \tag{26}$$

The area a_n is twice the area of the "upper hemisphere." Therefore, by (26),

$$a_n = 2 \int_{\mathcal{B}_{n-1}} \frac{dx_1 dx_2 \cdots dx_{n-1}}{\sqrt{1 - \sum_{i=1}^{n-1} x_i^2}}, \tag{27}$$

which is easily integrated to give

$$a_n = \frac{2\, \pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)} \tag{28}$$

using well-known properties of the *beta function*

$$B(p,q) := \int_0^1 (1-x)^{p-1}\ x^{q-1}\ dx,$$

and the *gamma function* $\quad \Gamma(p) := \displaystyle\int_0^\infty x^{p-1}\ e^{-x}\ dx.$

Example 7 (Radon transform). Let $\mathcal{H}_{\boldsymbol{\xi},p}$ be a hyperplane in $\mathbb{R}^n$ represented by

$$\mathcal{H}_{\boldsymbol{\xi},p} := \left\{ \mathbf{x} \in \mathbb{R}^n : \sum_{i=1}^n \xi_i\, x_i = p \right\} = \{\mathbf{x} : \langle \boldsymbol{\xi}, \mathbf{x} \rangle = p\}, \tag{29}$$

where $\xi_n \neq 0$ in the normal vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ of $\mathcal{H}_{\boldsymbol{\xi},p}$. (Such hyperplanes are called *nonvertical*.) Then $\mathcal{H}_{\boldsymbol{\xi},p}$ is given by

$$x_n \ := \ \frac{p}{\xi_n} - \sum_{i=1}^{n-1} \frac{\xi_i}{\xi_n}\, x_i, \tag{30}$$

which is of the form (22). The volume (23) is here

$$\operatorname{vol} J_\phi = \sqrt{1 + \sum_{i=1}^{n-1} \left(\frac{\xi_i}{\xi_n} \right)^2} = \frac{\|\boldsymbol{\xi}\|}{|\xi_n|}. \tag{31}$$

The Radon transform $(\mathbf{R}f)(\boldsymbol{\xi}, p)$ of a function $f : \mathbb{R}^n \to \mathbb{R}$ is its integral over the hyperplane $\mathcal{H}_{\boldsymbol{\xi},p}$ (see [4])

$$(\mathbf{R}f)(\boldsymbol{\xi}, p) \ := \int_{\{\mathbf{x}:\langle \boldsymbol{\xi}, \mathbf{x} \rangle = p\}} f(\mathbf{x})\, d\mathbf{x}. \tag{32}$$

Using (30) and (31), the Radon transform can be computed as an integral in $\mathbb{R}^{n-1}$:

$$(\mathbf{R}f)(\boldsymbol{\xi}, p) = \frac{\|\boldsymbol{\xi}\|}{|\xi_n|} \int_{\mathbb{R}^{n-1}} f\left(x_1, \ldots, x_{n-1}, \frac{p}{\xi_n} - \sum_{i=1}^{n-1} \frac{\xi_i}{\xi_n}\, x_i \right) dx_1 dx_2 \cdots dx_{n-1}. \tag{33}$$

See (A.15) and (A.17), Appendix A, for Radon transform in $\mathbb{R}^2$ and (A.20) for $\mathbb{R}^3$.

In tomography applications the Radon transforms $(\mathbf{R}f)(\boldsymbol{\xi}, p)$ are computed by the scanning equipment, so (33) is not relevant. The issue is the *inverse problem* of reconstructing f from its Radon transforms $(\mathbf{R}f)(\boldsymbol{\xi}, p)$ for all $\boldsymbol{\xi}, p$. The inverse Radon transform is also an integral (see, e.g., [4], [12]) and can be expressed analogously to (33), using the method of the next example.

Example 8. Consider an integral over $\mathbb{R}^n$:

$$\int_{\mathbb{R}^n} f(\mathbf{x})\, d\mathbf{x} = \int_{\mathbb{R}^n} f(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \cdots dx_n. \tag{34}$$

Since $\mathbb{R}^n$ is a union of (parallel) hyperplanes,

$$\mathbb{R}^n = \bigcup_{p=-\infty}^{\infty} \{\mathbf{x} : \langle \boldsymbol{\xi}, \mathbf{x} \rangle = p\}, \quad \text{where } \boldsymbol{\xi} \neq \mathbf{0}, \tag{35}$$

we can compute (34) iteratively: an integral over $\mathbb{R}^{n-1}$ (Radon transform), followed by an integral on $\mathbb{R}$,

$$\int_{\mathbb{R}^n} f(\mathbf{x})\, d\mathbf{x} = \int_{-\infty}^{\infty} \frac{dp}{\|\boldsymbol{\xi}\|}\, (\mathbf{R}f)(\boldsymbol{\xi}, p), \tag{36}$$

where $dp/\|\boldsymbol{\xi}\|$ is the differential of the distance along $\boldsymbol{\xi}$ (i.e., dp times the distance between the parallel hyperplanes $\mathcal{H}_{\boldsymbol{\xi},p}$ and $\mathcal{H}_{\boldsymbol{\xi},p+1}$). Combining (33) and (36) we get the integral of f on $\mathbb{R}^n$:

$$\begin{aligned}\int_{\mathbb{R}^n} & f(\mathbf{x})\, d\mathbf{x} \\ &= \frac{1}{|\xi_n|} \int_{-\infty}^{\infty} \left\{ \int_{\mathbb{R}^{n-1}} f\left(x_1, \ldots, x_{n-1}, \frac{p}{\xi_n} - \sum_{i=1}^{n-1} \frac{\xi_i}{\xi_n}\, x_i \right) dx_1 dx_2 \cdots dx_{n-1} \right\} dp;\end{aligned} \tag{37}$$

see (A.21), Appendix A, for integration over $\mathbb{R}^3$.

It is possible to derive (37) directly from the classical change-of-variables formula (1) by changing variables from $\{x_1, \ldots, x_{n-1}, x_n\}$ to $\{x_1, \ldots, x_{n-1}, p := \sum_{i=1}^n \xi_i x_i\}$ and using

$$\det\left(\frac{\partial(x_1, \ldots, x_{n-1}, x_n)}{\partial(x_1, \ldots, x_{n-1}, p)} \right) = \frac{1}{\xi_n}.$$

An advantage of our development is that it can be used recursively, i.e., the inner integral in (37) can again be expressed as an integral in $\mathbb{R}^{n-2}$ followed by an integral in $\mathbb{R}$, etc.

Example 9 (Fourier transform). In particular, the Fourier transform $(\mathbf{F}f)(\boldsymbol{\xi})$ of f is the integral

$$\begin{aligned}(\mathbf{F}f)(\boldsymbol{\xi}) &:= (2\,\pi)^{-n/2} \int_{\mathbb{R}^n} e^{-i\langle \boldsymbol{\xi}, \mathbf{x} \rangle}\, f(\mathbf{x})\, d\mathbf{x} \\ &= (2\,\pi)^{-n/2} \int_{\mathbb{R}^n} f\,(x_1, \ldots, x_n)\ \exp\left\{ -i \sum_{k=1}^{n} \xi_k\, x_k \right\} dx_1 dx_2 \cdots dx_n.\end{aligned} \tag{38}$$

For $\xi_n \neq 0$ we can compute (38) analogously to (37) as

$$(\mathbf{F}f)(\boldsymbol{\xi}) = \frac{(2\pi)^{-n/2}}{|\xi_n|} \int_{-\infty}^{\infty} e^{-ip} \left\{ \int_{\mathbb{R}^{n-1}} f\left(x_1, \ldots, x_{n-1}, \frac{p}{\xi_n} - \sum_{i=1}^{n-1} \frac{\xi_i}{\xi_n} x_i \right) \prod_{k=1}^{n-1} dx_k \right\} dp. \tag{39}$$

The Fourier transform of a function of n variables is thus computed as an integral over $\mathbb{R}^{n-1}$ followed by an integral on $\mathbb{R}$. The inverse Fourier transform of a function $g(\boldsymbol{\xi})$ is of the same form as (38),

$$(\mathbf{F}^{-1}g)(\mathbf{x}) := (2\,\pi)^{-n/2} \int_{\mathbb{R}^n} e^{i\langle \mathbf{x}, \boldsymbol{\xi} \rangle}\, g(\boldsymbol{\xi})\, d\boldsymbol{\xi}, \tag{40}$$

and can be computed as in (39).

Example 10 (the generalized Pythagorean theorem [10]). Consider an n-dimensional simplex

$$\boldsymbol{\Delta}_n \ := \ \left\{ (x_1, x_2, \ldots, x_n) \ : \ \sum_{i=1}^{n} a_i\, x_i \leq a_0,\ x_i \geq 0,\ i \in \overline{1, n} \right\}, \tag{41}$$

with all $a_j > 0\,,\ j \in \overline{0,n}$. We denote the $n+1$ faces of $\mathbf{\Delta}_n$ by

$$\mathcal{F}_0 := \left\{ (x_1, x_2, \ldots, x_n) \in \mathbf{\Delta}_n \,:\, \sum_{i=1}^n a_i\, x_i = a_0 \right\} \tag{42a}$$

$$\mathcal{F}_j := \{ (x_1, x_2, \ldots, x_n) \in \mathbf{\Delta}_n \,:\, x_j = 0 \}\,,\ j \in \overline{1,n}, \tag{42b}$$

and denote their areas by A_0, A_j, respectively. The generalized Pythagorean theorem (see [10]) states that

$$A_0^2 = \sum_{j=1}^n A_j^2. \tag{43}$$

We prove it here using the change-of-variables formula (2). For any $j \in \overline{1,n}$ we can represent the (largest) face $\mathcal{F}_0$ as $\mathcal{F}_0 = \phi^{\{j\}}(\mathcal{F}_j)$, where $\phi^{\{j\}} = (\phi_1^{\{j\}}, \ldots, \phi_n^{\{j\}})$ is

$$\begin{aligned} \phi_i^{\{j\}}(x_1, x_2, \ldots, x_n) &= x_i,\ i \neq j, \\ \phi_j^{\{j\}}(x_1, x_2, \ldots, x_n) &= \frac{a_0}{a_j} - \sum_{i \neq j} \frac{a_i}{a_j}\, x_i. \end{aligned}$$

The Jacobi matrix of $\phi^{\{j\}}$ is an $n \times (n-1)$ matrix with the ith unit vector in row $i \neq j$ and

$$\left(-\frac{a_1}{a_j}, -\frac{a_2}{a_j}, \ldots, -\frac{a_{j-1}}{a_j}, -\frac{a_{j+1}}{a_j}, \ldots, -\frac{a_{n-1}}{a_j}, -\frac{a_n}{a_j} \right)$$

in row j. The volume of the Jacobi matrix of $\phi^{\{j\}}$ is computed as

$$\operatorname{vol} J_{\phi^{\{j\}}} = \sqrt{1 + \sum_{i \neq j} \left(\frac{a_i}{a_j} \right)^2} = \sqrt{\frac{\sum_{i=1}^n a_i^2}{a_j^2}} = \frac{\|\mathbf{a}\|}{|a_j|},$$

where $\mathbf{a}$ is the vector $(a_1, \ldots, a_n)$. Therefore, the area of $\mathcal{F}_0$ is

$$A_0 = \int_{\mathcal{F}_j} \left(\frac{\|\mathbf{a}\|}{|a_j|} \right) \prod_{i \neq j} dx_i = \left(\frac{\|\mathbf{a}\|}{|a_j|} \right) A_j, \quad j \in \overline{1,n}, \tag{44}$$

$$\therefore \quad \frac{\sum_{j=1}^n A_j^2}{A_0^2} = \frac{\sum_{j=1}^n |a_j|^2}{\|\mathbf{a}\|^2}$$

and the generalized Pythagorean theorem (43) reduces to the ordinary Pythagorean theorem

$$\|\mathbf{a}\|^2 = \sum_{j=1}^n |a_j|^2.$$

Example 11. The simplex (41) with $a_j = 1$ for all $j \in \overline{0,n}$ is the n-dimensional *regular simplex*

$$\mathbf{\Delta}_n \ := \ \left\{ (x_1, x_2, \ldots, x_n) \,:\, \sum_{i=1}^n x_i \leq 1,\ x_i \geq 0,\ i \in \overline{1,n} \right\}. \tag{45}$$

The area A_0 of its largest face

$$\mathcal{F}_0 = \left\{ (x_1, \ldots, x_n) \,:\, \sum_{i=1}^{n} x_i = 1,\ x_i \geq 0, i \in \overline{1,n} \right\} \tag{46}$$

is given by (44) as

$$A_0 = \sqrt{n}\, V_{n-1},$$

where V_{n-1}, the volume of the $(n-1)$-dimensional regular simplex, is

$$V_{n-1} = \frac{1}{(n-1)!}.$$

See, e.g., [9, section 47]. Therefore,

$$A_0 = \frac{\sqrt{n}}{(n-1)!}. \tag{47}$$

Note that $A_0 \to 0$ as $n \to \infty$, although the side faces of the unit cube (of which the face $\mathcal{F}_0$ is a diagonal section) have areas 1.

Appendix A. Illustrations with Derive. The integrations of this paper can be done symbolically. We illustrate this for the symbolic package Derive [5], omitting details such as the commands (e.g., `Simplify`, `Approximate`) and settings (e.g., `Trigonometry:=Expand`) that are used to obtain these results.

The Jacobi matrix is computed by the function

$$\texttt{JACOBIAN}(u, \alpha) := \texttt{VECTOR(GRAD}(u_{m_}, \alpha), m_, \texttt{DIMENSION}(u)) \tag{A.1}$$

found in the Derive utility file `VECTOR.mth`.

For example, to compute the surface integral of Example 2 we define the surface as a function of r and θ:

$$Z(r, \theta) := \tag{A.2}$$

(see (14)). Then

$$\texttt{JACOBIAN}([r * \texttt{COS}(\theta), r * \texttt{SIN}(\theta), Z(r, \theta)], [r, \theta]) \tag{A.3}$$

gives (16),

$$\begin{bmatrix} \texttt{COS}(\theta) & -r * \texttt{SIN}(\theta) \\ \texttt{SIN}(\theta) & r * \texttt{COS}(\theta) \\ \dfrac{\partial Z(r,\theta)}{\partial r} & \dfrac{\partial Z(r,\theta)}{\partial \theta} \end{bmatrix}. \tag{A.4}$$

The volume of an $(n+1) \times n$ matrix of full rank is computed by

$$\texttt{VOL}(a) := \texttt{SQRT}(\texttt{SUM}(\texttt{DET}(\texttt{DELETE_ELEMENT}(a, k_))^2, k_, 1, \texttt{DIMENSION}(a))). \tag{A.5}$$

For example,

$$\texttt{VOL}(\texttt{JACOBIAN}([r * \texttt{COS}(\theta), r * \texttt{SIN}(\theta), Z(r, \theta)], [r, \theta])) \tag{A.6}$$

gives (17),

$$\sqrt{r^2 * \left(\frac{\partial Z(r,\theta)}{\partial r}\right)^2 + \left(\frac{\partial Z(r,\theta)}{\partial \theta}\right)^2 + r^2}. \tag{A.7}$$

In Example 3 the surface is symmetric about the z-axis and we use $Z(r)$ instead of $Z(r,\theta)$ in (A.6),

$$\texttt{VOL}(\texttt{JACOBIAN}([r * \texttt{COS}(\theta), r * \texttt{SIN}(\theta), Z(r)], [r, \theta])), \tag{A.8}$$

to get (19),

$$|r| * \sqrt{Z'(r)^2 + 1}. \tag{A.9}$$

For example, the volume associated with the surface $z = 1/r$

$$\texttt{VOL}(\texttt{JACOBIAN}([r * \texttt{COS}(\theta), r * \texttt{SIN}(\theta), 1/r], [r, \theta])) \quad \text{gives} \quad \frac{\sqrt{r^4+1}}{|r|}. \tag{A.10}$$

The surface area of $z = 1/r$ is 2π times the integral of (A.10). In particular, the surface area from $r = 1$ to $r = 100$

$$2 * \pi * \texttt{INT}\left[\frac{\texttt{SQRT}(r^4+1)}{|r|}, r, 1, 100\right] \quad \text{gives} \quad 3.14142524131 \times 10^4.$$

Consider now the Radon transform in $\mathbb{R}^2$. A line

$$a\,x + b\,y = p \tag{A.11}$$

can be solved for y

$$y = \frac{p}{b} - \frac{a}{b}\,x$$

if $b \neq 0$. Then

$$\texttt{VOL}\left(\texttt{JACOBIAN}\left(\left[x, \frac{p}{b} - \frac{a}{b}\,x\right], [x]\right)\right) \quad \text{gives} \quad \frac{\sqrt{a^2+b^2}}{|b|} \tag{A.12}$$

in agreement with (31). The Radon transform (33) is then computed by

$$\begin{aligned} &\texttt{RADON_AUX}(f, x, y, a, b, p) \\ &\quad := \texttt{INT}(\texttt{LIM}(f, y, p/b - a/b * x), x, -\infty, \infty) * \texttt{SQRT}(a^2+b^2)/|b|. \end{aligned} \tag{A.13}$$

For example, the Radon transform of $e^{-x^2-y^2}$,

$$\texttt{RADON_AUX}(\texttt{EXP}(-x^2-y^2), x, y, a, b, p) \quad \text{is} \quad \sqrt{\pi}\,\texttt{EXP}\left[-\frac{p^2}{a^2+b^2}\right]. \tag{A.14}$$

The function (A.13) requires $b \neq 0$. We modify it for lines with $b = 0$ (i.e., vertical lines) as follows:

$$\begin{aligned} &\texttt{RADON}(f, x, y, a, b, p) \\ &:= \texttt{IF}(b = 0, \texttt{INT}(\texttt{LIM}(f, x, p/a), y, -\infty, \infty), \texttt{RADON_AUX}(f, x, y, a, b, p)). \end{aligned} \tag{A.15}$$

For example, the Radon transform of $e^{-x^2-y^2}$ with respect to a vertical line

$$\texttt{RADON}(\texttt{EXP}(-x^2-y^2),x,y,a,0,p) \quad \text{gives} \quad \sqrt{\pi}\,\texttt{EXP}\left[-\frac{p^2}{a^2}\right],$$

which is (A.14) with $b=0$.

An alternative way is to normalize the coefficients of the lines (A.11), say, $a=\cos\alpha$ and $b=\sin\alpha$. The Radon transform is then

(A.16)
$$\begin{aligned}&\texttt{RADON_NORMAL_AUX}(f,x,y,\alpha,p)\\ &\qquad := \texttt{INT}(\texttt{LIM}(f,y,p/\texttt{SIN}(\alpha)-\texttt{COT}(\alpha)*x),x,-\infty,\infty)/\texttt{ABS}(\texttt{SIN}(\alpha)).\end{aligned}$$

provided $\alpha\neq 0$. The following function treats $\alpha=0$ separately:

(A.17)
$$\begin{aligned}&\texttt{RADON_NORMAL}(f,x,y,\alpha,p)\\ &:= \texttt{IF}(\alpha=0,\texttt{INT}(\texttt{LIM}(f,x,p),y,-\infty,\infty),\texttt{RADON_NORMAL_AUX}(f,x,y,\alpha,p))\end{aligned}$$

For example, the Radon transform of

$$f(x,y) := \begin{cases} 1 & \text{if } 0\le x,y\le 1,\\ 0 & \text{otherwise}\end{cases}$$

is computed by

(A.18)
$$\texttt{RADON_NORMAL}(\texttt{CHI}(0,x,1)*\texttt{CHI}(0,y,1),x,y,\alpha,p),$$

giving

(A.19)
$$\texttt{SIGN}(\texttt{COS}(\alpha)-p)*\left[\frac{p}{\texttt{SIN}(2\alpha)}-\frac{0.5}{\texttt{SIN}(\alpha)}\right]+\frac{|p|}{\texttt{SIN}(2\alpha)}.$$

The Radon transform in $\mathbb{R}^3$ is computed with respect to a plane

$$a\,x+b\,y+c\,z=p$$

as follows (ignoring the case $c=0$):

(A.20)
$$\begin{aligned}&\texttt{RADON_3}(f,x,y,z,a,b,c,p)\\ &\quad := \texttt{INT}(\texttt{INT}(\texttt{LIM}(f,z,p/c-a/c*x-b/c*y),x,-\infty,\infty),y,-\infty,\infty)\\ &\qquad *\texttt{SQRT}(a^2+b^2+c^2)/|c|.\end{aligned}$$

Finally, an integral over $\mathbb{R}^3$ is computed, using (37) as follows:

(A.21)
$$\begin{aligned}&\texttt{INT_3}(f,x,y,z,a,b,c)\\ &\quad := \texttt{INT}(\texttt{RADON_3}(f,x,y,z,a,b,c,p),p,-\infty,\infty)/\texttt{SQRT}(a^2+b^2+c^2).\end{aligned}$$

For example, the integral over $\mathbb{R}^3$ of $f(x,y,z)=e^{x+y+z-x^2-y^2-z^2}$,

$$\texttt{INT_3}(\texttt{EXP}(x+y+z-x^2-y^2-z^2),x,y,z,a,b,c) \quad \text{gives} \quad \pi^{3/2}\,e^{3/4}.$$

REFERENCES

[1] A. Ben-Israel, *A volume associated with $m \times n$ matrices*, Linear Algebra Appl., 167 (1992), pp. 87–111.

[2] M. Berger and B. Gostiaux, *Differential Geometry: Manifolds, Curves and Surfaces*, Grad. Texts in Math. 115, Springer-Verlag, New York, 1988 (translated by S. Levy).

[3] R. Courant, *Differential and Integral Calculus*, Interscience Publishers, New York, 1936 (translated by E. J. McShane).

[4] S. R. Deans, *The Radon Transform and Some of its Applications*, John Wiley, New York, 1983.

[5] *Derive for Windows*, Soft Warehouse, Honolulu, HI, 1996.

[6] H. Federer, *Geometric Measure Theory*, Springer-Verlag, New York, 1969.

[7] W. Fleming, *Functions of Several Variables*, 2nd ed., Springer-Verlag, New York, 1977.

[8] C. G. J. Jacobi, *De Determinantibus Functionalibus*, Crelle Journal für die reine und angewandte Mathematik, 22 (1841), pp. 319–359. Reprinted in *C. G. J. Jacobi's Gesammelte Werke*, Vol. 3, K. Weierstrass, ed., Berlin, 1884, pp. 393–438.

[9] M. G. Kendall, *A Course in the Geometry of n Dimensions*, Griffin's Statistical Monographs 8, Hafner, Riverside, NJ, 1961.

[10] S.-Y. T. Lin and Y.-F. Lin, *The n-dimensional Pythagorean theorem*, Linear and Multilinear Algebra, 26 (1990), pp. 9–13.

[11] C. Müller, *Spherical Harmonics*, Lecture Notes in Math. 17, Springer-Verlag, New York, 1966.

[12] F. Natterer, *The Mathematics of Computerized Tomography*, John Wiley, New York, 1986.

[13] M. Schreiber, *Differential Forms: A Heuristic Introduction*, Springer-Verlag, New York, 1977.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 1, pp. 313–326
© 1999 Society for Industrial and Applied Mathematics

ROW-WISE BACKWARD STABLE ELIMINATION METHODS FOR THE EQUALITY CONSTRAINED LEAST SQUARES PROBLEM*

ANTHONY J. COX† AND NICHOLAS J. HIGHAM‡

Abstract. It is well known that the solution of the equality constrained least squares (LSE) problem $\min_{Bx=d} \|b - Ax\|_2$ is the limit of the solution of the unconstrained weighted least squares problem

$$\min_x \left\| \begin{bmatrix} \mu d \\ b \end{bmatrix} - \begin{bmatrix} \mu B \\ A \end{bmatrix} x \right\|_2$$

as the weight μ tends to infinity, assuming that $[\,B^T \quad A^T\,]^T$ has full rank. We derive a method for the LSE problem by applying Householder QR factorization with column pivoting to this weighted problem and taking the limit analytically, with an appropriate rescaling of rows. The method obtained is a type of direct elimination method. We adapt existing error analysis for the unconstrained problem to obtain a row-wise backward error bound for the method. The bound shows that, provided row pivoting or row sorting is used, the method is well-suited to problems in which the rows of A and B vary widely in norm. As a by-product of our analysis, we derive a row-wise backward error bound of precisely the same form for the standard elimination method for solving the LSE problem. We illustrate our results with numerical tests.

Key words. constrained least squares problem, weighted least squares problem, Householder QR factorization, Gaussian elimination, elimination method, rounding error analysis, backward stability, row pivoting, row sorting, column pivoting

AMS subject classifications. 65F20, 65G05

PII. S0895479898335957

1. Introduction.

Consider the equality constrained least squares (LSE) problem

$$\min_{Bx=d} \|b - Ax\|_2, \qquad A \in \mathbb{R}^{m\times n}, \quad B \in \mathbb{R}^{p\times n}, \quad m + p \geq n \geq p. \tag{1.1}$$

We assume throughout this work that $\operatorname{rank}(B) = p$, which ensures that the constraint system is consistent, and also that $\operatorname{rank}\begin{bmatrix} B \\ A \end{bmatrix} = n$, which ensures that the LSE problem has a unique solution. Two commonly used methods for solving the LSE problem are based on QR factorization. The null space method (of which there are several variations [6]) begins by factorizing $B^T = QR$ to obtain a basis for the null space of B. The elimination method, on the other hand, uses QR factorization with column pivoting to factorize

$$B\Pi = Q\,[\,R_1 \quad R_2\,], \qquad R_1 \in \mathbb{R}^{p\times p} \text{ upper triangular, nonsingular.} \tag{1.2}$$

Partitioning $\Pi^T x = [\widetilde{x}_1^T,\ \widetilde{x}_2^T]^T$, $\widetilde{x}_1 \in \mathbb{R}^p$ and substituting the factorization (1.2) into the constraints yields

$$R_1\widetilde{x}_1 = Q^T d - R_2\widetilde{x}_2.$$

*Received by the editors March 19, 1998; accepted for publication (in revised form) by P. Van Dooren August 31, 1998; published electronically September 30, 1999.

http://www.siam.org/journals/simax/21-1/33595.html

†Ground Processing Department, Mail Point 31, Matra Marconi Space UK, Anchorage Road, Portsmouth, PO3 5PU, UK (anthonyj.cox@mmsuk.co.uk). The work of this author was supported by an Engineering and Physical Sciences Research Council Research Studentship.

‡Department of Mathematics, University of Manchester, Manchester, M13 9PL, UK (higham@ma.man.ac.uk, http://www.ma.man.ac.uk/~higham/). The work of this author was supported by Engineering and Physical Sciences Research Council grant GR/L76532.

By solving for $\widetilde{x}_1$ and partitioning $A\Pi = [\widetilde{A}_1,\ \widetilde{A}_2]$, $\widetilde{A}_1 \in \mathbb{R}^{m\times p}$ we reduce the LSE problem to the unconstrained problem

$$\min_{\widetilde{x}_2} \left\| (\widetilde{A}_2 - \widetilde{A}_1 R_1^{-1} R_2)\widetilde{x}_2 - (b - \widetilde{A}_1 R_1^{-1} Q^T d) \right\|_2 .$$

Solving this unconstrained problem by QR factorization completes the elimination method as originally presented by Björck and Golub [5] (see also [13, Chapter 21]). It is instructive to think of the method in terms of transformations on the matrix "B-over-A":

$$\begin{bmatrix} B \\ A \end{bmatrix} \begin{matrix} \\ {\scriptstyle m} \end{matrix} \begin{matrix} \overset{p}{} \\ \end{matrix} \!\!\begin{matrix} {\scriptstyle p} \\ \end{matrix}\!\! \begin{bmatrix} B_1 & B_2 \\ A_1 & A_2 \end{bmatrix} \to \begin{bmatrix} R_1 & R_2 \\ \widetilde{A}_1 & \widetilde{A}_2 \end{bmatrix} \to \begin{bmatrix} R_1 & R_2 \\ 0 & \widetilde{A}_2 - \widetilde{A}_1 R_1^{-1} R_2 \end{bmatrix} \to \begin{bmatrix} R_1 & R_2 \\ 0 & R_3 \\ 0 & 0 \end{bmatrix},$$

where $R_3 \in \mathbb{R}^{(n-p)\times(n-p)}$ is upper triangular. Note that the penultimate transformation is simply the annihilation of $\widetilde{A}_1$ by Gaussian elimination. We will refer to this method as the *EG method*.

The B-over-A matrix also arises in the method of weighting for solving the LSE problem, which is based on the observation that the LSE solution is the limit of the solution of the unconstrained problem

$$\min_x \left\| \begin{bmatrix} \mu d \\ b \end{bmatrix} - \begin{bmatrix} \mu B \\ A \end{bmatrix} x \right\|_2 \tag{1.3}$$

as the weight μ tends to infinity. Van Loan [18] describes an algorithm that solves (1.3) for a single weight and uses a refinement procedure to approximate the required limit. The algorithm is analyzed further by Barlow and Vemulapati [1], [3].

In this work we derive a method for the LSE problem by taking the limit $\mu \to \infty$ analytically rather than numerically. To be precise, we apply Householder QR factorization with column pivoting to (1.3), rescale the first p rows, and then take the limit $\mu \to \infty$ to obtain a matrix that we view as an update of B-over-A. The new method is an elimination method very similar, but not identical, to the EG method. We show that it satisfies a pleasing row-wise backward error bound when row pivoting or row sorting is used—a result that makes the method attractive when the rows of A and B vary widely in norm. We also show that a row-wise backward error bound of exactly the same form holds for the EG method. We give some numerical experiments to confirm the stability properties of the methods.

After this paper was submitted for publication, we learned that Reid [15] has obtained the method derived here by a different argument involving infinite weights. He shows that the algorithm of Powell and Reid [14], including its error analysis, may be applied to the weighted problem with implicit scaling, that is, with the weights stored separately. By holding inverse weights, the constrained case can be accommodated. He obtains a backward error bound similar to our Theorem 4.3 and also shows that iterative refinement may be applied. Reid points out that the method is equivalent to a method of Gulliksson and Wedin [9], [10], which is expressed in the language of "M-invariant reflections."

2. A single stage of the method. Define

$$C = \begin{bmatrix} B \\ A \end{bmatrix}, \quad f = \begin{bmatrix} d \\ b \end{bmatrix}, \quad C_\mu = \begin{bmatrix} \mu B \\ A \end{bmatrix}, \quad f_\mu = \begin{bmatrix} \mu d \\ b \end{bmatrix}, \tag{2.1}$$

where $\mu > 0$. We consider the first stage of Householder QR factorization applied to C_μ, examining what happens as $\mu \to \infty$. In outline, we write the transformed matrix as

$$C'_\mu = \begin{bmatrix} \mu B'_\mu \\ A'_\mu \end{bmatrix},$$

divide the first p rows of C'_μ by μ and take the limit as $\mu \to \infty$, obtaining a matrix

$$C' = \begin{bmatrix} B' \\ A' \end{bmatrix}$$

that is independent of μ. Then we view C' as the result of an update of C. We will show that all the elements of C' remain finite as $\mu \to \infty$. In this way, we set up a one-to-one correspondence between performing the first step of Householder QR factorization on C_μ,

$$\begin{bmatrix} \mu B \\ A \end{bmatrix} \to \begin{bmatrix} \mu B'_\mu \\ A'_\mu \end{bmatrix}, \tag{2.2}$$

and carrying out the update

$$\begin{bmatrix} B \\ A \end{bmatrix} \to \begin{bmatrix} B' \\ A' \end{bmatrix} \tag{2.3}$$

on the unscaled matrix C. Our aim is to determine the nature of this update.

We mention that Stewart [17] compares the equations for Householder QR factorization on C_μ with those for the EG method. While his analysis is related to ours, his aim (to understand the method of weighting) is different and he does not take limits.

We will denote by a_j, b_j, and c_j the jth columns of A, B, and C_μ, respectively (we will have no need to refer to the columns of C, so this slightly inconsistent notation should not cause any confusion); similarly, we write $C_\mu = (c_{ij})$. The first stage of Householder QR factorization applied to C_μ is

$$C'_\mu = C_\mu - 2v_\mu \frac{v_\mu^T C_\mu}{v_\mu^T v_\mu},$$

where

$$v_\mu = c_1 + \tau \|c_1\|_2 e_1, \qquad \tau = \operatorname{sign}(c_{11}), \tag{2.4}$$

which we can rewrite as

$$c'_{ij} = c_{ij} - 2v_\mu(i) \frac{v_\mu^T c_j}{v_\mu^T v_\mu}, \qquad i = 1\colon p+m, \quad j = 1\colon n. \tag{2.5}$$

We assume the use of column pivoting, so that columns are interchanged, if necessary, to ensure that

$$\|c_1\|_2 = \max_j \|c_j\|_2.$$

For notational simplicity, we assume that no interchanges are required by column pivoting. Our first concern is with the behavior of the multipliers

$$2\frac{v_\mu^T c_j}{v_\mu^T v_\mu}, \qquad j = 1\colon n,$$

which we will refer to as *Householder multipliers*, as $\mu \to \infty$. As a preliminary, we define ζ by

$$\|c_1\|_2 = \sqrt{\mu^2\|b_1\|_2^2 + \|a_1\|_2^2} = \mu\sqrt{\|b_1\|_2^2 + \frac{1}{\mu^2}\|a_1\|_2^2} = \mu\zeta,$$

and note that

$$\lim_{\mu\to\infty} \zeta = \|b_1\|_2.$$

We have

$$v_\mu^T c_j = c_1^T c_j + \tau\|c_1\|_2 c_{1j} = \mu^2 b_1^T b_j + a_1^T a_j + \mu^2\tau\zeta b_{1j}$$

and

$$v_\mu^T v_\mu = 2\|c_1\|_2(\|c_1\|_2 + |c_{11}|) = 2\mu^2\zeta(\zeta + |b_{11}|).$$

Hence

$$2\frac{v_\mu^T c_j}{v_\mu^T v_\mu} = \frac{b_1^T b_j + a_1^T a_j/\mu^2 + \tau\zeta b_{1j}}{\zeta(\zeta + |b_{11}|)},$$

which yields

$$\lim_{\mu\to\infty} 2\frac{v_\mu^T c_j}{v_\mu^T v_\mu} = \frac{b_1^T b_j + \tau\|b_1\|_2 b_{1j}}{\|b_1\|_2(\|b_1\|_2 + |b_{11}|)} = 2\frac{v_B^T b_j}{v_B^T v_B}, \tag{2.6}$$

where

$$v_B = b_1 + \tau\|b_1\|_2 e_1.$$

We see that, in the limit, A's contribution to the Householder multiplier $2v_\mu^T c_j/v_\mu^T v_\mu$ is lost.

Using (2.5) and (2.4) we have, for $i = 1\colon p$,

$$\begin{aligned} b'_{ij} := \lim_{\mu\to\infty} b_{ij}(\mu)' &= \lim_{\mu\to\infty} \frac{c'_{ij}}{\mu} \\ &= \lim_{\mu\to\infty} \left(b_{ij} - \frac{2v_\mu(i)}{\mu}\frac{v_\mu^T c_j}{v_\mu^T v_\mu} \right) \\ &= b_{ij} - 2v_B(i)\frac{v_B^T b_j}{v_B^T v_B}. \end{aligned}$$

This is just the first stage of Householder QR factorization applied to B, which corresponds to the application of the Householder update

$$C' = C - 2v_C\frac{v_C^T C}{v_C^T v_C}, \quad v_C = \begin{bmatrix} v_B \\ 0 \end{bmatrix} \tag{2.7}$$

to the unscaled matrix C. Moreover, $\|c_i\|_2/\|c_j\|_2 \to \|b_i\|_2/\|b_j\|_2$ so, in the limit, column pivoting based on the columns of C is equivalent to column pivoting based on the columns of B.

Now we turn to the rows of A. In the limit as $\mu \to \infty$, we have, from (2.4)–(2.6),

$$a'_{ij} = a_{ij} - 2a_{i1}\frac{v_B^T b_j}{v_B^T v_B},$$

which corresponds to the update

$$C' = C - 2v_A\frac{v_C^T C}{v_C^T v_C}, \quad v_A = \begin{bmatrix} 0 \\ a_1 \end{bmatrix}. \tag{2.8}$$

Combining (2.7) and (2.8) we see that, in the limit as $\mu \to \infty$, carrying out one step of Householder QR factorization on C_μ is equivalent to carrying out the outer product update

$$C' = PC = C - 2v\frac{v_C^T C}{v_C^T v_C}, \quad \text{where} \quad v = \begin{bmatrix} v_B \\ a_1 \end{bmatrix}, \tag{2.9}$$

on the unscaled matrix C. Note that the transformation matrix P is not a Householder matrix, because $v \neq v_C$.

It is natural to ask whether the equations we have derived are equivalent to those for the EG method described in section 1. Recall that the EG method QR factorizes B as in (1.2) and then eliminates the first p columns of $\widetilde{A} = A\Pi$ by Gaussian elimination. It is easily shown that the first row of $[\, R_1 \quad R_2 \,]$ is proportional to

$$[\, b_1^T b_1 \quad b_1^T b_2 \quad \dots \quad b_1^T b_n \,]. \tag{2.10}$$

The first column of $\widetilde{A}$ is therefore zeroed by performing one step of Gaussian elimination with this row as the pivot row. Using the fact that $v_B^T v_B = 2v_B^T b_1$, it is easy to show that for the new method we are effectively performing Gaussian elimination with the "virtual pivot row"

$$[\, v_B^T b_1 \quad v_B^T b_2 \quad \dots \quad v_B^T b_n \,], \tag{2.11}$$

which is not actually present in the matrix C. The vectors (2.10) and (2.11) are proportional only if B is upper trapezoidal, which shows that the new method and the EG method are mathematically different in general.

3. The complete method. By carrying out p eliminations of the type described in the previous section, we reduce C to the form

$$\begin{bmatrix} R_1 & R_2 \\ 0 & A'_2 \end{bmatrix},$$

where $R_1 \in \mathbb{R}^{p\times p}$ is upper triangular and nonsingular. We mention that in place of column pivoting some other pivoting strategy that keeps $R_1^{-1}R_2$ small could be used instead; cf. [2].

The remaining $\min(n-p, m-1)$ steps of the method consist of applying standard Householder QR factorization with column pivoting to the matrix A'_2, after which the whole of C has been reduced to upper trapezoidal form. By applying the same sequence of updates to f and solving the resulting triangular system, we obtain the solution to the LSE problem (1.1). We summarize our method as follows.

ALGORITHM EH *This algorithm solves the LSE problem* (1.1).

1. Let
$$C^{(1)} = \begin{bmatrix} B \\ A \end{bmatrix}, \qquad f^{(1)} = \begin{bmatrix} d \\ b \end{bmatrix}, \qquad q = p + m.$$
2. Stage I
for $k = 1\colon p$
Interchange the columns of $C^{(k)}$ so that $\|c_k^{(k)}(k\colon p)\|_2 = \max_{j \geq k} \|c_j^{(k)}(k\colon p)\|_2$.
Update
$$C^{(k+1)}(k\colon q, k\colon n) = C^{(k)}(k\colon q, k\colon n) - 2v_k(k\colon q)\frac{v_k(k\colon p)^T C^{(k)}(k\colon p, k\colon n)}{v_k(k\colon p)^T v_k(k\colon p)},$$
$$f^{(k+1)}(k\colon q) = f^{(k)}(k\colon q) - 2v_k(k\colon q)\frac{v_k(k\colon p)^T f^{(k)}(k\colon p)}{v_k(k\colon p)^T v_k(k\colon p)},$$
where
$$v_k(i) = \begin{cases} 0, & i = 1\colon k-1, \\ c_{kk}^{(k)} + \operatorname{sign}\left(c_{kk}^{(k)}\right)\left\|c_k^{(k)}(k\colon p)\right\|_2, & i = k, \\ c_{ik}^{(k)}, & i = k+1\colon q. \end{cases}$$
end
3. Stage II (standard Householder QR factorization with column pivoting)
for $k = p+1\colon \min(n, m+p-1)$
Interchange the columns of $C^{(k)}$ so that $\left\|c_k^{(k)}(k\colon q)\right\|_2 = \max_{j \geq k} \left\|c_j^{(k)}(k\colon q)\right\|_2$.
Update
$$C^{(k+1)}(k\colon q, k\colon n) = C^{(k)}(k\colon q, k\colon n) - 2v_k(k\colon q)\frac{v_k(k\colon q)^T C^{(k)}(k\colon q, k\colon n)}{v_k(k\colon q)^T v_k(k\colon q)},$$
$$f^{(k+1)}(k\colon q) = f^{(k)}(k\colon q) - 2v_k(k\colon q)\frac{v_k(k\colon q)^T f^{(k)}(k\colon q)}{v_k(k\colon q)^T v_k(k\colon q)},$$
where
$$v_k(i) = \begin{cases} 0, & i = 1\colon k-1, \\ c_{kk}^{(k)} + \operatorname{sign}\left(c_{kk}^{(k)}\right)\left\|c_k^{(k)}(k\colon q)\right\|_2, & i = k, \\ c_{ik}^{(k)}, & i = k+1\colon q. \end{cases}$$
end
4. Solve the triangular system $C^{(k+1)}(1\colon n, 1\colon n)y = f^{(k+1)}(1\colon n)$.
5. Obtain x by permuting y to take account of the column interchanges.

Algorithm EH implicitly computes a matrix factorization: ignoring column interchanges, we have
$$\underbrace{\begin{bmatrix} I & 0 \\ 0 & \widetilde{Q}_3 \end{bmatrix}}_{\text{Stage II}} \underbrace{\begin{bmatrix} \widetilde{Q}_1 & 0 \\ \widetilde{Q}_2 & I \end{bmatrix}}_{\text{Stage I}} \begin{bmatrix} B_1 & B_2 \\ A_1 & A_2 \end{bmatrix} = \begin{bmatrix} R_1 & R_2 \\ 0 & R_3 \\ 0 & 0 \end{bmatrix},$$
where $\widetilde{Q}_1$ and $\widetilde{Q}_3$ are orthogonal and $R_1 \in \mathbb{R}^{p\times p}$ and $R_3 \in \mathbb{R}^{(n-p)\times(n-p)}$ are upper triangular. Rewriting, we have
$$\begin{bmatrix} \widetilde{Q}_1 & 0 \\ \widetilde{Q}_3\widetilde{Q}_2 & \widetilde{Q}_3 \end{bmatrix} \begin{bmatrix} B_1 & B_2 \\ A_1 & A_2 \end{bmatrix} = \begin{bmatrix} R \\ 0 \end{bmatrix},$$

or, with $Q_1 = \widetilde{Q}_1^T$, $Q_2 = -\widetilde{Q}_2\widetilde{Q}_1^T$ and $Q_3 = \widetilde{Q}_3^T$,

$$\begin{array}{c} \\ p \\ m \end{array}\!\!\begin{array}{c} \begin{array}{cc} p & n-p \end{array} \\ \begin{bmatrix} B_1 & B_2 \\ A_1 & A_2 \end{bmatrix} \end{array} = \begin{array}{c} \\ p \\ m \end{array}\!\!\begin{array}{c} \begin{array}{cc} p & m \end{array} \\ \begin{bmatrix} Q_1 & 0 \\ Q_2 & Q_3 \end{bmatrix} \end{array} \begin{array}{c} \begin{array}{c} n \end{array} \\ \begin{bmatrix} R \\ 0 \end{bmatrix} \end{array} \begin{array}{l} \\ n \\ m-n+p \end{array}. \tag{3.1}$$

The EG method produces exactly the same factorization; the difference in the methods lies in the intermediate quantities computed. Note that another way to derive both the EG method and Algorithm EH is to substitute the factorization into the associated augmented system; this is done in [4] and [5] for the EG method.

4. Numerical stability. Now we consider the stability of Algorithm EH. First, we recall what is known about the stability of Householder QR factorization for solving the LS problem $\min_x \|b - Ax\|_2$. The method is normwise backward stable. Moreover, when column pivoting is used the following row-wise backward error result holds, which is of interest in situations in which the rows of A vary widely in norm. We employ the standard model of floating point arithmetic with unit roundoff u [12, Section 2.2], and define the constant

$$\widetilde{\gamma}_k = \frac{cku}{1 - cku},$$

where k is a positive integer and c is a small integer constant whose exact value is unimportant. We assume that the signs in the Householder vectors are chosen as in (2.4), which is the standard choice. For the other choice of sign, the following theorem is invalid (see [8] for more details).

THEOREM 4.1. *Let the LS problem* $\min_x \|b - Ax\|_2$*, where* $A \in \mathbb{R}^{m\times n}$ *is of full rank* n*, be solved using Householder QR factorization with column pivoting. Then the computed solution* $\widehat{x}$ *is the exact solution to*

$$\min_x \|(b + \Delta b) - (A + \Delta A)x\|_2,$$

where the perturbations satisfy

$$|\Delta a_{ij}| \le j^2\widetilde{\gamma}_m\alpha_i \max_s |a_{is}|, \qquad |\Delta b_i| \le n^2\widetilde{\gamma}_m\beta_i \max(\phi \max_j |a_{ij}|, |b_i|),$$

where

$$\alpha_i = \frac{\max_{j,k} |\widehat{a}_{ij}^{(k)}|}{\max_j |a_{ij}|}, \quad \beta_i = \frac{\max\bigl(\phi \max_{j,k} |\widehat{a}_{ij}^{(k)}|, |\widehat{b}_i^{(k)}|\bigr)}{\max(\phi \max_j |a_{ij}|, |b_i|)}, \quad \phi = \max_k \frac{\|\widehat{b}^{(k)}(k\colon m)\|_2}{\|\widehat{a}_k^{(k)}(k\colon m)\|_2}.$$

Proof. A similar result was originally proved under some additional assumptions by Powell and Reid [14]. The analysis was reworked by Cox and Higham [8]. The result as stated is a slightly improved version of the result in [8], with a minor error in the Δb bound corrected. □

Theorem 4.1 shows that, when column pivoting is used, the backward errors ΔA and Δb are bounded row-wise in terms of the element growth within each row, as measured by the α_i and β_i. Ideally, we would like the overall row-wise growth factor

$$\rho_{m,n} = \max_i\{\, \alpha_i, \beta_i \,\}$$

to be of order 1. In general, $\rho_{m,n}$ is unbounded. However, two techniques lead to a bounded $\rho_{m,n}$. Prior to carrying out the factorization, we can sort the rows so that

$$\|A(i,:)\|_\infty = \max_{j\geq i} \|A(j,:)\|_\infty, \qquad i = 1\colon m.$$

Alternatively, we can use row pivoting: at the kth stage of the factorization, after the column interchange has taken place, we interchange rows to ensure that

$$|a_{kk}^{(k)}| = \max_{i\geq k} |a_{ik}^{(k)}|.$$

The following result was obtained by Powell and Reid [14] for row pivoting and by Cox and Higham [8] for row sorting. As usual for growth factor bounds, this one assumes exact arithmetic; however, the growth of exact and computed quantities differs by only $O(u)$.

THEOREM 4.2. *With row pivoting or row sorting in Householder QR factorization with column pivoting applied to* $A \in \mathbb{R}^{m\times n}$,

$$\rho_{m,n} \leq \sqrt{m}(1+\sqrt{2})^{n-1}.$$

The bound of the theorem can be nearly attained, but $\rho_{m,n}$ is almost always small in practice.

The scalar ϕ in Theorem 4.1 is easily seen to be independent of the row ordering and so is beyond our control.

The limit technique used in the derivation of Algorithm EH can be used to obtain a row-wise backward error result. We apply Theorem 4.1 to $\min_x \|f_\mu - C_\mu x\|_2$ (see (2.1)) and then use a straightforward limit argument to deduce the following theorem. As a check on the plausibility of the theorem, we note that a key result used in the proof of Theorem 4.1 is that the Householder multipliers are all bounded by $\sqrt{2}$ (without column pivoting these multipliers are unbounded). For Stage I of Algorithm EH, the Householder multipliers

$$\eta_{k,j} = 2\frac{v_k(k\colon p)^T c_j^{(k)}(k\colon p)}{v_k^T(k\colon p)v_k(k\colon p)}$$

are precisely those determined by QR factorization with column pivoting applied to B, as we have already noted. Therefore $|\eta_{k,j}| \leq \sqrt{2}$. Stage II of Algorithm EH is standard QR factorization with column pivoting, so again the multipliers are bounded by $\sqrt{2}$.

THEOREM 4.3. *Let the LSE problem* (1.1) *be solved by Algorithm EH. Then the computed solution* $\widehat{x}$ *is the exact solution to*

$$\min_{(B+\Delta B)x=d+\Delta d} \|(b+\Delta b) - (A+\Delta A)x\|_2,$$

where, defining

$$\Delta C = \begin{bmatrix} \Delta B \\ \Delta A \end{bmatrix}, \qquad \Delta f = \begin{bmatrix} \Delta d \\ \Delta b \end{bmatrix},$$

we have

$$|\Delta c_{ij}| \leq j^2 \widetilde{\gamma}_{p+m}\alpha_i \max_s |c_{is}|, \qquad |\Delta f_i| \leq n^2\widetilde{\gamma}_{p+m}\beta_i \max(\phi \max_j |c_{ij}|, |f_i|),$$

with

$$\alpha_i = \frac{\max_{j,k} |\widehat{c}_{ij}^{(k)}|}{\max_j |c_{ij}|}, \quad \beta_i = \frac{\max\bigl(\phi \max_{j,k} |\widehat{c}_{ij}^{(k)}|, |\widehat{f}_i^{(k)}|\bigr)}{\max(\phi \max_j |c_{ij}|, |f_i|)},$$

$$\phi = \max\left\{ \max_{1\le k\le p} \frac{\|\widehat{f}^{(k)}(k\colon p)\|_2}{\|\widehat{c}_k^{(k)}(k\colon p)\|_2}, \max_{p+1\le k\le p+m} \frac{\|\widehat{f}^{(k)}(k\colon p+m)\|_2}{\|\widehat{c}_k^{(k)}(k\colon p+m)\|_2} \right\}.$$

Theorem 4.3 shows that Algorithm EH is row-wise backward stable provided that the row-wise growth factor $\rho_{m,n} = \max_i\{\,\alpha_i, \beta_i\,\}$ and the scalar ϕ are both of order 1. As for the LS problem, we can use row sorting or row pivoting in order to bound $\rho_{m,n}$, and we obtain a bound analogous to that in Theorem 4.2. For row sorting we sort the rows of A and B separately (that is, we do not sort the rows of C, which would change the LSE problem). For row pivoting, we interchange the rows of B in Stage I and the rows of A in Stage II.

The two stages of Algorithm EH can be implemented as a single loop with the help of a parameter that specifies the extent of the Householder vector, yielding the succinct pseudocode shown in Figure 4.1. The EG method and a variant of it based on modified Gram-Schmidt orthogonalization can be expressed in similarly concise fashions, as shown in [4], [5].

How does the stability of Algorithm EH compare with that of the EG method? The only published result for the EG method is one of Barlow and Handy [2], which bounds $\|C - \widehat{Q}\widehat{R}\|_2$ for the factorization (3.1) in terms of $\|C\|_2$; this result does not reveal the stability of the overall solution process or provide any information about row-wise stability.

From the remarks at the end of section 2 we know that if B is upper triangular, then Algorithm EH is identical to the EG method. We can therefore interpret the EG method as the two-stage process: (1) QR factorize B, using column pivoting, (2) apply Algorithm EH to the LSE problem with a triangular constraint matrix. Theorem 4.3 applies to stage (2), and an analogue of Theorem 4.1 for the QR factorization itself applies to Stage 1 [8] (essentially, we obtain $(B + \Delta B)\Pi = \widehat{Q}\widehat{R}$, with ΔB bounded in the same way as ΔA in Theorem 4.1). These two results can be combined, using the same techniques as in the proof of Theorem 4.1, to obtain a new result for the EG method.

THEOREM 4.4. *Theorem* 4.3 *holds also for the EG method.*

5. Numerical experiments. Backward errors of an approximate solution y to the LSE problem (1.1) can be defined by minimizing suitable measures of the quadruple $(\Delta A, \Delta b, \Delta B, \Delta d)$ over all perturbations ΔA, Δb, ΔB, and Δd for which y solves $\min_{(B+\Delta B)y=d+\Delta d} \|b + \Delta b - (A + \Delta A)y\|_2$. Ideally, we would verify the row-wise backward stability results in Theorems 4.3 and 4.4 by comparing the actual row-wise backward errors with the bounds for a range of problems. Unfortunately, no computable expression is known for any backward error (even just the normwise backward error) of an arbitrary approximate LSE solution. We therefore compute upper bounds for the backward errors using a technique developed in [7].

Given an approximate LSE solution y, suppose we have determined perturbations ΔB and Δd such that $(B + \Delta B)y = d + \Delta d$. Let ΔA_* and Δb_* be solutions of the problem

$$\min\{\, \|\,[\,\Delta A \quad \theta \Delta b\,]\,\|_F : y \text{ solves } \min_{(B+\Delta B)x=d+\Delta d} \|b + \Delta b - (A + \Delta A)x\|_2 \,\}, \tag{5.1}$$

This algorithm solves the LSE problem $\min_{Bx=d} \|b - Ax\|_2$*, where* $A \in \mathbb{R}^{m\times n}$*,* $B \in \mathbb{R}^{p\times n}$ *and* $m + p \geq n \geq p$*. A flag rp determines whether row pivoting is carried out.*

$C^{(1)} = \begin{bmatrix} B \\ A \end{bmatrix}$, $f^{(1)} = \begin{bmatrix} d \\ b \end{bmatrix}$, $top = p$, $q = p + m$.

for $k = 1\colon \min(n, m + p - 1)$

Permute columns so that $\|C^{(k)}(k\colon top, k)\|_2 = \max_{j\geq k} \|C^{(k)}(k\colon top, j)\|_2$.

If rp
Permute rows so that $|c_{kk}^{(k)}| = \max_{i=k\colon top} |c_{ik}^{(k)}|$.
Apply same row interchanges to $f^{(k)}$.
end

% Implicitly construct Householder-like matrix $I - \beta vv'^T$.
$v = C^{(k)}(k\colon q, k)$
$s = \operatorname{sign}(v_1)\|C^{(k)}(k\colon top, k)\|_2$
$v_1 = v_1 + s$
$\beta = 1/(s\, v_1)$

$C^{(k+1)}(k\colon q, k\colon n) = C^{(k)}(k\colon q, k\colon n)$
$\qquad - \beta\, v(v(1\colon top - k + 1)^T C^{(k)}(k\colon top, k\colon n))$
$f^{(k+1)}(k\colon q) = f^{(k)}(k\colon q) - \beta\, v(v(1\colon top - k + 1)^T f^{(k)}(k\colon top))$

% Once $k > p$ we do standard Householder QR factorization.
If $k = p$, $top = q$, end

end

Solve $C^{(k+1)}(1\colon n, 1\colon n)x = f^{(k+1)}(1\colon n)$ by substitution.
Permute x to undo the effect of column pivoting.

FIG. 4.1. *Concise pseudocode for Algorithm EH.*

where θ is a parameter. How to compute ΔA_* and Δb_* is explained in [7]. Here we are minimizing over ΔA and Δb for fixed ΔB and Δd instead of minimizing over all ΔA, Δb, ΔB and Δd. Therefore by taking an appropriate measure of the quadruple $(\Delta A_*, \Delta b_*, \Delta B, \Delta d)$ we obtain an upper bound for the corresponding backward error. We choose the perturbations

$$\Delta B_* = \frac{\|B\|_2\|y\|_2}{(\|B\|_2\|y\|_2 + \|d\|_2)} \frac{ry^T}{y^T y}, \quad \Delta d_* = \frac{-\|d\|_2}{\|B\|_2\|y\|_2 + \|d\|_2} r,$$

where $r = d - By$. These are the perturbations corresponding to the relative 2-norm backward error for the constraint system [16]; in other words, they minimize $\max(\|\Delta B\|_2/\|B\|_2, \|\Delta d\|_2/\|d\|_2)$. We define two backward error bounds in terms of

the perturbations ΔA_*, Δb_*, ΔB_* and Δd_*: the normwise backward error bound

$$\eta_N = \max\left\{ \frac{\|\Delta A_*\|_2}{\|A\|_2}, \frac{\|\Delta b_*\|_2}{\|b\|_2}, \frac{\|\Delta B_*\|_2}{\|B\|_2}, \frac{\|\Delta d_*\|_2}{\|d\|_2} \right\}$$

and the row-wise backward error bound

$$\eta_R = \max_i \frac{\|\Delta G(i,:)\|_2}{\|G(i,:)\|_2},$$

where

$$G = \begin{bmatrix} B & d \\ A & b \end{bmatrix}, \qquad \Delta G = \begin{bmatrix} \Delta B_* & \Delta d_* \\ \Delta A_* & \Delta b_* \end{bmatrix}.$$

In our experiments we took $\theta = 1$ in (5.1) since our chosen A, B, d, and f all have norms of order 1.

The experiments were performed in MATLAB, using simulated single precision arithmetic ($u = 2^{-24} \approx 5.96 \times 10^{-8}$). The backward error bounds were computed in double precision arithmetic.

We give results for four problems with $m = 16$, $n = 10$, $p = 6$. We denote by randn a matrix or vector from the normal(0,1) distribution and by randsvd(κ) a random matrix with 2-norm condition number κ and geometrically distributed singular values, generated by the routine `randsvd` in the Test Matrix Toolbox [11].

PROBLEM 1 $A =$ randn, $b =$ randn, $B =$ randn, $d =$ randn.

PROBLEM 2 $A =$ randsvd(10), $b =$ randn, $B =$ randsvd(10^4), $d =$ randn.

PROBLEM 3 $A =$ randsvd(10^6), $b =$ randn, $B =$ randsvd(10), $d =$ randn.

PROBLEM 4 $A =$ randsvd(10^4), $b =$ randn, $B =$ randsvd(10^4), $d =$ randn.

For each problem, a parameter tol $\in (0, 1]$ determines a row scaling applied to the original data:

$$[\,A \quad b\,] \leftarrow D_{m,\mathrm{tol}}\,[\,A \quad b\,], \quad [\,B \quad d\,] \leftarrow D_{p,\mathrm{tol}}\,[\,B \quad d\,],$$

where

$$D_{k,\mathrm{tol}} = \mathrm{diag}(\theta^k, \theta^{k-1}, \ldots, \theta, 1), \quad \theta^k = \mathrm{tol}.$$

The backward error results are given in Tables 5.1–5.3. We give results for Algorithm EH with no row interchanges and with row sorting; the results for row pivoting were very similar and so are omitted. We do not report results for the EG method because they were very similar to those for Algorithm EH. For comparison, we also give results for the null space method (implemented using generalized QR factorization, as in LAPACK). Results for tol = 1 are given only for Problem 1; for Problems 2 and 3 with tol = 1 the backward error bounds were again all of order u.

The main observations from the results are as follows.

1. The importance of row pivoting or row sorting in Algorithm EH for badly row scaled problems is shown by Tables 5.1–5.3, in which when tol is small η_R is much smaller when row sorting is used than when it is not. Row sorting produces a growth factor $\rho_{m,n}$ of order 1 in each case.
2. Tables 5.2 and 5.3 show that a large value of ϕ does not necessarily lead to a large row-wise backward error. Although ϕ is theoretically independent of the row ordering, it varies between Algorithm EH with and without row sorting in some of the tests as a result of roundoff: in these cases ϕ is determined by ratios of quantities that are of order u and have large relative error.

TABLE 5.1
Results for Problem 1 *(random normal A and B).*

tol = 1

	η_N	η_R	$\rho_{m,n}$	ϕ
Alg. EH (no row interchanges)	6.3e-8	6.4e-8	4.5e0	1.5e0
Alg. EH (row sorting)	4.8e-8	4.5e-8	2.6e0	1.5e0
Null space method	3.5e-8	3.9e-8		

tol = 10^{-7}

	η_N	η_R	$\rho_{m,n}$	ϕ
Alg. EH (no row interchanges)	9.6e-4	2.1e-1	2.5e7	3.8e0
Alg. EH (row sorting)	3.3e-8	4.3e-7	3.0e0	1.7e1
Null space method	5.1e-8	2.7e-7		

TABLE 5.2
Results for Problem 2 *($\kappa_2(A) = 10$, $\kappa_2(B) = 10^4$), $\mathrm{tol} = 10^{-7}$.*

	η_N	η_R	$\rho_{m,n}$	ϕ
Alg. EH (no row interchanges)	4.3e-4	1.1e-2	6.6e6	4.6e3
Alg. EH (row sorting)	2.4e-8	1.6e-7	2.7e0	1.4e4
Null space method	7.2e-8	2.2e-7		

3. The normwise backward error bound η_N is small in every case except for Algorithm EH without row interchanges in two instances. However, using numerical optimization by direct search we found alternative choices of ΔB_* and Δd_* for which $\eta_N \approx u$ in these two examples, confirming normwise backward stability. In the cases where Algorithm EH without row interchanges produced a large value of η_R, we were unable to reduce this value significantly using direct search.
4. The null space method gives very similar values of η_N and η_R to Algorithm EH with row sorting. It is surprising that η_R is small for the null space method when $\mathrm{tol} = 10^{-7}$ because our implementation does not incorporate any form of row interchanges. The reason for this excellent performance is not clear (it is not explained by the normwise backward error results in [6], for example).

Finally, although our concern in this work is with backward errors rather than forward errors, we carried out an experiment to show the effect of row interchanges on the forward error, $\|x - \widehat{x}\|_2/\|x\|_2$. For Problems 1 and 4 we computed forward errors by taking as the exact solution the one computed in double precision by Algorithm EH with row sorting. Table 5.4 shows that row sorting can greatly decrease the forward error when A and B have badly scaled rows.

6. Concluding remarks. We have used a limit argument to derive a method, Algorithm EH, for solving the LSE problem and have derived a row-wise backward error bound for the method by exploiting existing error analysis. The row-wise backward error is guaranteed to be small when the growth factor $\rho_{m,n}$ is small (which is usually the case with row pivoting or sorting) and when the scalar ϕ is small (ϕ can be arbitrarily large, and is independent of the row interchanges).

The method is closely related to the EG method, which dates back to the 1960s, and this relation enabled us to obtain a row-wise backward error bound for that

TABLE 5.3
Results for Problem 3 ($\kappa_2(A) = 10^6$, $\kappa_2(B) = 10$), tol $= 10^{-7}$.

	η_N	η_R	$\rho_{m,n}$	ϕ
Alg. EH (no row interchanges)	1.3e-8	2.8e-2	4.3e6	1.1e3
Alg. EH (row sorting)	2.8e-8	1.3e-7	2.2e0	1.2e3
Null space method	2.8e-8	7.3e-7		

TABLE 5.4
Forward errors $\|x - \widehat{x}\|_2/\|x\|_2$.

	Problem 1 tol $= 1$	Problem 1 tol $= 10^{-7}$	Problem 4 tol $= 1$	Problem 4 tol $= 10^{-7}$
Alg. EH (no row interchanges)	2.5e-7	6.6e-1	2.4e-5	1.3e0
Alg. EH (row sorting)	1.7e-7	1.2e-6	3.1e-6	2.1e-5
Null space method	1.2e-7	2.8e-6	4.9e-5	5.8e-6

method, too. Which method should be preferred? There is no difference between the two methods in terms of the backward error results or the numerical backward error bounds that we have evaluated. The methods also have similar operation counts. Algorithm EH can be coded slightly more concisely, since it carries out transformations on B-over-A as a whole while the EG method applies Householder transformations to B and separate elimination operations to A, and this difference should make Algorithm EH more efficient on high-performance computers.

REFERENCES

[1] J. L. BARLOW, *Error analysis and implementation aspects of deferred correction for equality constrained least squares problems*, SIAM J. Numer. Anal., 25 (1988), pp. 1340–1358.

[2] J. L. BARLOW AND S L. HANDY, *The direct solution of weighted and equality constrained least-squares problems*, SIAM J. Sci. Stat. Comput., 9 (1988), pp. 704–716.

[3] J. L. BARLOW AND U. B. VEMULAPATI, *A note on deferred correction for equality constrained least squares problems*, SIAM J. Numer. Anal., 29 (1992), pp. 249–256.

[4] Å. BJÖRCK, *Iterative refinement of linear least squares solutions* II, BIT, 8 (1968), pp. 8–30.

[5] Å. BJÖRCK AND G. H. GOLUB, *Iterative refinement of linear least squares solutions by Householder transformation*, BIT, 7 (1967), pp. 322–337.

[6] A. J. COX AND N. J. HIGHAM, *Accuracy and Stability of the Null Space Method for Solving the Equality Constrained Least Squares Problem*, BIT, 39 (1999), pp. 34–50.

[7] A. J. COX AND N. J. HIGHAM, *Backward Error Bounds for Constrained Least Squares Problems*, BIT, 39 (1999), pp. 210–227.

[8] A. J. COX AND N. J. HIGHAM, *Stability of Householder QR factorization for weighted least squares problems*, Numerical Analysis 1997, Proceedings of the 17th Dundee Biennial Conference, D. F. Griffiths, D. J. Higham, and G. A. Watson, eds., Pitman Res. Notes Math. Ser., 380, Addison Wesley Longman, Harlow, UK, 1998, pp. 57–73.

[9] M. GULLIKSSON, *Backward error analysis for the constrained and weighted linear least squares problem when using the weighted QR factorization*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 675–687.

[10] M. GULLIKSSON AND P. WEDIN, *Modifying the QR-decomposition to constrained and weighted linear least squares*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1298–1313.

[11] N. J. HIGHAM, *The Test Matrix Toolbox for* MATLAB *(version* 3.0*)*, Numerical Analysis Report No. 276, Manchester Centre for Computational Mathematics, Manchester, England, 1995.

[12] N. J. HIGHAM, *Accuracy and Stability of Numerical Algorithms*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1996.

[13] C. L. LAWSON AND R. J. HANSON, *Solving Least Squares Problems*, SIAM, Philadelphia, PA, 1995; revised republication of work first published by Prentice–Hall, 1974.

[14] M. J. D. Powell and J. K. Reid, *On applying Householder transformations to linear least squares problems*, in Proceedings of the IFIP Congress 1968, North-Holland, Amsterdam, The Netherlands, 1969, pp. 122–126.
[15] J. K. Reid, *Implicit scaling of linear least squares problems*, Report RAL-TR-98-027, Atlas Centre, Rutherford Appleton Laboratory, Didcot, Oxon, UK, 1998. 12 pp.
[16] J. L. Rigal and J. Gaches, *On the compatibility of a given solution with the data of a linear system*, J. Assoc. Comput. Mach., 14(3) (1967), pp. 543–548.
[17] G. W. Stewart, *On the weighting method for least squares problems with linear equality constraints*, BIT, 37 (1997), pp. 961–967.
[18] C. F. Van Loan, *On the method of weighting for equality-constrained least-squares problems*, SIAM J. Numer. Anal., 22 (1985), pp. 851–864.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 1, pp. 327–327
© 1999 Society for Industrial and Applied Mathematics

NOTE ON PERTURBATION THEORY FOR ALGEBRAIC RICCATI EQUATIONS*

M. KONSTANTINOV† AND P. PETKOV‡

Abstract. The expressions for the induced norms of two complex matrix operators, given in [J.-G. Sun, *SIAM J. Matrix Anal. Appl.*, 19 (1998), pp. 39–65], must be corrected. In this note we give the true values of these induced norms, which are involved in the perturbation analysis of matrix algebraic Riccati equations in the complex domain.

PII. S0895479898338111

In a recent paper [1] improvements of existing perturbation bounds for continuous- and discrete-time algebraic Riccati equations (CARE and DARE, respectively) are proposed. An important quantity here is the value of the induced norm of the operator $\mathbf{P}_c : \mathcal{C}^{n\times n} \to \mathcal{C}^{n\times n}$, defined from $\mathbf{P}_c N = \mathbf{L}_c^{-1}(XN + N^H X)$ for CARE and a similar operator $\mathbf{P}_d$ for DARE, where $\mathbf{L}_c$ is an invertible Lyapunov operator. (Note that the operators $\mathbf{P}$ are additive but not homogeneous and hence they are not linear over $\mathcal{C}$.)

The expressions (3.33) for the norm of $\mathbf{P}_c$ and (4.53) for the norm of $\mathbf{P}_d$ should be corrected as follows. (There are several ways to derive the proper expressions; the most instructive is to consider the realification of the corresponding operator.)

Let X_c be the positive definite solution of CARE (see (1.3)), and let T_c be the matrix (3.31) of $\mathbf{L}_c$. Let X_d and T_d be the corresponding matrices for DARE, defined from (1.5) and (4.51). Finally, let $K = X_d F(I + GX_d)^{-1}A$ be defined from (4.3). Denote

$$T_c^{-1}(I_n \otimes X_c) = \Psi_{c,0} + \jmath\Psi_{c,1}, T_c^{-1}(X_c^\top \otimes I_n)\Pi = \Theta_{c,0} + \jmath\Theta_{c,1},$$
$$T_d^{-1}(I_n \otimes K^H) = \Psi_{d,0} + \jmath\Psi_{d,1}, T_d^{-1}(K^\top \otimes I_n)\Pi = \Theta_{d,0} + \jmath\Theta_{d,1}.$$

Then the correct values for (3.33) and (4.53) are

$$p_e = \left\| \begin{bmatrix} \Psi_{e,0} + \Theta_{e,0} & \Theta_{e,1} - \Psi_{e,1} \\ \Psi_{e,1} + \Theta_{e,1} & \Psi_{e,0} - \Theta_{e,0} \end{bmatrix} \right\|_2; \quad e = c, d.$$

It may be observed that the operators $\mathbf{P}_e$ are somehow misfortunate in the complex case. In the report of Konstantinov et al. [2] there was a mistake in the vec representation of $\mathbf{P}_c$, which is corrected in the published version.

REFERENCES

[1] J.-G. SUN, *Perturbation theory for algebraic Riccati equations*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 39–65.

[2] M. KONSTANTINOV, P. PETKOV, D. GU, AND I. POSTLETHWAITE, *Perturbation Techniques for Linear Control Problems*, LUED report 95–7, Dept. Engineering, Leicester University, Leicester, UK, Feb. 1995; Ann. Inst. Archit. Genie Civil Sofia Fasc. II Math., 40 (1999), to appear.

*Received by the editors April 28, 1998; accepted for publication by P. Van Dooren September 1, 1998; published electronically September 30, 1999.

http://www.siam.org/journals/simax/21-1/33811.html

†University of Architecture and Civil Engineering, 1 Hr. Smirnenski Blv., 1421 Sofia, Bulgaria (mmk_fte@uacg.acad.bg).

‡Department of Automatics, Technical University of Sofia, 1756 Sofia, Bulgaria (php@mbox.digsys.bg).

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 1, pp. 328–354
© 1999 Society for Industrial and Applied Mathematics

ON THE SEQUENCE OF CONSECUTIVE POWERS OF A MATRIX IN A BOOLEAN ALGEBRA*

BART DE SCHUTTER† AND BART DE MOOR‡

Abstract. In this paper we consider the sequence of consecutive powers of a matrix in a Boolean algebra. We characterize the ultimate behavior of this sequence, we study the transient part of the sequence, and we derive upper bounds for the length of this transient part. We also indicate how these results can be used in the analysis of Markov chains and in max-plus algebraic system theory for discrete event systems.

Key words. Boolean algebra, Boolean matrices, transient behavior, Markov chains, max-plus algebra

AMS subject classifications. 06E99, 15A99, 16Y99

PII. S0895479897326079

1. Introduction.

In this paper we consider the sequence of consecutive powers of a matrix in a Boolean algebra. This sequence reaches a "cyclic" behavior after a finite number of terms. Even for more complex algebraic structures, such as the max-plus algebra (which has maximization and addition as its basic operations) this ultimate behavior has already been studied extensively by several authors (see, e.g., [1, 9, 13, 26] and the references therein). In this paper we completely characterize the ultimate behavior of the sequence of the consecutive powers of a matrix in a Boolean algebra. Furthermore, we also study the *transient* part of this sequence. More specifically, we give upper bounds for the length of the transient part of the sequence as a function of structural parameters of the matrix.

Our main motivation for studying this problem lies in the max-plus algebraic system theory for discrete event systems. Furthermore, our results can also be used in the analysis of the transient behavior of Markov chains.

This paper is organized as follows. In section 2 we introduce some of the notations and concepts from number theory, Boolean algebra, matrix algebra, and graph theory that will be used in the paper. In section 3 we characterize the ultimate behavior of the sequence of consecutive powers of a given matrix in a Boolean algebra, and we derive upper bounds for the length of the transient part of this sequence. In section 4 we briefly sketch how our results can be used in the analysis of Markov chains and in the max-plus algebraic system theory for discrete event systems. In this section we also explain why we have restricted ourselves to Boolean algebras in this paper and we indicate some of the phenomena that should be taken into account when extending our results to more general algebraic structures. Finally we present some conclusions in section 5.

*Received by the editors August 15, 1997; accepted for publication (in revised form) by V. Helmke August 3, 1998; published electronically October 8, 1999. This research was sponsored by the Concerted Action Project of the Flemish Community, GOA-MIPS, by the Belgian program on interuniversity attraction poles, IUAP P4-02 and IUAP P4-24, and by the ALAPEDES project of the European Community Training and Mobility of Researchers Program.

http://www.siam.org/journals/simax/21-1/32607.html

†Lab for Control Engineering, Faculty of Information Technology and Systems, Delft University of Technology, P.O. Box 5031, 2600 GA Delft, The Netherlands (b.deschutter@its.tudelft.nl).

‡ESAT/SISTA, K.U.Leuven, Kardinaal Mercierlaan 94, B-3001 Heverlee (Leuven), Belgium (bart.demoor@esat.kuleuven.ac.be). Research associate with the F. W. O. (Fund for Scientific Research, Flanders).

TABLE 2.1
The operations $\oplus$ and $\otimes$ for the Boolean algebra $(\{\mathtt{0},\mathtt{1}\},\oplus,\otimes)$.

$\oplus$	$\mathtt{0}$	$\mathtt{1}$
$\mathtt{0}$	$\mathtt{0}$	$\mathtt{1}$
$\mathtt{1}$	$\mathtt{1}$	$\mathtt{1}$

$\otimes$	$\mathtt{0}$	$\mathtt{1}$
$\mathtt{0}$	$\mathtt{0}$	$\mathtt{0}$
$\mathtt{1}$	$\mathtt{0}$	$\mathtt{1}$

2. Preliminaries.

2.1. Notation, definitions, and some lemmas from number theory. In this paper we use "vector" as a synonym for "column matrix." If a is a vector, then a_i is the ith component of a. If A is a matrix, then a_{ij} or $(A)_{ij}$ is the entry on the ith row and the jth column and $A_{\alpha\beta}$ is the submatrix of A obtained by removing all rows that are not indexed by the set α and all columns that are not indexed by the set β.

The set of the real numbers is denoted by $\mathbb{R}$, the set of the nonnegative integers by $\mathbb{N}$, and the set of the positive integers by $\mathbb{N}_0$.

If S is a set, then the number of elements of S is denoted by $\#S$. If γ is a set of positive integers, then the least common multiple of the elements of γ is denoted by $\operatorname{lcm}\gamma$ and the greatest common divisor of the elements of γ is denoted by $\gcd\gamma$.

If $x \in \mathbb{R}$, then $\lceil x \rceil$ is the smallest integer that is larger than or equal to x, and $\lfloor x \rfloor$ is the largest integer that is less than or equal to x.

LEMMA 2.1. *Let $p, q \in \mathbb{N}_0$ be coprime. The smallest integer n such that for any integer $m \geq n$ there exist two nonnegative integers α and β such that $m = \alpha p + \beta q$ is given by $n = (p-1)(q-1)$.*

Proof. See, e.g., the proof of Lemma 3.5.5 of [5]. □

Let $a_1, a_2, \ldots, a_n \in \mathbb{N}_0$ with $\gcd(a_1, a_2, \ldots, a_n) = 1$. We define $g(a_1, a_2, \ldots, a_n)$ to be the largest positive integer N for which the equation $a_1x_1 + a_2x_2 + \cdots + a_nx_n = N$ subject to $x_1, x_2, \ldots, x_n \in \mathbb{N}$ has *no* solution. From Lemma 2.1 it follows that $g(a,b) = (a-1)(b-1) - 1 = ab - a - b$. Although a formula exists for the case where $n = 3$, no general formulas are known for $n \geq 4$. However, some upper bounds have been proved [4, 11].

LEMMA 2.2. *If $a_1, a_2, \ldots, a_n \in \mathbb{N}_0$ with $a_1 < a_2 < \cdots < a_n$ and $\gcd(a_1, \ldots, a_n) = 1$, then $g(a_1, \ldots, a_n) \leq (a_1 - 1)(a_n - 1) - 1$.*

LEMMA 2.3. *If $a_1, a_2, \ldots, a_n \in \mathbb{N}_0$ with $a_1 < a_2 < \cdots < a_n$ and $\gcd(a_1, \ldots, a_n) = 1$, then we have $g(a_1, \ldots, a_n) \leq 2a_{n-1}\left\lfloor \frac{a_n}{n} \right\rfloor - a_n$.*

2.2. Boolean algebra. A Boolean algebra is an algebraic structure of the form $(\mathbb{B}, \oplus, \otimes)$ with $\mathbb{B} = \{\mathtt{0}, \mathtt{1}\}$ such that the operations $\oplus$ and $\otimes$ applied on $\mathtt{0}$ and $\mathtt{1}$ yield the results of Table 2.1, where $\oplus$ and $\otimes$ are associative and where $\otimes$ is distributive with respect to $\oplus$. The element $\mathtt{0}$ is called the Boolean zero element, $\mathtt{1}$ is called the Boolean identity element, $\oplus$ is called the Boolean addition, and $\otimes$ is called the Boolean multiplication.

Some examples of Boolean algebras are $(\{\text{false}, \text{true}\}, \text{or}, \text{and})$, $(\{0, \infty\}, \min, +)$, $(\{0,1\}, \max, \cdot)$, $(\{\emptyset, \mathbb{N}\}, \cup, \cap)$, $(\{0,1\}, \max, \min)$, $(\{-\infty, \infty\}, \max, \min)$, and so on (see [1, 15]). In this paper we shall use the following examples of Boolean algebra to transform known results from max-plus algebra and from nonnegative matrix algebra to Boolean algebra:

1. The Boolean algebra $(\{-\infty, 0\}, \max, +)$ is a subalgebra of the max-plus algebra $(\mathbb{R} \cup \{-\infty\}, \max, +)$.

2. The Boolean algebra $(\{0, \mathfrak{p}\}, +, \cdot)$, where $\mathfrak{p}$ stands for an arbitrary positive number,[1] can be considered as a Boolean restriction of nonnegative algebra.

A matrix with entries in $\mathbb{B}$ is called a Boolean matrix. The operations $\oplus$ and $\otimes$ are extended to matrices as follows. If $A, B \in \mathbb{B}^{m\times n}$, then we have

$$(A \oplus B)_{ij} \;=\; a_{ij} \oplus b_{ij}$$

for all i, j. If $A \in \mathbb{B}^{m\times p}$ and $B \in \mathbb{B}^{p\times n}$, then

$$(A \otimes B)_{ij} \;=\; \bigoplus_{k=1}^{p} a_{ik} \otimes b_{kj}$$

for all i, j. Note that these definitions resemble the definitions of the sum and the product of matrices in linear algebra but with $\oplus$ instead of $+$ and $\otimes$ instead of $\cdot$.

The $n \times n$ Boolean identity matrix is denoted by $\mathcal{I}_n$, the $m \times n$ Boolean zero matrix is denoted by $\mathcal{O}_{m\times n}$, and the $m \times n$ matrix all the entries of which are equal to $\mathbb{1}$ is denoted by $\mathcal{E}_{m\times n}$. If the dimensions of these matrices are not indicated, they should be clear from the context.

The Boolean matrix power of the matrix $A \in \mathbb{B}^{n\times n}$ is defined as follows:

$$A^{\otimes 0} = \mathcal{I}_n \qquad \text{and} \qquad A^{\otimes k} = A \otimes A^{\otimes k-1} \qquad \text{for } k = 1, 2, \ldots.$$

If we permute the rows or the columns of the Boolean identity matrix, we obtain a Boolean permutation matrix. If $P \in \mathbb{B}^{n\times n}$ is a Boolean permutation matrix, then we have $P \otimes P^T = P^T \otimes P = \mathcal{I}_n$. A matrix $R \in \mathbb{B}^{m\times n}$ is a Boolean upper triangular matrix if $r_{ij} = \mathbb{0}$ for all i, j with $i > j$.

2.3. Boolean algebra and graph theory. We assume that the reader is familiar with basic concepts of graph theory such as directed graph, path, (elementary) circuit, and so on (see, e.g., [1, 18, 27]). In this paper we shall use the definitions of [1] since they are well suited for our proofs. Sometimes these definitions differ slightly from the definitions adopted by other schools in the literature. The most important differences are as follows:

- In this paper we also consider *empty* paths, i.e., paths that consist of only one vertex and have length 0. However, unless it is explicitly specified, we always assume that paths have a nonzero length.
- The precedence graph of the matrix $A \in \mathbb{B}^{n\times n}$, denoted by $\mathcal{G}(A)$, is a directed graph with vertices 1, 2, ..., n and an arc $j \to i$ for each $a_{ij} \neq \mathbb{0}$. Note that vertex i is the *end point* of this arc.
- A directed graph is called strongly connected if for any two *different* vertices v_i, v_j there exists a path from v_i to v_j. Note that this implies that a graph consisting of one vertex (with or without a loop) is always strongly connected.
- A matrix is irreducible if its precedence graph is strongly connected. Since, according to the definition we use, a graph with only one vertex is always strongly connected, the 1×1 Boolean zero matrix $[\mathbb{0}]$ is irreducible. However, the 1×1 Boolean zero matrix $[\mathbb{0}]$ is the only Boolean zero matrix that is irreducible.

Let us now give a graph-theoretic interpretation of the Boolean matrix power. Let $A \in \mathbb{B}^{n\times n}$ and let $k \in \mathbb{N}_0$. Recall that there is an arc $j \to i$ in $\mathcal{G}(A)$ if and only if

[1] So $\mathfrak{p} + \mathfrak{p} = \mathfrak{p}$ and $\mathfrak{p} \cdot \mathfrak{p} = \mathfrak{p}$.

$a_{ij} = \mathbb{1}$. Since

$$\left(A^{\otimes^k}\right)_{ij} = \bigoplus_{i_1,i_2,\ldots,i_{k-1}} a_{ii_1} \otimes a_{i_1 i_2} \otimes \cdots \otimes a_{i_{k-1}j}$$

for all i, j and since $\mathbb{0}$ is absorbing for $\otimes$, $(A^{\otimes^k})_{ij}$ is equal to $\mathbb{1}$ if and only if there exists a path of length k from vertex j to vertex i in $\mathcal{G}(A)$.

A maximal strongly connected subgraph (m.s.c.s.) $\mathcal{G}_{\text{sub}}$, of a directed graph $\mathcal{G}$ is a strongly connected subgraph that is maximal, i.e., if we add an extra vertex (and some extra arcs) of $\mathcal{G}$ to $\mathcal{G}_{\text{sub}}$, then $\mathcal{G}_{\text{sub}}$ is no longer strongly connected.

A well-known result from matrix algebra states that any square matrix can be transformed into a block upper diagonal matrix with irreducible blocks by simultaneously reordering the rows and columns of the matrix (see, e.g., [1, 2, 5, 12, 17, 22] for the proof of this theorem and for its interpretation in terms of graph theory and Markov chains).

THEOREM 2.4. *If $A \in \mathbb{B}^{n\times n}$, then there exists a permutation matrix $P \in \mathbb{B}^{n\times n}$ such that the matrix $\hat{A} = P \otimes A \otimes P^T$ is a block upper triangular matrix of the form*

$$\hat{A} = \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} & \ldots & \hat{A}_{1l} \\ \mathcal{O} & \hat{A}_{22} & \ldots & \hat{A}_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{O} & \mathcal{O} & \ldots & \hat{A}_{ll} \end{bmatrix} \tag{1}$$

with $l \geq 1$ and where the matrices $\hat{A}_{11}$, $\hat{A}_{22}$, ..., $\hat{A}_{ll}$ are square and irreducible. The matrices $\hat{A}_{11}$, $\hat{A}_{22}$, ..., $\hat{A}_{ll}$ are uniquely determined to within simultaneous permutation of their rows and columns, but their ordering in (1) *is not necessarily unique.*

The form in (1) is called the Frobenius normal form of the matrix A. If A is irreducible, then there is only one block in (1) and then A is a Frobenius normal form of itself. Each diagonal block of $\hat{A}$ corresponds to an m.s.c.s. of the precedence graph of $\hat{A}$.

THEOREM 2.5. *If $A \in \mathbb{B}^{n\times n}$ is irreducible, then*

$$\exists k_0 \in \mathbb{N}, \exists c \in \mathbb{N}_0, \text{ such that for all } k \geq k_0 : A^{\otimes^{k+c}} = \lambda^{\otimes^c} \otimes A^{\otimes^k}, \tag{2}$$

where λ is equal to $\mathbb{1}$ if there exists a circuit in $\mathcal{G}(A)$, and equal to $\mathbb{0}$ otherwise.

Proof. See, e.g., [1, 7, 13]. □

The smallest c for which (2) holds is called the cyclicity [1], index of cyclicity [2], or index of *im*primitivity[2] [5, 12] of the matrix A. The cyclicity $c(A)$ of a matrix A is equal to the cyclicity of the precedence graph $\mathcal{G}(A)$ of A and can be computed as follows. The cyclicity of a strongly connected graph or of an m.s.c.s. is the greatest common divisor of the lengths of all the circuits of the given graph or m.s.c.s. If an m.s.c.s. or a graph contains no circuits then its cyclicity is equal to 0 by definition. The cyclicity of general graph is the least common multiple of the nonzero cyclicities of the m.s.c.s.'s of the given graph.

LEMMA 2.6. *If $A \in \mathbb{B}^{n\times n}$ is irreducible, then $c(A) \leq n$.*

Proof. Let $c = c(A)$. Since A is irreducible, $\mathcal{G}(A)$ contains only one m.s.c.s.

[2]We prefer to use the terms "cyclicity" and "index of cyclicity" in this paper to avoid confusion with the concept "index of primitivity" [2, 25] of a nonnegative matrix A, which is defined to be the least positive integer $\gamma(A)$ such that all the entries of $A^{\gamma(A)}$ are positive.

If $A = [\,\mathtt{0}\,]$, then we have $c = 0 \leq 1 = n$.

From now on we assume that $A \neq \mathcal{O}$. Since c is the greatest common divisor of the lengths of the (elementary) circuits in $\mathcal{G}(A)$, c is maximal if there is only one circuit and if this circuit has length n. In that case we have $c = n$. In the other cases, c will be less than n. Thus $c(A) \leq n$. □

LEMMA 2.7. *Let $A \in \mathbb{B}^{n\times n}$ be irreducible and let c be the cyclicity of A. Consider $i, j \in \{1, 2, \ldots, n\}$. If $c > 0$ and if there exists a (nonempty) path of length l_1 from j to i and a (nonempty) path of length l_2 from j to i, then there exists a (possibly negative) integer z such that $l_2 = l_1 + zc$.*

Proof. This lemma is a reformulation of Lemma 3.4.1 of [5] that states that if $\mathcal{G}$ is a strongly connected directed graph with cyclicity c, then for each pair of vertices j and i of $\mathcal{G}$, the lengths of the paths from j to i are congruent modulo c. □

Remark 2.8. Consider $A \in \mathbb{B}^{n\times n}$ and $i, j \in \{1, 2, \ldots, n\}$. Let l_{ij} be the length of the shortest path from vertex j to vertex i in $\mathcal{G}(A)$. Note that Lemma 2.7 does *not* imply that there exists a path of length $l_{ij} + kc$ from j to i for *every* $k \in \mathbb{N}$.

In the next section we discuss upper bounds for the integer k_0 that appears in Theorem 2.5. We also extend this theorem to Boolean matrices that are not irreducible.

3. Consecutive powers of a Boolean matrix. In this section we consider the sequence $\{A^{\otimes^k}\}_{k=1}^{\infty}$ where A is a Boolean matrix. First we consider matrices with a cyclicity that is equal to 0. Next we consider matrices with a cyclicity that is larger than or equal to 1. Here we shall make a distinction among four different cases depending on whether the given matrix is irreducible and whether its cyclicity is equal to 1 or larger than or equal to 1. Of course the last case that will be considered is the most general one, but for the other cases we can provide tighter upper bounds on the length of the transient part of the sequence $\{A^{\otimes^k}\}_{k=1}^{\infty}$ and that is why we consider four different cases.

If possible we also give examples of matrices for which the sequence of the consecutive matrix powers exhibits the longest possible transient behavior.

3.1. Boolean matrices with a cyclicity that is equal to 0.

LEMMA 3.1. *Let $A \in \mathbb{B}^{n\times n}$. If $c(A) = 0$, then we have $A^{\otimes^k} = \mathcal{O}_{n\times n}$ for all $k \geq n$.*

Proof. If the cyclicity of A is equal to 0, then there are no circuits in $\mathcal{G}(A)$, which means that there do not exist paths in $\mathcal{G}(A)$ with a length that is larger than or equal to n since in such paths at least one vertex would appear twice, which implies that such paths contain a circuit. Therefore, we have $A^{\otimes^k} = \mathcal{O}$ for all $k \geq n$. □

Example 3.2. If there exists a permutation matrix P such that $A \in \mathbb{B}^{n\times n}$ can be written as

$$\hat{A} = P \otimes A \otimes P^T = \begin{bmatrix} \mathtt{0} & \mathtt{1} & \mathtt{0} & \ldots & \mathtt{0} \\ \mathtt{0} & \mathtt{0} & \mathtt{1} & \ldots & \mathtt{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathtt{0} & \mathtt{0} & \mathtt{0} & \ldots & \mathtt{1} \\ \mathtt{0} & \mathtt{0} & \mathtt{0} & \ldots & \mathtt{0} \end{bmatrix},$$

then the upper bound of Lemma 3.1 is tight, that is, we have $A^{\otimes^k} \neq \mathcal{O}$ for $k = 1, 2, \ldots, n-1$ and $A^{\otimes^k} = \mathcal{O}$ for all $k \geq n$. The graph of the matrix $\hat{A}$ is represented

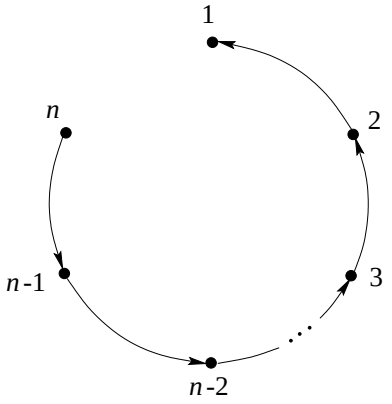


FIG. 3.1. *The precedence graph of the matrix $\hat{A}$ of Example* 3.2.

in Figure 3.1. Note that $c(A) = c(\hat{A}) = 0$ since $\mathcal{G}(\hat{A})$ contains no circuits and since the transformation from A to $\hat{A}$ corresponds to a simultaneous reordering of the rows and the columns of A (or of the vertices of $\mathcal{G}(A)$).

From now on we consider only matrices with a cyclicity that is larger than or equal to 1.

3.2. Boolean matrices with cyclicity 1.

THEOREM 3.3. *Let $A \in \mathbb{B}^{n\times n}$. If the cyclicity of A is equal to* 1 *and if A is irreducible, then we have $A^{\otimes^{k+1}} = A^{\otimes^k} = \mathcal{E}_{n\times n}$ for all $k \geq (n-1)^2 + 1$.*

Proof. This theorem can be considered as the Boolean equivalent of Theorem 4.14 of [2] or of Theorem 3.5.6 of [5]. Note that A cannot be equal to $[\mathtt{0}]$ because $c(A) = 1$. □

If more information about the structure of A is known (such as the number of diagonal entries that are equal to $\mathtt{1}$, the length of the shortest elementary circuit of $\mathcal{G}(A)$, or whether A is symmetrically nonnegative), other upper bounds for the length of the transient part of the sequence $\left\{A^{\otimes^k}\right\}_{k=1}^{\infty}$ where A is a Boolean matrix with cyclicity 1 can be found in section 2.4 of [2].

Example 3.4. If there exists a permutation matrix P such $A \in \mathbb{B}^{n\times n}$ can be written as

$$\hat{A} = P \otimes A \otimes P^T = \begin{bmatrix} \mathtt{0} & \mathtt{1} & \mathtt{0} & \dots & \mathtt{0} \\ \mathtt{0} & \mathtt{0} & \mathtt{1} & \dots & \mathtt{0} \\ \vdots & \vdots & \mathtt{0} & \ddots & \vdots \\ \mathtt{1} & \mathtt{0} & \mathtt{0} & \dots & \mathtt{1} \\ \mathtt{1} & \mathtt{0} & \mathtt{0} & \dots & \mathtt{0} \end{bmatrix}, \tag{3}$$

then the bound in Theorem 3.3 is tight: we have $A^{\otimes^k} = \mathcal{E}$ for all $k \geq (n-1)^2 + 1$ but $A^{\otimes^{(n-1)^2}} \neq \mathtt{1}$. Let us now show that the latter part of this statement indeed holds.

Since the transformation from A to $\hat{A} = P \otimes A \otimes P^T$ corresponds to a simultaneous reordering of the rows and columns of A, we may assume without loss of generality that P is the identity matrix. Thus $A = \hat{A}$. If $n = 2$, then we have $(n-1)^2 + 1 = 2$. Since

$$A = \begin{bmatrix} \mathtt{1} & \mathtt{1} \\ \mathtt{1} & \mathtt{0} \end{bmatrix} \quad \text{and} \quad A^{\otimes^2} = A^{\otimes^3} = \cdots = \mathcal{E},$$

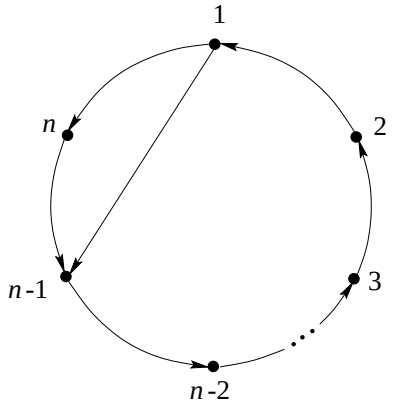


FIG. 3.2. *The precedence graph of the matrix $\hat{A}$ of Example* 3.4.

we indeed have $A^{\otimes^{(n-1)^2}} \neq \mathcal{E}$ if $n = 2$.

From now on we assume that $n > 2$. In Figure 3.2 we have drawn $\mathcal{G}(A)$. There are two elementary circuits in $\mathcal{G}(A)$: circuit $C_1 : n \to n-1 \to \cdots \to 1 \to n$ of length n and circuit $C_2 : n-1 \to n-2 \to 1 \to n-1$ of length $n-1$. Note that only the longest circuit passes through vertex n. Furthermore, if $n > 2$, then $\gcd(n-1, n) = 1$. Any circuit that passes through vertex n can be considered as a concatenation of α times C_1, a path from vertex n to a vertex t in C_2, β times C_2, and a path from t to n for some nonnegative integers α and β. The length of this circuit is equal to $n + \alpha n + \beta(n-1)$. By Lemma 2.1 the smallest integer N such that for any integer $p \geq N$ there exist nonnegative integers α and β such that $p = \alpha n + \beta(n-1)$ is given by $N = (n-1)(n-2)$. This implies that $(n-1)(n-2) - 1$ cannot be written as $\gamma n + \delta(n-1)$ with $\gamma, \delta \in \mathbb{N}$. This implies that there does not exist a circuit of length $n + (n-1)(n-2) - 1 = (n-1)^2$ that passes through vertex n. Hence,

$$\left(A^{\otimes^{(n-1)^2}}\right)_{nn} = \mathtt{o}$$

and thus $A^{\otimes^{(n-1)^2}} \neq \mathcal{E}$.

Let $A \in \mathbb{B}^{n\times n}$. If $\hat{A} = P \otimes A \otimes P^T$ is the Frobenius normal form of A, then we have $A = P^T \otimes \hat{A} \otimes P$. Hence,

$$A^{\otimes^k} = \left(P^T \otimes \hat{A} \otimes P\right)^{\otimes^k} = P^T \otimes \hat{A}^{\otimes^k} \otimes P$$

for all $k \in \mathbb{N}$. Therefore, we may consider without loss of generality the sequence $\left\{\hat{A}^{\otimes^k}\right\}_{k=1}^{\infty}$ instead of the sequence $\left\{A^{\otimes^k}\right\}_{k=1}^{\infty}$. Furthermore, since the transformation from A to $\hat{A}$ corresponds to a simultaneous reordering of the rows and columns of A (or to a reordering of the vertices of $\mathcal{G}(A)$), we have $c(A) = c(\hat{A})$.

THEOREM 3.5. *Let $\hat{A} \in \mathbb{B}^{n\times n}$ be a matrix of the form* (1)*, where the matrices $\hat{A}_{11}$, $\hat{A}_{22}$, …, $\hat{A}_{ll}$ are irreducible and such that $c(\hat{A}) = 1$. Define sets α_1, α_2, …, α_l such that $\hat{A}_{\alpha_i\alpha_j} = \hat{A}_{ij}$ for all i, j with $i \leq j$. Let $n_i = \#\alpha_i$ for all i. Define*

$$\lambda_i = \begin{cases} \mathtt{o} & \text{if } \hat{A}_{ii} = [\mathtt{o}], \\ \mathtt{1} & \text{otherwise} \end{cases}$$

for $i = 1, 2, \ldots, l$. *Define*

$$S_{ij} = \Big\{ \{i_0, i_1, \ldots, i_s\} \subseteq \{1, 2, \ldots, l\} \;\Big|\; i = i_0 < i_1 < \cdots < i_s = j \text{ and } \hat{A}_{i_r i_{r+1}} \neq \mathcal{O} \text{ for } r = 0, 1, \ldots, s-1 \Big\}$$

for all i, j *with* $i < j$.

Let $\lambda_{ii} = \lambda_i$ *and* $k_{ii} = (n_i - 1)^2 + 1$ *for* $i = 1, 2, \ldots, n$. *Define*

$$\Gamma_{ij} = \big\{ t \,\big|\, \exists \gamma \in S_{ij} \text{ such that } t \in \gamma \big\},$$

$$\lambda_{ij} = \begin{cases} \bigoplus\limits_{t \in \Gamma_{ij}} \lambda_t & \text{if } \Gamma_{ij} \neq \emptyset, \\ \mathfrak{o} & \text{otherwise}, \end{cases}$$

$$t_{ij} = \begin{cases} \arg\min \{n_t \,|\, t \in \Gamma_{ij} \text{ and } \lambda_{tt} \neq \mathfrak{o}\} & \text{if } \lambda_{ij} \neq \mathfrak{o}, \\ 0 & \text{otherwise}, \end{cases}$$

$$k_{ij} = \begin{cases} \sum\limits_{\substack{t \in \Gamma_{ij} \\ t \neq t_{ij}}} n_t + k_{t_{ij} t_{ij}} & \text{if } \lambda_{ij} \neq \mathfrak{o}, \\ \#\Gamma_{ij} & \text{if } \Gamma_{ij} \neq \emptyset \text{ and } \lambda_{ij} = \mathfrak{o}, \\ 1 & \text{if } \Gamma_{ij} = \emptyset \end{cases}$$

for all i, j *with* $i < j$. *Then we have for all* i, j *with* $i \leq j$*:*

$$\left(\hat{A}^{\otimes^k}\right)_{\alpha_i \alpha_j} = \begin{cases} \mathcal{E}_{n_i \times n_j} & \text{if } \lambda_{ij} \neq \mathfrak{o}, \\ \mathcal{O}_{n_i \times n_j} & \text{if } \lambda_{ij} = \mathfrak{o} \end{cases} \qquad \text{for all } k \geq k_{ij}. \tag{4}$$

For all i, j *with* $i > j$ *we have*

$$\left(\hat{A}^{\otimes^k}\right)_{\alpha_i \alpha_j} = \mathcal{O}_{n_i \times n_j} \qquad \text{for all } k \in \mathbb{N}. \tag{5}$$

Remark 3.6. Note that $\hat{A}_{ij}$ is an $n_i \times n_j$ matrix for all i, j.

Let us now give a graphical interpretation of the sets S_{ij} and Γ_{ij}.

Let C_i be the m.s.c.s. of $\mathcal{G}(\hat{A})$ that corresponds to $\hat{A}_{ii}$ for $i = 1, 2 \ldots, l$. Thus α_i is the set of vertices of C_i.

If $\{i_0 = i, i_1, \ldots, i_{s-1}, i_s = j\} \in S_{ij}$, then there exists a path from a vertex in C_{i_r} to a vertex in $C_{i_{r-1}}$ for each $r = 1, 2, \ldots, s$. Since each m.s.c.s. C_i of $\mathcal{G}(\hat{A})$ either is strongly connected or consists of only one vertex, this implies that there exists a path from a vertex in C_j to a vertex in C_i that passes through $C_{i_{s-1}}$, $C_{i_{s-2}}$, ..., C_{i_1}.

If $S_{ij} = \emptyset$, then there does not exist any path from a vertex in C_j to a vertex in C_i.

The set Γ_{ij} is the set of indices of the m.s.c.s.'s of $\mathcal{G}(\hat{A})$ through which some path from a vertex of C_j to a vertex of C_i passes.

If $S_{ij} \neq \emptyset$, then $C_{t_{ij}}$ is the smallest m.s.c.s. of $\mathcal{G}(A)$ that contains a circuit and through which some path from a vertex of C_j to a vertex of C_i passes.[3]

[3]Or more precisely: if $S_{ij} \neq \emptyset$, then $C_{t_{ij}}$ belongs to the set of the smallest m.s.c.s.'s of $\mathcal{G}(A)$ that contain a circuit and through which some path from a vertex of C_j to a vertex of C_i passes.

Proof of Theorem 3.5. Let C_i be the m.s.c.s. of $\mathcal{G}(\hat{A})$ that corresponds to $\hat{A}_{ii}$ for $i = 1, 2 \ldots, l$. Since $\hat{A}_{\alpha_i \alpha_j} = \mathcal{O}$ if $i > j$, there are no arcs from any vertex of C_j to a vertex in C_i. As a consequence, (5) holds if $i > j$.

Note that $c(\hat{A}_{ii}) = 1$ for all $i \in \{1, 2, \ldots, l\}$ with $\hat{A}_{ii} \neq [\mathfrak{o}]$ since $c(\hat{A}) = 1$ and since each $\hat{A}_{ii}$ corresponds to an m.s.c.s. of $\mathcal{G}(\hat{A})$.

If $l = 1$, then $\hat{A}$ is irreducible and then (4) holds by Theorem 3.3. It is easy to verify that (4) holds if $i = j$.

From now on we assume that $l > 1$ and $i < j$.

If $\Gamma_{ij} = \emptyset$, then there does not exist a path from a vertex in C_j to a vertex in C_i. Hence, $\left(\hat{A}^{\otimes^k}\right)_{\alpha_i \alpha_j} = \mathcal{O}$ for all $k \in \mathbb{N}$.

If $\Gamma_{ij} \neq \emptyset$ and $\lambda_{ij} = \mathfrak{o}$, then we have $\hat{A}_{tt} = [\mathfrak{o}]$ for all $t \in \Gamma_{ij}$. Thus there exist paths from a vertex in C_j to a vertex in C_i, but each path passes only through m.s.c.s.'s that consist of one vertex and contain no loop. Such a path passes through at most $\#\Gamma_{ij}$ of such m.s.c.s.'s (C_j and C_i included). This implies that there does not exist a path with a length that is larger than or equal to $\#\Gamma_{ij}$ from a vertex in C_j to a vertex in C_i. Hence, we have $\left(A^{\otimes^k}\right)_{\alpha_i \alpha_j} = \mathcal{O}$ for all $k \geq \#\Gamma_{ij}$.

From now on we assume that $l > 1$, $i < j$, $\Gamma_{ij} \neq \emptyset$, and $\lambda_{ij} \neq \mathfrak{o}$. Then there exists a set $\{i_0, i_1, \ldots, i_s\} \subseteq \{1, 2, \ldots, l\}$ with $t_{ij} \in \{i_0, i_1, \ldots, i_s\}$ and there exist indices $u_r \in \alpha_{i_r}$, $v_{r+1} \in \alpha_{i_{r+1}}$ for $r = 0, 1, \ldots, s-1$ such that $i = i_0 < i_1 < \cdots < i_s = j$ and $\hat{A}_{u_r v_{r+1}} \neq \mathfrak{o}$ for each r. Thus there exists an arc from vertex v_{r+1} of $C_{i_{r+1}}$ to vertex u_r of C_{i_r} for each $r \in \{0, 1, \ldots, s-1\}$. Select an arbitrary vertex u_s of $C_{i_s} = C_j$ and an arbitrary vertex v_0 of $C_{i_0} = C_i$. Note that $s \in \{1, 2, \ldots, j-i\}$. Let $r \in \{0, 1, \ldots, s\}$. Recall that the only Boolean zero matrix that is irreducible is the 1×1 Boolean zero matrix $[\mathfrak{o}]$. Now we distinguish between two cases:

- If $\hat{A}_{i_r i_r} = [\mathfrak{o}]$, then we have $n_{i_r} = 1$ and $u_r = v_r$. Thus in this case we could say that there exists an empty path of length $l_r = 0$ from vertex u_r to vertex v_r of C_{i_r}.
- On the other hand, if $\hat{A}_{i_r i_r} \neq [\mathfrak{o}]$, then there exists a (possibly empty) path of length $l_r \leq n_{i_r} - 1$ from vertex u_r to vertex v_r of C_{i_r} since $\mathcal{G}(\hat{A}_{i_r i_r})$ is strongly connected. If $u_r = v_r$, then this path is empty and has length 0.

Therefore for each $r \in \{0, 1, \ldots, s\}$ there exists a (possible empty) path of length $l_r \leq n_{i_r} - 1$ from vertex u_r to vertex v_r of C_{i_r}.

Let $\tilde{t} = i_{\tilde{r}} = t_{ij}$. Clearly, we have $\hat{A}_{i_{\tilde{r}} i_{\tilde{r}}} \neq [\mathfrak{o}]$. Since $\hat{A}_{i_{\tilde{r}} i_{\tilde{r}}}$ is irreducible and since $c(\hat{A}_{i_{\tilde{r}} i_{\tilde{r}}}) = 1$, it follows from Theorem 3.3 that there exists a path of length k from vertex $u_{\tilde{r}}$ to vertex $v_{\tilde{r}}$ of $C_{i_{\tilde{r}}}$ for any $k \geq k_{i_{\tilde{r}} i_{\tilde{r}}} = k_{\tilde{t}\tilde{t}}$. Note that $\#\Gamma_{ij} \geq s + 1$. Hence,

$$k_{ij} = \sum_{\substack{t \in \Gamma_{ij} \\ t \neq \tilde{t}}} n_t + k_{\tilde{t}\tilde{t}} = \sum_{\substack{t \in \Gamma_{ij} \\ t \neq \tilde{t}}} (n_t - 1) + (\#\Gamma_{ij} - 1) + k_{\tilde{t}\tilde{t}} \geq \sum_{\substack{r=0 \\ r \neq \tilde{r}}}^{s} l_r + s + k_{\tilde{t}\tilde{t}} \ .$$

Thus if we have an integer $k \geq k_{ij}$, we can decompose it as

$$k = l_0 + l_1 + \cdots + l_{\tilde{r}-1} + l_{\tilde{r}+1} + \cdots + l_s + s + k_{\tilde{t}\tilde{t}} + \tilde{k}$$

with $\tilde{k} \in \mathbb{N}$. By Theorem 3.3 there exists a path of length $k_{\tilde{t}\tilde{t}} + \tilde{k}$ from $u_{\tilde{r}}$ to $v_{\tilde{r}}$ in $C_{i_{\tilde{r}}}$ for each $\tilde{k} \in \mathbb{N}$. This implies that there exists a path from vertex u_s to vertex v_0 of length k in $\mathcal{G}(\hat{A})$. This path consists of the concatenation of paths of length l_r from vertex u_r to vertex v_r of C_{i_r} for $r = 0, 1, \ldots, \tilde{r}-1, \tilde{r}+1, \ldots, s$; paths of length 1 from vertex v_{r+1} of $C_{i_{r+1}}$ to vertex u_r of C_{i_r} for $r = 0, 1, \ldots, s-1$; and a path of

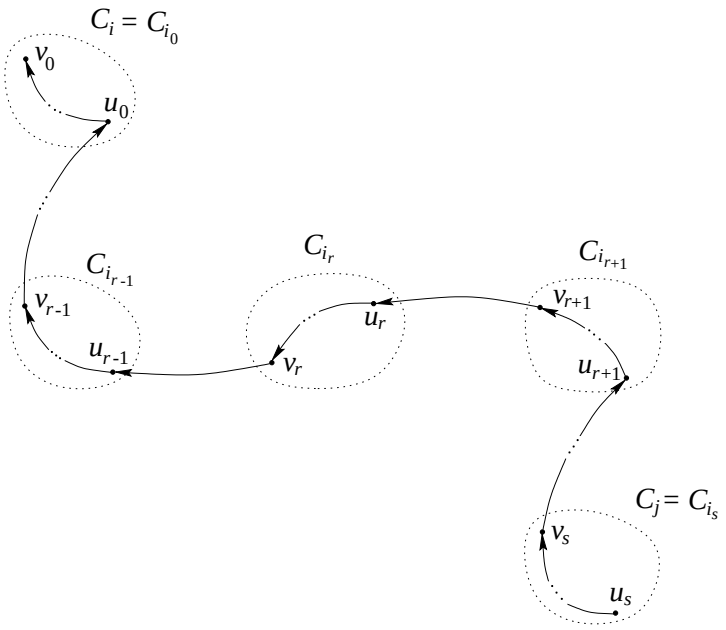


FIG. 3.3. *Illustration of the proof of Theorem* 3.5. *There exists a path from vertex u_s of m.s.c.s. C_j to vertex v_0 of m.s.c.s. C_i that passes through the m.s.c.s.'s $C_{i_{s-1}}$, $C_{i_{s-2}}$, ..., C_{i_1}.*

length $k_{\tilde{t}\tilde{t}} + \tilde{k}$ from vertex $u_{\tilde{r}}$ to vertex $v_{\tilde{r}}$ of $C_{i_{\tilde{r}}}$ (see Figure 3.3). This implies that $\left(\hat{A}^{\otimes^k}\right)_{v_0 u_s} = \mathbb{1}$ for all $k \geq k_{ij}$. Since u_s is an arbitrary vertex of C_j and since v_0 is an arbitrary vertex of C_i, this implies that $\left(\hat{A}^{\otimes^k}\right)_{\alpha_i \alpha_j} = \mathcal{E}$ for all $k \geq k_{ij}$. Thus (4) also holds if $\lambda_{ij} \neq \mathbb{0}$. □

Example 3.7. Consider the following matrix:

$$\hat{A} = \left[\begin{array}{ccc|c} \mathbb{0} & \mathbb{1} & \mathbb{0} & \mathbb{0} \\ \mathbb{1} & \mathbb{0} & \mathbb{1} & \mathbb{0} \\ \mathbb{1} & \mathbb{0} & \mathbb{0} & \mathbb{1} \\ \hline \mathbb{0} & \mathbb{0} & \mathbb{0} & \mathbb{0} \end{array}\right].$$

This matrix is in Frobenius normal form and its block structure is indicated by the vertical and horizontal lines. The precedence graph of $\hat{A}$ is represented in Figure 3.4. Using the notations and definitions of Theorem 3.5, we have $l = 2$,

$$\hat{A}_{11} = \left[\begin{array}{ccc} \mathbb{0} & \mathbb{1} & \mathbb{0} \\ \mathbb{1} & \mathbb{0} & \mathbb{1} \\ \mathbb{1} & \mathbb{0} & \mathbb{0} \end{array}\right], \quad \hat{A}_{12} = \left[\begin{array}{c} \mathbb{0} \\ \mathbb{0} \\ \mathbb{1} \end{array}\right], \text{ and } \quad \hat{A}_{22} = [\mathbb{0}].$$

Furthermore, $\alpha_1 = \{1, 2, 3\}$, $\alpha_2 = \{4\}$, $n_1 = 3$, $n_2 = 1$, $\lambda_1 = \lambda_{11} = \mathbb{1}$, $\lambda_2 = \lambda_{22} = \mathbb{0}$, $S_{12} = \big\{ \{1, 2\} \big\}$, $\Gamma_{12} = \{1, 2\}$, $\lambda_{12} = \mathbb{1}$, $t_{12} = 1$, $k_{11} = 5$, $k_{22} = 1$, and $k_{12} = 1 + 5 = 6$. We have

$$\hat{A}^{\otimes^2} = \left[\begin{array}{ccc|c} \mathbb{1} & \mathbb{0} & \mathbb{1} & \mathbb{0} \\ \mathbb{1} & \mathbb{1} & \mathbb{0} & \mathbb{1} \\ \mathbb{0} & \mathbb{1} & \mathbb{0} & \mathbb{0} \\ \hline \mathbb{0} & \mathbb{0} & \mathbb{0} & \mathbb{0} \end{array}\right], \quad \hat{A}^{\otimes^3} = \left[\begin{array}{ccc|c} \mathbb{1} & \mathbb{1} & \mathbb{0} & \mathbb{1} \\ \mathbb{1} & \mathbb{1} & \mathbb{1} & \mathbb{0} \\ \mathbb{1} & \mathbb{0} & \mathbb{1} & \mathbb{0} \\ \hline \mathbb{0} & \mathbb{0} & \mathbb{0} & \mathbb{0} \end{array}\right], \quad \hat{A}^{\otimes^4} = \left[\begin{array}{ccc|c} \mathbb{1} & \mathbb{1} & \mathbb{1} & \mathbb{0} \\ \mathbb{1} & \mathbb{1} & \mathbb{1} & \mathbb{1} \\ \mathbb{1} & \mathbb{1} & \mathbb{0} & \mathbb{1} \\ \hline \mathbb{0} & \mathbb{0} & \mathbb{0} & \mathbb{0} \end{array}\right],$$

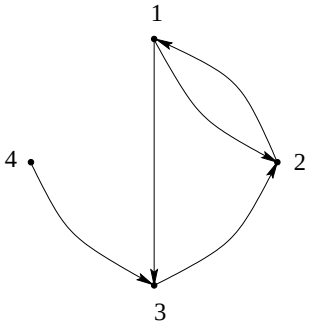


FIG. 3.4. *The precedence graph of the matrix $\hat{A}$ of Example* 3.7.

$$\hat{A}^{\otimes 5} = \left[\begin{array}{ccc|c} \mathbb{1} & \mathbb{1} & \mathbb{1} & \mathbb{1} \\ \mathbb{1} & \mathbb{1} & \mathbb{1} & \mathbb{1} \\ \mathbb{1} & \mathbb{1} & \mathbb{1} & \mathbb{0} \\ \hline \mathbb{0} & \mathbb{0} & \mathbb{0} & \mathbb{0} \end{array}\right], \quad \hat{A}^{\otimes 6} = \left[\begin{array}{ccc|c} \mathbb{1} & \mathbb{1} & \mathbb{1} & \mathbb{1} \\ \mathbb{1} & \mathbb{1} & \mathbb{1} & \mathbb{1} \\ \mathbb{1} & \mathbb{1} & \mathbb{1} & \mathbb{1} \\ \hline \mathbb{0} & \mathbb{0} & \mathbb{0} & \mathbb{0} \end{array}\right] = A^{\otimes 7} = \cdots.$$

Note that $\hat{A}_{11}$ is a matrix of the form (3). Therefore the smallest K_{11} for which $\hat{A}_{11}^{\otimes k} = \mathcal{E}$ for all $k \geq K_{11}$ is equal to $k_{11} = 5$. Furthermore, the smallest K_{12} such that $\left(\hat{A}^{\otimes k}\right)_{\alpha_1\alpha_2} = \mathcal{E}$ for all $k \geq K_{12}$ is equal to $k_{12} = 6$. Thus for the matrix $\hat{A}$ of this example all the bounds k_{ij} that appear in Theorem 3.5 are tight.

It is easy to verify that for a matrix of the form

$$\left[\begin{array}{ccccc|c} \mathbb{0} & \mathbb{1} & \mathbb{0} & \dots & \mathbb{0} & \mathbb{0} \\ \mathbb{0} & \mathbb{0} & \mathbb{1} & \dots & \mathbb{0} & \mathbb{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbb{1} & \mathbb{0} & \mathbb{0} & \dots & \mathbb{1} & \mathbb{0} \\ \mathbb{1} & \mathbb{0} & \mathbb{0} & \dots & \mathbb{0} & \mathbb{1} \\ \hline \mathbb{0} & \mathbb{0} & \mathbb{0} & \dots & \mathbb{0} & \mathbb{0} \end{array}\right]$$

all the bounds k_{ij} that appear in Theorem 3.5 are tight.

LEMMA 3.8. *Let $A \in \mathbb{B}^{n\times n}$ with $c(A) = 1$. Then we have $A^{\otimes k+1} = A^{\otimes k}$ for all $k \geq (n-1)^2 + 1$.*

Proof. If A is irreducible, then we have $A^{\otimes k+1} = A^{\otimes k}$ for all $k \geq (n-1)^2 + 1$ by Theorem 3.3.

So from now on we assume that A is not irreducible. Let $\hat{A} = P \otimes A \otimes P^T$ be the Frobenius normal form of A. Assume that $\hat{A}$ is of the form (1) where the $\hat{A}_{ii}$'s are square and irreducible. Let the numbers λ_{ij}, n_i, t_{ij}, k_{ij} and the sets α_i and Γ_{ij} be defined as in Theorem 3.5.

We have $k_{ii} = (n_i - 1)^2 + 1 \leq (n-1)^2 + 1$ for all i.

Let us now prove that $k_{ij} \leq (n-1)^2 + 1$ for all i, j with $i < j$. Consider indices $i, j \in \{1, 2, \dots, l\}$ with $i < j$.

- If $\Gamma_{ij} = \emptyset$, then we have $k_{ij} = 1 \leq (n-1)^2 + 1$ for all $n \in \mathbb{N}_0$.
- We have $n \leq (n-1)^2 + 1$ for all $n \in \mathbb{N}_0$. Thus if $\Gamma_{ij} \neq \emptyset$ and $\lambda_{ij} = \mathbb{0}$, then we have

$$k_{ij} \;=\; \#\Gamma_{ij} \;\leq\; j - i + 1 \;\leq\; n \;\leq\; (n-1)^2 + 1.$$

- Since $n_t \geq 1$ for each $t \in \{1, 2, \ldots, l\}$ and since $l \geq 1$, we have

$$\sum_{\substack{t,s=1 \\ t \neq s}}^{l} n_t n_s \geq \sum_{\substack{t,s=1 \\ t \neq s}}^{l} 1 \geq l^2 - l \geq l - 1.$$

Hence,

$$l \leq \sum_{\substack{t,s=1 \\ t \neq s}}^{l} n_t n_s + 1. \tag{6}$$

Therefore if $\lambda_{ij} \neq \mathfrak{o}$, then we have

$$\begin{aligned}
k_{ij} &= \sum_{\substack{t \in \Gamma_{ij} \\ t \neq t_{ij}}} n_t + k_{t_{ij} t_{ij}} \\
&\leq \sum_{t=1}^{l} k_{tt} \qquad \text{(since } k_{ii} = (n_i - 1)^2 + 1 \geq n_i \text{ for each } i\text{)} \\
&\leq \sum_{t=1}^{l} \left((n_t - 1)^2 + 1 \right) \\
&\leq \sum_{t=1}^{l} (n_t^2 - 2n_t + 2) \\
&\leq \sum_{t=1}^{l} n_t^2 - 2 \sum_{t=1}^{l} n_t + 2l \\
&\leq \sum_{t=1}^{l} n_t^2 - 2n + 2l \\
&\leq \sum_{t=1}^{l} n_t^2 - 2n + 2 \left(\sum_{\substack{t,s=1 \\ t \neq s}}^{l} n_t n_s + 1 \right) \qquad \text{(by (6))} \\
&\leq \left(\sum_{t=1}^{l} n_t^2 + 2 \sum_{\substack{t,s=1 \\ t \neq s}}^{l} n_t n_s \right) - 2n + 2 \\
&\leq \left(\sum_{t=1}^{l} n_t \right)^2 - 2n + 1 + 1 \\
&\leq n^2 - 2n + 1 + 1 \\
&\leq (n-1)^2 + 1.
\end{aligned}$$

Hence, $k_{ij} \leq (n-1)^2 + 1$ for all i, j with $i \leq j$. As a consequence, it follows from Theorem 3.5 that $\left(\hat{A}^{\otimes^{k+1}}\right)_{\alpha_i \alpha_j} = \left(\hat{A}^{\otimes^k}\right)_{\alpha_i \alpha_j}$ for all $k \geq (n-1)^2 + 1$ and for all $i, j \in \{1, 2, \ldots, l\}$. Hence, $\hat{A}^{\otimes^{k+1}} = \hat{A}^{\otimes^k}$ for all $k \geq (n-1)^2 + 1$. Since $A^{\otimes^k} = P^T \otimes \hat{A}^{\otimes^k} \otimes P$, this implies that $A^{\otimes^{k+1}} = A^{\otimes^k}$ for all $k \geq (n-1)^2 + 1$. □

3.3. Boolean matrices with a cyclicity that is larger than or equal to 1.

LEMMA 3.9. *Let $A \in \mathbb{B}^{n\times n}$ be an irreducible matrix with $c(A) \geq 2$ and let $i, j \in \{1, 2, \ldots, n\}$. Then there exists a (possibly empty) path P_{ij} from j to i in $\mathcal{G}(A)$ that passes through at least one vertex of each (elementary) circuit of $\mathcal{G}(A)$ and that has a length that is less than or equal to $\frac{n^2-1}{2}$.*

Proof. Since the cyclicity of A is larger than or equal to 2, there are no loops in $\mathcal{G}(A)$. Hence, A contains at least one circuit. Since A is irreducible, this implies that j has to belong to an elementary circuit of $\mathcal{G}(A)$. Since the length of any elementary circuit of $\mathcal{G}(A)$ is larger than or equal to 2, there exists a set $S = \{i_1 = j, i_2, \ldots, i_m\} \subseteq \{1, 2, \ldots, n\}$ with $m \leq \lceil \frac{n}{2} \rceil$ such that any (elementary) circuit of $\mathcal{G}(A)$ contains at least one vertex that belongs to S. Define $i_{m+1} = i$. Since $\mathcal{G}(A)$ is strongly connected there exists a (possibly empty) path P_k with length $l_k \leq n-1$ from vertex i_k to vertex i_{k+1} for each $k \in \{1, 2, \ldots, m\}$. Let l_k be the length of P_k. There exists a path P_{ij} from j to i that contains at least one vertex of each (elementary) circuit of $\mathcal{G}(A)$: this path consists of the concatenation of P_1, P_2, ..., P_m. If l_{ij} is the length of P_{ij}, then we have

$$l_{ij} = l_1 + \cdots + l_m \leq m(n-1) \leq \left(\frac{n+1}{2}\right)(n-1) \leq \frac{n^2-1}{2}. \qquad \square$$

Remark 3.10. Note that we could have derived an upper bound that is more tight in Lemma 3.9. The upper bound of Lemma 3.9 will be used in the proof of Theorem 3.11. However, in that proof we shall also use Lemmas 2.2 and 2.3, which also yield upper bounds, and therefore we do not refine the upper bound of Lemma 3.9.

THEOREM 3.11. *Let $A \in \mathbb{B}^{n\times n}$ be irreducible and let $c = c(A) > 0$. If we define*

$$k_{n,c} = \begin{cases} (n-1)^2 + 1 & \text{if } c = 1, \\ \max\left(n-1,\ \dfrac{n^2-1}{2} + \dfrac{n^2}{c} - 3n + 2c\right) & \text{if } c > 1, \end{cases} \tag{7}$$

then we have

$$A^{\otimes^{k+c}} = A^{\otimes^k} \text{ and } A^{\otimes^k} \oplus A^{\otimes^{k+1}} \oplus \cdots \oplus A^{\otimes^{k+c-1}} = \mathcal{E}_{n\times n} \quad \text{for all } k \geq k_{n,c}. \tag{8}$$

Proof. From Theorem 3.3 it follows that (8) holds if c is equal to 1. Furthermore, if the first part of (8) holds, then the second part also holds since A is irreducible.

From now on we assume that $c > 1$. Let $i, j \in \{1, 2, \ldots, n\}$.

Let C_1, C_2, ..., C_m be the elementary circuits of $\mathcal{G}(A)$. Let l_i be the length of C_i for $i = 1, 2, \ldots, m$. Since A is irreducible, we have $c = \gcd(l_1, l_2, \ldots, l_m)$. Hence, there exist positive integers w_1, w_2, ..., w_m such that $w_i\, c = l_i$ for each i and such that $\gcd(w_1, w_2, \ldots, w_m) = 1$.

First we consider the case where there is only one elementary circuit or where all the elementary circuits have the same length. Hence, $c = l_1$. Since A is irreducible, both i and j have to belong to some elementary circuit. We may assume without loss of generality that j belongs to C_1. Since A is irreducible there exist paths from vertex j to vertex i of $\mathcal{G}(A)$. Let P_{ij} be the shortest (possibly empty) path from j to i and let l_{ij} be the length of this path. We have $l_{ij} \leq n-1$. (Note that $l_{ij} = 0$ if i is equal to j.) For any integer $k \in \mathbb{N}$ there exists a path of length $l_{ij} + kc$ from j

to i: this path consists of k times C_1 followed by P_{ij}. Hence $\left(A^{\otimes^{l_{ij}+kc}}\right)_{ij} = \mathbb{1}$ for all $k \geq 0$. Let $l \in \mathbb{N}$ with $l \geq n-1$. Now there are two possibilities. If l can be written as $l = l_{ij} + kc$ for some $k \in \mathbb{N}$, then we have $\left(A^{\otimes^l}\right)_{ij} = \mathbb{1}$. If l cannot be written as $l = l_{ij} + kc$ for any $k \in \mathbb{N}$, then it follows from Lemma 2.7 that there does not exist a path from j to i and then we have $\left(A^{\otimes^l}\right)_{ij} = \mathbb{0}$.

This implies that (8) holds if all the elementary circuits of $\mathcal{G}(A)$ have the same length.

From now on we assume that there exist at least two elementary circuits in $\mathcal{G}(A)$ that have different lengths.

Since A is irreducible it follows from Lemma 3.9 that there exists a (possibly empty) path P_{ij} from vertex j to vertex i of $\mathcal{G}(A)$ with length $l_{ij} \leq \frac{n^2-1}{2}$ that passes through at least one vertex of each elementary circuit of $\mathcal{G}(A)$. For each circuit C_k we select one vertex v_k that belongs to the path P_{ij}. Let l be an integer that can be written as $l = l_{ij} + pc$ with $p \geq g(w_1, w_2, \ldots, w_m) + 1$. Since $\gcd(w_1, w_2, \ldots, w_m) = 1$, there exist nonnegative integers α_1, α_2, …, α_m such that $p = \alpha_1 w_1 + \alpha_2 w_2 + \cdots + \alpha_m w_m$. As a consequence, we have

$$l = l_{ij} + \alpha_1 w_1 c + \alpha_2 w_2 c + \cdots + \alpha_m w_m c = l_{ij} + \alpha_1 l_1 + \alpha_2 l_2 + \cdots + \alpha_m l_m.$$

Thus there exists a path of length l from j to i: this path consists of the concatenation of P_{ij} and α_k times the circuit C_k $(v_k \to \cdots \to v_k)$ for $k = 1, 2, \ldots, m$. Hence, $\left(A^{\otimes^{l_{ij}+pc}}\right)_{ij} = \mathbb{1}$ for all $p \geq g(w_1, w_2, \ldots, w_m) + 1$.

Let us now determine an upper bound for $g(w_1, w_2, \ldots, w_m)$. If $w_s = w_t$ for some $s, t \in \{1, 2, \ldots, m\}$ with $s \neq t$, then $g(w_1, w_2, \ldots, w_m) = g(w_1, w_2, \ldots, w_{s-1}, w_{s+1}, \ldots, w_m)$. Therefore, we may assume without loss of generality that all the w_i's are different and thus also that $w_1 < w_2 < \cdots < w_m$.

Since there are at least two elementary circuits in $\mathcal{G}(A)$ that have different lengths, we have $m \geq 2$. We have $w_k = \frac{l_k}{c} \leq \frac{n}{c}$ for all k. Hence, $w_m \leq \frac{n}{c}$ and $w_1 \leq \frac{n}{c} - (m-1) \leq \frac{n}{c} - 1$ since $w_1 < w_2 < \cdots < w_m$ and $m \geq 2$. As a consequence, we have

$$\begin{aligned} g(w_1, w_2, \ldots, w_m) &\leq (w_1 - 1)(w_m - 1) - 1 \qquad \text{(by Lemma 2.2)} \\ &\leq \left(\frac{n}{c} - 2\right)\left(\frac{n}{c} - 1\right) - 1 \\ &\leq \left(\frac{n}{c}\right)^2 - 3\frac{n}{c} + 1. \end{aligned} \tag{9}$$

If we define

$$K = \frac{n^2 - 1}{2} + \left(\left(\frac{n}{c}\right)^2 - 3\frac{n}{c} + 2\right) c,$$

then we have $l_{ij} + \big(g(w_1, w_2, \ldots, w_m) + 1\big)\, c \leq K$. Thus if we have an integer l that is larger than K, then either it can be written as $l = l_{ij} + pc$ with $p \geq g(w_1, w_2, \ldots, w_m) + 1$ and then $\left(A^{\otimes^l}\right)_{ij} = \mathbb{1}$ or it cannot be written as $l = l_{ij} + pc$ for any $p \in \mathbb{N}$ and then it follows from Lemma 2.7 that there does not exist a path of length l from j to i, i.e., $\left(A^{\otimes^l}\right)_{ij} = \mathbb{0}$. Note that $K \leq k_{n,c}$.

Hence, (8) also holds in this case. □

Remark 3.12. In the proof of Theorem 3.11 we also could have used Lemma 2.3 to determine an upper bound for $g(w_1, w_2, \ldots, w_m)$. We have $m \geq 2$ and thus also $w_m \geq 2$. Furthermore, $w_{m-1} \leq \frac{n}{c} - 1$. Hence,

$$\begin{aligned} g(w_1, w_2, \ldots, w_m) &\leq 2w_{m-1} \left\lfloor \frac{w_m}{m} \right\rfloor - w_m \\ &\leq 2\left(\frac{n}{c} - 1\right)\frac{n}{2c} - 2 \\ &\leq \left(\frac{n}{c}\right)^2 - \frac{n}{c} - 2. \end{aligned} \tag{10}$$

In the second part of the proof of Theorem 3.11 we have $c \geq 2$. Since A is irreducible, it follows from Lemma 2.6 that $c \leq n$. Hence, $1 \leq \frac{n}{c} \leq \frac{n}{2}$. It is easy to verify that the upper bound of (10) is less than the upper bound of (9) if $\frac{n}{c} < \frac{3}{2}$. However, if $\frac{n}{c} < \frac{3}{2}$, then we would have $w_{m-1} \leq \frac{1}{2}$, which is not possible. This implies that for combinations of n and c for which there are at least two elementary circuits in $\mathcal{G}(A)$ with different lengths, the upper bound of (9) is less than or equal to the upper bound of (10).

The Boolean sum of sequences is defined as follows. Consider sequences $g_i = \{(g_i)_k\}_{k=1}^{\infty}$ for $i = 1, 2, \ldots, m$ with $(g_i)_k \in \mathbb{B}$ for all i, k. The sequence $g = g_1 \oplus g_2 \oplus \cdots \oplus g_m$ is defined by $g_k = (g_1)_k \oplus (g_2)_k \oplus \cdots \oplus (g_m)_k$ for all $k \in \mathbb{N}_0$.

LEMMA 3.13. *Consider sequences $g_i = \{(g_i)_k\}_{k=1}^{\infty}$ for $i = 1, 2, \ldots, m$ with $(g_i)_k \in \mathbb{B}$ for all i, k. Suppose that for each $i \in \{1, 2, \ldots, m\}$ there exist integers $K_i, c_i \in \mathbb{N}_0$ such that*

$$(g_i)_{k+c_i} = (g_i)_k \qquad \text{for all } k \geq K_i. \tag{11}$$

If $K = \max_i K_i$ and

$$c = \begin{cases} 1 & \text{if } c_i = 1 \text{ and } (g_i)_{K_i} = \mathbb{1} \text{ for some } i \in \{1, 2, \ldots, m\}, \\ \operatorname{lcm}(c_1, c_2, \ldots, c_m) & \text{otherwise,} \end{cases}$$

then the sequence $g = g_1 \oplus g_2 \oplus \cdots \oplus g_m$ satisfies $g_{k+c} = g_k$ for all $k \geq K$.

Proof. Note that (11) implies that

$$(g_i)_{k+pc_i} = (g_i)_k \qquad \text{for all } k \geq K \geq K_i \text{ and for all } p \in \mathbb{N}. \tag{12}$$

First we assume that there exists an index $i \in \{1, 2, \ldots, m\}$ such that $c_i = 1$ and $(g_i)_{K_i} = \mathbb{1}$. Then we have $(g_i)_k = \mathbb{1}$ for all $k \geq K_i$ and thus also $g_k = \mathbb{1}$ for all $k \geq K \geq K_i$.

From now on we assume that there does not exist any index $i \in \{1, 2, \ldots, m\}$ such that $c_i = 1$ and $(g_i)_{K_i} = \mathbb{1}$.

Since $c = \operatorname{lcm}(c_1, c_2, \ldots, c_m)$ there exist positive integers $w_1, w_2, \ldots, w_m$ such that $c = w_i c_i$ for $i = 1, 2, \ldots, m$. Consider an integer $k \geq K$. We have

$$\begin{aligned} g_{k+c} &= \bigoplus_{i=1}^{m} (g_i)_{k+c} \\ &= \bigoplus_{i=1}^{m} (g_i)_{k+w_i c_i} \\ &= \bigoplus_{i=1}^{m} (g_i)_k \qquad \text{(by (12))} \\ &= g_k \ . \end{aligned}$$

□

Example 3.14. Consider the sequences

$$g_1 = \mathtt{1}, \mathtt{0}, \mathtt{1}, \mathtt{0}, \mathtt{1}, \mathtt{0}, \mathtt{1}, \mathtt{0}, \mathtt{1}, \mathtt{0}, \mathtt{1}, \mathtt{0}, \mathtt{1}, \mathtt{0}, \mathtt{1}, \mathtt{0}, \ldots,$$
$$g_2 = \mathtt{1}, \mathtt{0}, \mathtt{0}, \mathtt{1}, \mathtt{0}, \mathtt{0}, \mathtt{1}, \mathtt{0}, \mathtt{0}, \mathtt{1}, \mathtt{0}, \mathtt{0}, \mathtt{1}, \mathtt{0}, \mathtt{0}, \mathtt{1}, \ldots.$$

If we use the notation of Lemma 3.13, then we have $c_1 = 2$, $c_2 = 3$, and $K_1 = K_2 = 1$. Hence, $c = \operatorname{lcm}(2,3) = 6$ and $K = \max(K_1, K_2) = 1$. We have

$$g = g_1 \oplus g_2 = \mathtt{1}, \mathtt{0}, \mathtt{1}, \mathtt{1}, \mathtt{1}, \mathtt{0}, \mathtt{1}, \mathtt{0}, \mathtt{1}, \mathtt{1}, \mathtt{1}, \mathtt{0}, \mathtt{1}, \mathtt{0}, \mathtt{1}, \mathtt{1}, \ldots.$$

It is easy to verify that $g_{k+6} = g_k$ for all $k \geq 1$.

THEOREM 3.15. *Let $\hat{A} \in \mathbb{B}^{n\times n}$ be a matrix of the form* (1) *where the matrices $\hat{A}_{11}$, $\hat{A}_{22}$, …, $\hat{A}_{ll}$ are irreducible. Define sets α_1, α_2, …, α_l such that $\hat{A}_{\alpha_i\alpha_j} = \hat{A}_{ij}$ for all i, j with $i \leq j$. Let $n_i = \#\alpha_i$ and $c_{ii} = c_i = c(\hat{A}_{ii})$ for all i. Define*

$$\lambda_i = \begin{cases} \mathtt{0} & \text{if } \hat{A}_{ii} = [\mathtt{0}], \\ \mathtt{1} & \text{otherwise} \end{cases}$$

for $i = 1, 2, \ldots, l$. Define

$$S_{ij} = \big\{\, \{i_0, i_1, \ldots, i_s\} \subseteq \{1, 2, \ldots, l\} \;\big|\; i = i_0 < i_1 < \ldots < i_s = j \text{ and } \hat{A}_{i_r i_{r+1}} \neq \mathcal{O} \text{ for } r = 0, 1, \ldots, s-1 \,\big\}$$

for all i, j with $i < j$.

Let $\lambda_{ii} = \lambda_i$ and $k_{ii} = k_i = k_{n_i,c_i}$ for $i = 1, 2, \ldots, n$, where k_{n_i,c_i} is defined as in (7) *with $k_{n_i,0} = 0$ by definition.*

For each i, j with $i < j$ we define for each $\gamma \in S_{ij}$:

$$\delta_\gamma = \{t \in \gamma \,|\, \lambda_t \neq \mathtt{0}\}$$
$$c_\gamma = \begin{cases} \gcd\{c_t \,|\, t \in \delta_\gamma\} & \text{if } \delta_\gamma \neq \emptyset, \\ 1 & \text{otherwise.} \end{cases}$$

Define

$$\Gamma_{ij} = \big\{\, t \,\big|\, \exists \gamma \in S_{ij} \text{ such that } t \in \gamma \,\big\},$$
$$\Delta_{ij} = \big\{\, t \,\big|\, t \in \Gamma_{ij} \text{ and } \lambda_t \neq \mathtt{0} \,\big\},$$
$$\lambda_{ij} = \begin{cases} \mathtt{1} & \text{if } \Delta_{ij} \neq \emptyset, \\ \mathtt{0} & \text{otherwise,} \end{cases}$$
$$c_{ij} = \begin{cases} \operatorname{lcm}\{c_\gamma | \gamma \in S_{ij}\} & \text{if } \lambda_{ij} \neq \mathtt{0} \text{ and } c_\gamma \neq 1 \text{ for each } \gamma \in S_{ij} \text{ with } \delta_\gamma \neq \emptyset, \\ 1 & \text{otherwise,} \end{cases}$$
$$r_{ij} = \begin{cases} \arg\max\{n_t \,|\, t \in \Delta_{ij}\} & \text{if } \lambda_{ij} \neq \mathtt{0}, \\ 1 & \text{otherwise,} \end{cases}$$
$$k_{ij} = \begin{cases} \displaystyle\sum_{t\in\Delta_{ij}} k_{n_t,c_t} + \#\Gamma_{ij} - 1 + \max\left(0, \max_{\substack{\gamma\in S_{ij}\\ \delta_\gamma\neq\emptyset}} \left\{ \frac{n_{r_{ij}}^2}{c_\gamma} - 3n_{r_{ij}} + 2c_\gamma \right\}\right) & \text{if } \lambda_{ij} \neq \mathtt{0}, \\ \#\Gamma_{ij} & \text{if } \Gamma_{ij} \neq \emptyset \text{ and } \lambda_{ij} = \mathtt{0}, \\ 1 & \text{if } \Gamma_{ij} = \emptyset \end{cases}$$

for all i, j with $i < j$.

Then we have for all i, j with $i \leq j$:

$$\left(\hat{A}^{\otimes^{k+c_{ij}}}\right)_{\alpha_i\alpha_j} = \left(\hat{A}^{\otimes^{k}}\right)_{\alpha_i\alpha_j} \qquad \text{for all } k \geq k_{ij} \tag{13}$$

and

$$\left(\hat{A}^{\otimes^{k}} \oplus \hat{A}^{\otimes^{k+1}} \oplus \ldots \oplus \hat{A}^{\otimes^{k+c_{ij}-1}}\right)_{\alpha_i\alpha_j} = \begin{cases} \mathcal{E}_{n_i\times n_j} & \text{if } \lambda_{ij} \neq \mathfrak{o}, \\ \mathcal{O}_{n_i\times n_j} & \text{if } \lambda_{ij} = \mathfrak{o} \end{cases} \tag{14}$$

for all $k \geq k_{ij}$.

For all i, j with $i > j$ we have

$$\left(\hat{A}^{\otimes^{k}}\right)_{\alpha_i\alpha_j} = \mathcal{O}_{n_i\times n_j} \qquad \text{for all } k \in \mathbb{N}. \tag{15}$$

Proof. Let C_i be the m.s.c.s. of $\mathcal{G}(\hat{A})$ that corresponds to $\hat{A}_{ii}$ for $i = 1, 2, \ldots, l$.

Let $i, j \in \{1, 2, \ldots, l\}$. In the proof of Theorem 3.15 it has already been proved that (15) holds if $i > j$ and that (13) and (14) hold if $i \leq j$ and $\Gamma_{ij} = \emptyset$, or if $i \leq j$, $\Gamma_{ij} \neq \emptyset$ and $\lambda_{ij} = \mathfrak{o}$. Furthermore, (13) and (14) hold by Theorem 3.11 if $i = j$.

From now on we assume that $i < j$, $\Gamma_{ij} \neq \emptyset$, and $\lambda_{ij} \neq \mathfrak{o}$. Note that $i < j$ implies that $l > 1$.

Select an arbitrary vertex u of C_i and an arbitrary vertex v of C_j.

Since $\lambda_{ij} \neq \mathfrak{o}$ there exists at least one set $\gamma \in S_{ij}$ for which $\delta_\gamma \neq \emptyset$. Thus $\Delta_{ij} \neq \emptyset$ and k_{ij} is well defined. Note that $k_{ij} \geq \#\Gamma_{ij}$. If $\gamma_\delta = \emptyset$ for some set $\gamma \in S_{ij}$, then there do not exist paths from v to u of length $n \geq \#\Gamma_{ij}$ that correspond to γ. Therefore from now on we consider only sets $\gamma \in S_{ij}$ for which $\delta_\gamma \neq \emptyset$.

Let $\gamma = \{i_0, i_1, \ldots, i_s\} \in S_{ij}$ with $i = i_0 < i_1 < \cdots < i_s = j$. Since we assume that $\delta_\gamma \neq \emptyset$, we have $\hat{A}_{i_r i_r} \neq [\mathfrak{o}]$ for at least one index $i_r \in \gamma$. Assume that δ_γ is given by $\{j_0, j_1, \ldots, j_{\hat{s}}\}$. Define

$$\begin{aligned} \mathcal{S} = \big\{(U, V) \,\big|\, & U = \{u_0, u_1, \ldots, u_s\},\ V = \{v_0, v_1, \ldots, v_s\},\ u_s = v,\ v_0 = u, \text{ and} \\ & u_r \in \alpha_{i_r},\ v_{r+1} \in \alpha_{i_{r+1}} \text{ and } (\hat{A})_{u_r v_{r+1}} \neq \mathfrak{o} \text{ for } r = 0, 1, \ldots, s\big\}. \end{aligned}$$

Let $(U, V) \in \mathcal{S}$ with $U = \{u_0, u_1, \ldots, u_s\}$ and $V = \{v_0, v_1, \ldots, v_s\}$. Let $\mathcal{P}(\gamma, U, V)$ be the set of paths from v to u that pass through m.s.c.s. C_{i_r} for $r = 0, 1, \ldots, s$ and that enter C_{i_r} at vertex u_r for $r = 0, 1, \ldots, s-1$ and exit from C_{i_r} through vertex v_r for $r = 1, 2, \ldots, s$ (see also Figure 3.3). Let the sequences $\{(g_{\gamma,U,V})_k\}_{k=1}^\infty$ and $\{(g_\gamma)_k\}_{k=1}^\infty$ be defined by

$$(g_{\gamma,U,V})_k = \begin{cases} \mathfrak{1} & \text{if there exists a path of length } k \text{ that belongs to } \mathcal{P}(\gamma, U, V), \\ \mathfrak{o} & \text{otherwise,} \end{cases}$$

$$(g_\gamma)_k = \begin{cases} \mathfrak{1} & \text{if there exists a path of length } k \text{ that belongs to } \mathcal{P}(\gamma, U, V) \\ & \text{for some pair } (U, V) \in \mathcal{S}, \\ \mathfrak{o} & \text{otherwise.} \end{cases}$$

Define $c_\gamma = \gcd\{c_t | t \in \delta_\gamma\}$. Let us now show that the sequences $\{(g_{\gamma,U,V})_k\}_{k=1}^\infty$ and $\{(g_\gamma)_k\}_{k=1}^\infty$ satisfy

$$(g_{\gamma,U,V})_{k+c_\gamma} = (g_{\gamma,U,V})_k \quad \text{for all } k \geq k_{ij}, \tag{16}$$

$$(g_{\gamma,U,V})_k \oplus (g_{\gamma,U,V})_{k+1} \oplus \cdots \oplus (g_{\gamma,U,V})_{k+c_\gamma-1} = \mathbf{1} \quad \text{for all } k \geq k_{ij}, \tag{17}$$

$$(g_\gamma)_{k+c_\gamma} = (g_\gamma)_k \quad \text{for all } k \geq k_{ij}, \tag{18}$$

$$(g_\gamma)_k \oplus (g_\gamma)_{k+1} \oplus \cdots \oplus (g_\gamma)_{k+c_\gamma-1} = \mathbf{1} \quad \text{for all } k \geq k_{ij}. \tag{19}$$

Note that if (16) and (17) hold for each pair $(U,V) \in \mathcal{S}$, then (18) and (19) also hold. Therefore, we now show that (16) and (17) hold. Define $\tilde{u}_r = u_s$ and $\tilde{v}_r = v_s$ if $j_r = i_s$ for $r = 0, 1, \ldots, \hat{s}$. We consider three cases:

Case A. $c_{j_{\tilde{r}}} = 1$ for some $\tilde{r} \in \{0, 1, \ldots, \hat{s}\}$.

In this case we have $c_\gamma = 1$.

Let l_r be the length of the shortest (possibly empty) path from vertex $\tilde{u}_r$ to vertex $\tilde{v}_r$ of C_{j_r} for each $r \in \{0, 1, \ldots, \hat{s}\}$. We have $l_r \leq n_{j_r} - 1$ for each r. Since $\hat{A}_{j_{\tilde{r}} j_{\tilde{r}}}$ is irreducible and since $c_{j_{\tilde{r}}} = 1$, it follows from Theorem 3.3 that for any integer $p \geq (n_{j_{\tilde{r}}} - 1)^2 + 1$ there exists a path of length p from vertex $\tilde{u}_{\tilde{r}}$ to vertex $\tilde{v}_{\tilde{r}}$ of $C_{j_{\tilde{r}}}$. If we also take into account that there are s arcs of the form $v_{r+1} \to u_r$ for $r = 0, 1, \ldots, s-1$, then it follows that for any

$$k \geq \sum_{\substack{r=0 \\ r \neq \tilde{r}}}^{\hat{s}} (n_{j_r} - 1) + s + (n_{j_{\tilde{r}}} - 1)^2 + 1 \stackrel{\text{def}}{=} K_{\mathrm{A}}$$

there exists a path of length k that belongs to $\mathcal{P}(\gamma, U, V)$. Let us now show that $K_{\mathrm{A}} \leq k_{ij}$. Let $r \in \{0, 1, \ldots, \hat{s}\}$. If $c_{j_r} > 1$, then it follows from the definition of $k_{n_{j_r}, c_{j_r}}$ that $n_{j_r} - 1 \leq k_{n_{j_r}, c_{j_r}}$. Furthermore, if $c_{j_r} = 1$, then we have $n_{j_r} - 1 \leq (n_{j_r} - 1)^2 + 1 = k_{n_{j_r}, c_{j_r}}$. Since $c_{j_{\tilde{r}}} = 1$, we have $(n_{j_{\tilde{r}}} - 1)^2 + 1 = k_{n_{j_{\tilde{r}}}, c_{j_{\tilde{r}}}}$. Furthermore, $\delta_\gamma \subseteq \Delta_{ij}$ and $s \leq \#\Gamma_{ij} - 1$. Hence, $K_{\mathrm{A}} \leq k_{ij}$, which implies that (16) and (17) hold in this case.

Case B. $c_t = c_\gamma$ and $c_t \neq 1$ for all $t \in \delta_\gamma$.

Assume that $c_{j_{\tilde{r}}} = c_\gamma$ with $\tilde{r} \in \{0, 1, \ldots, \hat{s}\}$.

Let l_r be the length of the shortest (possibly empty) path from vertex $\tilde{u}_r$ to vertex $\tilde{v}_r$ of C_{j_r} for each $r \in \{0, 1, \ldots, \tilde{r}-1, \tilde{r}+1, \ldots, \hat{s}\}$. Thus $l_r \leq n_{j_r} - 1$ for each $r \neq \tilde{r}$. From Lemma 2.7 and from the proof of Theorem 3.11 it follows that there exists an integer $K_{\tilde{r}}$ with $k_{n_{j_{\tilde{r}}}, c_{j_{\tilde{r}}}} \leq K_{\tilde{r}} \leq k_{n_{j_{\tilde{r}}}, c_{j_{\tilde{r}}}} + c_{j_{\tilde{r}}} - 1$ such that there exist paths of length $K_{\tilde{r}} + p\, c_{j_{\tilde{r}}}$ from vertex $\tilde{u}_{\tilde{r}}$ to vertex $\tilde{v}_{\tilde{r}}$ of $C_{j_{\tilde{r}}}$ for any $p \in \mathbb{N}$, while there do not exist paths of length $K_{\tilde{r}} + p c_{j_{\tilde{r}}} + q$ from $\tilde{u}_{\tilde{r}}$ to $\tilde{v}_{\tilde{r}}$ for any $p \in \mathbb{N}$ and any $q \in \{1, 2, \ldots, c_{j_{\tilde{r}}} - 1\}$. Therefore if we define

$$K_{\mathrm{B}} = \sum_{\substack{r=0 \\ r \neq \tilde{r}}}^{\hat{s}} l_r + s + k_{n_{j_{\tilde{r}}}, c_{j_{\tilde{r}}}},$$

then it follows from Lemma 2.7 that for any $k \geq K_{\mathrm{B}}$ either there exists a path of length $k + p\, c_\gamma$ that belongs to $\mathcal{P}(\gamma, U, V)$ for each $p \in \mathbb{N}$, or there do not exist paths of length $k + p\, c_\gamma$ that belong to $\mathcal{P}(\gamma, U, V)$ for any $p \in \mathbb{N}$. It is easy to verify that $K_{\mathrm{B}} \leq k_{ij}$. Hence, (16) and (17) also hold in this case.

Case C. $c_{j_r} \neq 1$ for all $r \in \{0, 1, \ldots, \hat{s}\}$ and $c_{j_a} \neq c_{j_b}$ for some $a, b \in \{0, 1, \ldots, \hat{s}\}$.

From Lemma 2.7 and from the proof of Theorem 3.11 it follows that for each $r \in \{0, 1, \ldots, \hat{s}\}$ there exists an integer K_r with $k_{n_{j_r}, c_{j_r}} \leq K_r \leq k_{n_{j_r}, c_{j_r}} + c_{j_r} - 1$ such that there exist paths of length $K_r + p\, c_{j_r}$ from $\tilde{u}_r$ to $\tilde{v}_r$ for each $p \in \mathbb{N}$, while there do not exist paths of length $K_r + p\, c_{j_r} + q$ from $\tilde{u}_r$ to $\tilde{v}_r$ for any $p \in \mathbb{N}$ and for any $q \in \{1, 2, \ldots, c_{j_r} - 1\}$. This implies that there exist paths of length $K_0 +$

$K_1 + \cdots + K_{\hat{s}} + s + p_0 c_{j_0} + p_1 c_{j_1} + \cdots + p_{\hat{s}} c_{j_{\hat{s}}}$ that belong to $\mathcal{P}(\gamma, U, V)$ for each choice of $p_0, p_1, \ldots, p_{\hat{s}}$ such that $p_r \geq 0$ for each r. Define $K_\gamma = \sum_{r=0}^{\hat{s}} K_r + s$. Since $c_\gamma = \gcd\{c_t \,|\, t \in \delta_\gamma\}$ there exist positive integers $w_0, w_1, \ldots, w_{\hat{s}}$ such that $c_{j_r} = w_r\, c_\gamma$ for each $r \in \{0, 1, \ldots, \hat{s}\}$ and such that $\gcd(w_0, w_1, \ldots, w_{\hat{s}}) = 1$. Thus for any integer $q \geq g(w_0, w_1, \ldots, w_{\hat{s}}) + 1$ there exist nonnegative integers $\alpha_0, \alpha_1, \ldots, \alpha_{\hat{s}}$ such that $q = \alpha_0 w_0 + \alpha_1 w_1 + \cdots + \alpha_{\hat{s}} w_{\hat{s}}$.

Since $c_{j_a} \neq c_{j_b}$ we have $w_a \neq w_b$. Therefore, we may assume without loss of generality that $w_0 < w_1 < \cdots < w_{\hat{s}}$ with $\hat{s} \geq 2$. We have $w_{\hat{s}} = \frac{c_{j_{\hat{s}}}}{c_\gamma} \leq \frac{n_{j_{\hat{s}}}}{c_\gamma} \leq \frac{n_{r_{ij}}}{c_\gamma}$. Hence, $w_0 \leq \frac{n_{r_{ij}}}{c_\gamma} - 1$ since $w_0 < w_{\hat{s}}$. Furthermore, $w_0 \geq 1$ and thus $w_{\hat{s}} \geq 2$. Using a reasoning that is similar to the one used in the proof of Theorem 3.11 and Remark 3.12 we can show that

$$g(w_0, w_1, \ldots, w_{\hat{s}}) + 1 \leq \left(\frac{n_{r_{ij}}}{c_\gamma}\right)^2 - 3\,\frac{n_{r_{ij}}}{c_\gamma} + 2.$$

Thus if we define

$$K_{\mathrm{C}} = K_\gamma + c_\gamma \left(\left(\frac{n_{r_{ij}}}{c_\gamma}\right)^2 - 3\,\frac{n_{r_{ij}}}{c_\gamma} + 2\right),$$

then we have $K_{\mathrm{C}} \geq K_\gamma + c_\gamma\big(g(w_0, w_1, \ldots, w_{\hat{s}}) + 1\big)$. Let $k \in \mathbb{N}$ with $k \geq K_{\mathrm{C}}$. Now there are two possibilities:

- If k can be written as $K_\gamma + q\, c_\gamma$ with $q \geq g(w_0, w_1, \ldots, w_{\hat{s}}) + 1$ then we have

$$k = K_\gamma + (\alpha_0 w_0 + \alpha_1 w_1 + \cdots + \alpha_{\hat{s}} w_{\hat{s}}) c_\gamma = K_\gamma + \alpha_0\, c_{j_0} + \alpha_1\, c_{j_1} + \cdots + \alpha_{\hat{s}}\, c_{j_{\hat{s}}},$$

which implies that there exists a path of length k that belongs to $\mathcal{P}(\gamma, U, V)$.

- On the other hand, if k cannot be written as $K_\gamma + q\, c_\gamma$ for any $q \in \mathbb{N}$, then it follows from Lemma 2.7 that there does not exist a path of length k that belongs to $\mathcal{P}(\gamma, U, V)$.

Since $k_{ij} \geq K_{\mathrm{C}}$ this implies that (16) and (17) also hold in this case.

If we consider all possible paths from vertex v to vertex u of length $k \in \mathbb{N}$ with $k \geq \#\Gamma_{ij}$, then each of these paths corresponds to some set $\gamma \in S_{ij}$ with $\delta_\gamma \neq \emptyset$. Since $(\hat{A}^{\otimes^k})_{uv}$ is equal to $\mathbb{1}$ if and only if there exists a path of length k from v to u, we have

$$\left(\hat{A}^{\otimes^k}\right)_{uv} = \bigoplus_{\substack{\gamma \in S_{ij} \\ \delta_\gamma \neq \emptyset}} (g_\gamma)_k$$

if $k \geq \#\Gamma_{ij}$. Note that if $c_\gamma = 1$ and $\delta_\gamma \neq \emptyset$ then we have $(g_\gamma)_{k_{ij}} = \mathbb{1}$. Since each sequence $\{(g_\gamma)_k\}_{k=1}^{\infty}$ satisfies (18), it follows from Lemma 3.13 that

$$\left(\hat{A}^{\otimes^{k+c_{ij}}}\right)_{uv} = \left(\hat{A}^{\otimes^k}\right)_{uv} \qquad \text{for all } k \geq k_{ij}.$$

Furthermore, since each sequence $\{(g_\gamma)_k\}_{k=1}^{\infty}$ satisfies (19), we have

$$\left(\hat{A}^{\otimes^k} \,\oplus\, \hat{A}^{\otimes^{k+1}} \,\oplus\, \cdots \,\oplus\, \hat{A}^{\otimes^{k+c_{ij}-1}}\right)_{uv} = \mathbb{1} \qquad \text{for all } k \geq k_{ij}.$$

Thus (13) and (14) also hold if $\lambda_{ij} \neq \mathbb{0}$. □

Remark 3.16. Note that if $\gamma_1, \gamma_2 \in S_{ij}$ and $\gamma_1 \subseteq \gamma_2$, then we do not have to consider γ_1 when we are determining S_{ij}. Hence, we could have defined S_{ij} as the set of *maximal* subsets $\{i_0, i_1, \ldots, i_s\}$ of $\{1, 2, \ldots, l\}$ with $i = i_0 < i_1 < \cdots < i_s = j$ and $\hat{A}_{i_r i_{r+1}} \neq \mathcal{O}$ for $r = 0, 1, \ldots, s-1$.

Let us now give an example in which the various sets and indices that appear in the formulation of Theorem 3.15 are illustrated.

Example 3.17. Consider the matrix

$$A = \left[\begin{array}{c|cc|ccc|c} 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right].$$

This matrix is in Frobenius normal form and its block structure is indicated by the horizontal and vertical lines. The precedence graph of A is represented in Figure 3.5. We have

$$A^{\otimes 2} = \left[\begin{array}{ccc|ccc|c} 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \quad A^{\otimes 3} = \left[\begin{array}{ccc|ccc|c} 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right],$$

$$A^{\otimes 4} = \left[\begin{array}{ccc|ccc|c} 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \quad A^{\otimes 5} = \left[\begin{array}{ccc|ccc|c} 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right],$$

$$A^{\otimes 6} = \left[\begin{array}{ccc|ccc|c} 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \quad A^{\otimes 7} = \left[\begin{array}{ccc|ccc|c} 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right],$$

$$A^{\otimes 8} = \left[\begin{array}{ccc|ccc|c} 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \quad A^{\otimes 9} = \left[\begin{array}{ccc|ccc|c} 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right],$$

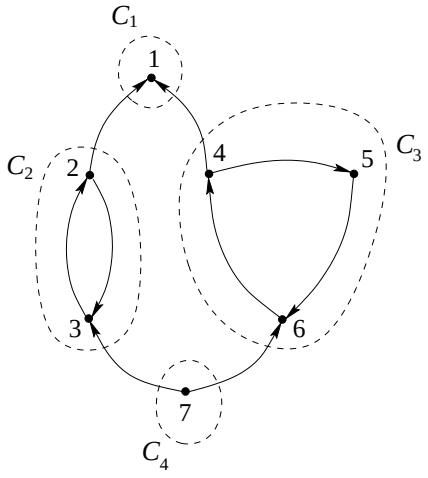


FIG. 3.5. *The precedence graph $\mathcal{G}(A)$ of the matrix A of Example* 3.17. *The subgraphs C_1, C_2, C_3, and C_4 are the m.s.c.s.'s of $\mathcal{G}(A)$.*

$$A^{\otimes^{10}} = \left[\begin{array}{c|cc|ccc|c} \mathtt{0} & \mathtt{0} & \mathtt{1} & \mathtt{1} & \mathtt{0} & \mathtt{0} & \mathtt{0} \\ \hline \mathtt{0} & \mathtt{1} & \mathtt{0} & \mathtt{0} & \mathtt{0} & \mathtt{0} & \mathtt{1} \\ \mathtt{0} & \mathtt{0} & \mathtt{1} & \mathtt{0} & \mathtt{0} & \mathtt{0} & \mathtt{0} \\ \hline \mathtt{0} & \mathtt{0} & \mathtt{0} & \mathtt{0} & \mathtt{0} & \mathtt{1} & \mathtt{0} \\ \mathtt{0} & \mathtt{0} & \mathtt{0} & \mathtt{1} & \mathtt{0} & \mathtt{0} & \mathtt{0} \\ \mathtt{0} & \mathtt{0} & \mathtt{0} & \mathtt{0} & \mathtt{1} & \mathtt{0} & \mathtt{1} \\ \hline \mathtt{0} & \mathtt{0} & \mathtt{0} & \mathtt{0} & \mathtt{0} & \mathtt{0} & \mathtt{0} \end{array}\right], \ldots.$$

Thus $A^{\otimes^{k+6}} = A^{\otimes^k}$ for all $k \geq 2$.

We have $\alpha_1 = \{1\}$, $\alpha_2 = \{2,3\}$, $\alpha_3 = \{4,5,6\}$, and $\alpha_4 = \{7\}$. Furthermore, $\lambda_1 = \lambda_4 = \mathtt{0}$ and $\lambda_2 = \lambda_3 = \mathtt{1}$. Let us now look at the sequence $\left\{\left(A^{\otimes^k}\right)_{\alpha_1\alpha_4}\right\}_{k=1}^{\infty}$. We have $S_{14} = \{\gamma_1, \gamma_2\}$ with $\gamma_1 = \{1,2,4\}$ and $\gamma_2 = \{1,3,4\}$. So $\delta_{\gamma_1} = \{2\}$, $c_{\gamma_1} = 2$, $\delta_{\gamma_2} = \{3\}$, and $c_{\gamma_2} = 3$. We have $\Gamma_{14} = \{1,2,3,4\}$, $\Delta_{14} = \{2,3\}$, $\lambda_{14} = \mathtt{1}$, $c_{14} = \mathrm{lcm}(2,3) = 6$, and $r_{14} = 3$. Hence,

$$\begin{aligned} k_{14} &= k_{2,2} + k_{3,3} + \#\Gamma_{14} - 1 + \max\left(0,\, \frac{n_3^2}{c_{\gamma_1}} - 3n_3 + 2\,c_{\gamma_1},\, \frac{n_3^2}{c_{\gamma_2}} - 3n_3 + 2\,c_{\gamma_2}\right) \\ &= \frac{3}{2} + 4 + 4 - 1 + \max\left(0,\, \frac{9}{2} - 9 + 4,\, \frac{9}{3} - 9 + 6\right) = \frac{17}{2}\,. \end{aligned}$$

Note that we indeed have $\left(A^{\otimes^{k+6}}\right)_{\alpha_1\alpha_4} = \left(A^{\otimes^k}\right)_{\alpha_1\alpha_4}$ for all $k \geq 9$.

LEMMA 3.18. *Consider m positive integers c_1, c_2, ..., c_m. Let $c = \mathrm{lcm}(c_1, c_2, \ldots, c_m)$. Consider r nonempty subsets α_1, α_2, ..., α_r of $\{1,2,\ldots,m\}$. Define $d_i = \gcd\{c_k \,|\, k \in \alpha_i\}$ for each i. If $d = \mathrm{lcm}(d_1, d_2, \ldots, d_r)$, then d is a divisor of c.*

Proof. We may assume without loss of generality that $c_i \neq c_j$ for all i, j with $i \neq j$. If d is a divisor of $\mathrm{lcm}(c_1, c_2, \ldots, c_{m-1})$, then it also is a divisor of $\mathrm{lcm}(c_1, c_2, \ldots, c_m)$. Therefore, we may assume without loss of generality that $\alpha_1 \cup \alpha_2 \cup \cdots \cup \alpha_r = \{1,2,\ldots,m\}$. If $\alpha_i \subseteq \alpha_j$, then d_j is a divisor of d_i and then $\mathrm{lcm}(d_1, d_2, \ldots, d_r) = \mathrm{lcm}(d_1, d_2, \ldots, d_{j-1}, d_{j+1}, \ldots, d_r)$, which implies that α_j is redundant and may be removed. If $d_i = d_j$, then α_j is also redundant and may be removed. It is easy

to verify that if we remove all redundant sets, then the resulting number of sets α_i is less than or equal to m (the worst cases being when $\alpha_i = \{1, 2, \ldots, m\} \setminus \{i\}$ for $i = 1, 2, \ldots, m$ or when $\alpha_i = \{i\}$ for $i = 1, 2, \ldots, m$). Hence, we may assume without loss of generality that $r \leq m$ and that $d_i \neq d_j$ for all i, j with $i \neq j$.

Since $r \leq m$, $\alpha_1 \cup \alpha_2 \cup \cdots \cup \alpha_r = \{1, 2, \ldots, m\}$ and $\alpha_i \not\subseteq \alpha_j$ for all i, j with $i \neq j$, we can select indices l_1, l_2, ..., l_r such that $l_i \in \alpha_i$ for $i = 1, 2, \ldots, r$ and $l_i \neq l_j$ for all i, j with $i \neq j$.

Since $d_i = \gcd\{c_k \,|\, k \in \alpha_i\}$, d_i is a divisor of c_{l_i} for each i.

Since all c_i's are different we have

$$c = \operatorname{lcm}(c_1, c_2, \ldots, c_m) = \frac{c_1\, c_2\, \ldots\, c_m}{\gcd(c_1, c_2, \ldots, c_m)}\,.$$

We also have

$$d = \frac{d_1\, d_2\, \ldots\, d_r}{\gcd(d_1, d_2, \ldots, d_r)}\,.$$

Since $\alpha_1 \cup \alpha_2 \cup \ldots \cup \alpha_r = \{1, 2, \ldots, m\}$ we have $\gcd(c_1, c_2, \ldots, c_m) = \gcd(d_1, d_2, \ldots, d_r)$. Therefore

$$c = \frac{c_1\, c_2\, \ldots\, c_m}{d_1\, d_2\, \ldots\, d_r}\, d.$$

Since d_i is a divisor of c_{l_i} for $i = 1, 2, \ldots, r$ there exist integers $w_i \in \mathbb{N}$ such that $c_{l_i} = w_i\, d_i$ for $i = 1, 2, \ldots, r$. If we define $\beta = \{1, 2, \ldots, m\} \setminus \{l_1, l_2, \ldots, l_r\}$, then we have

$$c = w_1\, w_2\, \ldots\, w_r \left(\prod_{i \in \beta} c_i \right) d,$$

where $\prod_{i \in \emptyset} c_i$ is equal to 1 by definition. Thus $c = wd$ for some integer $w \in \mathbb{N}$, which implies that d is a divisor of c. □

LEMMA 3.19. *Let $n \in \mathbb{N}$ and let $c \in \{0, 1, \ldots, n\}$. If $k_{n,c}$ is defined by* (7) *if $c > 0$ and if $k_{n,c} = 0$ if $c = 0$, then we have $k_{n,c} \leq (n-1)^2 + 1$.*

Proof. It is obvious that the lemma holds if $c = 0$ or if $c = 1$, so from now on we assume that $c > 1$.

Define $f(c) = \frac{n^2-1}{2} + \frac{n^2}{c} - 3n + 2c$. Let us now show that $f(c) \leq (n-1)^2 + 1$.

We have $\frac{df}{dc} = -\frac{n^2}{c^2} + 2$. Thus f reaches a local minimum in $c = \frac{n}{\sqrt{2}}$ and is decreasing if $1 < c < \frac{n}{\sqrt{2}}$ and increasing if $c > \frac{n}{\sqrt{2}}$.

Let us first consider the cases where n is equal to 2 or 3. If $n = 2$, then we have $c = 2$ and $f(c) = \frac{4-1}{2} + 2 - 6 + 4 = \frac{3}{2} \leq 2 = (2-1)^2 + 1$. If $n = 3$, then $\frac{n}{\sqrt{2}}$ does not belong to the interval $[2, 3]$ and then the maximal value of f in $[2, 3]$ is equal to $f(3) = \frac{9-1}{2} + \frac{9}{3} - 9 + 6 = \frac{9}{2} \leq 5 = (3-1)^2 + 1$.

From now on we assume that n is larger than or equal to 4. If $n \geq 4$, then $\frac{n}{\sqrt{2}}$ belongs to the interval $[2, n]$ and then the maximal value of f in $[2, n]$ is equal to

$$\begin{aligned}
\max(f(2), f(n)) &= \max\left(\frac{n^2-1}{2} + \frac{n^2}{2} - 3n + 4,\ \frac{n^2-1}{2} + n - 3n + 2n \right) \\
&= \max\left(n^2 - 3n + \frac{7}{2}\,,\ \frac{n^2}{2} - \frac{1}{2} \right) \\
&= n^2 - 3n + \frac{7}{2} \qquad (\text{since } n \geq 4) \\
&\leq (n-1)^2 + 1 \qquad (\text{since } n \geq 2).
\end{aligned}$$

Hence, $k_{n,c} \leq (n-1)^2 + 1$. □

THEOREM 3.20. *Let $A \in \mathbb{B}^{n\times n}$ and let c be the cyclicity of A. We have $A^{\otimes^{k+c}} = A^{\otimes^k}$ for all $k \geq 2n^2 - 3n + 2$.*

Proof. Let $\hat{A} = P \otimes A \otimes P^T$ be the Frobenius normal from of A where P is a permutation matrix. Assume that $\hat{A}$ is a matrix of the form (1) with $\hat{A}_{ii}$ irreducible for $i = 1, 2, \ldots, l$. Consider $i, j \in \{1, 2, \ldots, l\}$ with $i \leq j$. Let S_{ij}, Γ_{ij}, Δ_{ij}, λ_{ij}, c_{ij}, r_{ij}, and k_{ij} be defined as in Theorem 3.15. Let us now show that $k_{ij} \leq 2(n-1)^2 + n$. It is easy to verify that this holds if $i = j$, or if $i < j$ and $\lambda_{ij} = \mathfrak{o}$. Thus from now on we assume that $\lambda_{ij} \neq \mathfrak{o}$. Hence,

$$k_{ij} = \sum_{t\in\Delta_{ij}} k_{n_t,c_t} + \#\Gamma_{ij} - 1 + \max\left(0, \max_{\substack{\gamma\in S_{ij}\\ \delta_\gamma\neq\emptyset}} \left\{ \frac{n_{r_{ij}}^2}{c_\gamma} - 3n_{r_{ij}} + 2c_\gamma \right\}\right).$$

Since we have $k_{n_i,c_i} \leq (n_i - 1)^2 + 1$ for each $i \in \{1, 2, \ldots, l\}$ by Lemma 3.19, it follows from the proof of Lemma 3.8 that

$$\sum_{t\in\Delta_{ij}} k_{n_t,c_t} \leq (n-1)^2 + 1.$$

Furthermore, $\#\Gamma_{ij} \leq j - i + 1 \leq n$.

Let us now show that

$$\frac{n_{r_{ij}}^2}{c_\gamma} - 3n_{r_{ij}} + 2c_\gamma \leq (n-1)^2$$

for each $\gamma \in S_{ij}$ with $\delta_\gamma \neq \emptyset$. Let $\gamma \in S_{ij}$ with $\delta_\gamma \neq \emptyset$. From Lemma 2.6 it follows that $c_t \leq n_t$ for each $t \in \delta_\gamma$. Hence, $1 \leq c_\gamma = \gcd\{n_t \,|\, t \in \delta_\gamma\} \leq \max\{n_t \,|\, t \in \delta_\gamma\} \leq n_{r_{ij}}$.

Define $f(c) = \frac{n_{r_{ij}}^2}{c} - 3n_{r_{ij}} + 2c$. From the proof of Lemma 3.19 it follows that f is decreasing if $c < \frac{n_{r_{ij}}}{\sqrt{2}}$ and increasing if $c > \frac{n_{r_{ij}}}{\sqrt{2}}$.

If $n_{r_{ij}} = 1$, then we have $c_\gamma = 1$ and $f(c_\gamma) = 1 - 3 + 2 = 0 \leq 0 = (1-1)^2$.

If $n_{r_{ij}} > 1$, then the maximum value of f in the interval $[1, n_{r_{ij}}]$ is equal to

$$\begin{aligned}
\max(f(1), f(n_{r_{ij}})) &= \max\big(n_{r_{ij}}^2 - 3n_{r_{ij}} + 2,\, n_{r_{ij}} - 3n_{r_{ij}} + 2n_{r_{ij}}\big)\\
&= \max\big((n_{r_{ij}} - 1)^2 + 1 - n_{r_{ij}},\, 0\big)\\
&\leq (n_{r_{ij}} - 1)^2 \qquad\qquad (\text{since } n_{r_{ij}} \geq 1)\\
&\leq (n-1)^2.
\end{aligned}$$

Hence, $k_{ij} \leq 2(n-1)^2 + n = 2n^2 - 3n + 2$. Furthermore, c_{ij} is a divisor of c by Lemma 3.18. Hence, $\hat{A}^{\otimes^{k+c}} = \hat{A}^{\otimes^k}$ for all $k \geq 2n^2 - 3n + 2$. Since $A^{\otimes^k} = P^T \otimes \hat{A}^{\otimes^k} \otimes P$ for all $k \in \mathbb{N}$, this implies that $A^{\otimes^{k+c}} = A^{\otimes^k}$ for all $k \geq 2n^2 - 3n + 2$. □

4. Applications and extensions.

4.1. Markov chains. It is often possible to represent the behavior of a physical system by describing all the different states the system can occupy and by specifying how the system moves from one state to another at each time step. If the state space of the system is discrete and if the future evolution of the system depends only on the

current state of the system and not on past history, the system may be represented by a Markov chain. Markov chains can be used to describe a wide variety of systems and phenomena in domains such as diffusion processes, genetics, learning theory, sociology, and economics, [22].

A finite homogeneous Markov chain is a stochastic process with a finite number of states s_1, s_2, ..., s_n, where the transition probability to go from one state to another state depends only on the current state and is independent of the time step. We define an $n \times n$ matrix P such that p_{ij} is equal to the probability that the next state is s_i given that the current state is s_j. Note that $p_{ij} \geq 0$ for all i, j. We define a sequence of vectors $\{\theta(k)\}_{k=0}^{\infty}$ with $\theta(k) \in [0,1]^n$, where $\big(\theta(k)\big)_i$ is the probability that the system is in state s_i at time step k. If the initial probability vector $\theta(0)$ is given, the evolution of the system is described by

$$\theta(k+1) = P\theta(k) \qquad \text{for } k \in \mathbb{N}.$$

Hence,

$$\theta(k) = P^k\theta(0) \qquad \text{for } k \in \mathbb{N}.$$

Thus if we consider the Boolean algebra $(\{0, \mathfrak{p}\}, +, \cdot)$, where $\mathfrak{p}$ stands for an arbitrary positive number, and if we define a matrix $\tilde{P} \in \{0, \mathfrak{p}\}^{n\times n}$ such that $\tilde{p}_{ij} = 0$ if $p_{ij} = 0$ and $\tilde{p}_{ij} = \mathfrak{p}$ if $p_{ij} > 0$, then we can give the following interpretation to the Boolean matrix power $\tilde{P}^{\otimes^k}$. We can go from state s_j to state s_i in k steps if and only if $\big(P^k\big)_{ij} > 0$ or equivalently if $\big(\tilde{P}^{\otimes^k}\big)_{ij} = \mathfrak{p}$. As a consequence, the results of this paper can also be used to obtain upper bounds for the length of the transient behavior of a finite homogeneous Markov chain.

For more information on Markov chains and their applications the interested reader is referred to [2, 12, 20, 22] and the references therein.

4.2. Max-plus algebra. Our main motivation for studying sequences of consecutive powers of a matrix in a Boolean algebra lies in the max-plus algebraic system theory for discrete event systems. Typical examples of discrete event systems are flexible manufacturing systems, telecommunication networks, parallel processing systems, traffic control systems, and logistic systems. The class of the discrete event systems essentially consists of man-made systems that contain a finite number of resources (e.g., machines, communications channels, or processors) that are shared by several users (e.g., product types, information packets, or jobs), all of which contribute to the achievement of some common goal (e.g., the assembly of products, the end-to-end transmission of a set of information packets, or a parallel computation).

There are many modeling and analysis techniques for discrete event systems, such as queuing theory, (extended) state machines, max-plus algebra, formal languages, automata, temporal logic, generalized semi-Markov processes, Petri nets, perturbation analysis, and computer simulation (see [1, 6, 19, 23, 24] and the references cited therein). In general models that describe the behavior of a discrete event system are nonlinear in conventional algebra. However, there is a class of discrete event systems — the max-linear discrete event systems — that can be described by a model that is "linear" in the max-plus algebra [1, 7, 8]. The model of a max-linear discrete event system can be characterized by a triple of matrices (A, B, C), which are called the system matrices of the model.

One of the open problems in the max-plus algebraic system theory is the minimal realization problem, which consists of determining the system matrices of the model

of a max-linear discrete event system starting from its "impulse response"[4] such that the dimensions of the system matrices are as small as possible (see [1] for more information). In order to tackle the general minimal realization problem it is useful first to study a simplified version: the Boolean minimal realization problem, in which only models with Boolean system matrices are considered. The results of this paper on the length of the transient part of the sequence of consecutive powers of a matrix in a Boolean algebra can be used to obtain some results for the Boolean minimal realization problem in the max-plus algebraic system theory for discrete event systems [10]: they can be used to obtain a lower bound for the minimal system order (i.e., the smallest possible size of the system matrix A) and to prove that the Boolean minimal realization problem in the max-plus algebra is decidable (and can be solved in a time that is bounded from above by a function that is exponential in the minimal system order).

Both Boolean algebras and the max-plus algebra are special cases of a dioid (i.e., an idempotent semiring) [1, 16]. For applications of dioids in graph theory, generating languages, and automata theory the interested reader is referred to [14, 15, 16].

4.3. Extensions. In this paper we have restricted ourselves to Boolean algebras. In this section we give some examples that illustrate some of the phenomena that could occur when we want to extend our results to more general algebraic structures. In our examples we shall use the max-plus algebra $(\mathbb{R} \cup \{-\infty\}, \max, +)$, but for other extensions of Boolean algebras similar examples can be constructed.

In contrast to Boolean algebra (cf. Theorem 3.20) the sequence of consecutive powers of a matrix in a more general algebraic structure does not always reach a stationary or cyclic regime after a finite number of terms, as is shown by the following example.

Example 4.1. Consider the matrix

$$A = \begin{bmatrix} 0 & -\infty \\ -\infty & -1 \end{bmatrix}.$$

Since the kth max-plus algebraic power of A is given by

$$A^{\otimes^k} = \begin{bmatrix} 0 & -\infty \\ -\infty & -k \end{bmatrix}$$

for $k \in \mathbb{N}_0$, the sequence $\left\{A^{\otimes^k}\right\}_{k=1}^{\infty}$ does not reach a stationary or cyclic regime in a finite number of steps.

Note that the matrix A of Example 4.1 is not irreducible. However, if a matrix is irreducible, then it can be shown [1, 7, 13] that the sequence of consecutive max-plus algebraic powers of the given matrix always reaches a cyclic regime of the form (2) after a finite number of terms. However, even if the sequence of consecutive powers reaches a stationary regime, then in general the length of the transient part will not depend only on the size and the cyclicity of the matrix but also on the range and the resolution (i.e., on the size of the representation) of the finite elements of the matrix, as is shown by the following examples.

Example 4.2. Let $N \in \mathbb{N}$ and consider

$$A(N) = \begin{bmatrix} -1 & -N \\ 0 & 0 \end{bmatrix}.$$

[4]This is the output of the system when a certain standardized input sequence is applied to the system.

The matrix $A(N)$ is irreducible and has cyclicity 1 and its λ-value[5] is equal to 0. The kth max-plus algebraic power of $A(N)$ is given by

$$\big(A(N)\big)^{\otimes^k} = \begin{bmatrix} \max(-k,-N) & -N \\ 0 & 0 \end{bmatrix}$$

for each $k \in \mathbb{N}_0$. This implies that the smallest integer k_0 for which (2) holds is given by $k_0 = N$, i.e., k_0 depends on the range of the finite entries of $A(N)$.

Example 4.3. Let $\varepsilon > 0$ and consider the matrix

$$A(\varepsilon) = \begin{bmatrix} 0 & 0 \\ -1 & -\varepsilon \end{bmatrix}.$$

This matrix is irreducible and has cyclicity 1 and its λ-value is equal to 0. Since the kth max-plus algebraic power of $A(\varepsilon)$ is given by

$$\big(A(\varepsilon)\big)^{\otimes^k} = \begin{bmatrix} 0 & 0 \\ -1 & \max(-1,-k\varepsilon) \end{bmatrix},$$

the smallest integer k_0 for which (2) holds is $k_0 = \lceil \frac{1}{\varepsilon} \rceil$. This example—which has been inspired by the example on p. 152 of [1]—shows that in general the length of the transient part of the sequence $\{A^{\otimes^k}\}_{k=1}^{\infty}$ depends on the resolution of the finite entries of A.

5. Conclusions. In this paper we have studied the ultimate behavior of the sequence of consecutive powers of a matrix in a Boolean algebra, and we have derived some upper bounds for the length of the transient part of this sequence. The results that have been derived in this paper can be used in the analysis of the transient behavior of Markov chains and in the max-plus algebraic system theory for discrete event systems.

Topics for future research are the derivation of tighter upper bounds for the length of the transient part of the sequence of consecutive power of a matrix in a Boolean algebra, and extension of our results to more general algebraic structures such as the max-plus algebra.

Acknowledgments. The authors want to thank the anonymous reviewers for their useful comments and remarks and for pointing out the connection with Markov chains.

REFERENCES

[1] F. Baccelli, G. Cohen, G. Olsder, and J. Quadrat, *Synchronization and Linearity*, John Wiley, New York, 1992.
[2] A. Berman and R. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.
[3] J. Braker and G. Olsder, *The power algorithm in max algebra*, Linear Algebra Appl., 182 (1993), pp. 67–89.
[4] A. Brauer, *On a problem of partitions*, Amer. J. Math., 64 (1942), pp. 299–312.
[5] R. Brualdi and H. Ryser, *Combinatorial matrix theory*, in Encyclopedia of Mathematics and Its Applications, Vol. 39, Cambridge University Press, Cambridge, UK, 1991.

[5]For methods to compute the number λ that appears in Theorem 2.5 for a max-plus-algebraic matrix the reader is referred to [1, 3, 7, 21].

[6] C. Cassandras, S. Lafortune, and G. Olsder, *Introduction to the modelling, control and optimization of discrete event systems*, in Trends in Control: A European Perspective, A. Isidori, ed., Springer-Verlag, Berlin, 1995, pp. 217–291.
[7] G. Cohen, D. Dubois, J. Quadrat, and M. Viot, *A linear-system-theoretic view of discrete-event processes and its use for performance evaluation in manufacturing*, IEEE Trans. Automat. Control, 30 (1985), pp. 210–220.
[8] R. Cuninghame-Green, *Minimax Algebra*, Lecture Notes in Econom. and Math. Systems 166, Springer-Verlag, Berlin, 1979.
[9] B. De Schutter, *Max-Algebraic System Theory for Discrete Event Systems*, Ph.D. thesis, Faculty of Applied Sciences, K.U.Leuven, Leuven, Belgium, 1996.
[10] B. De Schutter, V. Blondel, R. de Vries, and B. De Moor, *On the Boolean minimal realization problem in the max-plus algebra*, Systems Control Lett., 35 (1998), pp. 69–78.
[11] P. Erdös and R. Graham, *On a linear Diophantine problem of Frobenius*, Acta Arith., 21 (1972), pp. 399–408.
[12] F. Gantmacher, *The Theory of Matrices*, Vol. 2, Chelsea Publishing Company, New York, 1959.
[13] S. Gaubert, *On rational series in one variable over certain dioids*, Tech. Rep. 2162, INRIA, Le Chesnay, France, 1994.
[14] M. Gondran and M. Minoux, *Eigenvalues and eigenvectors in semimodules and their interpretation in graph theory*, in Proceedings of the 9th International Mathematical Programming Symposium, Budapest, Hungary, 1976, pp. 333–348.
[15] M. Gondran and M. Minoux, *Graphs and Algorithms*, John Wiley, Chichester, 1984.
[16] M. Gondran and M. Minoux, *Linear algebra in dioids: A survey of recent results*, Ann. Discrete Math., 19 (1984), pp. 147–163.
[17] F. Harary, *A graph theoretic approach to matrix inversion by partitioning*, Numer. Math., 4 (1962), pp. 128–135.
[18] F. Harary, *Graph Theory*, Addison–Wesley, Reading, MA, 1969.
[19] Y. Ho, ed., *Discrete Event Dynamic Systems: Analyzing Complexity and Performance in the Modern World*, IEEE Press, Piscataway, NJ, 1992.
[20] D. Isaacson and R. Madsen, *Markov Chains, Theory and Applications*, John Wiley, New York, 1976.
[21] R. Karp, *A characterization of the minimum cycle mean in a digraph*, Discrete Math., 23 (1978), pp. 309–311.
[22] J. Kemeny, J. Snell, and A. Knapp, *Denumerable Markov Chains*, Springer-Verlag, New York, 1976.
[23] J. Peterson, *Petri Net Theory and the Modeling of Systems*, Prentice–Hall, Englewood Cliffs, NJ, 1981.
[24] Proceedings of the IEEE, 77 (1989). Special issue on Dynamics of Discrete Event Systems.
[25] R. Varga, *Matrix Iterative Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1962.
[26] M. Wang, Y. Li, and H. Liu, *On periodicity analysis and eigen-problem of matrix in max-algebra*, in Proceedings of the 1991 IFAC Workshop on Discrete Event System Theory and Applications in Manufacturing and Social Phenomena, Shenyang, China, 1991, pp. 44–48.
[27] R. Wilson, *Introduction to Graph Theory*, 2nd ed., Longman Group, London, 1979.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 357–361
© 1999 Society for Industrial and Applied Mathematics

A NOTE ON RELATIVE PERTURBATION BOUNDS*

TRISTAN LONDRÉ† AND NOAH H. RHEE†

Abstract. In this paper we provide an actual bound for the distance between the original and the perturbed right singular vector subspaces of a general matrix with full column rank. We also provide actual relative componentwise bounds for perturbed eigenvectors of a positive definite matrix.

Key words. perturbation theory, positive definite eigenproblem, singular value decomposition, right singular vector subspace

AMS subject classifications. 65F15, 15A18

PII. S0895479897331643

1. Introduction. Some relative perturbation bounds for eigenvectors and singular vectors have been studied in [2], [3], [4], [5], [6], [7], and [8]. In this note we provide an improved actual (nonasymptotic) bound for the perturbed right singular vector subspace of a matrix which is of full column rank. If the perturbation is reasonably small, this bound is always comparable to the result of [4], and it can be arbitrarily sharper. Mathias [5] has already provided actual componentwise bounds for relative accuracy of perturbed right singular vectors. We provide actual relative componentwise bounds for eigenvectors of a positive definite matrix.

Let $H = V\Lambda V^*$ and $\tilde{H} = H + \Delta H = \widetilde{V}\widetilde{\Lambda}\widetilde{V}^*$ be positive definite, where $V, \tilde{V}$ are unitary and $\Lambda, \tilde{\Lambda}$ are diagonal. We partition $V = \begin{pmatrix} V_1 & V_2 \end{pmatrix}$ and $\tilde{V} = \begin{pmatrix} \tilde{V}_1 & \tilde{V}_2 \end{pmatrix}$ so that $V_1, \tilde{V}_1 \in C^{n\times k}$. If $\sin\Theta(V_1, \tilde{V}_1)$ denotes the diagonal matrix whose entries are the sines of the canonical angles between the column spaces of V_1 and $\tilde{V}_1$, then it is well known ([4], for example) that

$$\left\|\sin\Theta\left(V_1, \tilde{V}_1\right)\right\| = \left\|V_1^*\tilde{V}_2\right\|$$

for any unitarily invariant norm. Using this notation, the next theorem, which we will use in sections 2 and 3, follows easily from the proof given by Mathias and Veselić in [6].

THEOREM 1.1. *Let $H = V\Lambda V^*$ and $\tilde{H} = H + \Delta H = \widetilde{V}\widetilde{\Lambda}\widetilde{V}^*$ be positive definite where $V, \tilde{V}$ are unitary and $\Lambda, \tilde{\Lambda}$ are diagonal with diagonal entries λ_i and $\widetilde{\lambda}_i$, respectively. (Here we don't assume any ordering of the eigenvalues.) Assume*

$$\eta \equiv \left\|H^{-1/2}\left(\Delta H\right)H^{-1/2}\right\|_2 < 1,$$

and write

$$\eta_F \equiv \left\|H^{-1/2}\left(\Delta H\right)H^{-1/2}\right\|_F.$$

For any i, j satisfying

$$\chi\left(\lambda_i, \widetilde{\lambda}_j\right) > 0,$$

*Received by the editors December 17, 1997; accepted for publication (in revised form) by I. Ipsen December 7, 1998; published electronically October 26, 1999.
http://www.siam.org/journals/simax/21-2/33164.html
†Department of Mathematics and Statistics, University of Missouri at Kansas City, Kansas City, MO 64110 (tlondre@cctr.umkc.edu, rhee@cctr.umkc.edu).

where $\chi(a,b) \equiv |a-b|/\sqrt{ab}$, *we have*

$$|s_{ij}| \leq \frac{|y_{ij}|}{\chi\left(\lambda_i, \widetilde{\lambda}_j\right)}, \tag{1.1}$$

where $S \equiv V^*\tilde{V}$ *and* $Y \equiv \Lambda^{-1/2}\left[V^*(\Delta H)\tilde{V}\right]\tilde{\Lambda}^{-1/2}$ *satisfies*

$$\|Y\|_2 \leq \frac{\eta}{\sqrt{1-\eta}} \qquad \textit{and} \qquad \|Y\|_F \leq \frac{\eta_F}{\sqrt{1-\eta}}. \tag{1.2}$$

Also, if

$$\xi_\chi \equiv \min_{1\leq i\leq k, k+1\leq j\leq n} \chi\left(\lambda_i, \tilde{\lambda}_j\right) > 0,$$

then

$$\left\|\sin\Theta\left(V_1, \tilde{V}_1\right)\right\|_F = \left\|V_1^*\tilde{V}_2\right\|_F \leq \frac{1}{\sqrt{1-\eta}} \cdot \frac{\eta_F}{\xi_\chi}. \tag{1.3}$$

Remark. The bound given by (1.3) is identical to the result of [4, Theorem 4.3], but the method of proof used by Mathias and Veselić is different.

2. Right singular vector subspace bound. We will now apply Theorem 1.1 to the estimation of a bound for the distance between the perturbed and the original right singular vector subspaces. Mathias gives a first-order version of this bound for the case of a single perturbed right singular vector in [5, Theorem 2.8].

THEOREM 2.1. *Let* $G \in C^{m\times n}$ *be of full column rank and partition it as follows:*

$$G = (U_1, U_2)\begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} V_1^* \\ V_2^* \end{pmatrix},$$

where $U_1 \in C^{m\times k}, V_1 \in C^{n\times k}, \Sigma_1 = \operatorname{diag}(\sigma_1, \ldots, \sigma_k)$ *and* $\Sigma_2 = \operatorname{diag}(\sigma_{k+1}, \ldots, \sigma_n)$. *Let* $\tilde{G} = G + \Delta G$ *be partitioned and* $\widetilde{\sigma}_i$ *be defined in a similar way. Assume*

$$\delta \equiv \left\|(\Delta G)G^\dagger\right\|_2 < \frac{1}{3} \quad \textit{and} \quad \xi_{\hat{\chi}} \equiv \min_{1\leq i\leq k, k+1\leq j\leq n} \chi\left(\sigma_i^2, \widetilde{\sigma}_j^2\right) > 0.$$

Then

$$\left\|\sin\Theta\left(V_1, \tilde{V}_1\right)\right\|_F \leq \frac{1}{\sqrt{1-3\delta}} \cdot \frac{2\delta_F + \delta_F^2}{\xi_{\hat{\chi}}},$$

where $\delta_F \equiv \left\|(\Delta G)G^\dagger\right\|_F$.

Proof. Let $\Sigma = \Sigma_1 \oplus \Sigma_2, H \equiv G^*G = V\Sigma^2V^*$, and $\tilde{H} \equiv (G+\Delta G)^*(G+\Delta G)$. If we write $\tilde{H} \equiv H + \Delta H$, then $\Delta H \equiv G^*\Delta G + \Delta G^*G + \Delta G^*\Delta G$. Note that $\delta \equiv \left\|\Delta G G^\dagger\right\|_2 = \left\|\Delta G V\Sigma^{-1}\right\|_2$. Since $H^{-1/2} = V\Sigma^{-1}V^*$ one can easily show that

$$\eta_F \equiv \left\|H^{-1/2}(\Delta H)H^{-1/2}\right\|_F \leq 2\delta_F + \delta_F^2,$$

and (using the assumption on δ) that

$$\eta \equiv \left\|H^{-1/2}(\Delta H)H^{-1/2}\right\|_2 \leq 2\delta + \delta^2 < 3\delta < 1.$$

Since the right singular vectors of G are also the eigenvectors of H and the squares of the singular values of G are also the eigenvalues of H, applying Theorem 1.1 gives the result. ∎

In [4, Theorem 4.9] Li gives the bound

$$\left\|\sin\Theta\left(V_1,\tilde{V}_1\right)\right\|_F \leq \frac{2-\delta}{2-2\delta}\cdot\frac{\delta_F}{\xi_\chi}$$

for the square case, under the assumptions that

$$\delta \equiv \left\|\Delta G G^{-1}\right\|_2 < 1 \qquad \text{and} \qquad \xi_\chi \equiv \min_{1\leq i\leq k, k+1\leq j\leq n} \chi\left(\sigma_i,\widetilde{\sigma}_j\right) > 0.$$

The most useful comparison will be the case where $\delta \ll 1$. For ease of analysis we compare first-order bounds. Notice that our bound and Li's bound can be written as

$$\left\|\sin\Theta\left(V_1,\tilde{V}_1\right)\right\|_F \leq \frac{2\delta_F}{\xi_{\hat{\chi}}} + O(\delta_F^2) \quad \text{and} \quad \left\|\sin\Theta\left(V_1,\tilde{V}_1\right)\right\|_F \leq \frac{\delta_F}{\xi_\chi} + O(\delta_F^2),$$

respectively. Note that

$$\frac{2\delta_F}{\xi_{\hat{\chi}}} \leq \frac{\delta_F}{\xi_\chi},$$

because $2\chi(a,b) \leq \chi\left(a^2,b^2\right)$ for any positive a, b. So if δ is reasonably small, our bound is always comparable with Li's. In fact, since it is possible for $\xi_{\hat{\chi}}$ to be arbitrarily bigger than ξ_χ, our bound can be arbitrarily sharper. The following example demonstrates this.

Example. Consider the 4×4 matrix $G = \operatorname{diag}\left(1, 2, 10^{-6}, 2\cdot 10^6\right)$ and the perturbed matrix

$$G + \Delta G = \begin{pmatrix} 1 & 10^{-4} & 10^{-10} & 100 \\ 0 & 2 & 10^{-10} & 100 \\ 0 & 0 & 10^{-6} & 100 \\ 0 & 0 & 0 & 2\cdot 10^6 \end{pmatrix}.$$

Here $\delta = 1.6378\cdot 10^{-4}$ and $\delta_F = 1.7320\cdot 10^{-4}$. We will let $V_1\left(\tilde{V}_1\right)$ be the matrix of the right singular vectors corresponding to the second and third biggest singular values of G $(G+\Delta G)$. The actual value of $\left\|\sin\Theta\left(V_1,\tilde{V}_1\right)\right\|_F$ is $1.2499\cdot 10^{-10}$. Our bound is $3.4652\cdot 10^{-10}$ and Li's bound is $1.7321\cdot 10^{-7}$ up to the accuracy shown.

3. Componentwise relative bounds. In this section we will find relative componentwise bounds for a perturbed eigenvector of a positive definite matrix. To our knowledge these results are the only actual componentwise bounds available.

THEOREM 3.1. *Let $H = V\Lambda V^*$ and $\tilde{H} = H + \Delta H = \tilde{V}\tilde{\Lambda}\tilde{V}^*$ be positive definite, where $V, \tilde{V}$ are unitary, $\Lambda = \operatorname{diag}(\lambda_1,\ldots,\lambda_n)$ with $\lambda_1 \geq \cdots \geq \lambda_n$, and $\tilde{\Lambda} = \operatorname{diag}\left(\widetilde{\lambda}_1,\ldots,\widetilde{\lambda}_n\right)$ with $\widetilde{\lambda}_1 \geq \cdots \geq \widetilde{\lambda}_n$. Assume*

$$\eta \equiv \left\|H^{-1/2}(\Delta H)H^{-1/2}\right\|_2 < 1$$

and

$$\eta\left(1 + \frac{1}{2\sqrt{1-\eta}}\right) < \min_{k\neq j}\rho\left(\lambda_k,\lambda_j\right), \tag{3.1}$$

where $\rho(a,b) \equiv |a-b|/(a+b)$. *Then for all* i

$$|v_{ij} - \widetilde{v}_{ij}| \leq \left(\sqrt{n-1} + .5\right) \frac{\eta}{(1-\eta)\left[\min_{k\neq j} \rho(\lambda_k, \lambda_j) - \eta\right]} \overline{v}_{ij},$$

where

$$\overline{v}_{ij} \equiv \min\left\{ \frac{1}{\sqrt{\lambda_n(A)}} \sqrt{\frac{\lambda_j}{h_{ii}}}, \sqrt{\lambda_1(A)} \sqrt{\frac{h_{ii}}{\lambda_j}} \right\},$$

$H = DAD$, *and* D *is a diagonal matrix chosen so that* A *has a unit diagonal.*

Proof. In this proof S and Y are defined as in Theorem 1.1. Also, for any matrix A, the symbol a_j will denote the jth column of A. Mathias has shown in [5, Corollary 3.3] that $|V| \leq \overline{V} \equiv (\overline{v}_{ij})$. For convenience, let $R \equiv I - S$. We can see that

$$|v_j - \widetilde{v}_j| = |V(e_j - s_j)| \leq \overline{V}\,|e_j - s_j| = \overline{V}\,|r_j|, \tag{3.2}$$

where e_j is the jth unit vector. One can verify that the condition (3.1) implies that

$$\frac{\eta}{\sqrt{1-\eta}} \cdot \frac{1}{\min_{k\neq j} \chi\left(\lambda_k, \widetilde{\lambda}_j\right)} < 1, \tag{3.3}$$

using $2\rho(a,b) \leq \chi(a,b)$ and Theorem 2.3 and the inequality (1.1) in [5]. This also implies that $\chi\left(\lambda_k, \widetilde{\lambda}_j\right) > 0$ when $k \neq j$. So for $k \neq j$, (1.1) from Theorem 1.1 tells us that

$$|r_{kj}| = |s_{kj}| \leq \frac{|y_{kj}|}{\chi\left(\lambda_k, \widetilde{\lambda}_j\right)}.$$

When $k = j$, since we may assume that s_{jj} is positive, we can use (1.1), (1.2), the fact that $1 - \sqrt{1-x} \leq x$ if $0 \leq x \leq 1$, and (3.3) to get

$$|r_{jj}| = 1 - s_{jj} = 1 - \sqrt{1 - \sum_{k\neq j} |s_{kj}|^2} \leq \frac{\eta^2}{1-\eta} \cdot \frac{1}{\min_{k\neq j}\left[\chi\left(\lambda_k, \widetilde{\lambda}_j\right)\right]^2}.$$

Thus taking the ith component from (3.2) gives

$$\begin{aligned} |v_{ij} - \widetilde{v}_{ij}| &\leq \sum_{k=1}^{n} \overline{v}_{ik}\,|r_{kj}| \\ &\leq \frac{\eta^2}{1-\eta} \cdot \frac{\overline{v}_{ij}}{\min_{k\neq j}\left[\chi\left(\lambda_k, \widetilde{\lambda}_j\right)\right]^2} + \sum_{k\neq j} \overline{v}_{ik} \frac{|y_{kj}|}{\chi\left(\lambda_k, \widetilde{\lambda}_j\right)}. \end{aligned} \tag{3.4}$$

Using (3.3) above and Theorem 2.3 and the inequality (1.1) in [5], the first term can be bounded as

$$\begin{aligned} \frac{\eta^2}{1-\eta} \cdot \frac{\overline{v}_{ij}}{\min_{k\neq j}\left[\chi\left(\lambda_k, \widetilde{\lambda}_j\right)\right]^2} &\leq \frac{\eta}{\sqrt{1-\eta}} \cdot \frac{\overline{v}_{ij}}{\min_{k\neq j} \chi\left(\lambda_k, \widetilde{\lambda}_j\right)} \\ &\leq \frac{\eta}{\sqrt{1-\eta}} \cdot \frac{\overline{v}_{ij}}{2 \cdot \min_{k\neq j} \rho\left(\lambda_k, \widetilde{\lambda}_j\right)} \\ &\leq \frac{1}{2} \cdot \frac{\eta}{(1-\eta)\left[\min_{k\neq j} \rho(\lambda_k, \lambda_j) - \eta\right]} \overline{v}_{ij}. \end{aligned} \tag{3.5}$$

We begin to bound the second term by following Mathias [5, Lemma 3.4]:

$$\begin{aligned}\sum_{k\neq j}\overline{v}_{ik}\frac{|y_{kj}|}{\chi\left(\lambda_k,\widetilde{\lambda}_j\right)} &= \sum_{k\neq j}\min\left\{\frac{1}{\sqrt{\lambda_n(A)}}\sqrt{\frac{\lambda_k}{h_{ii}}},\sqrt{\lambda_1(A)}\sqrt{\frac{h_{ii}}{\lambda_k}}\right\}\frac{|y_{kj}|}{\chi\left(\lambda_k,\widetilde{\lambda}_j\right)}\\ &\leq \max_{k\neq j}\frac{\lambda_k+\tilde{\lambda}_j}{\left|\lambda_k-\tilde{\lambda}_j\right|}\min\left\{\frac{1}{\sqrt{\lambda_n(A)}}\sqrt{\frac{\tilde{\lambda}_j}{h_{ii}}},\sqrt{\lambda_1(A)}\sqrt{\frac{h_{ii}}{\tilde{\lambda}_j}}\right\}\sum_{k\neq j}|y_{kj}|\\ &\leq \frac{\sqrt{n-1}}{\min\limits_{k\neq j}\rho\left(\lambda_k,\widetilde{\lambda}_j\right)}\cdot\min\left\{\frac{1}{\sqrt{\lambda_n(A)}}\sqrt{\frac{\tilde{\lambda}_j}{h_{ii}}},\sqrt{\lambda_1(A)}\sqrt{\frac{h_{ii}}{\tilde{\lambda}_j}}\right\}\sqrt{\sum_{k\neq j}|y_{kj}|^2}.\end{aligned}$$

Using (1.2) above, Theorem 2.3 and the inequality (1.1) in [5], the fact that $\sqrt{1+\eta}\leq 1/\sqrt{1-\eta}$, and the definition of $\overline{v}_{ij}$, the second term can be bounded by

$$\sqrt{n-1}\cdot\frac{\eta}{(1-\eta)\left[\min\limits_{k\neq j}\rho\left(\lambda_k,\lambda_j\right)-\eta\right]}\overline{v}_{ij}. \tag{3.6}$$

So it follows from (3.4), (3.5), and (3.6) that

$$|\tilde{v}_{ij}-v_{ij}|\leq\left(\sqrt{n-1}+.5\right)\frac{\eta}{(1-\eta)\left[\min\limits_{k\neq j}\rho\left(\lambda_k,\lambda_j\right)-\eta\right]}\overline{v}_{ij}$$

for all i, which proves the theorem. □

Relative componentwise bounds for right singular vectors can easily be obtained from this theorem in a manner similar to the derivation of Theorem 2.1 from Theorem 1.1; however, Mathias [5] has already provided essentially the same result.

Acknowledgments. We thank the anonymous referees whose suggestions helped improve this note.

REFERENCES

[1] J. Barlow and J. Demmel, *Computing accurate eigensystems of scaled diagonally dominant matrices*, SIAM J. Numer. Anal., 27 (1990), pp. 762–791.

[2] J. Demmel and K. Veselić, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1204–1245.

[3] S. C. Eisenstat and I. C. F. Ipsen, *Relative perturbation techniques for singular value problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1972–1988.

[4] R.-C. Li, *Relative Perturbation Theory:* (II) *Eigenspace and Singular Subspace Variations*, LAPACK working notes 85, available from http://www.netlib.org/lapack/lawns/lawn85.ps (1996).

[5] R. Mathias, *Spectral perturbation bounds for positive definite matrices*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 959–980.

[6] R. Mathias and K. Veselić, *A relative perturbation bound for positive definite matrices*, Linear Algebra Appl., 270 (1998), pp. 315–321.

[7] I. Slapničar and K. Veselić, *Perturbations of the eigenprojections of a factorized Hermitian matrix*, Linear Algebra Appl., 218 (1995), pp. 273–280.

[8] K. Veselić and I. Slapničar, *Floating-point perturbations of Hermitian matrices*, Linear Algebra Appl., 195 (1993), pp. 81–116.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 362–378
© 1999 Society for Industrial and Applied Mathematics

BOUNDS ON THE EXTREME EIGENVALUES OF REAL SYMMETRIC TOEPLITZ MATRICES*

A. MELMAN†

Abstract. We derive upper and lower bounds on the smallest and largest eigenvalues, respectively, of real symmetric Toeplitz matrices. The bounds are first obtained for positive-definite matrices and then extended to the general real symmetric case. They are computed as the roots of rational and polynomial approximations to spectral, or secular, equations for the symmetric and antisymmetric parts of the spectrum; this leads to separate bounds on the even and odd eigenvalues. We also present numerical results.

Key words. Toeplitz matrix, eigenvalues, bounds, spectral equation, secular equation, rational approximation

AMS subject classifications. 65F15, 15A18

PII. S0895479898339621

1. Introduction.

The study of eigenvalues of Toeplitz matrices continues to be of interest, due to the occurrence of these matrices in a host of applications (see [4] for a good overview) including linear prediction, a well-known problem in digital signal processing.

In this work we present improved bounds on the extreme eigenvalues of real symmetric positive-definite Toeplitz matrices and describe their extension to matrices that are not positive definite. The computation of the smallest eigenvalue of such matrices was considered in, e.g., [8, 18, 21, 22, 25, and 27], whereas bounds were studied in [11, 16, and 26]. Among the latter, the best currently available bounds seem to have been obtained in [11]. In our approach, which is similar to the one used in aforementioned papers, the eigenvalues of the matrix are computed as the roots of a one-dimensional rational function. The extreme eigenvalues can then be bounded by computing bounds on the roots of the aforementioned equation, often called a *spectral equation*, or *secular equation* (see [14]).

In [11] the bounds are obtained by using a Taylor series expansion for the secular equation. We propose to improve this in two ways: first by considering "better" secular equations (of a similar rational nature) and second by considering rational approximations to the secular function, rather than a Taylor series, which is an inappropriate approximation for a rational function. As an added advantage of our different equations, we obtain separate bounds on the even and odd eigenvalues.

To put matters in perspective, we note that these "better" equations are hinted at in [11] without being explicitly stated; they also appear in [9] in an equivalent form that is less suitable for computation. No applications of these equations were considered in either paper. More recently, in [24] and [27], these same equations were derived in a form suitable for applications. In [24], they were applied to the study of spectral evolution as a function of one parameter, whereas in [25] and [27], they were used to compute the smallest eigenvalue. In [18], an equation such as one of

*Received by the editors May 26, 1998; accepted for publication (in revised form) by L. Reichel January 13, 1999; published electronically October 26, 1999.
http://www.siam.org/journals/simax/21-2/33962.html

†Department of Computer Science, SCCM, Stanford University, Stanford, CA 94305-9025 (melman@sccm.stanford.edu). The author is on leave from Ben-Gurion University, Beer-Sheva, Israel.

ours is derived in a different way, without taking into account the spectral structure of the submatrices, thereby obscuring key properties of the equation. It is used there to compute the smallest even eigenvalue.

The idea of a rational approximation for secular equations is not new. In a different context, it was already used in, e.g., [5] and many other references, the most relevant to this work being [21, 25, and 27]. However, such rational approximations seem to be rarely considered beyond the first order. We use techniques from [23] to analyze higher-order rational approximations, which we prove to be better than polynomial ones of the same order. The combination of the equations for the even and odd spectra with higher-order rational approximations to compute bounds on the extreme eigenvalues results in a significant improvement over existing bounds.

The paper is organized as follows: Section 1 contains definitions, a brief overview of the properties of Toeplitz matrices, and basic results for a class of rational functions. In section 2 we develop spectral equations, and in section 3 we consider the approximations which lead, in section 4, to the bounds on the extreme eigenvalues. Finally, we present numerical results in section 5. In sections 2 and 3, we have included a summary of parts of [24] to improve readability and to make the paper as self-contained as possible.

2. Preliminaries. A symmetric matrix $T \in \mathbb{R}^{(n,n)}$ is said to be *Toeplitz* if its elements T_{ij} satisfy $T_{ij} = \mathbf{t}_{|j-i|}$ for some vector $\mathbf{t} = (t_0, \ldots, t_{n-1})^T \in \mathbb{R}^n$. Many early results about such matrices can be found in, e.g., [3, 6, and 9].

Toeplitz matrices are *persymmetric*; i.e., they are symmetric about their southwest-northeast diagonal. For such a matrix T, this is the same as requiring that $JT^TJ = T$, where J is a matrix with ones on its southwest-northeast diagonal and zeros everywhere else (the exchange matrix). It is easy to see that the inverse of a persymmetric matrix is also persymmetric. A matrix that is both symmetric and persymmetric is called *doubly symmetric*.

The identity matrix is denoted by I throughout this paper. As with J, we will not specifically indicate its dimension, which is assumed to be clear from the context.

A *symmetric vector* v is defined as a vector satisfying $Jv = v$ and an *antisymmetric vector* w as one that satisfies $Jw = -w$. If these vectors are eigenvectors, then their associated eigenvalues are called *even* and *odd*, respectively. Drawing on results in [3], it was shown in [6] that, given a real symmetric Toeplitz matrix T of order n, there exists an orthonormal basis for $\mathbb{R}^n$, composed of $n - \lfloor n/2 \rfloor$ symmetric and $\lfloor n/2 \rfloor$ antisymmetric eigenvectors of T, where $\lfloor \alpha \rfloor$ denotes the integral part of α. In the case of simple eigenvalues, this is easy to see from the fact that if $Tu = \lambda u$, then $T(Ju) = \lambda(Ju)$, because $JTJ = T$ and $J^2 = I$. Therefore u and Ju must be proportional, and therefore u must be an eigenvector of J, which means that either $Ju = u$ or $Ju = -u$. Finally, we note that for $\alpha \in \mathbb{R}$, the matrix $(T - \alpha I)$ is symmetric and Toeplitz, whenever T is.

We now state two lemmas, the proofs of which can be found in [23]. They concern the relation between a class of rational functions, which will be considered later, and their approximations.

LEMMA 2.1. *Let $g(\lambda)$ be a strictly positive and twice continuously differentiable real function, defined on some interval $K \subset \mathbb{R}$. With γ a nonzero integer, consider the real function of $\lambda : a(b-\lambda)^\gamma$, where the parameters a and b are such that it interpolates g up to first order at a point $\bar{\lambda} \in K$ with $g'(\bar{\lambda}) \neq 0$. If*

$$\frac{1-\gamma}{\gamma} g'^2(\lambda) + g(\lambda) g''(\lambda)$$

is positive (negative) for all $\lambda \in K$, then for all λ such that $a(b-\lambda)^\gamma \geq 0$, the interpolant $a(b-\lambda)^\gamma$ lies below (above) the function $g(\lambda)$. ∎

LEMMA 2.2. *The function*

$$g(\lambda) = \sum_{j=1}^{m} \alpha_j \, (\beta_j - \lambda)^\gamma,$$

with m a positive and γ a nonzero integer and the α_j's nonnegative, satisfies

$$\frac{1-\gamma}{\gamma} g'^2(\lambda) + g(\lambda) g''(\lambda) \geq 0$$

for all λ such that for all $j : \beta_j - \lambda > 0$. ∎

The following lemma is closely related to Lemma 2.2 for $\lambda = 0$.

LEMMA 2.3. *Define*

$$R_p \triangleq \sum_{j=1}^{m} \frac{\alpha_j}{\beta_j^p},$$

where $\alpha_j, \beta_j > 0$ $(j = 1, \ldots, m)$, and p a positive integer with $p \geq 2$. Then the following inequality holds:

$$R_p^2 \leq R_{p-1} R_{p+1}.$$

The inequality holds as an equality if and only if $\beta_1 = \beta_2 = \cdots = \beta_m$.

Proof. Define the m-dimensional vectors a and b by $a_j = \sqrt{\alpha_j}/\beta_j^{\frac{p-1}{2}}$ and $b_j = \sqrt{\alpha_j}/\beta_j^{\frac{p+1}{2}}$, so that $\|a\|^2 = R_{p-1}$, $\|b\|^2 = R_{p+1}$, and $|a^T b|^2 = R_p^2$. The proof then follows from the Cauchy–Schwarz inequality, which states that $|a^T b|^2 \leq \|a\|^2 \|b\|^2$ with equality if and only if $\frac{a_1}{b_1} = \frac{a_2}{b_2} = \cdots = \frac{a_m}{b_m}$. ∎

3. Spectral equations. In this section we derive various spectral, or secular, equations for the eigenvalues of a real symmetric Toeplitz matrix. Several of these results are not new, even though some were not explicitly stated elsewhere or appear in a form less suitable for computation.

Let us consider the following partition of a symmetric Toeplitz matrix T :

$$T = \begin{pmatrix} t_0 & t^T \\ t & Q \end{pmatrix},$$

where $t = (t_1, \ldots, t_{n-1})^T$ and Q is an $(n-1) \times (n-1)$ symmetric Toeplitz matrix, generated by the vector $(t_0, \ldots, t_{n-2})^T$. Then the following well-known theorem (see, e.g., [8]) holds.

THEOREM 3.1. *The eigenvalues of T that are not shared with those eigenvalues of Q, whose associated eigenspaces are not entirely contained in $\{t\}^\perp$, are given by the solutions of the equation*

$$-t_0 + \lambda + t^T (Q - \lambda I)^{-1} t = 0. \qquad \blacksquare \tag{3.1}$$

We define the function $\phi(\lambda)$ by

$$\phi(\lambda) \triangleq -t_0 + \lambda + t^T (Q - \lambda I)^{-1} t. \tag{3.2}$$

Equation (3.1) is equivalent to

$$-t_0 + \lambda + \sum_{i=1}^{p} \frac{c_i^2}{\omega_i - \lambda} = 0, \tag{3.3}$$

where $\{\omega_i\}_{i=1}^p$ are the p eigenvalues of Q for which the associated eigenspace U_{ω_i} is not entirely contained in the subspace $\{t\}^\perp$, i.e., for which $U_{\omega_i} \not\subset \{t\}^\perp$(see, e.g., [11] for more details). Denote the orthonormal vectors which form a basis for U_{ω_i} by $\{u_j^{(i)}\}_{j=1}^{d_i}$, where d_i is the dimension of U_{ω_i}. Then the scalars c_i^2 are given by $c_i^2 = \sum_{j=1}^{d_i} (t^T u_j^{(i)})^2$. The rational function in (3.1) or (3.3) has p simple poles, dividing the real axis into $p+1$ intervals on each of which it is monotonically increasing from $-\infty$ to $+\infty$. The $p+1$ solutions $\{\lambda_j\}_{j=1}^{p+1}$ of (3.3) therefore satisfy

$$\lambda_1 < \omega_1 < \lambda_2 < \omega_2 < \cdots < \omega_p < \lambda_{p+1};$$

i.e., the eigenvalues ω_i strictly interlace the eigenvalues λ_j, which is known as Cauchy's interlacing theorem. These results are well known and we refer to, e.g., [8] and [11].

A positive-definite matrix T will therefore certainly have an eigenvalue in the interval $(0, \omega_1)$. An upper bound can then be found by approximating the function in (3.1) at $\lambda = 0$ in such a way that the approximation always lies below that function and by subsequently computing the root of this approximation. This is the approach used in [11], where the approximations are the Taylor polynomials. However, such polynomials are inadequate for rational functions, and we shall return to this matter after deriving additional spectral equations.

It would appear that our previous partition of T is inappropriate, given the persymmetry of Toeplitz matrices. We therefore consider the following, more natural, partition for a matrix that is both symmetric and persymmetric:

$$T = \begin{pmatrix} t_0 & \tilde{t}^T & t_{n-1} \\ \tilde{t} & G & J\tilde{t} \\ t_{n-1} & (J\tilde{t})^T & t_0 \end{pmatrix},$$

where $\tilde{t} = (t_1, \ldots, t_{n-2})^T$ and G is an $(n-2) \times (n-2)$ symmetric Toeplitz matrix, generated by the vector $(t_0, \ldots, t_{n-3})^T$. This is the partition used in [24] and [27]. It is also used in [11, Theorem 4], but without exploiting it to compute or bound eigenvalues. In what follows, we denote the even and odd eigenvalues of T by λ_i^e and λ_i^o and the even and odd eigenvalues of G by μ_i and ν_i, respectively. We then have the following theorem, which yields two equations: one for even and one for odd eigenvalues of T. It can be found in [24] and, in a slightly different form, in [27].

THEOREM 3.2. *The even eigenvalues of T that are not shared with those even eigenvalues of G, whose associated eigenspaces are not entirely contained in $\{\tilde{t}\}^\perp$, are the solutions of*

$$-t_0 - t_{n-1} + \lambda + \tilde{t}^T (G - \lambda I)^{-1} (\tilde{t} + J\tilde{t}) = 0, \tag{3.4}$$

whereas the odd eigenvalues of T that are not shared with those odd eigenvalues of G, whose associated eigenspaces are not entirely contained in $\{\tilde{t}\}^\perp$, are the solutions of

$$-t_0 + t_{n-1} + \lambda + \tilde{t}^T (G - \lambda I)^{-1} (\tilde{t} - J\tilde{t}) = 0. \tag{3.5}$$

Proof. The proof is based on finding the conditions under which $(T - \lambda I)x = 0$ has a nontrivial solution for x. These conditions take the form of a factorable equation, which then leads directly to (3.4) and (3.5). For more details, we refer to [24]. ∎

To gain a better understanding of (3.4) and (3.5), let us assume for a moment that all eigenvalues of G are simple (the general case does not differ substantially) and denote an orthonormal basis of $\mathbb{R}^{n-2}$, composed of orthonormal eigenvalues of G, by $\{v_i\}_{i=1}^{r} \bigcup \{w_j\}_{j=1}^{s}$, where $r = n-2-\lfloor (n-2)/2 \rfloor$, $s = \lfloor (n-2)/2 \rfloor$, (v_i, μ_i) are (symmetric eigenvector, even eigenvalue) pairs, and (w_j, ν_j) are (antisymmetric eigenvector, odd eigenvalue) pairs. With $\tilde{t} = \sum_{i=1}^{r} a_i v_i + \sum_{j=1}^{s} b_j w_j$, this means that

$$J\tilde{t} = \sum_{i=1}^{r} a_i (Jv_i) + \sum_{j=1}^{s} b_j (Jw_j) = \sum_{i=1}^{r} a_i v_i - \sum_{j=1}^{s} b_j w_j.$$

Once more, exploiting the orthonormality of the eigenvectors yields

$$\tilde{t}^T (G - \lambda I)^{-1} (\tilde{t} + J\tilde{t}) = 2 \sum_{i=1}^{r} \frac{a_i^2}{\mu_i - \lambda}.$$

Analogously, we obtain

$$\tilde{t}^T (G - \lambda I)^{-1} (\tilde{t} - J\tilde{t}) = 2 \sum_{j=1}^{s} \frac{b_j^2}{\nu_j - \lambda}.$$

Equations (3.4) and (3.5) now become

$$-t_0 - t_{n-1} + \lambda + 2 \sum_{i=1}^{r} \frac{a_i^2}{\mu_i - \lambda} = 0, \tag{3.6}$$

$$-t_0 + t_{n-1} + \lambda + 2 \sum_{j=1}^{s} \frac{b_j^2}{\nu_j - \lambda} = 0, \tag{3.7}$$

which shows that the rational functions in each of (3.4) and (3.5) are of the same form as the function in (3.1). It is also clear that T will certainly have an even eigenvalue on $(0, \mu_1)$ and an odd one on $(0, \nu_1)$. These equations were also hinted at in [11] without, however, being derived or stated in an explicit way. The meaning of Theorem 3.2 is therefore that those even and odd eigenvalues of G, whose associated eigenspaces are not completely contained in $\{\tilde{t}\}^{\perp}$, interlace, respectively, the even and odd eigenvalues of T that are not shared with those eigenvalues of G. This result was obtained in [9] in a different way, along with equivalent forms of (3.6) and (3.7). However, the use of determinants there makes them less suitable for applications.

Finally, because of the orthonormality of the eigenvectors, (3.4) and (3.5) can be written in a more symmetric way, as shown in the following two equations, which at the same time define the functions $\phi^e(\lambda)$ and $\phi^o(\lambda)$:

$$\phi^e(\lambda) \triangleq -t_0 - t_{n-1} + \lambda + \frac{1}{2} (\tilde{t} + J\tilde{t})^T (G - \lambda I)^{-1} (\tilde{t} + J\tilde{t}) = 0, \tag{3.8}$$

$$\phi^o(\lambda) \triangleq -t_0 + t_{n-1} + \lambda + \frac{1}{2} (\tilde{t} - J\tilde{t})^T (G - \lambda I)^{-1} (\tilde{t} - J\tilde{t}) = 0. \tag{3.9}$$

We note that (3.8) was also obtained in [18], where it was used to compute the smallest eigenvalue which was known in advance to be even. However, the derivation of the equation is quite different, concentrating exclusively on the smallest eigenvalue and

disregarding the spectral structure of the submatrices of T, which obscures important properties of that equation.

To evaluate the functions $\phi(\lambda)$, $\phi^e(\lambda)$, and $\phi^o(\lambda)$, and their derivatives, as we will need to do later on, we need to compute expressions of the form $s^T S^{-k} s$ for a positive integer k, where S is the real symmetric Toeplitz matrix defined by $(s_0, s_1, \ldots, s_{n-1})$ and where $s = (s_1, \ldots, s_n)$. In this work, we will use the *Levinson–Durbin algorithm* (LDA). The original references for this algorithm are [13] and [20], but an excellent overview of this and other Toeplitz-related algorithms can be found in [15]. Let us start with $k = 1$. In this case we have to solve $Sw = -s$, where the minus sign in the right-hand side is by convention. This system of linear equations is called the Yule–Walker (YW) system, and the LDA solves this problem recursively in $2n^2 + \mathcal{O}(n)$ flops, where we define one flop as in [15], namely, a multiplication/division or an addition/subtraction. Because of the persymmetry of S, once the YW equations are solved, the solution to $Sy = -Js$ is given by $y = Jw$.

After solving the YW system, we have obtained $w = -S^{-1}s$. The expression $s^T S^{-2} s$ can then be evaluated as $\|S^{-1}s\|_2^2$ in $\mathcal{O}(n)$ flops. To compute higher-order derivatives, we use a decomposition of S, supplied by the LDA itself in the process of solving the YW system. Denoting by $w^{(\ell)}$ the solution to the ℓth-dimensional YW subsystem, obtained in the course of the LDA algorithm, this decomposition is given by $U^T S U = D$, where D is a diagonal matrix and U is the upper triangular $n \times n$ matrix, whose ℓth column is given by $(Jw^{(\ell-1)}, 1)^T$; i.e., all its diagonal elements are equal to one. This result is due to [7]. We calculate $s^T S^{-3} s$ as follows:

$$s^T S^{-3} s = (S^{-1}s)^T S^{-1}(S^{-1}s) = w^T U D^{-1} U^T w = (U^T w)^T D^{-1}(U^T w).$$

This computation costs $n^2 + \mathcal{O}(n)$ flops. To evaluate $s^T S^{-4} s$, we compute first $S^{-2}s$ as $UD^{-1}U^T w$, which requires, once again, $n^2 + \mathcal{O}(n)$ flops, and then compute $\|S^{-2}s\|_2^2$ with an additional $\mathcal{O}(n)$ flops. Roughly speaking, we can say that, for each increase of k by one, we need to execute an additional n^2 flops. We mention here that $4n^2 + \mathcal{O}(n)$ flops are required to compute the second derivative of both the even and odd spectral functions. This doubling of the number of flops does not take place for the second and third derivatives.

Of course, other, potentially faster, algorithms could be substituted for the classical LDA, such as the superfast Toeplitz solvers (see, e.g., [1] and [2]) or a version of the LDA which exploits the symmetry of the solution (see, e.g., [10, 12]). However, this influences only the complexity of computing our bounds and not the bounds themselves, which are the focus of this work.

4. Approximations. As we mentioned before, our bounds will be obtained by the roots of approximations to the secular equations. In the case of the smallest eigenvalue, those approximations will be shown to be dominated by the spectral function, so that their root will provide an upper bound on the smallest eigenvalue. Bounds for the largest eigenvalue will be based on the bounds for the smallest eigenvalue of a different matrix. In the derivation of the bounds, we will assume that the matrices are positive definite, even though a slight modification suffices to extend our results to general symmetric matrices. All this will be explained in section 5.

We shall now construct approximations to our spectral equations. These will be of two types: rational and polynomial, each of which will be of three kinds: first, second, and third order.

Throughout this section we will consider approximations, obtained by interpola-

tion at $\lambda = 0$, for a function g of the form

$$g(\lambda) = \sum_{j=1}^{m} \frac{\alpha_j}{\beta_j - \lambda}, \tag{4.1}$$

where $\alpha_j > 0$, $0 < \beta_1 < \beta_2 < \cdots < \beta_m$, and $m \geq 2$. Functions of this form occur in all the equations considered in this paper. We note that g has simple poles at the β_j's and is a positive, monotonically increasing convex function (or $g, g', g'' > 0$) on the interval $(-\infty, \beta_1)$. Our results are applicable for interpolation at a point $\bar{\lambda} \in (-\infty, \beta_1)$, different from zero, simply by translating the origin to that point.

We will denote the approximations of first, second, and third order by ρ_1, ρ_2, ρ_3 for the rational ones and by τ_1, τ_2, τ_3 for the polynomial ones, respectively. The polynomial approximations are nothing but the Taylor polynomials of degree 1, 2, and 3. We now define these approximations but warn that some of the parameters are defined "locally"; i.e., the same letter may have different meanings in different contexts, when no confusion is possible.

(1) *First-order rational.* A function $\rho_1(\lambda) \stackrel{\triangle}{=} a/(b-\lambda)$ such that $\rho_1(0) = g(0)$ and $\rho_1'(0) = g'(0)$. The coefficients a and b are easily computed to be $a = g^2(0)/g'(0)$ and $b = g(0)/g'(0)$. Since

$$b = \frac{g(0)}{g'(0)} = \frac{\sum_{j=1}^{m}(\alpha_j/\beta_j)}{\sum_{j=1}^{m}(\alpha_j/\beta_j^2)} = \sum_{j=1}^{m} \left(\frac{\alpha_j/\beta_j^2}{\sum_{j=1}^{m}(\alpha_j/\beta_j^2)} \right) \beta_j, \tag{4.2}$$

which is a strictly convex combination of the β_j's, this means that $\beta_1 < b < \beta_m$. We also have $a > 0$, so that ρ_1 is a positive, monotonically increasing convex function on $(-\infty, \beta_1)$.

(2) *Second-order rational.* A function $\rho_2(\lambda) \stackrel{\triangle}{=} a + b/(c - \lambda)$ such that $\rho_2(0) = g(0)$, $\rho_2'(0) = g'(0)$, and $\rho_2''(0) = g''(0)$. The coefficients a, b, and c are then given by $a = g(0) - 2g'^2(0)/g''(0)$, $b = 4g'^3(0)/g''^2(0)$, and $c = 2g'(0)/g''(0)$. It is clear immediately that $b > 0$. From Lemma 2.2 with $\gamma = -1$, we have that $a > 0$ as well. Since

$$c = \frac{2g'(0)}{g''(0)} = \frac{\sum_{j=1}^{m}(\alpha_j/\beta_j^2)}{\sum_{j=1}^{m}(\alpha_j/\beta_j^3)} = \sum_{j=1}^{m} \left(\frac{\alpha_j/\beta_j^3}{\sum_{j=1}^{m}(\alpha_j/\beta_j^3)} \right) \beta_j, \tag{4.3}$$

we have once again obtained a strictly convex combination of the β_j's, so that $\beta_1 < c < \beta_m$. The approximation ρ_2 is therefore positive, monotonically increasing, and convex on $(-\infty, \beta_1)$. Since $\beta_1 < \beta_2 < \cdots < \beta_m$, we also have (in the notation of Lemma 2.3) that $R_2^2 < R_1 R_3$, or $R_2/R_3 < R_1/R_2$, which is the same as saying that $2g'(0)/g''(0) < g(0)/g'(0)$. This means that the pole of ρ_2 lies closer to β_1 than the pole of ρ_1.

(3) *Third-order rational.* A function $\rho_3(\lambda) \stackrel{\triangle}{=} a/(b - \lambda) + c/(d - \lambda)$ such that $\rho_3(0) = g(0)$, $\rho_3'(0) = g'(0)$, $\rho_3''(0) = g''(0)$, and $\rho_3'''(0) = g'''(0)$. For convenience, we temporarily leave out the argument of g and its derivatives, i.e., $g \equiv g(0)$, and we denote $v = 1/b$ and $w = 1/d$. To compute the coefficients of ρ_3, we then have to find a solution for the following system of equations in a, c, v, and w:

$$av + cw = g, \tag{4.4}$$

$$av^2 + cw^2 = g', \tag{4.5}$$

$$2av^3 + 2cw^3 = g'', \tag{4.6}$$

$$6av^4 + 6cw^4 = g''', \tag{4.7}$$

establish that $v, w \neq 0$, and set $b = 1/v$, $d = 1/w$.

We first show that $acvw \neq 0$. To see this, assume that $av = 0$. Since $g > 0$, this would mean that $cw \neq 0$, so that (4.4)–(4.6) imply $g'/g = g''/2g'$ or $R_2^2 = R_1 R_3$, in the notation of Lemma 2.3. This same lemma then shows that this is possible only if all β_j's are equal, which contradicts our assumption that $\beta_1 < \beta_2 < \cdots < \beta_m$. Therefore, $av \neq 0$. Similarly, we find that $cw \neq 0$, so that a, c, v, and w are all nonzero.

Furthermore, neither $a/(b-\lambda)$ nor $c/(d-\lambda)$ can be equal to any of the terms $\alpha_j/(\beta_j - \lambda)$ appearing in $g(\lambda)$. If this were the case, e.g., for $a/(b-\lambda)$, then the effect on (4.4)–(4.7) would be to set $a = 0$ and to leave out one term in the expressions for g, g', g'', g'''. An argument analogous to the one in the previous paragraph then leads, once again, to a contradiction.

We also claim that $v \neq w$. To show this, assume that $v = w$. Equation (4.4) then yields $a + c \neq 0$ and (4.4)–(4.6) once again produce $R_2^2 = R_1 R_3$, which, as we already saw, is impossible.

From Cramer's rule, applied to (4.4)–(4.7), we have

$$a = \frac{\begin{vmatrix} g & w \\ g' & w^2 \end{vmatrix}}{\begin{vmatrix} v & w \\ v^2 & w^2 \end{vmatrix}} = \frac{\begin{vmatrix} g' & w^2 \\ g'' & 2w^3 \end{vmatrix}}{\begin{vmatrix} v^2 & w^2 \\ 2v^3 & 2w^3 \end{vmatrix}} = \frac{\begin{vmatrix} g'' & 2w^3 \\ g''' & 6w^4 \end{vmatrix}}{\begin{vmatrix} 2v^3 & 2w^3 \\ 6v^4 & 6w^4 \end{vmatrix}}.$$

Since $v, w \neq 0$ and $v \neq w$, these expressions are well defined and are equivalent to

$$a = \frac{gw - g'}{(w-v)v} = \frac{2g'w - g''}{2(w-v)v^2} = \frac{3g''w - g'''}{6(w-v)v^3}. \tag{4.8}$$

Expression (4.8) then leads to

$$v = \frac{1}{2}\left(\frac{2g'w - g''}{gw - g'}\right), \tag{4.9}$$

$$v = \frac{1}{3}\left(\frac{3g''w - g'''}{2g'w - g''}\right). \tag{4.10}$$

Analogously, by considering c instead of a, we obtain

$$c = \frac{g' - gv}{(w-v)w} = \frac{g'' - 2g'v}{2(w-v)w^2} = \frac{g''' - 3g''v}{6(w-v)w^3}, \tag{4.11}$$

and

$$w = \frac{1}{2}\left(\frac{2g'v - g''}{gv - g'}\right), \tag{4.12}$$

$$w = \frac{1}{3}\left(\frac{3g''v - g'''}{2g'v - g''}\right). \tag{4.13}$$

Because $a, c \neq 0$, (4.8) and (4.11) imply that neither v nor w can be equal to g'/g, $g''/2g'$, or $g'''/3g''$, so that the above identities are well defined.

Equations (4.9) and (4.10) then yield, after some algebra,

$$v + w \stackrel{\triangle}{=} \eta = \frac{gg''' - 3g'g''}{3gg'' - 6g'^2}, \tag{4.14}$$

$$vw \stackrel{\triangle}{=} \zeta = \frac{g'g''' - \frac{3}{2}g''^2}{3gg'' - 6g'^2}. \tag{4.15}$$

From Lemma 2.3, we have

$$3gg'' - 6g'^2 = 6\left(\frac{1}{2}gg'' - g'^2\right) = 6\left(R_1R_3 - R_2^2\right) > 0,$$

so that the denominators in (4.14) and (4.15) are well defined. This means that v and w are the solutions for x of the quadratic equation $x^2 - \eta x + \zeta = 0$. Since $v, w \neq 0$, b and d are then determined by $b = 1/v$ and $d = 1/w$. Our equations are symmetric in v and w, so let us choose v as the larger of the two. This makes b the smaller of the two poles of ρ_3.

Since no two β_j's are the same, Lemma 2.3 with $p = 2, 3$ yields $g'''/3g'' > g''/2g' > g'/g$, so that $v + w > 0$ and $vw > 0$, which in turn means that $v, w > 0$. As a direct consequence of (4.9), (4.10), (4.12), and (4.13), we then have that $v > g'''/3g''$ and $w < g'/g$ or $b < 3g''/g'''$ and $d > g/g'$, since we designated b as the smaller of the two poles of ρ_3. Using these inequalities in (4.8) and (4.11) then shows that $a, c > 0$.

Finally, we claim that $b > \beta_1$ and $d < \beta_m$, so that both poles of ρ_3 lie in the interval (β_1, β_m). We first note that, because of (4.2) and (4.3), we also have that $b < 3g''/g''' < 2g'/g'' < \beta_m$ and $d > g'/g > \beta_1$. Let us prove that $b > \beta_1$. From (4.12) and (4.13), we have

$$\frac{1}{2}\left(\frac{\sum_{j=1}^m \frac{2\alpha_j}{\beta_j^2}v - \sum_{j=1}^m \frac{2\alpha_j}{\beta_j^3}}{\sum_{j=1}^m \frac{\alpha_j}{\beta_j}v - \sum_{j=1}^m \frac{\alpha_j}{\beta_j^2}}\right) = \frac{1}{3}\left(\frac{\sum_{j=1}^m \frac{6\alpha_j}{\beta_j^3}v - \sum_{j=1}^m \frac{6\alpha_j}{\beta_j^4}}{\sum_{j=1}^m \frac{2\alpha_j}{\beta_j^2}v - \sum_{j=1}^m \frac{2\alpha_j}{\beta_j^3}}\right).$$

This can be rewritten as follows:

$$\frac{\sum_{j=1}^m \frac{\alpha_j(v-\beta_j^{-1})}{\beta_j^2}}{\sum_{j=1}^m \frac{\alpha_j(v-\beta_j^{-1})}{\beta_j}} = \frac{\sum_{j=1}^m \frac{\alpha_j(v-\beta_j^{-1})}{\beta_j^3}}{\sum_{j=1}^m \frac{\alpha_j(v-\beta_j^{-1})}{\beta_j^2}}. \tag{4.16}$$

Assume now that $v \geq \beta_1^{-1}$ and therefore also $v \geq \beta_j^{-1}$ for $j = 1, 2, \ldots, m$. We can then define two m-dimensional vectors ζ and η by

$$\zeta_j = \frac{\left(\alpha_j\left(v - \beta_j^{-1}\right)\right)^{\frac{1}{2}}}{\beta_j^{\frac{1}{2}}} \quad \text{and} \quad \eta_j = \frac{\left(\alpha_j\left(v - \beta_j^{-1}\right)\right)^{\frac{1}{2}}}{\beta_j^{\frac{3}{2}}},$$

so that (4.16) becomes $|\zeta^T\eta|^2 = \|\zeta\|^2\|\eta\|^2$, which by the Cauchy–Schwarz inequality cannot hold when $\beta_1 < \beta_2 < \cdots < \beta_m$. The conclusion therefore is that $v < \beta_1^{-1}$, or $b > \beta_1$. Using (4.9) and (4.10), we find entirely analogously that $d < \beta_m$.

Putting all of the above together, one obtains that ρ_3 is a positive, monotonically increasing convex function on the interval $(-\infty, \beta_1)$, with its poles on the interval (β_1, β_m). Since $b < 3g''/g''' < 2g'/g''$, its smallest pole lies between β_1 and the pole of ρ_2.

(4) *First-order polynomial.* A function $\tau_1(\lambda) \triangleq a + b\lambda$ such that $\tau_1(0) = g(0)$ and $\tau_1'(0) = g'(0)$. The coefficients a and b are easily computed to be $a = g(0)$ and $b = g'(0)$, which are both positive. The function τ_1 is therefore a linear and increasing function everywhere.

(5) *Second-order polynomial.* A function $\tau_2(\lambda) \triangleq a + b\lambda + c\lambda^2$ such that $\tau_2(0) = g(0)$, $\tau_2'(0) = g'(0)$, and $\tau_2''(0) = g''(0)$. The coefficients a, b, and c are given by $a = g(0)$, $b = g'(0)$, and $c = g''(0)/2$, and they are all positive. The function τ_2 is therefore an increasing and convex function for $\lambda \geq 0$.

(6) *Third-order polynomial.* A function $\tau_3(\lambda) \triangleq a + b\lambda + c\lambda^2 + d\lambda^3$ such that $\tau_2(0) = g(0)$, $\tau_2'(0) = g'(0)$, $\tau_2''(0) = g''(0)$, and $\tau_2'''(0) = g'''(0)$. The coefficients a, b, c, and d are $a = g(0)$, $b = g'(0)$, $c = g''(0)/2$, and $d = g'''(0)/6$, and they are all positive. The function τ_3 is therefore an increasing and convex function for $\lambda \geq 0$.

THEOREM 4.1. *The following inequalities hold on the interval* $[0, \beta_1)$:

$$\tau_i(\lambda) \leq \rho_i(\lambda) \leq g(\lambda) \qquad (i = 1, 2, 3), \tag{4.17}$$

$$\rho_1(\lambda) \leq \rho_2(\lambda) \leq \rho_3(\lambda) \leq g(\lambda), \tag{4.18}$$

$$\tau_1(\lambda) \leq \tau_2(\lambda) \leq \tau_3(\lambda) \leq g(\lambda). \tag{4.19}$$

Proof. We start with the remark that some inequalities will be proved on the larger interval $(-\infty, \beta_1)$. Let us begin with inequalities (4.17). The function $\rho_1(\lambda)$ is a first-order rational approximation to $g(\lambda)$ at $\lambda = 0$. It is then immediate from Lemmas 2.1 and 2.2 with $\gamma = -1$ that $\rho_1(\lambda) \leq g(\lambda)$ on $(-\infty, \beta_1)$. The linear approximation $\tau_1(\lambda)$ to $g(\lambda)$ at $\lambda = 0$ is also the linear approximation to $\rho_1(\lambda)$ at that same point. Since $\rho_1(\lambda)$ is a convex function on $(-\infty, \beta_1)$, $\tau_1(\lambda)$ must lie below it on the same interval. This concludes the proof for $i = 1$.

For $i = 2$, we have that $\rho_2(\lambda) \equiv a + b/(c - \lambda)$ approximates $g(\lambda)$ up to second order. This is the same as saying that $b/(c-\lambda)^2$ approximates $g'(\lambda)$ up to first order. Lemmas 2.1 and 2.2 with $\gamma = -2$ then yield that $b/(c - \lambda)^2 \leq g'(\lambda)$, from which it follows, with $\lambda \in [0, \beta_1)$, that

$$\int_0^\lambda \frac{b}{(c-\sigma)^2} \mathrm{d}\sigma \leq \int_0^\lambda g'(\sigma) \mathrm{d}\sigma.$$

Integrating, we obtain $b/(c - \lambda) - b/c \leq g(\lambda) - g(0)$. Adding and subtracting a in the left-hand side and using the function value interpolation condition then yields $\rho_2(\lambda) \leq g(\lambda)$ on $[0, \beta_1)$. Furthermore, $\tau_2(\lambda)$ interpolates both $g(\lambda)$ and $\rho_2(\lambda)$ at $\lambda = 0$ up to second order. This means that $2c\lambda + b$ interpolates $\rho_2'(\lambda)$ up to first order at $\lambda = 0$. Because $\rho_2''' > 0$ on $(-\infty, \beta_1)$, ρ_2' is convex on that interval. We then have that $2c\lambda + b \leq \rho_2'(\lambda)$ on that same interval and we integrate back, as we did before, to obtain the desired inequality for $\lambda \in [0, \beta_1)$. This concludes the proof for $i = 2$.

For $i = 3$, we consider the function

$$h(\lambda) \triangleq g(\lambda) - \rho_3(\lambda) = g(\lambda) - \frac{a}{b - \lambda} - \frac{c}{d - \lambda}.$$

We recall that $b, d \in (\beta_1, \beta_m)$, that $b < d$, and that neither $a/(b-\lambda)$ nor $c/(d-\lambda)$ can be equal to any of the terms $\alpha_j/(\beta_j - \lambda)$ of $g(\lambda)$. We then have the following cases.

Case 1. Neither b nor d are equal to any of the poles β_j of $g(\lambda)$. This means that the function $h(\lambda)$ must have a total of $m+1$ roots. There are four roots at $\lambda = 0$ and

an additional $m-3$ roots on the $m-3$ intervals of the form (β_k, β_{k+1}) that do not contain either b or d. This argument also shows that it would be impossible for both b and d to lie together in any such interval, as this would add at least $m-2$ roots, bringing the total number of roots to at least $m+2$.

Case 2. Either b or d (but not both) is equal to one of the poles of $g(\lambda)$. This means that the function $h(\lambda)$ must have a total of m roots. There are four roots at $\lambda = 0$ and an additional $m-4$ roots on the $m-4$ intervals of the form (β_k, β_{k+1}) that do not contain either b or d in their closure. This argument also shows that it would be impossible for both b and d to lie together in the closure of any such interval, as this would add at least $m-3$ roots, bringing the total number of roots to at least $m+1$.

Case 3. Both b and d are equal to (different) poles of $g(\lambda)$. This means that the function $h(\lambda)$ must have a total of $m-1$ roots. There are four roots at $\lambda = 0$ and an additional $m-5$ roots on the $m-5$ intervals of the form (β_k, β_{k+1}) that do not have either b or d as one of their endpoints. This argument also shows that it would be impossible for any such interval to have both b and d as their endpoints, as this would add at least $m-4$ roots, bringing the total number of roots to at least m.

In each of the above three cases, $h(\lambda)$ cannot have roots on the interval $[0, \beta_1)$ other than at $\lambda = 0$ and must therefore have the same sign throughout that interval. Since $h(\lambda) \to +\infty$ as $\lambda \to \beta_1^-$, $h(\lambda)$ must be positive on that interval, which means that $\rho_3(\lambda) \leq g(\lambda)$ on $[0, \beta_1)$.

Turning now to $\tau_3(\lambda)$, we have, similar to the case $i = 2$, that $\tau_3(\lambda)$ also interpolates $\rho_3(\lambda)$ up to third order at $\lambda = 0$, which is equivalent to $2c + 6d\lambda$ interpolating ρ_3'' up to first order at $\lambda = 0$. Because $\rho_3^{(iv)} > 0$ on $(-\infty, \beta_1)$, $\rho_3''(\lambda)$ is convex on that same interval, which means that on that interval $2c + 6d\lambda \leq \rho_3''(\lambda)$. Integrating back twice, we obtain once again our inequality for $\lambda \in [0, \beta_1)$. This concludes the proof for the inequalities in (4.17).

Let us now consider inequalities (4.18), starting with the inequality between ρ_1 and ρ_2. We first note that ρ_1 approximates ρ_2 up to first order. We also have

$$-2\rho_2'^2(\lambda) + \rho_2(\lambda)\rho_2''(\lambda) = -2\left(\frac{b}{c-\lambda}\right)'^2 + \left(\frac{b}{c-\lambda}\right)\left(\frac{b}{c-\lambda}\right)'' + a\left(\frac{b}{c-\lambda}\right)''. \tag{4.20}$$

Taking into account that a and b are positive and that $c > \beta_1 > 0$, we easily see that the last term in the right-hand side of (4.20) is positive on $(-\infty, \beta_1)$, whereas the sum of the first two terms is positive on that same interval because of Lemma 2.2 with $\gamma = -1$. Lemma 2.1, with $\gamma = -1$, then shows that $\rho_1(\lambda) \leq \rho_2(\lambda)$ on $(-\infty, \beta_1)$.

For the inequality between ρ_2 and ρ_3, it suffices to note that ρ_2 approximates ρ_3 up to second order and that ρ_3 is a function of the same form and with the same general properties as g. An argument analogous to the one used to prove that $\rho_2(\lambda) \leq g(\lambda)$ then also yields that $\rho_2(\lambda) \leq \rho_3(\lambda)$ for $\lambda \in [0, \beta_1)$, and this concludes the proof for inequality (4.18).

Inequalities (4.19) all follow by an analogous argument to the one used to prove the inequalities between $\tau_i(\lambda)$ $(i = 1, 2, 3)$ and $g(\lambda)$, after observing that τ_1 is the first-order approximation to τ_2, which itself is the second-order approximation to τ_3. This concludes the proof of the theorem. □

5. Bounds. We now finally derive our bounds on the extreme eigenvalues and we start with the smallest eigenvalue. We first consider matrices that are positive definite. Upper bounds are then obtained by computing the roots of the various

approximations at $\lambda = 0$ to the secular equations $\phi(\lambda)$, $\phi^e(\lambda)$, and $\phi^o(\lambda)$, which were defined in (3.2), (3.8), and (3.9). As we have shown before, all these equations are of the form

$$-\alpha + \lambda + g(\lambda) = 0, \tag{5.1}$$

where $g(\lambda)$ is of the form defined in (4.1) and $\alpha \in \mathbb{R}$. Their approximations are obtained by replacing $g(\lambda)$ with the various approximations that were described in the previous section.

We define the following:

$$f(\lambda) \triangleq t^T(Q - \lambda I)^{-1}t,$$

$$f^e(\lambda) \triangleq \frac{1}{2}(\tilde{t} + J\tilde{t})^T(G - \lambda I)^{-1}(\tilde{t} + J\tilde{t}),$$

$$f^o(\lambda) \triangleq \frac{1}{2}(\tilde{t} - J\tilde{t})^T(G - \lambda I)^{-1}(\tilde{t} - J\tilde{t}),$$

where all quantities are as previously defined. Once again, all these functions are of the same form as the function g, defined in (4.1).

The bounds obtained by replacing $f(\lambda)$ in $\phi(\lambda)$ with ρ_1, ρ_2, and ρ_3 will be denoted r_1, r_2, and r_3, respectively. Those bounds obtained by replacing $f(\lambda)$ with τ_1, τ_2, and τ_3 will be denoted p_1, p_2, and p_3, respectively. As an example, this means that r_1 is the root of the equation

$$-t_0 + \lambda + \frac{g^2(0)/g'(0)}{g(0)/g'(0) - \lambda} = 0,$$

which is obtained by solving a quadratic equation. To compute r_3 and p_3, a cubic equation needs to be solved, which can either be accomplished in closed form or by an iterative method such as Newton's method (at negligible cost, compared with the computation of $g(0)$).

This notation is now applied to $\phi^e(\lambda)$ and $\phi^o(\lambda)$ as well; i.e., the bounds on the smallest even eigenvalue, obtained by approximating $f^e(\lambda)$ in $\phi^e(\lambda)$, will be denoted by r_1^e, r_2^e, and r_3^e for the rational approximations and p_1^e, p_2^e, and p_3^e for the polynomial approximations. For the odd eigenvalues, approximating $f^o(\lambda)$ in $\phi^o(\lambda)$ yields the bounds r_1^o, r_2^o, r_3^o and p_1^o, p_2^o, p_3^o for the rational and polynomial approximations, respectively.

If one uses the LDA for the computation of the spectral equations and their derivatives as was discussed in section 3, then the computational cost for first-, second-, and third-order bounds is $2n^2 + \mathcal{O}(n)$, $3n^2 + \mathcal{O}(n)$, and $4n^2 + \mathcal{O}(n)$ flops, respectively, and is $\mathcal{O}(n^2)$ in storage space for the higher-order bounds (higher than one). As was mentioned before, the computational cost for the second-order bounds, when using the even and odd spectral functions, is doubled. To avoid this, one could start by computing the first-order bounds and then compute the second-order bound only for that equation which yielded the lower first-order bound. Of course, there is no guarantee that this would indeed yield the best second-order bound. However, even though we did not include it in the section on numerical experiments, we did verify that such a procedure would not significantly alter our results. A lower computational cost might be achieved by using methods other than the LDA, as was mentioned at the end of section 3.

One of the advantages of the rational approximations is that, contrary to polynomial approximations, they always generate bounds that are guaranteed not to exceed the largest pole of f, f^e, or f^o, whichever applies, as is obvious from their properties, regardless of how badly behaved the matrix is. In addition to providing separate bounds on the even and odd eigenvalues, the approximations to the functions $\phi^e(\lambda)$ and $\phi^o(\lambda)$ should be more accurate than those for the function $\phi(\lambda)$, since now only roughly half of the terms appear in the function to be interpolated. There is also the additional benefit that both the smallest and the largest roots are now farther removed from the nearest singularity in the equation so that once again an improved approximation can be expected. All this is borne out by our numerical experiments.

Better upper bounds can be obtained if a positive lower bound is known on the smallest eigenvalue. The only difference in that case is that the approximations are performed at that lower bound, rather than at $\lambda = 0$. As was mentioned before, all our results can be applied in this case to the same spectral equations but with the origin translated to the lower bound.

Before presenting numerical comparisons in the next section, we will first establish a theoretical result. We denote the smallest even and odd eigenvalues of T by λ^e_{min} and λ^o_{min}, respectively, and its smallest eigenvalue by λ_{min}. We note that $\lambda_{min} = \min\{\lambda^e_{min}, \lambda^o_{min}\}$. The theoretical comparisons between the various bounds are then given in the following theorem.

THEOREM 5.1. *The upper bounds on the smallest eigenvalue of T satisfy*

$$
\begin{aligned}
&\lambda_{min} \le r_i \le p_i \qquad (i = 1, 2, 3),\\
&\lambda^e_{min} \le r^e_i \le p^e_i \qquad (i = 1, 2, 3),\\
&\lambda^o_{min} \le r^o_i \le p^o_i \qquad (i = 1, 2, 3),\\
&\lambda_{min} \le r_3 \le r_2 \le r_1,\\
&\lambda^e_{min} \le r^e_3 \le r^e_2 \le r^e_1,\\
&\lambda^o_{min} \le r^o_3 \le r^o_2 \le r^o_1,\\
&\lambda_{min} \le p_3 \le p_2 \le p_1,\\
&\lambda^e_{min} \le p^e_3 \le p^e_2 \le p^e_1,\\
&\lambda^o_{min} \le p^o_3 \le p^o_2 \le p^o_1.
\end{aligned}
$$

Proof. The proof follows immediately from the properties of the approximations that define the bounds, which were proved in Theorem 4.1. ∎

This theorem shows that the bounds, obtained by rational approximations, are always superior to those obtained by polynomial approximations, which should not be surprising, as the functions they approximate are themselves rational. It also confirms the intuitive result that, as the order of the approximations increases, then so does the accuracy of the bounds. This also means that the bounds obtained in [11] (as far as we were able to verify, the best currently available), which are all based on polynomial approximations and correspond to our p_1, p_2, and p_3, are inferior to those produced by rational approximations.

Let us now consider lower bounds on the largest eigenvalue. The largest eigenvalue of T can be bounded from below, given an upper bound δ on it. This can be accomplished by translating the origin in the spectral equations to δ, replacing the resulting new variable by its opposite, and multiplying the equation by (-1), thus obtaining the exact same type of spectral equation for the matrix $(\delta I - T)$, which is positive definite. Computing an upper bound on the smallest eigenvalue of this new matrix then leads

to the desired bound on the largest eigenvalue, since $\lambda_{min}(\delta I - T) = \delta - \lambda_{max}(T)$. One possible value for δ is the Frobenius norm of T, defined as (see [15])

$$\|T\|_F = \left(\sum_{i,j=1}^{n} T_{ij}^2 \right)^{1/2},$$

which for a Toeplitz matrix can be computed in $\mathcal{O}(n)$ flops. An exact analogue of Theorem 5.1 holds for the maximal eigenvalues.

To conclude this section we briefly consider the fact that the same procedure for obtaining bounds for real symmetric positive-definite matrices can be used for general real symmetric matrices as well, provided that a lower bound on the smallest eigenvalue is available. Any known lower bound can be used (see, e.g., [11] or [16]), or one could be obtained by a process where a trial value is iteratively lowered until it falls below the smallest eigenvalue. Calling such a trial value ϵ, Sylvester's law of inertia can then be applied to the decomposition of $(T - \epsilon I)$, obtained in the course of the LDA algorithm (see section 3), to determine its position relative to the smallest eigenvalue of T. Such a procedure is extensively described and used in, e.g., [8] and [17], and we refer to these papers for further details.

6. Numerical results. In this section we will test our methods on three classes of positive semidefinite matrices. For each class and for each of the dimensions $n = 100, 200, 400$, we have run 200 experiments to examine the quality of the bounds on the smallest eigenvalue and 200 separate experiments doing the same for the maximal eigenvalue. Tables 6.1–6.6 report the average values (with their standard deviations in parentheses) of the bound-to-eigenvalue ratio for the smallest eigenvalue and eigenvalue-to-bound ratio for the largest eigenvalue. The closer this ratio is to one, the better the bound. Each entry in the table has a left and right part, separated by a slash.

TABLE 6.1
Bound-to-eigenvalue ratio for the minimal eigenvalue of CVL matrices.

Method	Dimension		
	100	200	400
Rational1	2.37(0.84) / 3.29(1.27)	3.70(1.98) / 5.22(3.07)	5.47(3.03) / 7.73(4.62)
Taylor1	2.82(1.08) / 3.93(1.57)	4.33(2.30) / 6.13(3.49)	6.43(3.57) / 9.10(5.40)
Rational2	1.14(0.19) / 1.26(0.24)	1.15(0.19) / 1.27(0.26)	1.18(0.20) / 1.30(0.29)
Taylor2	1.58(0.39) / 1.93(0.46)	1.93(0.53) / 2.39(0.70)	2.38(0.70) / 2.95(0.86)
Rational3	1.03(0.06) / 1.08(0.11)	1.03(0.08) / 1.08(0.12)	1.04(0.07) / 1.09(0.10)
Taylor3	1.29(0.23) / 1.49(0.26)	1.46(0.29) / 1.70(0.35)	1.67(0.33) / 1.96(0.39)

TABLE 6.2
Eigenvalue-to-bound ratio for the maximal eigenvalue of CVL matrices.

Method	Dimension		
	100	200	400
Rational1	1.34(0.09) / 1.41(0.10)	1.47(0.11) / 1.54(0.13)	1.61(0.13) / 1.67(0.14)
Taylor1	2.80(0.42) / 2.96(0.41)	3.19(0.46) / 3.35(0.49)	3.59(0.41) / 3.71(0.41)
Rational2	1.31(0.08) / 1.37(0.09)	1.44(0.11) / 1.51(0.12)	1.59(0.12) / 1.64(0.13)
Taylor2	2.63(0.37) / 2.79(0.37)	3.06(0.44) / 3.22(0.46)	3.49(0.40) / 3.61(0.40)
Rational3	1.07(0.04) / 1.11(0.04)	1.12(0.06) / 1.16(0.06)	1.17(0.04) / 1.21(0.04)
Taylor3	2.48(0.33) / 2.63(0.32)	2.94(0.41) / 3.09(0.43)	3.40(0.38) / 3.51(0.38)

TABLE 6.3
Bound-to-eigenvalue ratio for the minimal eigenvalue of KMS matrices.

Method	Dimension		
	100	200	400
Rational1	1.49(0.30) / 1.49(0.30)	1.49(0.28) / 1.49(0.28)	1.45(0.26) / 1.45(0.26)
Taylor1	1.68(0.32) / 1.68(0.32)	1.66(0.29) / 1.66(0.29)	1.66(0.28) / 1.66(0.28)
Rational2	1.33(0.16) / 1.33(0.16)	1.33(0.15) / 1.33(0.15)	1.31(0.14) / 1.31(0.14)
Taylor2	1.49(0.19) / 1.49(0.19)	1.50(0.18) / 1.50(0.18)	1.48(0.17) / 1.48(0.17)
Rational3	1.20(0.10) / 1.20(0.10)	1.20(0.09) / 1.20(0.09)	1.19(0.09) / 1.19(0.09)
Taylor3	1.39(0.14) / 1.39(0.14)	1.40(0.13) / 1.40(0.13)	1.39(0.12) / 1.39(0.12)

TABLE 6.4
Eigenvalue-to-bound ratio for the maximal eigenvalue of KMS matrices.

Method	Dimension		
	100	200	400
Rational1	1.36(0.19) / 1.37(0.19)	1.42(0.21) / 1.42(0.20)	1.43(0.24) / 1.43(0.24)
Taylor1	3.86(3.00) / 3.95(3.16)	5.13(5.17) / 5.23(5.41)	5.27(6.44) / 5.30(6.47)
Rational2	1.33(0.18) / 1.33(0.17)	1.39(0.19) / 1.39(0.19)	1.41(0.22) / 1.41(0.22)
Taylor2	3.28(2.06) / 3.35(2.15)	4.41(3.76) / 4.48(3.86)	4.86(5.45) / 4.86(5.45)
Rational3	1.13(0.07) / 1.14(0.07)	1.16(0.07) / 1.16(0.07)	1.17(0.08) / 1.17(0.08)
Taylor3	2.88(1.52) / 2.93(1.56)	3.89(2.88) / 3.93(2.91)	4.49(4.55) / 4.49(4.55)

TABLE 6.5
Bound-to-eigenvalue ratio for the minimal eigenvalue of UNF matrices.

Method	Dimension		
	100	200	400
Rational1	1.63(0.34) / 1.92(0.40)	2.04(0.43) / 2.41(0.48)	2.48(0.53) / 2.86(0.54)
Taylor1	3.18(0.89) / 3.82(0.94)	4.39(0.96) / 5.02(0.98)	5.36(1.03) / 5.89(0.97)
Rational2	1.17(0.13) / 1.28(0.17)	1.29(0.19) / 1.43(0.23)	1.41(0.20) / 1.59(0.270
Taylor2	1.91(0.45) / 2.25(0.51)	2.51(0.52) / 2.94(0.59)	3.12(0.68) / 3.55(0.71)
Rational3	1.05(0.07) / 1.09(0.08)	1.03(0.05) / 1.06(0.07)	1.06(0.05) / 1.10(0.07)
Taylor3	1.50(0.27) / 1.70(0.30)	1.84(0.31) / 2.11(0.36)	2.20(0.40) / 2.49(0.43)

TABLE 6.6
Eigenvalue-to-bound ratio for the maximal eigenvalue of UNF matrices.

Method	Dimension		
	100	200	400
Rational1	1.30(0.08) / 1.34(0.09)	1.35(0.09) / 1.38(0.09)	1.39(0.08) / 1.41(0.08)
Taylor1	1.81(0.19) / 1.87(0.20)	1.85(0.17) / 1.89(0.18)	1.87(0.16) / 1.90(0.16)
Rational2	1.29(0.08) / 1.33(0.09)	1.34(0.09) / 1.37(0.09)	1.38(0.08) / 1.40(0.08)
Taylor2	1.78(0.18) / 1.84(0.19)	1.83(0.17) / 1.87(0.17)	1.86(0.15) / 1.89(0.16)
Rational3	1.12(0.04) / 1.15(0.04)	1.15(0.04) / 1.18(0.04)	1.19(0.04) / 1.22(0.05)
Taylor3	1.75(0.17) / 1.80(0.18)	1.81(0.17) / 1.85(0.17)	1.85(0.15) / 1.88(0.15)

The left part pertains to the use of ϕ^e and ϕ^o (i.e., the bound is obtained by taking the minimum of the bounds on the even and odd extreme eigenvalues), whereas the right part represents the use of ϕ. The polynomial approximation-based bounds are denoted by "Taylor," followed by the order of the approximation, and the rational ones by "Rational." Let us now list those three classes of matrices.

(1) *CVL matrices*. These are the matrices defined in [8] (whence their name) as

$$T = \mu \sum_{k=1}^{n} \xi_k T_{2\pi\theta_k},$$

where n is the dimension of T, μ is such that $T_{kk} = 1$, $k = 1, \ldots, n$, and

$$(T_\theta)_{ij} = \cos(\theta(i-j)).$$

These matrices are positive semidefinite of rank two. We generated random matrices of this kind by taking the value of θ to be uniformly distributed on $(0, 1)$.

(2) *KMS matrices*. These are the Kac–Murdock–Szegö matrices (see [19]), defined as $T_{ij} = \nu^{|i-j|}$, where $0 < \nu < 1$, $i, j = 1, \ldots, n$, and n is the dimension of the matrix. These matrices are positive definite and are characterized by the fact that their even and odd eigenvalues lie extremely close together. Random matrices of this kind were generated by taking the value of ν to be uniformly distributed on (0, 1).

(3) *UNF matrices*. We define UNF matrices by first defining a random vector v of length n whose components are uniformly distributed on $(-10, 10)$. We then modify that vector by adding to its first component 1.1 times the absolute value of the smallest eigenvalue of the Toeplitz matrix generated by v. Finally, the vector v is normalized by dividing it by its first component, provided that it is different from zero. The Toeplitz matrix generated by this normalized vector is then called a UNF matrix. From their construction, these matrices are positive semidefinite.

Theoretically, some of the matrices that we generated might be singular, although we never encountered this situation in practice.

The experiments clearly show that exploiting the even and odd spectra yields better bounds. The magnitude of the improvement diminishes the closer the even and odd eigenvalues are lying together, as is obviously true for the KMS matrices. The superiority of rational bounds is also clearly demonstrated, both in their smaller average ratios and smaller standard deviations. They may yield a ratio of up to three times smaller than polynomial ones and in many cases, lower-order rational bounds are better than higher-order polynomial ones. This is especially true for larger matrix dimensions. These results also confirm our previous remark that the bounds obtained in [11] are inferior to rational approximation-based bounds.

Although we did not report results on bounds for the even and odd eigenvalues separately, we did verify that they are virtually identical to those obtained for the smallest eigenvalue proper.

All experiments were run in MATLAB on an Intel Pentium II 233MHz machine.

We conclude that computing separate, rational approximation-based bounds on the even and odd spectra leads to a significant improvement over existing bounds.

REFERENCES

[1] G. S. Ammar and W. B. Gragg, *The Generalized Schur Algorithm for the Superfast Solution of Toeplitz Systems*, in Rational Approximation and Its Applications in Mathematics and Physics, J. Gilewicz, M. Pindor, and W. Siemaszko, eds., Lecture Notes in Math. 1237, Springer-Verlag, Berlin, 1987, pp. 315–330.

[2] G. S. Ammar and W. B. Gragg, *Numerical experience with a superfast Toeplitz solver*, Linear Algebra Appl., 121 (1989), pp. 185–206.

[3] A. L. Andrew, *Eigenvectors of certain matrices*, Linear Algebra Appl., 7 (1973), pp. 151–162.

[4] J. R. Bunch, *Stability of methods for solving Toeplitz systems of equations*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 349–364.

[5] J. R. Bunch, C. P. Nielsen, and D. C. Sorensen, *Rank-one modification of the symmetric eigenproblem*, Numer. Math., 31 (1978), pp. 31–48.
[6] A. Cantoni and F. Butler, *Eigenvalues and eigenvectors of symmetric centrosymmetric matrices*, Linear Algebra Appl., 13 (1976), pp. 275–288.
[7] G. Cybenko, *The numerical stability of the Levinson–Durbin algorithm for Toeplitz systems of equations*, SIAM J. Sci. Statist. Comput., 1 (1980), pp. 303–319.
[8] G. Cybenko and C. Van Loan, *Computing the minimum eigenvalue of a symmetric positive definite Toeplitz matrix*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 123–131.
[9] P. Delsarte and Y. Genin, *Spectral properties of finite Toeplitz matrices*, in Mathematical Theory of Networks and Systems, Proceedings MTNS-83 International Symposium, Beer-Sheva, Israel, 1983, pp. 194–213.
[10] P. Delsarte and Y. Genin, *The split Levinson algorithm*, IEEE Trans. Acoust. Speech Signal Process., ASSP-34 (1986), pp. 470–478.
[11] A. Dembo, *Bounds on the extreme eigenvalues of positive definite Toeplitz matrices*, IEEE Trans. Inform. Theory, 34 (1988), pp. 352–355.
[12] C. J. Demeure, *Bowtie factors of Toeplitz matrices by means of split algorithms*, IEEE Trans. Acoust. Speech Signal Process., 34 (1989), pp. 1601–1603.
[13] J. Durbin, *The fitting of time series model*, Rev. Inst. Int. Stat., 28 (1960), pp. 233–243.
[14] G. H. Golub, *Some modified matrix eigenvalue problems*, SIAM Rev., 15 (1973), pp. 318–334.
[15] G. Golub and C. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, London, 1989.
[16] D. Hertz, *Simple bounds on the extreme eigenvalues of Toeplitz matrices*, IEEE Trans. Inform. Theory, 38 (1992), pp. 175–176.
[17] Y. H. Hu and S. Y. Kung, *A Toeplitz eigensystem solver*, IEEE Trans. Acoust. Speech Signal Process., 34 (1985), pp. 1264–1271.
[18] D. Huang, *Symmetric solutions and eigenvalue problems of Toeplitz systems*, IEEE Trans. Signal Process., 40 (1992), pp. 3069–3074.
[19] M. Kac, W. L. Murdock, and G. Szegö, *On the eigenvalues of certain Hermitian forms*, J. Rat. Mech. Anal., 2 (1953), pp. 767–800.
[20] N. Levinson, *The Wiener RMS* (*root mean square*) *error criterion in filter design and prediction*, J. Math. Phys., 25 (1947), pp. 261–278.
[21] W. Mackens and H. Voss, *The minimum eigenvalue of a symmetric positive-definite Toeplitz matrix and rational Hermitian interpolation*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 521–534.
[22] W. Mackens and H. Voss, *A projection method for computing the minimum eigenvalue of a symmetric positive definite Toeplitz matrix*, Linear Algebra Appl., 275–276 (1998), pp. 401–415.
[23] A. Melman, *A unifying convergence analysis of second-order methods for secular equations*, Math. Comp., 66 (1997), pp. 333–344.
[24] A. Melman, *Spectral functions for real symmetric Toeplitz matrices*, J. Comput. Appl. Math., 98 (1998), pp 233–243.
[25] A. Melman, *Extreme Eigenvalues of Real Symmetric Toeplitz Matrices*, Manuscript SCCM 98–11, Computer Science Department, SCCM Program, Stanford University, Stanford CA, 1998; available online at http://www-sccm.stanford.edu/reports.html.
[26] D. Slepian and H. J. Landau, *A note on the eigenvalues of Hermitian matrices*, SIAM J. Math. Anal., 9 (1978), pp. 291–297.
[27] H. Voss, *Symmetric schemes for computing the minimum eigenvalue of a symmetric Toeplitz matrix*, Linear Algebra Appl., 287 (1999), pp. 337–357.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 379–381
© 1999 Society for Industrial and Applied Mathematics

LIDSKII'S THEOREM VIA NONSMOOTH ANALYSIS*

A. S. LEWIS†

Abstract. Lidskii's theorem on eigenvalue perturbation is proved via a nonsmooth mean value theorem.

Key words. Lidskii's theorem, eigenvalue optimization, nonsmooth analysis, Clarke generalized gradient

AMS subject classifications. Primary, 90C31, 15A18; Secondary, 49K40, 26B05

PII. S0895479898338676

One of the central tools for studying perturbation theory for the eigenvalues of symmetric matrices is Lidskii's theorem. This states that any matrices Z and Y in $\mathbf{S}^n$, the Euclidean space of $n \times n$ real symmetric matrices (with trace inner product), satisfy

$$\lambda(Z+Y) - \lambda(Z) \prec \lambda(Y), \tag{0.1}$$

where $\lambda(Z) \in \mathbf{R}^n$ is the column vector of eigenvalues of Z written by multiplicity and in decreasing order. The symbol $\prec$ denotes *majorization*: $u \prec v$ for vectors u and v in $\mathbf{R}^n$ means

$$u \in \operatorname{conv}\{Pv : P \in \mathbf{P}^n\},$$

where $\mathbf{P}^n$ is the group of $n \times n$ permutation matrices and "conv" denotes convex hull. A standard separation argument (see [6], for example) shows this inclusion is equivalent to the condition

$$x^T u \leq \bar{x}^T \bar{v} \quad \text{for all } x \in \mathbf{R}^n,$$

where $\bar{x}$ is the vector in $\mathbf{R}^n$ with the same components as x arranged in decreasing order. Lidskii's theorem is, for example, one of the unifying themes of the recent book [1].

This note approaches Lidskii's theorem via nonsmooth analysis, using the *Clarke generalized gradient* (see [2]). For a real, locally Lipschitz function g defined close to a point p in a Euclidean space, we can define the generalized gradient $\partial g(p)$ as the convex hull of the set of cluster points of gradients of g at points near p in a set of full measure [2, Thm. 2.5.1]. Thus for a smooth function g, the generalized gradient coincides with the usual gradient $g'(p)$.

For a smooth function g, the classical mean value theorem states that, given points q and r, there is a point p on the line segment between q and r satisfying the equation

$$g(q) - g(r) = \langle g'(p), q - r \rangle.$$

*Received by the editors May 13, 1998; accepted for publication (in revised form) by M. Overton December 16, 1998; published electronically October 26, 1999. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada.
http://www.siam.org/journals/simax/21-2/33867.html

†Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1 (aslewis@orion.uwaterloo.ca, http://orion.uwaterloo.ca/∼aslewis).

If g is merely locally Lipschitz, the Lebourg mean value theorem [2, Thm. 2.3.7] generalizes this result, asserting the existence of an element s of $\partial g(p)$ satisfying the equation

$$g(q) - g(r) = \langle s, q - r \rangle.$$

We apply this mean value theorem to derive Lidskii's theorem from a powerful variational result about eigenvalues. This variational result calculates the Clarke generalized gradient of the function $f \circ \lambda$ for an arbitrary locally Lipschitz *permutation-invariant* function $f : \mathbf{R}^n \to \mathbf{R}$ (that is, $f(Px) = f(x)$ for all matrices P in $\mathbf{P}^n$ and vectors x in $\mathbf{R}^n$). Specifically, the result (see [4]) considers diagonalizations of any matrix X in $\mathbf{S}^n$,

$$X = U^T(\operatorname{Diag} x)U, \quad U \in \mathbf{O}^n, \quad x \in \mathbf{R}^n, \tag{0.2}$$

where $\mathbf{O}^n$ is the group of $n \times n$ orthogonal matrices, and states

$$\partial(f \circ \lambda)(X) = \{U^T(\operatorname{Diag} y)U : (0.2) \text{ holds and } y \in \partial f(x)\}. \tag{0.3}$$

The approach of this note is certainly not the simplest proof of Lidskii's theorem, formula (0.3) being a relatively difficult result (see also [5]). However, the approach is delightfully transparent, and it reveals the increasing possibilities of applying nonsmooth analytic techniques in matrix analysis.

Lidskii's theorem (0.1) is equivalent to the inequality

$$w^T(\lambda(Z + Y) - \lambda(Z)) \leq \bar{w}^T\lambda(Y) \text{ for all } w \in \mathbf{R}^n. \tag{0.4}$$

Fix w and consider the (nonconvex) locally Lipschitz, permutation-invariant function defined by

$$f(x) = w^T\bar{x}.$$

Note the relationship

$$(f \circ \lambda)(X) = w^T\lambda(X) \text{ for all } X \in \mathbf{S}^n.$$

Whenever x has distinct components (a subset of $\mathbf{R}^n$ of full measure), f is differentiable at x with gradient $f'(x) = Pw$ for some matrix P in $\mathbf{P}^n$. Hence at *any* point x in $\mathbf{R}^n$ we have the inclusion

$$\partial f(x) \subset \operatorname{conv}\{Pw : P \in \mathbf{P}^n\}, \tag{0.5}$$

by our definition of the generalized gradient. (There are more elementary although less concise ways to see this inclusion.)

By the Lebourg mean value theorem applied to $f \circ \lambda$, there is a matrix X in $\mathbf{S}^n$ (between Z and $Z + Y$) and a matrix V in $\partial(f \circ \lambda)(X)$ satisfying

$$w^T(\lambda(Z + Y) - \lambda(Z)) = \operatorname{tr}(VY) \leq \lambda(V)^T\lambda(Y),$$

the last inequality following from von Neumann's trace theorem (see [3, Eq. (7.4.14)]). But now we can apply the generalized gradient formulae (0.3) and (0.5) to deduce

$$\lambda(V) \in \operatorname{conv}\{Pw : P \in \mathbf{P}^n\}$$

(or in other words, $\lambda(V) \prec \bar{w}$). Hence $\lambda(V)^T\lambda(Y) \leq \bar{w}^T\lambda(Y)$, and Lidskii's theorem (0.4) follows.

Acknowledgment. The author thanks the School of Operations Research and Industrial Engineering at Cornell University, where this work was completed.

REFERENCES

[1] R. Bhatia, *Matrix Analysis*, Grad. Texts in Math. 169, Springer-Verlag, New York, 1997.
[2] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983; reprinted as Classics Appl. Math. 5, SIAM, Philadelphia, 1990.
[3] R. A. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.
[4] A. S. Lewis, *Derivatives of spectral functions*, Math. Oper. Res., 6 (1996), pp. 576–588.
[5] A. S. Lewis, *Nonsmooth analysis of eigenvalues*, Math. Programming, 84 (1999), pp. 1–24.
[6] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and its Application*, Academic Press, New York, 1979.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 382–389
© 1999 Society for Industrial and Applied Mathematics

SECOND LOGARITHMIC DERIVATIVE OF A COMPLEX MATRIX IN THE CHEBYSHEV NORM*

L. KOHAUPT†

Abstract. The second logarithmic derivative $\mu_\infty^{(2)}[A]$ of a complex $n \times n$ matrix A in the Chebyshev norm is defined as the second right derivative of $\|\Phi(t)\|_\infty = \|e^{At}\|_\infty$ at $t = 0$, where $\|\cdot\|_\infty$ denotes the operator norm corresponding to the norm $\|\cdot\|_\infty$ in $\mathbb{C}^n$. The obtained formula is illustrated by a numerical example. The result may be of interest in applications such as stability theory.

Key words. logarithmic derivative, first logarithmic derivative, second logarithmic derivative, complex matrix, Chebyshev norm, matrix measure, fundamental matrix, evolution, local regularity

AMS subject classifications. 10M20, 34D, 35E05, 40G10, 46G05, 47A30

PII. S0895479897316072

1. Introduction.

For a given complex $n \times n$ matrix A, consider the initial-value problem $\dot{x}(t) = A\,x(t),\ x(0) = x_0$. It is well known that the solution can be expressed in the form $x(t) = \Phi(t)\,x_0$, where $\Phi(t) = e^{At}$ is the so-called *fundamental matrix* or *evolution*.

For many purposes, e.g., in error estimates, the operator norm $\|\Phi(t)\| = \|e^{At}\|$ of $\Phi(t) = e^{At}$ corresponding to a vector norm $\|\cdot\|$ in $\mathbb{C}^n$ is needed. An upper bound on $\|\Phi(t)\|$ is given by $e^{\mu^{(1)}[A]t}$, where $\mu^{(1)}[A]$ is the (first) logarithmic derivative of the matrix A in a matrix operator norm. The quantity $\mu^{(1)}[A]$ is first defined and investigated in [5]. For the inequality $\|\Phi(t)\| \le e^{\mu^{(1)}[A]t},\ t \ge 0$, see [5, Formula (12), pp. 59–60] and [1, p. 59]. In [1] and [3], $\mu^{(1)}[A]$ is called *matrix measure* and it is applied there to bound solutions of ordinary differential equations; it is also used for other purposes.

Formulae for the first logarithmic derivative $\mu_p^{(1)}[A]$ in the p-norm with $p \in \{1, 2, \infty\}$ are derived in [5, Lemma 4, pp. 59–60]; these formulae are also stated without proof in [1, p. 41] and [2, p. 11].

In the present paper, the second logarithmic derivative $\mu_\infty^{(2)}[A]$ is defined and an associated formula is derived. The case $p = 1$ can be treated by interchanging the roles of the row index and column index of the matrix A; this case is therefore omitted. The case $p = 2$ has not yet been resolved.

The paper is structured as follows. In section 2, it is shown that a real analytic function $\hat{\Phi}(\cdot)$ exists, defined on some neighborhood $[0, t_0]$ of $t = 0$, such that $\hat{\Phi}(\cdot)$ and $\|\Phi(\cdot)\|_\infty$ coincide on $[0, t_0]$. In section 3, $\mu_\infty^{(2)}[A]$ is defined and an explicit formula is derived for it. In section 4, the result obtained in section 3 is illustrated by a numerical example.

2. Local regularity.

In this section, it is shown that the mapping $t \mapsto \|\Phi(t)\|_\infty$, $0 \le t \le t_0$, is real analytic for sufficiently small $t_0 > 0$, or, more precisely, that a real analytic function $\hat{\Phi}(\cdot)$ exists, defined on some neighborhood $[0, t_0]$ of $t = 0$, such that

*Received by the editors February 3, 1997; accepted for publication (in revised form) by U. Helmke February 19, 1999; published electronically October 26, 1999.
http://www.siam.org/journals/simax/21-2/31607.html
†Prager Str. 9, D-10779 Berlin, Germany (kohaupt@tfh-berlin.de).

$\hat{\Phi}(\cdot)$ and $\|\Phi(\cdot)\|_\infty$ coincide on $[0,t_0]$. In this article, when we speak of the real analytic function $t \mapsto \|\Phi(t)\|_\infty$, $0 \le t \le t_0$, we mean the mapping $t \mapsto \hat{\Phi}(t)$, $0 \le t \le t_0$.

It follows that the kth logarithmic derivative $\mu_\infty^{(k)}[A] := D_+^k \|\Phi(0)\|_\infty$ (with $k \in \mathbb{N}$) exists, where $D_+^k \|\Phi(0)\|_\infty$ denotes the kth right derivative of the mapping $t \mapsto \|\Phi(t)\|_\infty$ at $t = 0$.

In the next section, we calculate the second logarithmic derivative of a complex matrix A, $\mu_\infty^{(2)}[A]$, explicitly, where the continuous differentiability of the function $t \mapsto \|\Phi(t)\|_\infty$, $0 \le t \le t_0$, up to the second order is used.

As a preparation to the proof of Lemma 2.2 and Theorem 3.1, we need the following lemma.

LEMMA 2.1. *Let $n \in \mathbb{N}$ and $a_i \in \mathbb{R}$, $i = 1, \dots, n$. Further, let $t_0 > 0$ and $t \mapsto b_i(t)$, $0 \le t \le t_0$, be real continuous functions with $b_i(0) = 0$, $i = 1, \dots, n$. Moreover, let $I_0 = \{1, \dots, n\}$ and I_1 be the set of all indices $i_1 \in I_0$, where a_i attains its maximum, i.e.,*

$$I_1 := \left\{ i_1 \in I_0 \,|\, a_{i_1} = \max_{i \in I_0} a_i \right\}.$$

Then, there exists a number $t_1 > 0$ with $t_1 \le t_0$ such that

$$\max_{i \in I_0} \{a_i + b_i(t)\} = \max_{i \in I_0} a_i + \max_{i \in I_1} b_i(t), \quad 0 \le t \le t_1.$$

Proof. Case 1. $I_1 = I_0$. Here, for $0 \le t \le t_0$,

$$\max_{i \in I_0} \{a_i + b_i(t)\} = \max_{i \in I_0} \left\{ \max_{i \in I_0} a_i + b_i(t) \right\} = \max_{i \in I_0} a_i + \max_{i \in I_0} b_i(t) = \max_{i \in I_0} a_i + \max_{i \in I_1} b_i(t).$$

Case 2. $I_1 \neq I_0$. In this case,

$$\max_{i \in I_0 \setminus I_1} a_i < \max_{i \in I_1} a_i.$$

Thus, there exists a number $t_1 > 0$ with $t_1 \le t_0$ such that

$$\max_{i \in I_0 \setminus I_1} \{a_i + b_i(t)\} < \max_{i \in I_1} \{a_i + b_i(h)\}, \quad 0 \le t \le t_1.$$

Consequently, for $0 \le t \le t_1$,

$$\max_{i \in I_0} \{a_i + b_i(t)\} = \max_{i \in I_1} \{a_i + b_i(t)\} = \max_{i \in I_1} \left\{ \max_{i \in I_0} a_i + b_i(t) \right\} = \max_{i \in I_0} a_i + \max_{i \in I_1} b_i(t). \qquad \square$$

LEMMA 2.2. *There exists a number $t_0 > 0$ and a function $t \mapsto \hat{\Phi}(t)$, which is real analytic on $[0, t_0]$, such that $\hat{\Phi}(t) = \|\Phi(t)\|_\infty$ for all $t \in [0, t_0]$.*

Proof. We have

$$\|\Phi(t)\|_\infty = \max_{i=1,\dots,n} \sum_{j=1}^{n} |\Phi_{ij}(t)| = \max_{i=1,\dots,n} \Psi_i(t) \tag{1}$$

with

$$\Psi_i(t) = \sum_{j=1}^{n} |\Phi_{ij}(t)| = \sum_{j=1}^{n} \sqrt{[\operatorname{Re} \Phi_{ij}(t)]^2 + [\operatorname{Im} \Phi_{ij}(t)]^2}, \; i = 1, \dots, n,$$

where the range of t is restricted to $t \geq 0$. Further,

$$\Phi_{ij}(t) = \sum_{k=0}^{\infty} (A^k)_{ij} \frac{t^k}{k!}, \quad i,j = 1,\dots,n,$$

so that the functions $t \mapsto \operatorname{Re} \Phi_{ij}(t)$ and $t \mapsto \operatorname{Im} \Phi_{ij}(t)$ are real analytic for $t \geq 0$. Therefore, if we choose $t_1 > 0$ sufficiently small, the function $t \mapsto \Psi_i(t),\ 0 \leq t \leq t_1$, also is real analytic. Consequently, there exist real numbers $\lambda_i^{(k)}$ such that

$$\Psi_i(t) = \sum_{k=0}^{\infty} \lambda_i^{(k)} \frac{t^k}{k!}, \quad 0 \leq t \leq t_1; \quad i = 1,\dots,n,$$

where

$$\lambda_i^{(0)} = 1, \quad i = 1,\dots,n,$$

and

$$\lambda_i^{(k)} = D_+^k \Psi_i(0), \quad k = 1,2,\dots, \quad i = 1,\dots,n.$$

Thus, we have

$$\|\Phi(t)\|_\infty = \max_{i=1,\dots,n} \left\{ \sum_{k=0}^{\infty} \lambda_i^{(k)} \frac{t^k}{k!} \right\}, \quad 0 \leq t \leq t_1.$$

Let $I_0 := \{1,\dots,n\}$ and I_k be the set of all indices $i_k \in I_{k-1}$, where $\lambda_i^{(k)}$ attains its maximum, i.e.,

$$I_k := \left\{ i_k \in I_{k-1} \,|\, \lambda_{i_k}^{(k)} = \max_{i \in I_{k-1}} \lambda_i^{(k)} \right\}, \quad k = 1,2,\dots.$$

Then,

$$I_0 \supset I_1 \supset \cdots \supset I_k \supset \cdots. \tag{2}$$

Since I_0 contains only a finite number of elements, there is an index N such that

$$I_i = I_N, \quad i \geq N. \tag{3}$$

Choose any indices $i_k \in I_k,\ k = 0,1,\dots,N$. Then, by a repeated application of Lemma 2.1, we obtain a number $t_2 \in (0,t_1]$ such that

$$\|\Phi(t)\|_\infty = \max_{i=1,\dots,n} \Psi_i(t) = \max_{i \in I_0} \left\{ \sum_{k=0}^{\infty} \lambda_i^{(k)} \frac{t^k}{k!} \right\} = \sum_{k=0}^{N-1} \lambda_{i_k}^{(k)} \frac{t^k}{k!} + \sum_{k=N}^{\infty} \lambda_{i_N}^{(k)} \frac{t^k}{k!} =: \hat{\Phi}(t),$$

$0 \leq t \leq t_2$. Consequently, the assertion follows with $t_0 := t_2$. □

Remark. Lemma 2.2 is illustrated by the following example:

$$A := \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \Phi(t) = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}, \quad \|\Phi(t)\|_\infty = 1 + |t|, \quad \hat{\Phi}(t) = 1 + t.$$

Thus, $\|\Phi(\cdot)\|_\infty$ is not real analytic at $t = 0$, but it coincides with a real analytic function $\hat{\Phi}(\cdot)$ on $[0,t_0]$ for every $t_0 > 0$.

3. Second logarithmic derivative in the Chebyshev norm. In [5], the *first logarithmic derivative* of a complex $n \times n$ matrix $A = (A_{ij})$ in the p-norm is defined by the right-hand limit

$$\mu_p^{(1)}[A] = D_+^1 \|\Phi(0)\|_p := \lim_{\Delta t \to 0+} \frac{\|\Phi(\Delta t)\|_p - \|\Phi(0)\|_p}{\Delta t} = \lim_{\Delta t \to 0+} \frac{\|E + A\Delta t\|_p - 1}{\Delta t}, \tag{4}$$

where E is the $n \times n$ identity matrix.

For $p \in \{1, 2, \infty\}$, in [5, pp. 59–60, Lemma 4] the following explicit formulae are stated and proved:

$$\mu_\infty^{(1)}[A] = \max_{i=1,\dots,n} \left\{ \operatorname{Re} A_{ii} + \sum_{\substack{j=1 \\ j \neq i}}^{n} |A_{ij}| \right\}, \tag{5}$$

$$\mu_1^{(1)}[A] = \max_{j=1,\dots,n} \left\{ \operatorname{Re} A_{jj} + \sum_{\substack{i=1 \\ i \neq j}}^{n} |A_{ij}| \right\}, \tag{6}$$

$$\mu_2^{(1)}[A] = \lambda_{\max} \left(\frac{A^* + A}{2} \right), \tag{7}$$

where $\lambda_{\max}(\frac{A^*+A}{2})$ denotes the largest eigenvalue of $\frac{A^*+A}{2}$. These formulae can also be found (without proof) in [1, p. 41] and [2, p. 11].

The *second logarithmic derivative* in the Chebyshev norm is defined by the right-hand limit

$$\mu_\infty^{(2)}[A] = D_+^2 \|\Phi(0)\|_\infty := \lim_{\Delta t \to 0+} \frac{D_+^1 \|\Phi(\Delta t)\|_\infty - D_+^1 \|\Phi(0)\|_\infty}{\Delta t}, \tag{8}$$

where the existence of this limit follows from section 2.

To prove Theorem 3.1, we do not use (8), but we compute the coefficient of $\frac{t^2}{2!}$ in the power series expansion of $\|\Phi(t)\|_\infty$ to obtain a formula for $\mu_\infty^{(2)}[A]$. This seems to be more convenient than the direct use of formula (8).

Now, define the following functionals on $\mathbb{C}^{n \times n}$:

$$\nu_{ij}[A] := \begin{cases} \dfrac{\operatorname{Re} A_{ij} \operatorname{Re} (A^2)_{ij} + \operatorname{Im} A_{ij} \operatorname{Im} (A^2)_{ij}}{|A_{ij}|}, & A_{ij} \neq 0, \\ |(A^2)_{ij}|, & A_{ij} = 0, \end{cases}$$

$$\lambda_i^{(1)}[A] := \operatorname{Re} A_{ii} + \sum_{\substack{j=1 \\ j \neq i}}^{n} |A_{ij}|, \tag{9}$$

$$\lambda_i^{(2)}[A] := \operatorname{Re} (A^2)_{ii} + (\operatorname{Im} A_{ii})^2 + \sum_{\substack{j=1 \\ i \neq j}}^{n} \nu_{ij}[A] \tag{10}$$

and observe that the first logarithmic derivative is given by

$$\mu_\infty^{(1)}[A] = \max_{i=1,\ldots,n} \lambda_i^{(1)}[A].$$

Then, we obtain the following theorem.

THEOREM 3.1. *Let $A \in \mathbb{C}^{n\times n}$ and I_1 be the set of all indices $i_1 \in I_0 := \{1,\ldots,n\}$ where $\lambda_i^{(1)}[A]$ attains its maximum, i.e.,*

$$I_1 := \left\{ i_1 \in I_0 \,|\, \lambda_{i_1}^{(1)}[A] = \max_{i\in I_0} \lambda_i^{(1)}[A] \right\}.$$

Then, the second logarithmic derivative of A is given by

$$\mu_\infty^{(2)}[A] = \max_{i\in I_1} \lambda_i^{(2)}[A],$$

where $\lambda_i^{(1)}[A]$ and $\lambda_i^{(2)}[A]$ are given by (9) *and* (10)*, respectively.*

Proof. We obtain

$$\begin{aligned} \|\Phi(t)\|_\infty &= \left\| E + A\,t + \frac{1}{2}\,A^2\,t^2 + O(t^3) \right\|_\infty \\ &= \max_{i=1,\ldots,n} \left\{ \left| 1 + A_{ii}\,t + \frac{1}{2}\,(A^2)_{ii}\,t^2 + O(t^3) \right| \right. \\ &\qquad \left. + \sum_{\substack{j=1\\ j\neq i}}^{n} \left| A_{ij}\,t + \frac{1}{2}\,(A^2)_{ij}\,t^2 + O(t^3) \right| \right\}. \end{aligned} \tag{11}$$

Let $i \in \{1,\ldots,n\}$ be fixed, $t > 0$ be sufficiently small, and

$$\begin{aligned} U_i &:= \left| 1 + A_{ii}\,t + \frac{1}{2}\,(A^2)_{ii}\,t^2 + O(t^3) \right| \\ &\quad + \sum_{\substack{j=1\\ j\neq i}}^{n} \left| A_{ij}\,t + \frac{1}{2}\,(A^2)_{ij}\,t^2 + O(t^3) \right|. \end{aligned} \tag{12}$$

Then,

$$U_i = u_{ii}^{(2)}(A) + \sum_{\substack{j=1\\ j\neq i}}^{n} u_{ij}^{(2)}(A) \tag{13}$$

with

$$\begin{aligned} u_{ii}^{(2)}(A) &:= \left\{ \left[1 + \operatorname{Re} A_{ii}\,t + \frac{1}{2}\operatorname{Re}(A^2)_{ii}\,t^2 + O(t^3) \right]^2 \right. \\ &\quad \left. + \left[\operatorname{Im} A_{ii}\,t + \frac{1}{2}\operatorname{Im}(A^2)_{ii}\,t^2 + O(t^3) \right]^2 \right\}^{\frac{1}{2}} \end{aligned} \tag{14}$$

and

$$u_{ij}^{(2)}(A) := \left\{ \left[\operatorname{Re} A_{ij} + \frac{1}{2} \operatorname{Re} (A^2)_{ij}\, t + O(t^2) \right]^2 + \left[\operatorname{Im} A_{ij} + \frac{1}{2} \operatorname{Im} (A^2)_{ij}\, t + O(t^2) \right]^2 \right\}^{\frac{1}{2}} t, \tag{15}$$

$i \neq j,\ j \in \{1, \ldots, n\}$.

To evaluate (14) and (15), the relation

$$(1+x)^{\frac{1}{2}} = 1 + \frac{1}{2}\, x - \frac{1}{8}\, x^2 + O(x^3), \tag{16}$$

$x \in \mathbb{R},\ |x| \leq 1$, is used. For sufficiently small $t > 0$, we obtain

$$u_{ii}^{(2)}(A) = 1 + \{\operatorname{Re} A_{ii}\}\, t + \frac{1}{2} \{\operatorname{Re} (A^2)_{ii} + (\operatorname{Im} A_{ii})^2\}\, t^2 + O(t^3). \tag{17}$$

To determine $u_{ij}^{(2)}(A)$ in (15), we consider two cases.

Case 1. $A_{ij} \neq 0$, where $i \neq j$.

Here, we get

$$u_{ij}^{(2)}(A) = |A_{ij}|\, t + \frac{1}{2} \left\{ \frac{\operatorname{Re} A_{ij} \operatorname{Re} (A^2)_{ij} + \operatorname{Im} A_{ij} \operatorname{Im} (A^2)_{ij}}{|A_{ij}|} \right\} t^2 + O(t^3). \tag{18}$$

Case 2. $A_{ij} = 0$, where $i \neq j$.

From (15), we infer

$$u_{ij}^{(2)}(A) = \frac{1}{2}\, |(A^2)_{ij}|\, t^2 + O(t^3). \tag{19}$$

Thus, we find

$$U_i = 1 + U_i^{'}\, t + \frac{1}{2}\, U_i^{''}\, t^2 + O(t^3) \tag{20}$$

with

$$U_i^{'} = \lambda_i^{(1)}[A] \tag{21}$$

and

$$U_i^{''} = \lambda_i^{(2)}[A], \tag{22}$$

where $\lambda_i^{(1)}[A]$ and $\lambda_i^{(2)}[A]$ are defined by (9) and (10), respectively.

Now, let I_2 be the set of all indices $i_2 \in I_1$, where $U_i^{''}$ attains its maximum, i.e.,

$$I_2 := \left\{ i_2 \in I_1 \,|\, U_{i_2}^{''} = \max_{i \in I_1} U_i^{''} \right\}. \tag{23}$$

Then, by a repeated application of Lemma 2.1, for sufficiently small $t > 0$ and arbitrary indices $i_1 \in I_1$ and $i_2 \in I_2$ we obtain

$$\begin{aligned} \|\Phi(t)\|_\infty = \left\| E + A\, t + \frac{1}{2}\, A^2\, t^2 + O(t^3) \right\|_\infty &= \max_{i=1,\ldots,n} U_i = \max_{i \in I_0} U_i \\ &= 1 + U_{i_1}^{'}\, t + \frac{1}{2}\, U_{i_2}^{''}\, t^2 + O(t^3) \end{aligned} \tag{24}$$

since $I_1 \supset I_2$. From (24), it follows that

$$\mu_{2,\infty}[A] = U_{i_2}^{''} = \max_{i \in I_1} U_i^{''} \tag{25}$$

for every index $i_2 \in I_2$. This is the asserted formula for $\mu_\infty^{(2)}[A]$. □

Remark. The formula for the first logarithmic derivative, $\mu_\infty^{(1)}[A]$, as stated in (5), is given by the coefficient of t in the power series expansion of $\|\Phi(t)\|_\infty$. This coefficient reads $\max_{i \in I_0} U_i^{'}$ with $U_i^{'}$ from (21). □

In the special case of *real matrices*, a unified formula for all logarithmic derivatives can be given. To see this, define the following sign functionals:

$$s_{ij}^{(1)}(A) := \operatorname{sgn}(A_{ij}) \tag{26}$$

and

$$s_{ij}^{(k)}(A) := \begin{cases} \operatorname{sgn}(A_{ij}), & A_{ij} \neq 0, \\ \operatorname{sgn}((A^2)_{ij}), & A_{ij} = 0, \quad (A^2)_{ij} \neq 0, \\ \operatorname{sgn}((A^3)_{ij}), & A_{ij} = 0, \quad (A^2)_{ij} = 0, \quad (A^3)_{ij} \neq 0, \\ \quad \vdots & \\ \operatorname{sgn}((A^k)_{ij}), & (A^l)_{ij} = 0, \quad l = 1, \ldots, k-1 \end{cases} \tag{27}$$

for $k = 2, 3, \ldots$. The relation (27) can also be written as

$$s_{ij}^{(k)}(A) = \begin{cases} s_{ij}^{(k-1)}(A), & s_{ij}^{(k-1)}(A) \neq 0, \\ \operatorname{sgn}((A^k)_{ij}), & s_{ij}^{(k-1)}(A) = 0 \end{cases}$$

for $k = 2, 3, \ldots$. With these sign functionals, define the further functionals

$$\lambda_i^{(k)}[A] = (A^k)_{ii} + \sum_{\substack{j=1 \\ j \neq i}}^{n} s_{ij}^{(k)}(A)\,(A^k)_{ij}, \tag{28}$$

$k = 1, 2, \ldots$. Then, the logarithmic derivatives for *real matrices* read as follows.

LEMMA 3.2. *Let $A \in \mathbb{R}^{n \times n}$, $I_0 = \{1, \ldots, n\}$, and I_k be the set of all indices $i_k \in I_{k-1}$ where $\lambda_i^{(k)}[A]$ from (28) attains its maximum, i.e.,*

$$I_k := \left\{ i_k \in I_{k-1} \,|\, \lambda_{i_k}^{(k)}[A] = \max_{i \in I_{k-1}} \lambda_i^{(k)}[A] \right\}, \tag{29}$$

$k = 1, 2, \ldots$. Then, the logarithmic derivatives are given by

$$\mu_\infty^{(k)}[A] = \max_{i \in I_{k-1}} \lambda_i^{(k)}[A], \tag{30}$$

$k = 1, 2, \ldots$. □

The proof of Lemma 3.2 is left to the reader.

4. Numerical example. We take matrix A from [4, p. 114, Example 6.1], that is,

$$A = \begin{bmatrix} 2 & -i & 0 \\ i & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}.$$

Then,

$$\mu_\infty^{(1)}[A] = \max\{2+1, 2+1, 3\} = 3$$

and

$$I_1 = \{1, 2, 3\}.$$

Further,

$$A^2 = \begin{bmatrix} 5 & -4i & 0 \\ 4i & 5 & 0 \\ 0 & 0 & 9 \end{bmatrix}.$$

Thus,

$$\mu_\infty^{(2)}[A] = \max\left\{5 + \frac{0\cdot 0 + (-1)(-4)}{1} + 0,\ 5 + \frac{0\cdot 0 + 1\cdot 4}{1} + 0,\ 9 + 0 + 0\right\}$$

$$= \max\{9, 9, 9\} = 9$$

and

$$I_2 = I_1 = \{1, 2, 3\}. \qquad \square$$

Remark. If the element A_{22} of A is replaced by $A_{22} = 1$, we obtain the same values for the first and second logarithmic derivatives as before, but the index sets are different; namely, they read

$$I_1 = \{1, 3\}, \qquad I_2 = \{3\}. \qquad \square$$

Acknowledgment. The author would like to thank the referees for their valuable suggestions.

REFERENCES

[1] W.A. Coppel, *Stability and Asymptotic Behavior of Differential Equations*, D.C. Heath and Company, Boston, 1965.
[2] G. Dahlquist, *Stability and error bounds in the numerical integration of ordinary differential equations*, Kungl. Tekn. Högsk. Handl. Stockholm, 130 (1959).
[3] Ch.A. Desoer and H. Haneda, *The measure of a matrix as a tool to analyze computer algorithms for circuit analysis*, IEEE Trans. Circuit Theory, 19 (1972), pp. 480–486.
[4] R.T. Gregory and D.L. Karney, *A Collection of Matrices for Testing Computational Algorithms*, Wiley-Interscience, New York, 1969.
[5] S.M. Lozinskiĭ, *Error estimates for the numerical integration of ordinary differential equations*, I. Izv. Vysš. Učebn. Zaved. Matematika, 5(6) (1958), pp. 52–90 (in Russian).

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 390–395
© 1999 Society for Industrial and Applied Mathematics

NONSINGULARITY OF THE DIFFERENCE OF TWO OBLIQUE PROJECTORS*

JÜRGEN GROß† AND GÖTZ TRENKLER†

Abstract. For two real projectors $\mathbf{P}$ and $\mathbf{Q}$ of the same order it is shown that the difference $\mathbf{P}-\mathbf{Q}$ is nonsingular if and only if the column spaces of $\mathbf{P}$ and $\mathbf{Q}$ are complementary and the row spaces of $\mathbf{P}$ and $\mathbf{Q}$ are complementary. Moreover, it is demonstrated that nonsingularity of $\mathbf{P}-\mathbf{Q}$ implies nonsingularity of $\mathbf{P}+\mathbf{Q}$, $\mathbf{I}-\mathbf{QP}$, and $\mathbf{P}+\mathbf{Q}-\mathbf{QP}$.

Key words. oblique projector, rank, complementary subspaces

AMS subject classifications. 15A03, 15A24

PII. S0895479897320277

1. Notation and preliminaries. For an $m\times n$ real matrix $\mathbf{A}$, the symbols $\mathbf{A}'$, $\mathcal{R}(\mathbf{A})$, $\mathcal{N}(\mathbf{A})$, and $\mathrm{rk}(\mathbf{A})$ will stand for the transpose, the column space, the null space, and the rank, respectively, of $\mathbf{A}$. An $n\times n$ matrix $\mathbf{P}$ is called an (oblique) projector if $\mathbf{P}=\mathbf{P}^2$, or equivalently $\mathcal{N}(\mathbf{P})=\mathcal{R}(\mathbf{I}_n-\mathbf{P})$, in which case $\mathbf{P}$ projects on $\mathcal{R}(\mathbf{P})$ along $\mathcal{N}(\mathbf{P})$. If $\mathbf{P}$ projects on $\mathcal{R}(\mathbf{P})$ along $\mathcal{R}(\mathbf{P})^{\perp}$, then $\mathbf{P}$ is an orthogonal projector. It is readily verified that $\mathbf{P}$ is an orthogonal projector if and only if $\mathbf{P}=\mathbf{P}^2=\mathbf{P}'$.

Throughout this paper, we will repeatedly refer to Marsaglia and Styan [5] for results on equalities for ranks of matrices. An often-applied result is [5, Corollary 6.2], which reads

$$\mathrm{rk}(\mathbf{B})=\mathrm{rk}(\mathbf{AB})+\dim[\mathcal{N}(\mathbf{A})\cap\mathcal{R}(\mathbf{B})]$$

for matrices $\mathbf{A}$ and $\mathbf{B}$ of appropriate dimension. Moreover, we will use the identities $\mathcal{N}(\mathbf{P})=\mathcal{R}(\mathbf{I}_n-\mathbf{P})$ and $\mathrm{rk}(\mathbf{I}_n-\mathbf{P})=n-\mathrm{rk}(\mathbf{P})$ for a projector $\mathbf{P}$ without explicitly mentioning it each time.

2. Rank of the difference. Our first result gives the rank of the difference $\mathbf{P}-\mathbf{Q}$ of two projectors $\mathbf{P}$ and $\mathbf{Q}$, which enables us to establish nonsingularity of $\mathbf{P}-\mathbf{Q}$ as an easy consequence.

THEOREM 1. *Let* $\mathbf{P}$ *and* $\mathbf{Q}$ *be two* $n\times n$ *projectors. Then*

$$\mathrm{rk}(\mathbf{P}-\mathbf{Q})=\mathrm{rk}\begin{pmatrix}\mathbf{P}\\ \mathbf{Q}\end{pmatrix}+\mathrm{rk}(\mathbf{P},\mathbf{Q})-\mathrm{rk}(\mathbf{P})-\mathrm{rk}(\mathbf{Q}) \tag{2.1}$$

$$=n-\dim[\mathcal{N}(\mathbf{P})\cap\mathcal{N}(\mathbf{Q})]-\dim[\mathcal{R}(\mathbf{P})\cap\mathcal{R}(\mathbf{Q})]. \tag{2.2}$$

Proof. From Marsaglia and Styan [5, Theorem 8] we conclude that

$$\mathrm{rk}(\mathbf{P}-\mathbf{Q})=\mathrm{rk}\begin{pmatrix}\mathbf{0} & \mathbf{P} & -\mathbf{Q}\\ \mathbf{P} & \mathbf{P} & \mathbf{0}\\ -\mathbf{Q} & \mathbf{0} & -\mathbf{Q}\end{pmatrix}-\mathrm{rk}(\mathbf{P})-\mathrm{rk}(\mathbf{Q}).$$

*Received by the editors May 7, 1997; accepted for publication (in revised form) by G. Styan October 31, 1998; published electronically October 26, 1999. This research was supported by Deutsche Forschungsgemeinschaft under grant Tr 253/2-3.

http://www.siam.org/journals/simax/21-2/32027.html

†Department of Statistics, University of Dortmund, Vogelpothsweg 87, D-44221 Dortmund, Germany (gross@amadeus.statistik.uni-dortmund.de, trenkler@amadeus.statistik.uni-dortmund.de).

In view of

$$\begin{pmatrix} \mathbf{I}_n & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_n & \mathbf{0} \\ -\mathbf{Q} & \mathbf{0} & \mathbf{I}_n \end{pmatrix} \begin{pmatrix} \mathbf{0} & \mathbf{P} & -\mathbf{Q} \\ \mathbf{P} & \mathbf{P} & \mathbf{0} \\ -\mathbf{Q} & \mathbf{0} & -\mathbf{Q} \end{pmatrix} \begin{pmatrix} -\mathbf{I}_n & -\mathbf{P} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\mathbf{I}_n \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \mathbf{P} & \mathbf{Q} \\ \mathbf{P} & \mathbf{0} & \mathbf{0} \\ \mathbf{Q} & \mathbf{0} & \mathbf{0} \end{pmatrix},$$

(2.1) follows. Moreover, by the identities

$$\mathrm{rk}\begin{pmatrix} \mathbf{P} \\ \mathbf{Q} \end{pmatrix} - \mathrm{rk}(\mathbf{P}) = \mathrm{rk}[\mathbf{Q}(\mathbf{I}_n - \mathbf{P})] = n - \mathrm{rk}(\mathbf{P}) - \dim[\mathcal{N}(\mathbf{P}) \cap \mathcal{N}(\mathbf{Q})]$$

and

$$\mathrm{rk}(\mathbf{P}, \mathbf{Q}) - \mathrm{rk}(\mathbf{Q}) = \mathrm{rk}[(\mathbf{I}_n - \mathbf{Q})\mathbf{P}] = \mathrm{rk}(\mathbf{P}) - \dim[\mathcal{R}(\mathbf{P}) \cap \mathcal{R}(\mathbf{Q})]$$

(cf. Marsaglia and Styan [5, Theorem 19 and Corollary 6.2]) (2.2) follows from (2.1). □

As it will be easily seen from the above Theorem 1, $\mathbf{P} - \mathbf{Q}$ is nonsingular if and only if the column spaces of $\mathbf{P}$ and $\mathbf{Q}$ are complementary and the row spaces of $\mathbf{P}$ and $\mathbf{Q}$ are complementary.

COROLLARY 1. *Let* $\mathbf{P}$ *and* $\mathbf{Q}$ *be two* $n \times n$ *projectors. Then the following statements are equivalent*:

(i) $\mathrm{rk}(\mathbf{P} - \mathbf{Q}) = n$,

(ii) $\mathcal{R}(\mathbf{P}) \cap \mathcal{R}(\mathbf{Q}) = \{\mathbf{0}\}$ *and* $\mathcal{N}(\mathbf{P}) \cap \mathcal{N}(\mathbf{Q}) = \{\mathbf{0}\}$,

(iii) $\mathcal{R}(\mathbf{P}) \oplus \mathcal{R}(\mathbf{Q}) = \mathbb{R}^n$ *and* $\mathcal{R}(\mathbf{P}') \oplus \mathcal{R}(\mathbf{Q}') = \mathbb{R}^n$.

Proof. The equivalence between (i) and (ii) follows immediately from (2.2). Clearly (iii) implies $\mathcal{R}(\mathbf{P}) \cap \mathcal{R}(\mathbf{Q}) = \{\mathbf{0}\}$ and $[\mathcal{R}(\mathbf{P}') + \mathcal{R}(\mathbf{Q}')]^{\perp} = \{\mathbf{0}\}$, i.e., (ii). On the other hand, if (ii) holds, then from the equivalence between (ii) and (i) we have $\mathrm{rk}(\mathbf{P}' - \mathbf{Q}') = n$, and thus, in addition to (ii), we obtain $\mathcal{R}(\mathbf{P}') \cap \mathcal{R}(\mathbf{Q}') = \{\mathbf{0}\}$ and $\mathcal{N}(\mathbf{P}') \cap \mathcal{N}(\mathbf{Q}') = \{\mathbf{0}\}$. But then (iii) also holds. □

When $\mathbf{P}$ and $\mathbf{Q}$ are orthogonal projectors, then from condition (iii) in Corollary 1 we note that $\mathrm{rk}(\mathbf{P} - \mathbf{Q}) = n$ if and only if $\mathcal{R}(\mathbf{P}) \oplus \mathcal{R}(\mathbf{Q}) = \mathbb{R}^n$. This fact has been established by Ipsen and Meyer [4] and in connection with Hilbert spaces by Buckholtz [1].

From condition (iii) of Corollary 1 it is clear that nonsingularity of the difference $\mathbf{P} - \mathbf{Q}$ entails existence of the oblique projector on $\mathcal{R}(\mathbf{P})$ along $\mathcal{R}(\mathbf{Q})$ as well as existence of the oblique projector on $\mathcal{R}(\mathbf{P}')$ along $\mathcal{R}(\mathbf{Q}')$. Computation of these projectors is quite easy, as shown by the following corollary.

COROLLARY 2. *Let* $\mathbf{P}$ *and* $\mathbf{Q}$ *be two* $n \times n$ *projectors. If* $\mathbf{P} - \mathbf{Q}$ *is nonsingular, then* $\mathbf{P}(\mathbf{P} - \mathbf{Q})^{-1}$ *is the projector on* $\mathcal{R}(\mathbf{P})$ *along* $\mathcal{R}(\mathbf{Q})$ *and* $\mathbf{P}'(\mathbf{P}' - \mathbf{Q}')^{-1}$ *is the projector on* $\mathcal{R}(\mathbf{P}')$ *along* $\mathcal{R}(\mathbf{Q}')$.

Proof. The identity $\mathbf{P}(\mathbf{P} - \mathbf{Q})^{-1}(\mathbf{P} - \mathbf{Q}) = \mathbf{P}$ may be reformulated as

$$\mathbf{P}(\mathbf{P} - \mathbf{Q})^{-1}\mathbf{P} - \mathbf{P} = \mathbf{P}(\mathbf{P} - \mathbf{Q})^{-1}\mathbf{Q}.$$

By noting that $\mathcal{R}(\mathbf{P}') \cap \mathcal{R}(\mathbf{Q}') = \{\mathbf{0}\}$, this identity is seen to hold if and only if $\mathbf{P}(\mathbf{P} - \mathbf{Q})^{-1}\mathbf{P} = \mathbf{P}$, and $\mathbf{P}(\mathbf{P} - \mathbf{Q})^{-1}\mathbf{Q} = \mathbf{0}$. Thus $\mathbf{P}(\mathbf{P} - \mathbf{Q})^{-1}$ is the projector on $\mathcal{R}(\mathbf{P})$ along $\mathcal{R}(\mathbf{Q})$; cf. Trenkler [7, Theorem 7(iii)]. Similarly $\mathbf{P}'(\mathbf{P}' - \mathbf{Q}')^{-1}$ is seen to be the projector on $\mathcal{R}(\mathbf{P}')$ along $\mathcal{R}(\mathbf{Q}')$. □

The following result shows that the difference of two commuting projectors can only be nonsingular when the projectors add up to identity. In other words, for a given

projector $\mathbf{P}$, there exists only one projector $\mathbf{Q}$ with the property that $\mathbf{Q}$ commutes with $\mathbf{P}$ and $\mathbf{P}-\mathbf{Q}$ is nonsingular, namely, $\mathbf{Q}=\mathbf{I}_n-\mathbf{P}$.

COROLLARY 3. *Let* $\mathbf{P}$ *and* $\mathbf{Q}$ *be two commuting* $n\times n$ *projectors. Then the following statements are equivalent*:

(i) $\mathrm{rk}(\mathbf{P}-\mathbf{Q})=n$,

(ii) $\mathbf{P}-\mathbf{Q}$ *is an involution, i.e.,* $(\mathbf{P}-\mathbf{Q})^2=\mathbf{I}_n$,

(iii) $\mathbf{P}+\mathbf{Q}=\mathbf{I}_n$.

Proof. Since projectors are diagonalizable, it is clear that $\mathbf{P}$ and $\mathbf{Q}$ commute if and only if they are simultaneously diagonalizable. Then, however, the equivalence between the statements follows immediately. □

3. Rank of the sum. As in the previous section, we will now derive the rank of the sum $\mathbf{P}+\mathbf{Q}$ of two projectors $\mathbf{P}$ and $\mathbf{Q}$. From our derivations, it will become obvious that nonsingularity of $\mathbf{P}-\mathbf{Q}$ implies nonsingularity of $\mathbf{P}+\mathbf{Q}$.

THEOREM 2. *Let* $\mathbf{P}$ *and* $\mathbf{Q}$ *be two* $n\times n$ *projectors. Then*

$$\mathrm{rk}(\mathbf{P}+\mathbf{Q})=\mathrm{rk}(\mathbf{P})+\mathrm{rk}[(\mathbf{I}_n-\mathbf{P})\mathbf{Q}(\mathbf{I}_n-\mathbf{P})] \tag{3.1}$$

$$=n-\dim\left(\mathcal{N}[(\mathbf{I}_n-\mathbf{P})\mathbf{Q}]\cap\mathcal{N}(\mathbf{P})\right), \tag{3.2}$$

where $\mathcal{N}[(\mathbf{I}_n-\mathbf{P})\mathbf{Q}]=\mathcal{N}(\mathbf{Q})\oplus[\mathcal{R}(\mathbf{P})\cap\mathcal{R}(\mathbf{Q})]$.

Proof. From Marsaglia and Styan [5, Theorem 8] we see that

$$\mathrm{rk}(\mathbf{P}+\mathbf{Q})=\mathrm{rk}\begin{pmatrix}\mathbf{0}&\mathbf{P}&\mathbf{Q}\\ \mathbf{P}&\mathbf{P}&\mathbf{0}\\ \mathbf{Q}&\mathbf{0}&\mathbf{Q}\end{pmatrix}-\mathrm{rk}(\mathbf{P})-\mathrm{rk}(\mathbf{Q}).$$

In view of

$$\begin{pmatrix}\mathbf{I}_n&\mathbf{0}&\mathbf{0}\\ \mathbf{0}&\mathbf{I}_n&\mathbf{0}\\ \mathbf{Q}&\mathbf{0}&\mathbf{I}_n\end{pmatrix}\begin{pmatrix}\mathbf{0}&\mathbf{P}&\mathbf{Q}\\ \mathbf{P}&\mathbf{P}&\mathbf{0}\\ \mathbf{Q}&\mathbf{0}&\mathbf{Q}\end{pmatrix}\begin{pmatrix}\mathbf{I}_n&-\mathbf{P}&\mathbf{0}\\ \mathbf{0}&\mathbf{I}_n&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathbf{I}_n\end{pmatrix}=\begin{pmatrix}\mathbf{0}&\mathbf{P}&\mathbf{Q}\\ \mathbf{P}&\mathbf{0}&\mathbf{0}\\ \mathbf{Q}&\mathbf{0}&\mathbf{2Q}\end{pmatrix},$$

we obtain

$$\mathrm{rk}(\mathbf{P}+\mathbf{Q})=\mathrm{rk}\begin{pmatrix}\mathbf{0}&\mathbf{X}\\ \mathbf{Y}&\mathbf{S}\end{pmatrix}-\mathrm{rk}(\mathbf{P})-\mathrm{rk}(\mathbf{Q}),$$

where

$$\mathbf{X}=(\mathbf{P},\mathbf{Q}),\mathbf{Y}=\begin{pmatrix}\mathbf{P}\\ \mathbf{Q}\end{pmatrix}\quad\text{and}\quad\mathbf{S}=\begin{pmatrix}\mathbf{0}&\mathbf{0}\\ \mathbf{0}&2\mathbf{Q}\end{pmatrix}.$$

By using

$$\mathbf{S}^-=\begin{pmatrix}\mathbf{0}&\mathbf{0}\\ \mathbf{0}&(1/2)\mathbf{Q}\end{pmatrix}$$

as a generalized inverse for $\mathbf{S}$, i.e., $\mathbf{S}\mathbf{S}^-\mathbf{S}=\mathbf{S}$, and by applying Marsaglia and Styan [5, Theorem 19, equation (8.4)], it follows that

$$\mathrm{rk}(\mathbf{P}+\mathbf{Q})=\mathrm{rk}(\mathbf{S})+\mathrm{rk}(\mathbf{U})+\mathrm{rk}(\mathbf{V})+\mathrm{rk}(\mathbf{W})-\mathrm{rk}(\mathbf{P})-\mathrm{rk}(\mathbf{Q}),$$

where

$$\mathbf{U} = \mathbf{X}(\mathbf{I}_{2n} - \mathbf{S}^{-}\mathbf{S}) = (\mathbf{P}, \mathbf{0}), \qquad \mathbf{V} = (\mathbf{I}_{2n} - \mathbf{S}\mathbf{S}^{-}) \begin{pmatrix} \mathbf{P} \\ \mathbf{Q} \end{pmatrix} = \begin{pmatrix} \mathbf{P} \\ \mathbf{0} \end{pmatrix}$$

and

$$\mathbf{W} = (\mathbf{I}_n - \mathbf{U}\mathbf{U}^{-})\mathbf{X}\mathbf{S}^{-}\mathbf{Y}(\mathbf{I}_n - \mathbf{V}^{-}\mathbf{V}) = (1/2)(\mathbf{I}_n - \mathbf{P})\mathbf{Q}(\mathbf{I}_n - \mathbf{P})$$

for the choices $\mathbf{U}^{-} = (\mathbf{P}', \mathbf{0})'$ and $\mathbf{V}^{-} = (\mathbf{P}, \mathbf{0})$. In view of $\mathrm{rk}(\mathbf{S}) = \mathrm{rk}(\mathbf{Q})$ and $\mathrm{rk}(\mathbf{U}) = \mathrm{rk}(\mathbf{V}) = \mathrm{rk}(\mathbf{P})$, (3.1) follows. Since, in view of Marsaglia and Styan [5, Corollary 6.2],

$$\mathrm{rk}(\mathbf{I}_n - \mathbf{P}) - \mathrm{rk}[(\mathbf{I}_n - \mathbf{P})\mathbf{Q}(\mathbf{I}_n - \mathbf{P})] = \dim \left(\mathcal{N}[(\mathbf{I}_n - \mathbf{P})\mathbf{Q}] \cap \mathcal{N}(\mathbf{P}) \right),$$

(3.2) follows from (3.1). The inclusion $\mathcal{N}(\mathbf{Q}) \oplus [\mathcal{R}(\mathbf{P}) \cap \mathcal{R}(\mathbf{Q})] \subseteq \mathcal{N}[(\mathbf{I}_n - \mathbf{P})\mathbf{Q}]$ has, for example, been proven in Groß and Trenkler [3, Lemma 2(b)]. The reverse inclusion is also easy. □

We may now give a characterization of the nonsingularity of the sum $\mathbf{P} + \mathbf{Q}$ of two projectors $\mathbf{P}$ and $\mathbf{Q}$ in terms of disjointedness of subspaces similar to the characterization of nonsingularity of $\mathbf{P} - \mathbf{Q}$ in Corollary 1(ii).

COROLLARY 4. *Let* $\mathbf{P}$ *and* $\mathbf{Q}$ *be two* $n \times n$ *projectors. Then the following statements are equivalent*:

(i) $\mathrm{rk}(\mathbf{P} + \mathbf{Q}) = n$,

(ii) $\mathcal{R}(\mathbf{P}) \cap \mathcal{R}[\mathbf{Q}(\mathbf{I}_n - \mathbf{P})] = \{\mathbf{0}\}$ *and* $\mathcal{N}(\mathbf{P}) \cap \mathcal{N}(\mathbf{Q}) = \{\mathbf{0}\}$.

Proof. By using $\mathcal{N}(\mathbf{P}) = \mathcal{R}(\mathbf{I}_n - \mathbf{P})$, it follows from Marsaglia and Styan [5, Corollary 6.2] that the condition $\mathcal{N}(\mathbf{P}) \cap \mathcal{N}(\mathbf{Q}) = \{\mathbf{0}\}$ is equivalent to $\mathrm{rk}[\mathbf{Q}(\mathbf{I}_n - \mathbf{P})] = \mathrm{rk}(\mathbf{I}_n - \mathbf{P})$. Similarly, the condition $\mathcal{R}(\mathbf{P}) \cap \mathcal{R}[\mathbf{Q}(\mathbf{I}_n - \mathbf{P})] = \{\mathbf{0}\}$ is seen to be equivalent to $\mathrm{rk}[(\mathbf{I}_n - \mathbf{P})\mathbf{Q}(\mathbf{I}_n - \mathbf{P})] = \mathrm{rk}[\mathbf{Q}(\mathbf{I}_n - \mathbf{P})]$. Hence, statement (ii) implies $\mathrm{rk}[(\mathbf{I}_n - \mathbf{P})\mathbf{Q}(\mathbf{I}_n - \mathbf{P})] + \mathrm{rk}(\mathbf{P}) = n$, which in view of (3.1) from Theorem 2 gives (i). Conversely, from Theorem 2, statement (i) implies $\mathrm{rk}[(\mathbf{I}_n - \mathbf{P})\mathbf{Q}(\mathbf{I}_n - \mathbf{P})] = \mathrm{rk}(\mathbf{I}_n - \mathbf{P})$ and equivalently $\mathcal{N}[(\mathbf{I}_n - \mathbf{P})\mathbf{Q}] \cap \mathcal{N}(\mathbf{P}) = \{\mathbf{0}\}$. Since the latter obviously entails $\mathcal{N}(\mathbf{Q}) \cap \mathcal{N}(\mathbf{P}) = \{\mathbf{0}\}$, i.e., $\mathrm{rk}[\mathbf{Q}(\mathbf{I}_n - \mathbf{P})] = \mathrm{rk}(\mathbf{I}_n - \mathbf{P})$, it also follows $\mathrm{rk}[(\mathbf{I}_n - \mathbf{P})\mathbf{Q}(\mathbf{I}_n - \mathbf{P})] = \mathrm{rk}[\mathbf{Q}(\mathbf{I}_n - \mathbf{P})]$, showing that (i) implies (ii). □

Note that for two orthogonal projectors $\mathbf{P}$ and $\mathbf{Q}$ we always have $\mathrm{rk}[(\mathbf{I}_n - \mathbf{P})\mathbf{Q}(\mathbf{I}_n - \mathbf{P})] = \mathrm{rk}[\mathbf{Q}(\mathbf{I}_n - \mathbf{P})]$, and therefore condition (ii) of Corollary 4 would reduce to $\mathcal{N}(\mathbf{P}) \cap \mathcal{N}(\mathbf{Q}) = \{\mathbf{0}\}$, i.e., $\mathcal{R}(\mathbf{P}) + \mathcal{R}(\mathbf{Q}) = \mathbb{R}^n$.

By comparing the respective conditions (ii) of Corollaries 1 and 4, it is clear that nonsingularity of $\mathbf{P} - \mathbf{Q}$ implies nonsingularity of $\mathbf{P} + \mathbf{Q}$. The fact that nonsingularity of the sum of two projectors does *not* imply nonsingularity of the difference can be seen, for example, from

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & a \\ 0 & 1 & b \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{Q} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

for arbitrary scalars a and b. Obviously nonsingularity of $\mathbf{P} + \mathbf{Q}$ implies neither disjointness of $\mathcal{R}(\mathbf{P})$ and $\mathcal{R}(\mathbf{Q})$ nor disjointness of $\mathcal{R}(\mathbf{P}')$ and $\mathcal{R}(\mathbf{Q}')$, both conditions being necessary for nonsingularity of $\mathbf{P} - \mathbf{Q}$.

4. Rank of other transforms. In this section we give the rank of the transforms $\mathbf{I}_n - \mathbf{QP}$ and $\mathbf{P} + \mathbf{Q} - \mathbf{QP}$ of two projectors $\mathbf{P}$ and $\mathbf{Q}$. As it will turn out, the rank of $\mathbf{P} + \mathbf{Q} - \mathbf{QP}$ coincides with the rank of $\mathbf{P} + \mathbf{Q}$. Moreover nonsingularity of $\mathbf{P} + \mathbf{Q} - \mathbf{QP}$ and $\mathbf{I}_n - \mathbf{QP}$ are seen to be equivalent to nonsingularity of $\mathbf{P} - \mathbf{Q}$.

THEOREM 3. *Let* $\mathbf{P}$ *and* $\mathbf{Q}$ *be two* $n \times n$ *projectors. Then*

$$\operatorname{rk}(\mathbf{I}_n - \mathbf{QP}) = \operatorname{rk}[(\mathbf{I}_n - \mathbf{P}) + (\mathbf{I}_n - \mathbf{Q})], \tag{4.1}$$

$$\operatorname{rk}(\mathbf{P} + \mathbf{Q} - \mathbf{QP}) = \operatorname{rk}(\mathbf{P} + \mathbf{Q}). \tag{4.2}$$

Proof. From (3.1) in Theorem 2,

$$\begin{aligned}\operatorname{rk}[(\mathbf{I}_n - \mathbf{P}) + (\mathbf{I}_n - \mathbf{Q})] &= \operatorname{rk}(\mathbf{I}_n - \mathbf{P}) + \operatorname{rk}[\mathbf{P}(\mathbf{I}_n - \mathbf{Q})\mathbf{P}] \\ &= \operatorname{rk}(\mathbf{I}_n - \mathbf{P}) + \operatorname{rk}[\mathbf{P}(\mathbf{I}_n - \mathbf{QP})].\end{aligned}$$

On the other hand,

$$\operatorname{rk}[\mathbf{P}(\mathbf{I}_n - \mathbf{QP})] = \operatorname{rk}(\mathbf{I}_n - \mathbf{QP}) - \dim[\mathcal{N}(\mathbf{P}) \cap \mathcal{R}(\mathbf{I}_n - \mathbf{QP})];$$

cf. Marsaglia and Styan [5, Corollary 6.2]. Since obviously $\mathcal{N}(\mathbf{P}) \subseteq \mathcal{R}(\mathbf{I}_n - \mathbf{QP})$, we have

$$\dim[\mathcal{N}(\mathbf{P}) \cap \mathcal{R}(\mathbf{I}_n - \mathbf{QP})] = \dim[\mathcal{N}(\mathbf{P})] = \operatorname{rk}(\mathbf{I}_n - \mathbf{P}),$$

which shows (4.1). Equation (4.2) follows from (4.1) by noting that $\mathbf{P} + \mathbf{Q} - \mathbf{QP} = \mathbf{I}_n - (\mathbf{I}_n - \mathbf{Q})(\mathbf{I}_n - \mathbf{P})$. □

We may now give a characterization of the nonsingularity of $\mathbf{P} - \mathbf{Q}$ in terms of the transforms $\mathbf{I}_n - \mathbf{QP}$ and $\mathbf{P} + \mathbf{Q} - \mathbf{QP}$.

COROLLARY 5. *Let* $\mathbf{P}$ *and* $\mathbf{Q}$ *be two* $n \times n$ *projectors. Then the following statements are equivalent*:

(i) $\operatorname{rk}(\mathbf{P} - \mathbf{Q}) = n$,

(ii) $\operatorname{rk}(\mathbf{I}_n - \mathbf{QP}) = n$ *and* $\operatorname{rk}(\mathbf{P} + \mathbf{Q} - \mathbf{QP}) = n$.

Proof. As mentioned before, by comparing the corresponding conditions (ii) in Corollaries 1 and 4, it is clear that (i) implies $\operatorname{rk}(\mathbf{P} + \mathbf{Q}) = n$, where $\operatorname{rk}(\mathbf{P} + \mathbf{Q}) = \operatorname{rk}(\mathbf{P} + \mathbf{Q} - \mathbf{QP})$ by Theorem 3. Since $\operatorname{rk}(\mathbf{P} - \mathbf{Q}) = \operatorname{rk}[(\mathbf{I}_n - \mathbf{P}) - (\mathbf{I}_n - \mathbf{Q})]$, (i) implies $\operatorname{rk}[(\mathbf{I}_n - \mathbf{P}) + (\mathbf{I}_n - \mathbf{Q})] = n$, where $\operatorname{rk}[(\mathbf{I}_n - \mathbf{P}) + (\mathbf{I}_n - \mathbf{Q})] = \operatorname{rk}(\mathbf{I}_n - \mathbf{QP})$ from Theorem 3. Hence (i) implies (ii). Conversely, if (ii) is satisfied, then in view of $\operatorname{rk}(\mathbf{P}+\mathbf{Q}) = \operatorname{rk}(\mathbf{P}+\mathbf{Q}-\mathbf{QP})$ we obtain $\mathcal{N}(\mathbf{P})\cap\mathcal{N}(\mathbf{Q}) = \{\mathbf{0}\}$ from Corollary 4. Similarly in view of $\operatorname{rk}(\mathbf{I}_n-\mathbf{QP}) = \operatorname{rk}[(\mathbf{I}_n-\mathbf{P})+(\mathbf{I}_n-\mathbf{Q})]$ we obtain $\mathcal{N}(\mathbf{I}_n-\mathbf{P})\cap\mathcal{N}(\mathbf{I}_n-\mathbf{Q}) = \{\mathbf{0}\}$. Hence by Corollary 1(ii), statement (ii) implies (i). □

In addition to Corollary 2, we may state the following result.

COROLLARY 6. *Let* $\mathbf{P}$ *and* $\mathbf{Q}$ *be two* $n\times n$ *projectors. If* $\mathbf{P}-\mathbf{Q}$ *is nonsingular, then the projector on* $\mathcal{R}(\mathbf{P})$ *along* $\mathcal{R}(\mathbf{Q})$ *may be written in the form* $(\mathbf{I}_n - \mathbf{QP})^{-1}(\mathbf{I}_n - \mathbf{Q})$ *or in the form* $\mathbf{P}(\mathbf{P} + \mathbf{Q} - \mathbf{QP})^{-1}$.

Proof. Since $\mathbf{P}(\mathbf{P} - \mathbf{Q})^{-1}(\mathbf{P} - \mathbf{QP}) = \mathbf{P}$ and $\mathbf{P}(\mathbf{P} - \mathbf{Q})^{-1}\mathbf{Q} = \mathbf{0}$, cf. the proof of Corollary 2, we obtain $\mathbf{P}(\mathbf{P} - \mathbf{Q})^{-1}(\mathbf{P} + \mathbf{Q} - \mathbf{QP}) = \mathbf{P}$, such that by Corollary 5 we have $\mathbf{P}(\mathbf{P} - \mathbf{Q})^{-1} = \mathbf{P}(\mathbf{P} + \mathbf{Q} - \mathbf{QP})^{-1}$, which by Corollary 2 is the projector on $\mathcal{R}(\mathbf{P})$ along $\mathcal{R}(\mathbf{Q})$. □

Corollary 6 generalizes a result by Greville [2], who has shown the assertion for orthogonal projectors $\mathbf{P}$ and $\mathbf{Q}$. Recall that for orthogonal projectors the condition $\operatorname{rk}(\mathbf{P} - \mathbf{Q}) = n$ is equivalent to $\mathcal{R}(\mathbf{P}) \oplus \mathcal{R}(\mathbf{Q}) = \mathbb{R}^n$, which in turn is necessary and sufficient for existence of an (oblique) projector on $\mathcal{R}(\mathbf{P})$ along $\mathcal{R}(\mathbf{Q})$.

Acknowledgments. The authors are grateful to an anonymous referee for delivering rank formulas for transforms of projectors, which improved the presentation of the main results considerably.

REFERENCES

[1] D. Buckholtz, *Inverting the difference of Hilbert space projections*, Amer. Math. Monthly, 104 (1997), pp. 60–61.

[2] T.N.E. Greville, *Solutions of the matrix equation $XAX = X$, and relations between oblique and orthogonal projectors*, SIAM J. Appl. Math., 26 (1974), pp. 828–832.

[3] J. Groß and G. Trenkler, *On the product of oblique projectors*, Linear and Multilinear Algebra, 44 (1998), pp. 247–260.

[4] I.C.F. Ipsen and C.D. Meyer, *The angle between subspaces*, Amer. Math. Monthly, 102 (1995), pp. 904–911.

[5] G. Marsaglia and G.P.H. Styan, *Equalities and inequalities for ranks of matrices*, Linear and Multilinear Algebra, 2 (1974), pp. 269–292.

[6] C.R. Rao and S.K. Mitra, *Generalized Inverse of Matrices and Its Applications*, Wiley, New York, 1971.

[7] G. Trenkler, *Characterizations of oblique and orthogonal projectors*, in Proceedings of the International Conference on Linear Statistical Inference LINSTAT '93, Kluwer Academic Publishers, Dordrecht, 1994, pp. 255–270.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 396–417
© 1999 Society for Industrial and Applied Mathematics

TRANSFORMATION OF FAMILIES OF MATRICES TO NORMAL FORMS AND ITS APPLICATION TO STABILITY THEORY*

ALEXEI A. MAILYBAEV†

Abstract. Families of matrices smoothly depending on a vector of parameters are considered. Arnold [*Russian Math. Surveys*, 26 (1971), pp. 29–43] and Galin [*Uspekhi Mat. Nauk*, 27 (1972), pp. 241–242] have found and listed normal forms of families of complex and real matrices (miniversal deformations), to which any family of matrices can be transformed in the vicinity of a point in the parameter space by a change of basis, smoothly dependent on a vector of parameters, and by a smooth change of parameters. In this paper a constructive method of determining functions describing a change of basis and a change of parameters, transforming an arbitrary family to the miniversal deformation, is suggested. Derivatives of these functions with respect to parameters are determined from a recurrent procedure using derivatives of the functions of lower orders and derivatives of the family of matrices. Then the functions are found as Taylor series. Examples are given. The suggested method allows using efficiently miniversal deformations for investigation of different properties of matrix families. This is shown in the paper where tangent cones (linear approximations) to the stability domain at the singular boundary points are found.

Key words. family of matrices, normal form, versal deformation, stability, tangent cone

AMS subject classifications. 15A21, 15A90, 58F36

PII. S0895479898338378

1. Introduction.

The paper deals with the study of families of complex and real matrices (a family of matrices is a matrix-valued function $A(p)$ holomorphically (smoothly) depending on a vector p of complex (real) parameters). It is known that from all families of matrices we can select such families that transformation of any family to them can be carried out by a change of basis, smoothly depending on parameters, and by a smooth change of parameters [1, 2, 3]. Such families are called versal deformations. Versal deformations with the minimum possible number of parameters are called miniversal deformations or normal forms. Miniversal deformations of complex matrices have been found by Arnold [1, 2, 3]. He showed that the miniversal deformation, to which a family of matrices $A(p)$ can be transformed in the vicinity of a point $p = p_0$, is determined by the Jordan form of the matrix $A(p_0)$. Miniversal deformations of real matrices have been studied by Galin [8].

The important part of the transformation of a family of matrices to the normal form is to find functions describing a change of basis, depending on parameters, and a change of parameters. Determination of these functions is necessary for effective use of normal forms, but no general algorithm for finding them is available yet. Efforts of construction of the transformation functions required in a specific case have been done in [5, 12]. In the case when the matrix $A_0 = A(p_0)$ has only nonderogatory eigenvalues, the problem of construction of transformation to the normal form has been discussed in [13]. However, the method of finding the transformation functions suggested in [13] can be used only for matrices of small dimensions or in the case when the matrix A_0 has only one nonderogatory eigenvalue. Otherwise it leads to a

*Received by the editors May 6, 1998; accepted for publication (in revised form) by M. Overton February 22, 1999; published electronically November 2, 1999.
http://www.siam.org/journals/simax/21-2/33837.html
†Institute of Mechanics, Moscow State Lomonosov University, Michurinsky pr. 1, 117192 Moscow, Russia (mailybaev@inmech.msu.su).

complicated system of nonlinear equations containing parameters of the family and parameters of the normal form.

Another approach has been suggested in [14], where the transformation functions have been found at a specific value of the parameter vector p by solving a nonlinear implicit matrix equation using secant, updating secant, or using Newton's method. Such a numerical approach, though useful for finding functions in a finite domain (functions are calculated at several points and then approximated), is not efficient in the case under consideration. This is connected with the local character of miniversal deformations, which are defined in the vicinity of a point in the parameter space and used for investigation of local properties of matrix families (singularities, bifurcations, etc.). For local analysis, information about derivatives of the transformation functions with respect to parameters is required. Computation of these derivatives by means of approximate methods of [14] is very time-consuming and leads to numerical errors (especially when the size of the matrix A is not small). In [14] also a method for computation of derivatives of the transformation functions for a case of one-parameter matrix family has been suggested. This family had a special form allowing transformation to a normal form which was simpler than a miniversal deformation.

In the present paper a constructive method of finding functions describing the transformation of any given family of matrices to the normal form is suggested. The functions are found as Taylor series. Derivatives of the functions with respect to parameters, to determine coefficients of Taylor series, are found from a recurrent procedure using the derivatives of lower orders of these functions and derivatives of the family at the initial value of the vector of parameters. The recurrent procedure is convenient for numerical implementation, since it consists of elementary arithmetic operations. Proofs are based on properties of centralizers of matrices transformed to the Jordan form.

Transformation of a family of matrices to the normal form has great significance for applications. It allows studying specific families (normal forms), which consist usually of sparse matrices depending on parameters in a simple way, and then obtaining results for arbitrary families of matrices. Miniversal deformations (without knowledge about transformation to them) have been used in [1, 2, 3] for classification of singularities of bifurcation diagrams, decrement diagrams, and boundaries of the stability domains in the case of two and three parameters. In [10] miniversal deformations have been used for finding tangent cones (linear approximations) to the stability domain at singular boundary points up to a diffeomorphism. All previous results were qualitative. To obtain quantitative results we need information about the transformation. In [4] specific linear combinations of first derivatives of the functions describing a change of parameters have been found and used to obtain necessary conditions for the stable perturbations of matrices. Other applications of miniversal deformations can be found in [6, 7], where a bound for a distance to a less generic matrix pencil and backward error bound in the procedure of calculation of polynomial roots are obtained.

The method of finding functions of transformation, suggested in this paper, allows using miniversal deformations for quantitative study of different properties of matrices depending on parameters. Efficiency of the method for applications is illustrated by the problem of finding tangent cones (or parts of tangent cones) to the stability domain for singular boundary points. All types of singular boundary points are considered. The obtained results represent an extension of the paper [11], where, using another approach, tangent cones for singular boundary points of the stability domain of two- and three-parameter generic families of matrices have been found.

The paper is organized as follows. In section 2 the method for determining derivatives of functions describing the transformation of a family of complex matrices to miniversal deformation is formulated and proved. Section 3 discusses the case of families of real matrices. Section 4 uses results of the previous sections to find tangent cones to the stability domain at singular boundary points. The conclusion gives a brief overview of the results obtained and discusses possibilities for their application to different problems.

2. Transformation of families of complex matrices to normal forms. For the following presentation we need definitions given by Arnold [1, 2], [3, p. 237].

DEFINITION 2.1. *A family of complex matrices is a holomorphic mapping* $A : P \longrightarrow \mathbf{C}^{n^2}$, *where* P *is the vicinity of some point* p_0 *in the parameter space* $\mathbf{C}^m$ *and* $\mathbf{C}^{n^2}$ *is the space of* $n \times n$ *complex matrices.*

Let us consider an arbitrary family of matrices $A(p),\ p = (p_1, \dots, p_m)^T \in \mathbf{C}^m$. Denote by λ_i the eigenvalues of the matrix $A_0 = A(p_0)$ and let $n_1(\lambda_i) \geq n_2(\lambda_i) \geq \cdots$ be the dimensions of the Jordan blocks belonging to λ_i, beginning with the largest one. Let the Jordan form J of the matrix A_0 be given by

$$A_0 = C_0 J C_0^{-1}, \tag{2.1}$$

where

$$J = \begin{pmatrix} J_1 & & \\ & J_2 & \\ & & \ddots \end{pmatrix},$$

$$J_i = \begin{pmatrix} J_i^1 & & \\ & J_i^2 & \\ & & \ddots \end{pmatrix},$$

and the Jordan block

$$J_i^k = \begin{pmatrix} \lambda_i & 1 & & & \\ & \cdot & \cdot & & \\ & & \cdot & \cdot & \\ & & & \cdot & 1 \\ & & & & \lambda_i \end{pmatrix}$$

has dimension $n_k(\lambda_i)$.

Denote by X_{ij}^{kl} the block of a square $n \times n$ matrix X, which is situated on the intersection of the rows corresponding to the block J_i^k and columns corresponding to the block J_j^l; see, for example, Figure 2.1. Denote elements of each block X_{ij}^{kl} by $x_{ij}^{kl}(r,s),\ r = 1, \dots, n_k(\lambda_i),\ s = 1, \dots, n_l(\lambda_j)$, where r, s are the numbers of a row and a column in X_{ij}^{kl} corresponding to $x_{ij}^{kl}(r,s)$. In the case $i = j$ we will leave only one number in the subscript, e.g., X_i^{kl}, $x_i^{kl}(r,s)$. The block of the matrix X which corresponds to the block J_i of the Jordan form J will be denoted by X_i.

DEFINITION 2.2. *A versal deformation of a matrix* A_0 *is a family of matrices* $A'(p'),\ p' \in \mathbf{C}^d$ *such that any family* $A(p),\ A(p_0) = A_0$ *in the vicinity of the point* $p = p_0$ *can be represented in the form*

$$A(p) = C(p) A'(\varphi(p)) C^{-1}(p), \tag{2.2}$$

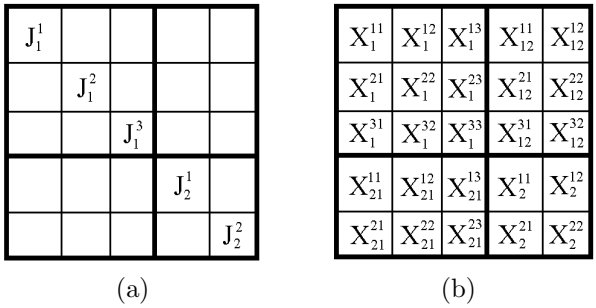


FIG. 2.1. *Block decomposition of a square matrix X according to the Jordan form J.*

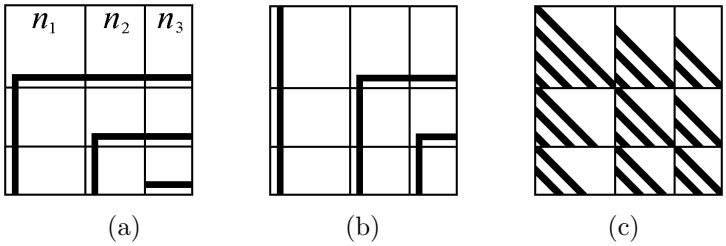


FIG. 2.2. *Location of parameters in blocks of a miniversal deformation.*

where $p' = \varphi(p)$, $\varphi(p_0) = \mathbf{0}$ is a holomorphic mapping from the vicinity of the point $p = p_0$ in the parameter space $\mathbf{C}^m$ into the vicinity of the origin of coordinate in the space $\mathbf{C}^d$; $C(p)$ is a family of nonsingular matrices describing a change of basis. A versal deformation is said to be miniversal if the dimension d of the parameter space is the smallest possible for a versal deformation [1, 2], [3, p. 237].

It is shown in [1, 2], [3, p. 238], that a miniversal deformation is determined by the Jordan form of the matrix A_0 and can be chosen in the form of a sum $J+B(p')$, where $B(p') = \operatorname{Diag}(B_1, B_2, \ldots)$ is a family of block-diagonal matrices with blocks $B_i(p')$ corresponding to the blocks J_i of the matrix J. There are different ways of choosing the blocks $B_i(p')$. Three possible types are shown in Figure 2.2, where the block $B_i(p')$ has all zeros except at the places indicated. In cases (a) and (b), the indicated places are filled by different components of the vector p' (parameters of the miniversal deformation); in case (c) each slanted segment is filled by a corresponding component of the vector p'. Other types of blocks are also possible, e.g., blocks whose elements are all zero, except for one element on each slanted segment in Figure 2.2(c) which is an independent parameter. In all cases, each block B_i^{kl} contains exactly $n_l(\lambda_i)$, if $k \le l$, and $n_k(\lambda_i)$, if $k > l$, different components of p' (one on each slanted segment in Figure 2.2(c)). Denote these components by $b_i^{kl}(t)$, $t = 1, \ldots, \min(n_k(\lambda_i), n_l(\lambda_i))$ numbered inside each block from the top and from the right (see examples for the case of two Jordan blocks of the size 3 and 2 in Figure 2.3). Components $b_i^{kl}(t)$ compose the vector of parameters of the miniversal deformation p' of the dimension $d = \sum_i (n_1(\lambda_i) + 3n_2(\lambda_i) + \cdots)$ [1, 2], [3, p. 237].

DEFINITION 2.3. *Define a generalized trace of a $k \times l$ matrix F denoted by $\operatorname{tr}^{(t)} F$ as a sum of elements on that diagonal of F which is $(\min(k, l) - t)$ positions above the left lower element*

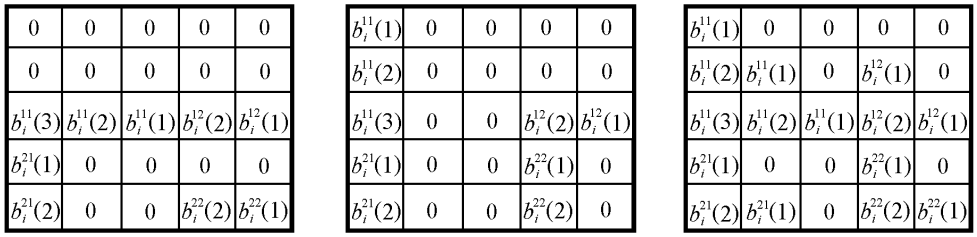


FIG. 2.3. *Numbering of parameters of a miniversal deformation.*

$$\operatorname{tr}^{(t)} F = \sum_{r=t}^{\min(k,l)} f(k - \min(k,l) + r,\ r - t + 1). \tag{2.3}$$

In the case of a square matrix ($k = l$), Definition 2.3 coincides with the definition of a generalized trace given in [4]. In particular, a generalized trace $\operatorname{tr}^{(\min(k,l))} F$ is equal to the left lower element of the matrix F, and a generalized trace $\operatorname{tr}^{(1)} F$ of a square matrix F coincides with the ordinary trace $\operatorname{tr} F$, which is equal to the sum of elements on the main diagonal. In the case of an $n \times n$ matrix F a generalized trace $\operatorname{tr}^{(t)}$ of a block F_i^{kl} is equal to the sum of elements on that slanted segment which is $(t-1)$ positions below the longest one in this block; see Figure 2.2,

$$\operatorname{tr}^{(t)} F_i^{kl} = \sum_{r=t}^{n'} f_i^{kl}(n_k(\lambda_i) - n' + r,\ r - t + 1), \quad n' = \min(n_k(\lambda_i), n_l(\lambda_i)).$$

For blocks of the matrix $B(p')$ we have $\operatorname{tr}^{(t)} B_i^{kl} = \alpha_i^{kl}(t) b_i^{kl}(t)$, where $\alpha_i^{kl}(t) = 1$, if B_i is a block of the type (a) or (b), and $\alpha_i^{kl}(t) = n' - t + 1$, if B_i is a block of the type (c); see Figure 2.2.

Denote by D^h the derivative

$$D^h = \frac{\partial^{|h|}}{\partial p_1^{h^1} \cdots \partial p_m^{h^m}}$$

taken at the point $p = p_0$, where $h = (h^1, \ldots, h^m)$, $h^i \geq 0$, is a vector whose components h^i denote the order of the partial derivative with respect to p_i ($h^i = 0$ means that the derivative with respect to p_i is not taken), $|h| = h^1 + \cdots + h^m$ is the order of the derivative.

Transformation of the family of matrices $A(p)$ to the normal form (miniversal deformation) (2.2) includes finding the family $A'(p')$ corresponding to the Jordan form J of the matrix A_0, families $C(p)$, $C^{-1}(p)$, and holomorphic mapping $p' = \varphi(p)$ (functions $b_i^{kl}(t)(p)$). To obtain $C(p)$, $C^{-1}(p)$, and $b_i^{kl}(t)(p)$, it is sufficient to determine derivatives $D^h C$, $D^h C^{-1}$, and $D^h b_i^{kl}(t)$ for all h. Then the desired functions are found locally as Taylor series (by virtue of holomorphy).

THEOREM 2.4. *Let $A(p)$ be a family of matrices. Then the following expressions are valid for derivatives D^h, $|h| > 0$ of the functions $C(p)$, $C^{-1}(p)$, $b_i^{kl}(t)(p)$ transforming the family of matrices $A(p)$ to the miniversal deformation $A'(p')$ described above* (2.2).

1°.

$$D^h b_i^{kl}(t) = \frac{\operatorname{tr}^{(t)} F_i^{kl}}{\alpha_i^{kl}(t)}, \tag{2.4}$$

$$\begin{aligned} F = C_0^{-1}\Bigg(D^h A - \sum_{\substack{h_1+h_2+h_3=h \\ h_1,h_2,h_3\neq h}} C_h^{h_1h_2h_3} D^{h_1}C\, D^{h_2}A'\, D^{h_3}C^{-1} \\ + A_0 \sum_{\substack{h_1+h_2=h \\ h_1,h_2\neq h}} C_h^{h_1h_2} D^{h_1}C\, D^{h_2}C^{-1} \Bigg) C_0, \end{aligned} \tag{2.5}$$

$$C_h^{h_1h_2h_3} = \prod_{i=1}^{m} \frac{h^i!}{h_1^i!h_2^i!h_3^i!}, \quad C_h^{h_1h_2} = \prod_{i=1}^{m} \frac{h^i!}{h_1^i!h_2^i!}.$$

2°.

$$D^h C = C_0 X, \tag{2.6}$$

$$x_{ij}^{kl}(r,s) = \sum_{r_1=r}^{n_k(\lambda_i)} \sum_{s_1=1}^{s} \frac{(-1)^{r_1-r}}{(\lambda_i-\lambda_j)^{r_1-r+s-s_1+1}}\, C_{r_1-r+s-s_1}^{r_1-r} g_{ij}^{kl}(r_1,s_1),\ i\neq j, \tag{2.7}$$

$$C_n^k = \frac{n!}{k!(n-k)!},$$

$$x_i^{kl}(r,s) = \begin{cases} 0, & k\le l,\ r=1, \\ \displaystyle\sum_{\substack{s_1=s'(r,s) \\ r'(r,s)}} g_i^{kl}(r-s+s_1-1,s_1), & k\le l,\ r\neq 1, \\ \displaystyle -\sum_{r_1=r} g_i^{kl}(r_1,s-r+r_1+1), & k>l,\ s\neq n_l(\lambda_i), \\ 0, & k>l,\ s=n_l(\lambda_i), \end{cases} \tag{2.8}$$

$$s'(r,s) = \max(1, s-r+2), \quad r'(r,s) = \min(n_k(\lambda_i), r-s+n_l(\lambda_i)-1),$$

$$G = D^h A' - F. \tag{2.9}$$

3°.

$$D^h C^{-1} = -C_0^{-1} \sum_{\substack{h_1+h_2=h \\ h_2\neq h}} C_h^{h_1h_2} D^{h_1}C\, D^{h_2}C^{-1}. \tag{2.10}$$

The zero order derivatives D^h, $h=\mathbf{0}$ of functions C, C^{-1}, A' (values of functions at $p=p_0$) are $D^0C = C_0$, $D^0C^{-1} = C_0^{-1}$, $D^0A' = J$. Theorem 2.4 allows finding

derivatives D^h of the order $|h| > 0$ of functions C, C^{-1}, and $b_i^{kl}(t)$, using only derivatives of these functions of a lower order and the derivative $D^h A$. Thus, (2.4)–(2.10) represent a recurrent procedure for determining the derivatives $D^h C$, $D^h C^{-1}$, and $D^h b_i^{kl}(t)$ for any h using derivatives $D^{h_1} A$ of the order $|h_1| < |h|$ and $h_1 = h$.

If all derivatives $D^h C$, $D^h C^{-1}$, and $D^h b_i^{kl}(t)$ of the order $|h| \le s$ are determined, the desired functions can be found approximately as Taylor series containing terms with derivatives D^h of the order $|h| \le s$

$$C(p) = \sum_{0 \le |h| \le s} D^h C \prod_{r=1}^{m} \frac{(p_r - p_{0r})^{h^r}}{h^r!} + o(\|p - p_0\|^s),$$

$$C^{-1}(p) = \sum_{0 \le |h| \le s} D^h C^{-1} \prod_{r=1}^{m} \frac{(p_r - p_{0r})^{h^r}}{h^r!} + o(\|p - p_0\|^s),$$

$$b_i^{kl}(t)(p) = \sum_{0 \le |h| \le s} D^h b_i^{kl}(t) \prod_{r=1}^{m} \frac{(p_r - p_{0r})^{h^r}}{h^r!} + o(\|p - p_0\|^s).$$

Here $\|p\|$ is the norm in $\mathbf{C}^m$. Denoting these approximations by $\widetilde{C}(p)$, $\widetilde{C}^{-1}(p)$, and $\widetilde{\varphi}(p)$ we can represent the family of matrices $A(p)$ in the form $A(p) = \widetilde{C}(p) A'(\widetilde{\varphi}(p)) \widetilde{C}^{-1}(p) + o(\|p - p_0\|^s)$; that is, the recurrent procedure allows transforming an arbitrary given family of matrices $A(p)$ to the normal form in the vicinity of the point $p = p_0$ with the accuracy up to small terms of an arbitrary order.

Note that Burke and Overton [4] obtained the formulae for calculation specific sums of the first order derivatives of functions $b_i^{kl}(t)(p)$ in the case of one parameter $p \in \mathbf{R}$. These formulae can be derived from expressions (2.4) and (2.5).

2.1. Example. As an example let us consider a one-parameter family of 3×3 matrices $A(z)$ which has at the point $z = 0$ eigenvalues $\lambda_1 = 0$, $\lambda_2 = 2i$ with corresponding Jordan blocks of the order $n_1(\lambda_1) = 2$, $n_1(\lambda_2) = 1$, respectively,

$$A(z) = \begin{pmatrix} 2i + z & -8z & z^2 \\ 2iz & 4iz & 1 - iz^2 \\ 2z^2 & -2z & z^2 \end{pmatrix}.$$

The Jordan form of the matrix $A_0 = A(0)$ and matrices C_0, C_0^{-1} have the form

$$J = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2i \end{pmatrix}, \quad C_0 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad C_0^{-1} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$

For the first iteration ($h = 1$, $D^h = d/dz$) we have

$$F = \begin{pmatrix} f_1^{11}(1,1) & f_1^{11}(1,2) & f_{12}^{11}(1,1) \\ f_1^{11}(2,1) & f_1^{11}(2,2) & f_{12}^{11}(2,1) \\ f_{21}^{11}(1,1) & f_{21}^{11}(1,2) & f_2^{11}(1,1) \end{pmatrix} = \begin{pmatrix} 4i & 0 & 2i \\ -2 & 0 & 0 \\ -8 & 0 & 1 \end{pmatrix},$$

$$G = \begin{pmatrix} -4i & 0 & -2i \\ 0 & 4i & 0 \\ 8 & 0 & 0 \end{pmatrix}, \quad X = \begin{pmatrix} 0 & 0 & 1 \\ -4i & 0 & 0 \\ -4i & -2 & 0 \end{pmatrix}.$$

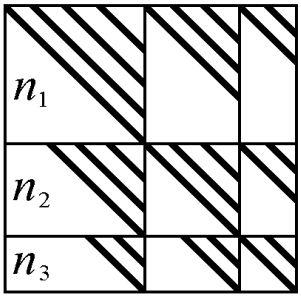


FIG. 2.4. *Block of a matrix commuting with the Jordan form J.*

These matrices are used in (2.4), (2.6), and (2.10) to determine the first derivatives of the desired functions. Then these derivatives are used in the second iteration to find the derivatives $D^h = d^2/dz^2$. As a result, the normal form $A'(p')$ of the family $A(z)$ and, determined by Theorem 2.4, the functions $p' = \varphi(z)$, $C(z)$, $C^{-1}(z)$ have the form

$$A'(p') = \begin{pmatrix} 0 & 1 & 0 \\ b_1(2) & b_1(1) & 0 \\ 0 & 0 & 2i + b_2(1) \end{pmatrix}, \quad p' = (b_1(1), b_1(2), b_2(1))^T,$$

$$b_1(1) = 4iz + 9z^2 + o(z^2), \quad b_1(2) = -2z + o(z^2), \quad b_2(1) = z - 8z^2 + o(z^2),$$

$$C(z) = \begin{pmatrix} -4iz + (2 - 2i)z^2 & -2z - (5 + 1.5i)z^2 & 1 \\ 1 & 0 & z + (2 + 0.5i)z^2 \\ -4iz - 8z^2 & 1 + 5iz^2 & 0 \end{pmatrix} + o(z^2),$$

$$C^{-1}(z) = \begin{pmatrix} -z - (2 + 0.5i)z^2 & 1 - 4iz^2 & -2z^2 \\ -4iz^2 & 4iz + 8z^2 & 1 - 5iz^2 \\ 1 - 4iz^2 & 4iz - (2 - 10i)z^2 & 2z + (5 + 1.5i)z^2 \end{pmatrix} + o(z^2).$$

2.2. Proof of the theorem.

DEFINITION 2.5. *The centralizer of a square matrix X is the set of all matrices commuting with X. The notation is as follows:*

$$Z_X = \{Y : [X, Y] = 0\},$$

where $[X, Y] = XY - YX$ [1, 2, 3].

A centralizer Z_J of the Jordan form J consists of block-diagonal matrices $Y = \text{Diag}\,(Y_1, Y_2, \ldots)$ whose blocks Y_i correspond to the blocks J_i and have the form shown in Figure 2.4 [1, 2, 3, 7]. In Figure 2.4 every slanted segment denotes a sequence of equal numbers, while other elements are all zeros. The centralizer Z_J is a plane in the space of all matrices $\mathbf{C}^{n^2}$ which passes through the zero and identity matrix.

LEMMA 2.6. *For any $n \times n$ matrix X and any i, k, l, t the following equality is valid:*

$$\text{tr}^{(t)}\left([J, X]_i^{kl}\right) = 0.$$

Proof. For any matrix $Y \in Z_J$ we have

$$\begin{aligned} &\operatorname{tr}([J,X]Y) = \operatorname{tr}(JXY - XJY) = \\ &\operatorname{tr}(YJX - JYX) = -\operatorname{tr}([J,Y]X) = 0. \end{aligned} \tag{2.11}$$

Consider the matrix Y which has units on the tth slanted segment (beginning from the largest one) in the block Y_i^{lk} (see Figure 2.4), while the other components are all zeros. Then, using (2.11), we obtain $0 = \operatorname{tr}([J,X]Y) = \operatorname{tr}^{(t)}\big([J,X]_i^{kl}\big)$. ∎

LEMMA 2.7. *For any $n \times n$ matrix G, which satisfies the condition $\operatorname{tr}^{(t)} G_i^{kl} = 0$ for all i,k,l,t, the equation*

$$[J,X] = G \tag{2.12}$$

has the unique solution (2.7), (2.8), *satisfying*

$$\begin{aligned} x_i^{kl}(1,s) &= 0, \qquad k \le l, \\ x_i^{kl}(s, n_l(\lambda_i)) &= 0, \quad k > l, \end{aligned} \tag{2.13}$$

for all i,k,l,s.

Proof. In the case when G is the zero matrix a set of solutions of (2.12) is the centralizer Z_J. To satisfy conditions (2.13) the matrix $X \in Z_J$ must have zeros on the sth slanted segment (beginning from the smallest one for $k \le l$ and from the largest one for $k > l$) of the block X_i^{kl}; see Figure 2.4. Hence, the equation $[J,X] = 0$ has the unique solution $X = 0$ satisfying all conditions (2.13). Due to the linearity of equation (2.12) and conditions (2.13), if two matrices X_1 and X_2 are solutions of (2.12) and (2.13) for nonzero matrix G, then the matrix $X = X_1 - X_2$ satisfies the equations $[J,X] = 0$ and (2.13). Consequently, $X = X_1 - X_2 = 0$; that is, if a solution of (2.12) and (2.13) exists, it is unique.

By substituting the explicit form of J into (2.12) we obtain equations on the elements of the matrix X

$$(\lambda_i - \lambda_j) x_{ij}^{kl}(r,s) + x_{ij}^{kl}(r+1,s) - x_{ij}^{kl}(r,s-1) = g_{ij}^{kl}(r,s), \tag{2.14}$$

where, instead of all elements x_{ij}^{kl} with coordinates $(r+1,s)$ or $(r,s-1)$, which don't belong to the block X_{ij}^{kl}, we take zeros. The fact that (2.7) and (2.8) is a solution to equations (2.12) and (2.13) can be verified by substitution of expressions (2.7) and (2.8) into (2.13) and (2.14). After substitution, equations (2.14) for $i = j$, $r = n_k(\lambda_i)$, $s = n_l(\lambda_i) - t + 1$, $k \le l$ and for $i = j$, $r = t$, $s = 1$, $k > l$ take the form $\operatorname{tr}^{(t)} G_i^{kl} = 0$. They are satisfied by virtue of the conditions of the lemma. Other equations are satisfied identically. ∎

Proof of Theorem 2.4. The proof is divided into three parts according to the items of the theorem. Each part is denoted by a corresponding item number.

Item 3°. Let us take the derivative $D^h, |h| > 0$ from both sides of the equality $C(p)C^{-1}(p) = I$, where I is the identity matrix:

$$\sum_{h_1+h_2=h} C_h^{h_1 h_2} D^{h_1} C \, D^{h_2} C^{-1} = 0, \quad C_h^{h_1 h_2} = \prod_{i=1}^{m} \frac{h^i!}{h_1^i! h_2^i!}.$$

After expressing $C_0 D^h C^{-1}$ in terms of the other summands and multiplying both sides of the equation from the left by C_0^{-1}, we will prove item 3° of the theorem.

Item 1°. Let us take the derivative D^h from both sides of expression (2.2)

$$D^hA = \sum_{h_1+h_2+h_3=h} C_h^{h_1h_2h_3} D^{h_1}C\, D^{h_2}A' D^{h_3}C^{-1}, \quad C_h^{h_1h_2h_3} = \prod_{i=1}^{m} \frac{h^i!}{h_1^i!h_2^i!h_3^i!}.$$

Expressing $C_0D^hA'C_0^{-1}$ in terms of other summands and then multiplying both sides of the equation by C_0^{-1} and C_0 from left and from right, respectively, we obtain

$$D^hA' = C_0^{-1}D^hAC_0 - C_0^{-1} \sum_{\substack{h_1+h_2+h_3=h \\ h_2\neq h}} C_h^{h_1h_2h_3} D^{h_1}C\, D^{h_2}A' D^{h_3}C^{-1}\, C_0.$$

Using equation (2.10), proved above for the term containing D^hC^{-1}, we get

$$D^hA' = [J, C_0^{-1}D^hC] + F, \tag{2.15}$$

where the matrix F is defined in (2.5). Using Lemma 2.6 and taking into account that $\mathrm{tr}^{(t)}(D^hA_i'^{kl}) = \mathrm{tr}^{(t)}(D^hB_i^{kl}) = \alpha_i^{kl}(t)D^hb_i^{kl}(t)$ for $|h| > 0$, we will prove item 1° of the theorem.

Item 2°. Rearranging (2.15), we obtain

$$[J, C_0^{-1}D^hC] = G, \tag{2.16}$$

where a matrix G is defined in (2.9).

Note that $C_0^{-1}D^hC = D^h(C_0^{-1}C)$ is a derivative of the family of matrices $C_J = C_0^{-1}C$, $C_J(p_0) = I$, which describe a change of basis transforming the family $A_J(p) = C_0^{-1}A(p)C_0$ to the normal form, that is,

$$A_J(p) = C_0^{-1}A(p)C_0 = C_0^{-1}C(p)A'(\varphi(p))C^{-1}(p)C_0 = C_J(p)A'(\varphi(p))C_J^{-1}(p).$$

At the point $p = p_0$ the matrix $A_J(p_0) = C_0^{-1}A_0C_0 = J$ is in the Jordan form. In the proof of the versality theorem [1, 2], [3, pp. 238–241] it was shown that in this case the family C_J can be chosen lying on an arbitrary smooth surface of the dimension $n^2 - d$ in the space of all matrices which passes through the identity matrix and transversal at this point to the centralizer Z_J. For each such a surface the family C_J can be chosen in the unique way. In particular, the plane $I + \Pi$, where Π is the set of all matrices X satisfying conditions (2.13), has all these properties. The planes Z_J and $I + \Pi$ are transversal, because the only matrix from Z_J which satisfies conditions (2.13) (intersection of the planes) is the identity matrix and the sum of dimensions of the planes is equal to the dimension of the space n^2. If we choose the plane $I + \Pi$ for determining C_J, we obtain that all derivatives $D^hC_J = C_0^{-1}D^hC = X$ must satisfy conditions (2.13), which can be regarded as normality conditions. Then the formulae of item 2° follow from Lemma 2.7 applied to (2.16) (conditions $\mathrm{tr}^{(t)}G_i^{kl} = \mathrm{tr}^{(t)}(D^hA_i'^{kl}) - \mathrm{tr}^{(t)}F_i^{kl} = \alpha_i^{kl}(t)D^hb_i^{kl}(t) - \mathrm{tr}^{(t)}F_i^{kl} = 0$ are fulfilled by virtue of equality (2.4)). □

Note that instead of normality conditions (2.13) one can take the same amount of any other linearly independent homogeneous equations such that the only matrix from Z_J which satisfies them is the zero matrix. Theorem 2.4 and Lemma 2.7 will remain valid if (2.8) is changed in accordance with the new normality conditions by solving (2.14) for $i = j$, which determines elements of the blocks X_i^{kl}.

The case of families of complex matrices $A(p)$ depending on a vector of real parameters $p \in \mathbf{R}^m$ is the same as considered above except that the holomorphic dependence on parameters is changed to the smooth one. In this case, Theorem 2.4 is valid and allows finding derivatives D^h of functions C, C^{-1}, $b_i^{kl}(t)$ of the order $|h| \le s$, where s is the degree of smoothness of the family $A(p)$.

3. Transformation of families of real matrices to normal forms. Let us consider a family of real matrices $A : P \longrightarrow \mathbf{R}^{n^2}$, $P \subset \mathbf{R}^m$ smoothly depending on a vector of real parameters $p \in \mathbf{R}^m$ and determined in the vicinity of a point $p = p_0$. Let the matrix $A_0 = A(p_0)$ have real eigenvalues λ_i, $i = 1, \ldots, q$ and complex eigenvalues λ_i, $\overline{\lambda_i}$, $i = q+1, q+2, \ldots$ (overline denotes the complex conjugate operation). Then the Jordan form of the matrix A_0 is given by $J = \mathrm{Diag}\,(J_1, \ldots, J_q, J_{q+1}, \overline{J_{q+1}}, J_{q+2}, \overline{J_{q+2}}, \ldots)$, where the blocks J_i correspond to the eigenvalues λ_i and have the form $J_i = \mathrm{Diag}\,(J_i^1, J_i^2, \ldots)$; J_i^k are the Jordan blocks of dimensions $n_k(\lambda_i)$ beginning with the largest one.

Denote by X^{kl}_{ijA}, X^{kl}_{ijB}, X^{kl}_{ijC}, X^{kl}_{ijD} blocks of an arbitrary matrix X of the dimension $n \times n$ which are situated on the intersection of rows and columns belonging to the blocks J_i^k and J_j^l, J_i^k and $\overline{J_j^l}$, $\overline{J_i^k}$ and J_j^l, $\overline{J_i^k}$ and $\overline{J_j^l}$, respectively. Denote elements of these blocks by $x^{kl}_{ijA}(r,s)$, $x^{kl}_{ijB}(r,s)$, $x^{kl}_{ijC}(r,s)$, and $x^{kl}_{ijD}(r,s)$. In the case $i = j$ we will leave only one number in the subscript, e.g., X^{kl}_{iA}, $x^{kl}_{iA}(r,s)$.

Denote by $\mathrm{Rl}\,Y$ a decomplexification of a complex matrix Y [3, pp. 244–245]

$$\mathrm{Rl}\,Y = \begin{pmatrix} \mathrm{Re}\,Y & -\mathrm{Im}\,Y \\ \mathrm{Im}\,Y & \mathrm{Re}\,Y \end{pmatrix}.$$

Define a matrix $J_R = \mathrm{Diag}\,(J_1, \ldots, J_q, \mathrm{Rl}\,J_{q+1}, \mathrm{Rl}\,J_{q+2}, \ldots)$. The matrices J_R and J are connected by the relation $J_R = RJR^{-1}$, where $R = \mathrm{Diag}\,(I_1, \ldots, I_q, I'_{q+1}, I'_{q+2}, \ldots)$, $R^{-1} = \overline{R} = \mathrm{Diag}\,(I_1, \ldots, I_q, \overline{I'_{q+1}}, \overline{I'_{q+2}}, \ldots)$. The blocks I'_j have the form

$$I'_j = \frac{1+i}{2}\begin{pmatrix} I_j & -i\,I_j \\ -i\,I_j & I_j \end{pmatrix}, \quad i = \sqrt{-1},$$

where I_j is the identity matrix of the same dimension as J_j. Let us choose the matrix C_0 transforming A_0 to the Jordan form J (2.1) such that the matrices $C_{R0} = C_0 R^{-1}$ and $C_{R0}^{-1} = RC_0^{-1}$ are real. For this purpose it is sufficient to choose the matrix C_0 so that its columns, passing through the blocks corresponding to real eigenvalues J_j, $j = 1, \ldots, q$, are real, and columns, passing through the blocks corresponding to complex conjugate eigenvalues J_j and $\overline{J_j}$, $j = q+1, q+2, \ldots$, are complex conjugate. This is possible due to the reality of A_0. Then each column of the matrix C_{R0} will be equal to the sum of the real and imaginary parts of the corresponding column of the matrix C_0. Using C_{R0} the matrix A_0 can be transformed to the form J_R

$$A_0 = C_0 J C_0^{-1} = C_{R0} J_R C_{R0}^{-1}.$$

It is shown in [8], [3, pp. 244–245] that the real miniversal deformation $A'_R(p')$, $p' \in \mathrm{R}^d$, to which an arbitrary family of real matrices $A(p)$, $A(p_0) = A_0$ can be transformed by a smooth change of basis $C_R(p)$ and a smooth change of parameters $p' = \varphi(p)$, $\varphi(p_0) = \mathbf{0}$, can be chosen in the form $A'_R(p') = J_R + B_R(p')$, $B_R = \mathrm{Diag}\,(B_1, \ldots, B_q, \mathrm{Rl}\,B_{q+1}, \mathrm{Rl}\,B_{q+2}, \ldots)$. Here B_j is a block corresponding to the block J_j whose elements are all zero, except at the places indicated in Figure 2.3 (a, b, or c). In the blocks B_j, $j = 1, \ldots, q$, the indicated places are filled by independent parameters $b_j^{kl}(t)_a$, and in the blocks B_j, $j = q+1, q+2, \ldots$, by complex numbers $b_j^{kl}(t) = b_j^{kl}(t)_a + ib_j^{kl}(t)_b$, where $b_j^{kl}(t)_a$ and $b_j^{kl}(t)_b$ are independent parameters. Numbering of parameters $b_j^{kl}(t)_a$, $b_j^{kl}(t)_b$ is carried out in the same way as in section 2. Let us introduce families of matrices $C = C_R R$, $C^{-1} = R^{-1}C_R^{-1}$,

$A' = J + B$, and $B = \operatorname{Diag}(B_1, \dots, B_q, B_{q+1}, \overline{B_{q+1}}, B_{q+2}, \overline{B_{q+2}}, \dots)$. Using expressions $A'_R = J_R + B_R = R(J+B)R^{-1} = RA'R^{-1}$, we obtain

$$A(p) = C_R(p)A'_R(\varphi(p))C_R^{-1}(p) = C(p)A'(\varphi(p))C^{-1}(p). \tag{3.1}$$

The family $A'(p')$ has the same structure as the complex miniversal deformation in section 2. This family as well as $A'_R(p')$ is a miniversal deformation to which any family of real matrices $A(p)$, $A(p_0) = A_0$ can be transformed in the vicinity of the point $p = p_0$, but it and the families C, C^{-1} contain complex elements. For derivatives D^h of the functions $p' = \varphi(p)$, $C(p)$, and $C^{-1}(p)$ transforming the family $A(p)$ to the form $A'(p')$ the formulae of Theorem 2.4 are valid if we assume $b_j^{kl}(t) = b_j^{kl}(t)_a$ for $j \le q$ and $b_j^{kl}(t) = b_j^{kl}(t)_a + ib_j^{kl}(t)_b$ for $j > q$. Using the relations $C_R = CR^{-1}$, $C_R^{-1} = RC^{-1}$ and Theorem 2.4, analogous formulae can be found for the derivatives D^hC_R, $D^hC_R^{-1}$, $D^hb_j^{kl}(t)_a$, and $D^hb_j^{kl}(t)_b$.

THEOREM 3.1. *Let $A(p)$, $p \in \mathbf{R}^n$ be a family of real matrices. Then the following formulae are valid for derivatives D^h, $|h| > 0$ of the functions $C_R(p)$, $C_R^{-1}(p)$, $b_j^{kl}(t)_a(p)$, and $b_j^{kl}(t)_b(p)$ transforming the family $A(p)$ to the real miniversal deformation $A'_R(p')$ described above* (3.1)

1°.

$$D^hb_j^{kl}(t)_a = \operatorname{Re}\frac{\operatorname{tr}^{(t)}\left[(R^{-1}F_RR)_{jA}^{kl}\right]}{\alpha_j^{kl}(t)}, \quad D^hb_j^{kl}(t)_b = \operatorname{Im}\frac{\operatorname{tr}^{(t)}\left[(R^{-1}F_RR)_{jA}^{kl}\right]}{\alpha_j^{kl}(t)},$$

$$F_R = C_{R0}^{-1}\Bigg(D^hA - \sum_{\substack{h_1+h_2+h_3=h\\ h_1,h_2,h_3\ne h}} C_h^{h_1h_2h_3}D^{h_1}C_R\,D^{h_2}A'_R\,D^{h_3}C_R^{-1} + A_0\sum_{\substack{h_1+h_2=h\\ h_1,h_2\ne h}} C_h^{h_1h_2}D^{h_1}C_R\,D^{h_2}C_R^{-1}\Bigg)C_{R0}.$$

2°.

$$D^hC_R = C_{R0}RXR^{-1},$$
$$G = R^{-1}(D^hA'_R - F_R)R.$$

3°.

$$D^hC_R^{-1} = -C_{R0}^{-1}\sum_{\substack{h_1+h_2=h\\ h_2\ne h}} C_h^{h_1h_2}D^{h_1}C_R\,D^{h_2}C_R^{-1}.$$

Here elements of the blocks X_{iA}^{kl} and X_{iD}^{kl} are expressed in terms of elements of the blocks G_{iA}^{kl} and G_{iD}^{kl}, respectively, by formulae (2.8), *and elements of other blocks of the matrix X are expressed in terms of elements of the corresponding blocks of the matrix G by formulae* (2.7). *In the last case it is necessary to take into account that the complex conjugate eigenvalues λ_i, $\overline{\lambda_i}$ are considered independently (for example, to obtain the formulae for elements of the block X_{ijC}^{kl}, it is necessary to replace in* (2.7) *$(\lambda_i - \lambda_j)$ by $(\overline{\lambda_i} - \lambda_j)$ and $g_{ij}^{kl}(r_1, s_1)$ by $g_{ijC}^{kl}(r_1, s_1)$).*

Proof. The proof is based on the connection of the real and complex miniversal deformations (3.1). First let us consider a structure of the complex miniversal

deformation $A'(p')$ more thoroughly. As shown above, the matrix A' has a block-diagonal form with blocks A'_{jA}, $j \le q$ and A'_{jA}, A'_{jD}, $j > q$ corresponding to the real and complex eigenvalues $\lambda_j \in \mathbf{R}$, $j \le q$ and λ_j, $\overline{\lambda_j} \in \mathbf{C}$, $j > q$, respectively. The blocks A'_{jA} depend on parameters $b_j^{kl}(t) = b_j^{kl}(t)_a$ for $j \le q$ and on parameters $b_j^{kl}(t) = b_j^{kl}(t)_a + ib_j^{kl}(t)_b$ for $j > q$. The blocks $A'_{jD} = \overline{A'_{jA}}$, $j > q$ depend on parameters $\overline{b_j^{kl}(t)} = b_j^{kl}(t)_a - ib_j^{kl}(t)_b$. Thus, we have a connection between the real and complex miniversal deformations, as well as transformation functions in the form

$$A' = R^{-1}A'_R R, \quad C = C_R R, \quad C^{-1} = R^{-1}C_R^{-1}, \tag{3.2}$$

$$b_j^{kl}(t) = \begin{cases} b_j^{kl}(t)_a, & j \le q, \\ b_j^{kl}(t)_a + ib_j^{kl}(t)_b, & j > q. \end{cases} \tag{3.3}$$

The formulae of theorem 2.4 are valid for derivatives of functions $b_j^{kl}(t)(p)$, $C(p)$, and $C^{-1}(p)$. Combining these formulae with equalities (3.2) and (3.3), we can obtain expressions for derivatives of $b_j^{kl}(t)_a$, $b_j^{kl}(t)_b$, C_R, and C_R^{-1}.

Substituting relations (3.2) into formulae (2.5), (2.6), and (2.9), we obtain

$$F = R^{-1}F_R R, \tag{3.4}$$

$$D^h C_R = C_{R0} R X R^{-1}, \tag{3.5}$$

$$G = R^{-1}(D^h A'_R - F_R)R. \tag{3.6}$$

It follows from expressions (3.3) and Theorem 2.4 that the derivatives $D^h b_j^{kl}(t)_a$ and $D^h b_j^{kl}(t)_b$ can be found by taking the real and imaginary parts of the right-hand side of relation (2.4) and using expression (3.4) for the matrix F. This proves item 1° of the theorem.

Item 2° follows directly from relations (3.5), (3.6), and Theorem 2.4. The fact that the matrix $D^h C_R$ calculated by these formulae will be real can be proved in the following way. Elements of the matrix $G = R^{-1}G_R R$, where $G_R = D^h A'_R - F_R$ is a real matrix, have the properties $g_{ijA}^{kl}(r,s) = \overline{g_{ijD}^{kl}(r,s)}$, $g_{ijB}^{kl}(r,s) = \overline{g_{ijC}^{kl}(r,s)}$. It can be shown that elements of the matrix X obtained from such a matrix G have the same properties. It remains to note that then the matrix RXR^{-1} and, consequently, the matrix $D^h C_R = C_{R0}RXR^{-1}$ will be real.

For the proof of remaining item 3° of the theorem it is sufficient to substitute the relations $C = C_R R$ and $C^{-1} = R^{-1}C_R^{-1}$ into expression (2.10). □

For derivatives of the zero order $h = \mathbf{0}$ we have $D^0 A'_R = J_R$, $D^0 C_R = C_{R0}$, and $D^0 C_R^{-1} = C_{R0}^{-1}$. The recurrent procedure described in Theorem 3.1 allows finding derivatives D^h of the real functions C_R, C_R^{-1}, φ, transforming the family of matrices $A(p)$ to the real miniversal deformation $A'_R(p')$, for any h. Using these derivatives, the real family $A(p)$ can be transformed to the normal form with the accuracy up to $o(\|p - p_0\|^s)$, where s is the degree of smoothness of the family.

3.1. Example. Let us consider a one-parameter family of real matrices $A(x)$ of the order 4×4 having at the point $x = 0$ the pair of complex conjugate eigenvalues

$\lambda = 1 + i$, $\overline{\lambda} = 1 - i$ with corresponding Jordan blocks of the order $n(\lambda) = 2$:

$$A(x) = \begin{pmatrix} 1 - 4x^2 & 1 + x & 2x^2 & 1 \\ -1 + 2x & 1 - x & 1 - 2x^2 & 3x + 2x^2 \\ 2x^2 & 2x^2 & 1 - 2x & -1 + x \\ -x & 3x & 1 + 4x^2 & 1 + x \end{pmatrix}.$$

The normal forms $A'(p')$ and $A'_R(p')$ of this family and the function $p' = \varphi(x)$ determined using the recurrent procedure of Theorem 3.1 (up to $o(x^2)$) have the form

$$A'_R(p') = \begin{pmatrix} 1 & 1 & -1 & 0 \\ b(2)_a & 1 + b(1)_a & -b(2)_b & -1 - b(1)_b \\ 1 & 0 & 1 & 1 \\ b(2)_b & 1 + b(1)_b & b(2)_a & 1 + b(1)_a \end{pmatrix},$$

$$A'(p') = \begin{pmatrix} 1 + i & 1 & 0 & 0 \\ b(2) & 1 + i + b(1) & 0 & 0 \\ 0 & 0 & 1 - i & 1 \\ 0 & 0 & \overline{b(2)} & 1 - i + \overline{b(1)} \end{pmatrix},$$

$$p' = (b(1)_a,\ b(1)_b,\ b(2)_a,\ b(2)_b)^T,$$

$$b(1) = b(1)_a + ib(1)_b = -x - 2x^2 + i\Big(-x - \frac{21}{8}x^2\Big) + o(x^2),$$

$$b(2) = b(2)_a + ib(2)_b = -\frac{x}{2} + \frac{25}{8}x^2 + i\Big(\frac{3}{2}x - \frac{7}{4}x^2\Big) + o(x^2).$$

4. Application to stability theory. As an application of miniversal deformations to the stability theory, let us consider a problem of investigation of the stability domain of a family of real $n \times n$ matrices $A(p)$, $p \in \mathbf{R}^m$ in the vicinity of a singular boundary point. The stability domain of a family of matrices $A(p)$ is a set of values of the vector of parameters p at which all eigenvalues λ of the matrix $A(p)$ have negative real parts $\operatorname{Re}\lambda < 0$. A boundary of the stability domain is characterized by matrices $A(p)$ having pure imaginary eigenvalues $\operatorname{Re}\lambda = 0$, while for all other eigenvalues $\operatorname{Re}\lambda < 0$. Nonsingular points of the stability boundary are characterized by the existence of only one simple zero eigenvalue $\lambda = 0$ or one pair of simple complex conjugate eigenvalues $\lambda = \pm i\omega$ on the imaginary axis. Otherwise the boundary point is singular [2], [3, pp. 255–256]. A linear approximation of the stability domain at the boundary point is described by a tangent cone introduced by Levantovskii [10].

DEFINITION 4.1. *A tangent cone to the stability domain at the boundary point is a set of direction vectors* $e = dp/d\varepsilon\big|_{\varepsilon=0}$ *of the curves* $p(\varepsilon)$, $\varepsilon > 0$, *which start at this point and lie in the stability domain.*

A concept of the tangent cone can be illustrated by a simple example; see Figure 4.1. Here the tangent cone to the stability domain $S = \{p = (p_1, p_2)^T :\ 2p_2 > p_1 + p_1^2,\ 2p_1 + p_1^2 > p_2\}$ at the origin has the form of an angle $K = \{e = (e_1, e_2)^T :$

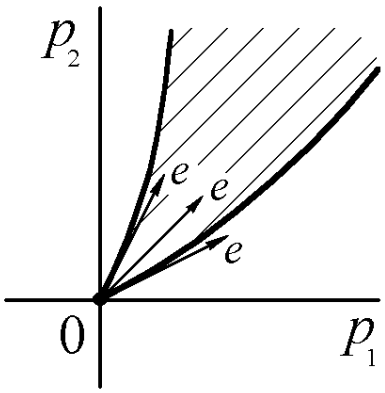


FIG. 4.1. *A tangent cone to the stability domain.*

$2e_2 \geq e_1$, $2e_1 \geq e_2\}$. Thus, the tangent cone consists of all vectors directed into the stability domain or tangent to its boundary.

In [10], using the miniversal deformation tangent cones have been described for some types of singular boundary points up to a linear diffeomorphism. In this section using the miniversal deformations and Theorem 2.4 expressions for tangent cones in the parameter space are found for all types of boundary points. It is shown that for determining the tangent cone we need only the Jordan form of the matrix A and its first derivatives with respect to parameters at the boundary point under consideration.

Let $p = p_0$ be a boundary point of the stability domain and let $\lambda_0 = 0$, $\lambda_j = i\omega_j \neq 0$, $\overline{\lambda_j} = -i\omega_j$, $j = 1, \ldots, k$ be pure imaginary eigenvalues of the matrix $A(p_0)$ (i is the imaginary unit).

THEOREM 4.2. *Assume that each pure imaginary eigenvalue λ_0, λ_j, $\overline{\lambda_j}$, $j = 1, \ldots, k$ of the matrix $A(p_0)$ has exactly one corresponding Jordan block of the dimension n_j, $j = 0, \ldots, k$. If the system of vectors $f(l_1)$, $g_j(l_2)$, $h_j(l_3) \in \mathbf{R}^m$, $l_1 = 1, \ldots, n_0$, $l_2 = 1, \ldots, n_j$, $l_3 = 2, \ldots, n_j$, $j = 1, \ldots, k$, is linearly independent, where*

$$
\begin{aligned}
f(l_1) &= \left(\mathrm{tr}^{(l_1)}[F(1)_0^{11}], \ldots, \mathrm{tr}^{(l_1)}[F(m)_0^{11}]\right)^T, \\
g_j(l_2) &= \left(\mathrm{Re}\,\mathrm{tr}^{(l_2)}[F(1)_j^{11}], \ldots, \mathrm{Re}\,\mathrm{tr}^{(l_2)}[F(m)_j^{11}]\right)^T, \\
h_j(l_3) &= \left(\mathrm{Im}\,\mathrm{tr}^{(l_3)}[F(1)_j^{11}], \ldots, \mathrm{Im}\,\mathrm{tr}^{(l_3)}[F(m)_j^{11}]\right)^T, \\
F(r) &= C_0^{-1}\frac{\partial A}{\partial p_r}C_0,
\end{aligned}
\tag{4.1}
$$

and derivatives are taken at $p = p_0$, then the tangent cone to the stability domain of the family $A(p)$ at the point $p = p_0$ has the form

$$
\begin{aligned}
K \;=\; \{e \in \mathbf{R}^m :\; & \\
& (f(1), e) \leq 0,\ (f(2), e) \leq 0,\ (f(3), e) = \cdots = (f(n_0), e) = 0, \\
& (g_j(1), e) \leq 0,\ (g_j(2), e) \leq 0,\ (g_j(3), e) = \cdots = (g_j(n_j), e) = 0, \\
& (h_j(2), e) = \cdots = (h_j(n_j), e) = 0,\ j = 1, \ldots, k\ \}.
\end{aligned}
\tag{4.2}
$$

Here $(x, y) = x^1y^1 + \cdots + x^my^m$ is the scalar product in $\mathbf{R}^m$. In the case when zero

is not an eigenvalue of the matrix $A(p_0)$ *the expressions containing the vectors* $f(l)$ *have to be excluded from* (4.2).

For the proof of the theorem we need the following lemmas.

LEMMA 4.3. *The tangent cone to the stability domain of the family of polynomials* $x^n - a_1x^{n-1} - \cdots - a_n$ *at the point* $a_1 = \cdots = a_n = 0$ *in the parameter space* $p = (a_1, \ldots, a_n)^T \in \mathbf{R}^n$ *is equal to the set* $\{a_1 \le 0,\ a_2 \le 0,\ a_3 = \cdots = a_n = 0\}$.

LEMMA 4.4. *The tangent cone to the stability domain of the family of polynomials* $z^n - (a_1 + ib_1)z^{n-1} - \cdots - (a_n + ib_n)$ *at the point* $a_1 = b_1 = \cdots = a_n = b_n = 0$ *in the parameter space* $p = (a_1, b_1, \ldots, a_n, b_n)^T \in \mathbf{R}^{2n}$ *is equal to the set* $\{a_1 \le 0,\ b_1 \in \mathbf{R},\ a_2 \le 0,\ b_2 = a_3 = \cdots = a_n = b_n = 0\}$.

A polynomial is said to be stable if real parts of all its roots are negative. Lemmas 4.3 and 4.4 have been formulated and proved in [10].

LEMMA 4.5. *For any* $\alpha_1^j \le 0$, $\alpha_2^j \le 0$, $j = 0, \ldots, k$, *there exists a curve* $q_1(\varepsilon)$, $q_1(0) = \mathbf{0}$ *in the space* $q_1 = (a_{l_1}^{j_1},\ b_{l_2}^{j_2},\ l_2 \ne 1)^T$ *of all coefficients (except* $b_1^1, \ldots, b_1^k$*) of the polynomials*

$$\begin{gathered} x^{n_0} - a_1^0 x^{n_0-1} - \cdots - a_{n_0}^0, \\ z^{n_j} - (a_1^j + ib_1^j)z^{n_j-1} - \cdots - (a_{n_j}^j + ib_{n_j}^j), \quad j = 1, \ldots, k, \end{gathered} \tag{4.3}$$

such that all polynomials (4.3) *are stable along the curve at* $\varepsilon > 0$, *the vector of coefficients* $q_2 = (b_1^1, \ldots, b_1^k)^T$ *is related with* q_1 *and* ε *by an arbitrary given smooth function* $q_2 = Q_2(q_1, \varepsilon)$, $Q_2(\mathbf{0}, 0) = \mathbf{0}$, *and the following conditions are fulfilled:*

$$\begin{gathered} \frac{da_1^j}{d\varepsilon} = \alpha_1^j,\ \frac{da_2^j}{d\varepsilon} = \alpha_2^j,\ \frac{da_3^j}{d\varepsilon} = \cdots = \frac{da_{n_j}^j}{d\varepsilon} = 0,\ j = 0, \ldots, k, \\ \frac{db_2^j}{d\varepsilon} = \cdots = \frac{db_{n_j}^j}{d\varepsilon} = 0,\ j = 1, \ldots, k. \end{gathered} \tag{4.4}$$

Proof. Let us consider polynomials of the form

$$\begin{gathered} (x + \varepsilon^2)^{n_0-2}\big(x^2 + (-\alpha_1^0\varepsilon + \varepsilon^2)x - \alpha_2^0\varepsilon + \varepsilon^2\big), \\ (z + \varepsilon^2)^{n_j-2}\big(z^2 + (-(\alpha_1^j + ib_1^j)\varepsilon + \varepsilon^2)z - \alpha_2^j\varepsilon + \varepsilon^2\big). \end{gathered} \tag{4.5}$$

These polynomials are stable at $\varepsilon > 0$ and their coefficients, obtained after performing multiplication, satisfy conditions (4.4). Expressing these coefficients through $q_2 = (b_1^1\varepsilon, \ldots, b_1^k\varepsilon)^T$, ε and then substituting the obtained expressions into the relation $q_2 = Q_2(q_1, \varepsilon)$, we obtain $q_2 = Q_2'(q_2, \varepsilon)$. Moreover, the matrix of the derivatives $\partial Q_2'/\partial q_2$ is the zero matrix at $\varepsilon = 0$. Hence, by the implicit function theorem the vector q_2 is a smooth function $q_2 = Q_2''(\varepsilon)$ in the vicinity of $\varepsilon = 0$. If we substitute the function $q_2 = (b_1^j)^T = Q_2''(\varepsilon)$ into (4.5), the coefficients of polynomials (4.5) will form the desired curve $q_1 = q_1(\varepsilon)$. □

Proof of Theorem 4.2. Using expression (2.2) the characteristic equation of the matrix $A(p)$ in the vicinity of the point p_0 can be written in the form

$$\det(A(p) - \lambda E) = \det(C(p)A'(\varphi(p))C^{-1}(p) - \lambda E) = \det(A'(\varphi(p)) - \lambda E) = 0,$$

where $A'(p')$ is the miniversal deformation of the matrix $A(p_0)$. Hence, stability of the matrix $A(p)$ is equivalent to stability of the matrix $A'(\varphi(p))$ (normal form) and, by virtue of the block-diagonal structure of A', it is equivalent to simultaneous stability of

all its blocks. The blocks corresponding to eigenvalues of the matrix A_0 with negative real parts are stable at the point p_0 and, consequently, in the vicinity of this point. Thus, stability in the vicinity of p_0 is characterized by stability of the blocks A_0', A_j', $\overline{A_j'}$, $j = 1, \dots, k$, corresponding to pure imaginary eigenvalues or, since the blocks A_j' and $\overline{A_j'}$ are stable or unstable simultaneously (because of complex conjugacy; see section 3), by stability of the blocks A_0', A_j', $j = 1, \dots, k$. Choosing the blocks A_0', A_j' as the blocks of the type (a) (see Figure 2.2), and taking into account that there is only one Jordan block corresponding to each eigenvalue, their characteristic equations take the form

$$\begin{aligned} &\det(A_0' - \lambda E) = \lambda^{n_0} - b_0^{11}(1)_a \lambda^{n_0-1} - \dots - b_0^{11}(n_0)_a = 0, \\ &\det(A_j' - \lambda E) = \eta^{n_j} - (b_j^{11}(1)_a + i\, b_j^{11}(1)_b)\eta^{n_j-1} - \dots \\ &\qquad \dots - (b_j^{11}(n_j)_a + i\, b_j^{11}(n_j)_b) = 0, \\ &\qquad \eta = \lambda - i\,\omega_j, \quad j = 1, \dots, k, \end{aligned} \tag{4.6}$$

where $b_0^{11}(l_1)_a$, $b_j^{11}(l_2)_a$, and $b_j^{11}(l_3)_b$ are parameters of the miniversal deformation smoothly depending on p. By Theorem 2.4, and using expressions (4.1), we get

$$\nabla b_0^{11}(l_1)_a = f(l_1), \quad \nabla b_j^{11}(l_2)_a = g_j(l_2), \quad \nabla b_j^{11}(l_3)_b = h_j(l_3), \tag{4.7}$$

$$\nabla = \left(\frac{\partial}{\partial p_1}, \dots, \frac{\partial}{\partial p_m} \right)^T.$$

Let us consider an arbitrary smooth curve $p = p(\varepsilon)$ with a direction $e = dp/d\varepsilon$. Coefficients of polynomials (4.6) along this curve have the form

$$\begin{aligned} b_0^{11}(l_1)_a(\varepsilon) &= (f(l_1), e)\varepsilon + O(\varepsilon^2), \\ b_j^{11}(l_2)_a(\varepsilon) &= (g_j(l_2), e)\varepsilon + O(\varepsilon^2), \\ b_j^{11}(l_2)_b(\varepsilon) &= (h_j(l_2), e)\varepsilon + O(\varepsilon^2). \end{aligned} \tag{4.8}$$

Applying Lemmas 4.3 and 4.4 to polynomials (4.6) and taking into account expressions (4.8), we obtain that conditions (4.2) are necessary for the curve $p(\varepsilon)$ to lie in the stability domain.

Let us prove that conditions (4.2) are also sufficient. Denote by q a vector whose components are $b_0^{11}(l_1)_a$, $b_j^{11}(l_2)_a$, and $b_j^{11}(l_3)_b$, and by q_1, q_2 the parts of q so that $q_2 = (b_1^{11}(1)_b, \dots, b_k^{11}(1)_b)^T$ and q_1 consists of $d' = n_0 + 2n_1 + \dots + 2n_k - k$ remaining components of the vector q. The vectors q, q_1, and q_2 are connected with p by smooth functions $q = Q(p)$, $q_1 = Q_1(p)$, and $q_2 = Q_2(p)$, which are parts of the mapping $p' = \varphi(p)$. The Jacobian dQ_1/dp is a $d' \times m$ matrix whose rows $f(l_1)$, $g_j(l_2)$, $h_j(l_3)$, $l_3 \neq 1$ form, according to the conditions of the theorem, a linear independent system. Let p_1 be the vector consisting of d' components of the vector p, which corresponds to a linearly independent system of columns of this Jacobian, and denote by p_2 the vector consisting of the remaining components of p. Then $\det[dQ_1/dp_1] \neq 0$ at the point $p = p_0$, and by the implicit function theorem $p_1 = P_1(q_1, p_2)$ in the vicinity of the point $p = p_0$, $q_1 = \mathbf{0}$. By substituting this relation into $q_2 = Q_2(p_1, p_2)$, we get $q_2 = Q_2'(q_1, p_2)$. Let us consider an arbitrary direction e satisfying conditions (4.2) and assume that $p_2 = e_2\varepsilon$, where e_1 and e_2 are the parts of the vector e corresponding to p_1 and p_2, respectively. Then $q_2 = Q_2'(q_1, e_2\varepsilon) = Q_2''(q_1, \varepsilon)$. By Lemma 4.5 there exists

a curve $q_1 = q_1(\varepsilon)$ such that $dq_1/d\varepsilon = (dQ_1/dp)e$, $q_2 = Q_2''(q_1, \varepsilon)$ and all polynomials (4.6) are stable at $\varepsilon > 0$ (conditions (4.4) of the lemma are fulfilled due to (4.2), (4.8)). Then the curve $p = p(\varepsilon)$, $\varepsilon > 0$ such that $p_1 = P_1(q_1(\varepsilon), e_2\varepsilon)$, $p_2 = e_2\varepsilon$ lies in the stability domain of the family $A(p)$ and has the direction e, which was chosen as an arbitrary direction from the set (4.2). □

Theorem 4.2 covers a case when all pure imaginary eigenvalues of $A(p_0)$ are non-derogatory (have only one corresponding Jordan block). In the general case, when some pure imaginary eigenvalues of $A(p_0)$ have several corresponding Jordan blocks, the tangent cone to the stability domain is no longer a product of lines and half-lines as in (4.2). Investigation of this case is much more complicated. However, using an approach presented in the proof of Theorem 4.2, it is possible to find sets of directions belonging to the tangent cone for a boundary point of an arbitrary type.

THEOREM 4.6. *Let $\lambda_0 = 0$, λ_j, $\overline{\lambda_j}$, $j = 1, \dots, k$ be pure imaginary eigenvalues of the matrix $A(p_0)$ and let each eigenvalue λ_j, $j = 0, \dots, k$ have corresponding Jordan blocks of the dimensions $n_1(\lambda_j) \geq n_2(\lambda_j) \geq \cdots$. Assume that the system of vectors $f^{rs}(l_1)$, $g_j^{rs}(l_2)$, $h_j^{rs}(l_3) \in \mathbf{R}^m$ for $r \geq s$ and $l_3 \neq 1$ at $r = s$ is linearly independent, where*

$$
\begin{aligned}
f^{rs}(l_1) &= \Big(\operatorname{tr}^{(l_1)}[F(1)_0^{rs}], \dots, \operatorname{tr}^{(l_1)}[F(m)_0^{rs}]\Big)^T, \\
g_j^{rs}(l_2) &= \Big(\operatorname{Re}\operatorname{tr}^{(l_2)}[F(1)_j^{rs}], \dots, \operatorname{Re}\operatorname{tr}^{(l_2)}[F(m)_j^{rs}]\Big)^T, \\
h_j^{rs}(l_3) &= \Big(\operatorname{Im}\operatorname{tr}^{(l_3)}[F(1)_j^{rs}], \dots, \operatorname{Im}\operatorname{tr}^{(l_3)}[F(m)_j^{rs}]\Big)^T, \\
F(r) &= C_0^{-1}\frac{\partial A}{\partial p_r}C_0,
\end{aligned}
\tag{4.9}
$$

and the derivatives are taken at $p = p_0$. Then the set

$$
\begin{aligned}
K_1 = \{ & e \in \mathbf{R}^m : \\
& (f^{rr}(1), e) \leq 0,\ (f^{rr}(2), e) \leq 0,\ (f^{rr}(3), e) = \cdots = (f^{rr}(n_r(\lambda_0)), e) = 0, \\
& (g_j^{rr}(1), e) \leq 0,\ (g_j^{rr}(2), e) \leq 0,\ (g_j^{rr}(3), e) = \cdots = (g_j^{rr}(n_r(\lambda_j)), e) = 0, \\
& (h_j^{rr}(2), e) = \cdots = (h_j^{rr}(n_r(\lambda_j)), e) = 0, \\
& (f^{rs}(l_1), e) = (g_j^{rs}(l_2), e) = (h_j^{rs}(l_2), e) = 0,\ r > s,\ j = 1, \dots, k\ \}
\end{aligned}
\tag{4.10}
$$

belongs to the tangent cone to the stability domain at the point p_0. Analogously, if the system of vectors (4.9) for $r \leq s$ and $l_3 \neq 1$ at $r = s$ is linearly independent, then the set

$$
\begin{aligned}
K_2 = \{ & e \in \mathbf{R}^m : \\
& (f^{rr}(1), e) \leq 0,\ (f^{rr}(2), e) \leq 0,\ (f^{rr}(3), e) = \cdots = (f^{rr}(n_r(\lambda_0)), e) = 0, \\
& (g_j^{rr}(1), e) \leq 0,\ (g_j^{rr}(2), e) \leq 0,\ (g_j^{rr}(3), e) = \cdots = (g_j^{rr}(n_r(\lambda_j)), e) = 0, \\
& (h_j^{rr}(2), e) = \cdots = (h_j^{rr}(n_r(\lambda_j)), e) = 0, \\
& (f^{rs}(l_1), e) = (g_j^{rs}(l_2), e) = h_j^{rs}(l_2), e) = 0,\ r < s,\ j = 1, \dots, k,\ \}
\end{aligned}
\tag{4.11}
$$

belongs to the tangent cone.

Proof. Characteristic equations of the blocks A'_j in this case cannot be written explicitly as in (4.6). However, if we consider curves $p = p(\varepsilon)$ lying on the surface determined by the equations

$$b_0^{rs}(l_1)_a(p) = b_j^{rs}(l_2)_a(p) = b_j^{rs}(l_2)_b(p) = 0, \quad r > s, \quad j = 1, \dots, k, \tag{4.12}$$

the blocks A'_j will have block-triangular form, and their characteristic equations along these curves will take the form

$$\det(A'_0 - \lambda E) = \prod_r \left(\lambda^{n_r(\lambda_0)} - b_0^{rr}(1)_a \lambda^{n_r(\lambda_0)-1} - \dots - b_0^{rr}(n_r(\lambda_0))_a\right) = 0$$

$$\det(A'_j - \lambda E) = \prod_r \big(\eta^{n_r(\lambda_j)} - (b_j^{rr}(1)_a + i\, b_j^{rr}(1)_b)\eta^{n_r(\lambda_j)-1} - \dots \tag{4.13}$$

$$\dots - (b_j^{rr}(n_r(\lambda_j))_a + i\, b_j^{rr}(n_r(\lambda_j))_b)\big) = 0, \quad \eta = \lambda - i\,\omega_j, \quad j = 1, \dots, k.$$

By Theorem 2.4 and using expressions (4.9), we obtain

$$\nabla b_0^{rs}(l_1)_a = f^{rs}(l_1), \quad \nabla b_j^{rs}(l_2)_a = g_j^{rs}(l_2), \quad \nabla b_j^{rs}(l_3)_b = h_j^{rs}(l_3).$$

Since the system of the vectors $f^{rs}(l_1)$, $g_j^{rs}(l_2)$, $h_j^{rs}(l_2)$, $r > s$ is linearly independent, by the implicit function theorem (4.12) can be solved with respect to a part of variables $p_1 = P_1(p_2)$, where p_1 and p_2 are parts of the vector p. Then the coefficients of polynomials (4.13) depend only on the vector of parameters p_2, and the direction of any curve $p_1 = P_1(p_2(\varepsilon))$, $p_2 = p_2(\varepsilon)$ satisfies all the conditions standing in the last row of (4.10). For the proof of the theorem it is sufficient to find a curve $p_2 = p_2(\varepsilon)$ having an arbitrary chosen direction e_2 such that the vector e consisting of $e_1 = (dP_1/dp_2)e_2$ and e_2 satisfies all conditions (4.10), and all polynomials (4.13) are stable along this curve (conditions in the last row of (4.10) are satisfied automatically due to the choice of e_1). This problem coincides with the problem solved in the proof of Theorem 4.2.

To prove that the set (4.11) belongs to the tangent cone, one should consider curves lying in the surface

$$b_0^{rs}(l_1)_a(p) = b_j^{rs}(l_2)_a(p) = b_j^{rs}(l_2)_b(p) = 0, \quad r < s, \quad j = 1, \dots, k.$$

Then the proof is carried out analogously. ∎

The relations determining sets (4.10) and (4.11) are sufficient conditions that a direction belongs to the tangent cone. Note that Burke and Overton [4] obtained necessary conditions for the stable perturbations of the matrix in the general case, which can be used as necessary conditions that a direction belongs to the tangent cone.

In the nonderogatory case, sets (4.10) and (4.11) coincide and are equal to the tangent cone by Theorem 4.2. Theorems 4.2 and 4.6 give a constructive method for determining the tangent cone to the stability domain at the boundary point. For this purpose we need only information about the matrix A and its first derivatives with respect to parameters at the point under consideration.

In the case of Theorem 4.2 for determining the tangent cone, we need to transform the matrix $A_0 = A(p_0)$ to the Jordan form (2.1) and find the derivatives $\partial A/\partial p_r$ at the point under consideration. Note that components of vectors (4.1) are sums of products of some row of the matrix C_0^{-1} by the matrix $\partial A/\partial p_r$ and by some column of the matrix C_0. Moreover, only the columns and rows of the matrices C_0 and C_0^{-1},

corresponding to the blocks of pure imaginary eigenvalues, are used. This fact can be used to reduce the amount of calculations for determining the tangent cone.

It is well known that the columns of the matrix C_0 corresponding to the block A'_j (in the nonderogatory case) are the eigenvector and associated vectors $u_j^0, \dots, u_j^{n_j-1}$ (also called generalized eigenvectors) of the eigenvalue λ_j

$$\begin{aligned} A_0 u_j^0 &= \lambda_j u_j^0, \\ A_0 u_j^1 &= \lambda_j u_j^1 + u_j^0, \\ &\vdots \\ A_0 u_j^{n_j-1} &= \lambda_j u_j^{n_j-1} + u_i^{n_j-2}. \end{aligned} \tag{4.14}$$

Then the rows of the matrix C_0^{-1} corresponding to the same block are determined as the left eigenvector and associate vectors $v_j^{n_j-1}, \dots, v_j^0$, taken in reverse order,

$$\begin{aligned} v_j^0 A_0 &= \lambda_j v_j^0, \\ v_j^1 A_0 &= \lambda_j v_j^1 + v_j^0, \\ &\vdots \\ v_j^{n_j-1} A_0 &= \lambda_j v_j^{n_j-1} + v_i^{n_j-2}, \end{aligned} \tag{4.15}$$

and satisfying the normalization conditions

$$v_j^0 u_j^{n_j-1} = 1, \quad v_j^l u_j^{n_j-1} = 0, \quad l = 1, \dots, n_j - 1. \tag{4.16}$$

Using these properties of the matrices C_0 and C_0^{-1} expressions (4.1) can be written in the form

$$\begin{aligned} f(l_1) &= \left(\sum_{t=0}^{n_0-l_1} v_0^t \frac{\partial A}{\partial p_1} u_0^{n_0-l_1-t}, \dots, \sum_{t=0}^{n_0-l_1} v_0^t \frac{\partial A}{\partial p_m} u_0^{n_0-l_1-t} \right)^T, \\ g_j(l_2) &= \left(\operatorname{Re} \sum_{t=0}^{n_j-l_2} v_j^t \frac{\partial A}{\partial p_1} u_j^{n_j-l_2-t}, \dots, \operatorname{Re} \sum_{t=0}^{n_j-l_2} v_j^t \frac{\partial A}{\partial p_m} u_j^{n_j-l_2-t} \right)^T, \\ h_j(l_3) &= \left(\operatorname{Im} \sum_{t=0}^{n_j-l_3} v_j^t \frac{\partial A}{\partial p_1} u_j^{n_j-l_3-t}, \dots, \operatorname{Im} \sum_{t=0}^{n_j-l_3} v_j^t \frac{\partial A}{\partial p_m} u_j^{n_j-l_3-t} \right)^T. \end{aligned} \tag{4.17}$$

According to (4.17), for determining the vectors $f(l_1)$, $g_j(l_2)$, and $h_j(l_3)$ we need to know only right and left eigenvectors and associated vectors corresponding to pure imaginary eigenvalues and satisfying normalization conditions (4.16), and the derivatives $\partial A/\partial p_r$, $r = 1, \dots, m$ at the point $p = p_0$.

Analogous expressions can be obtained for vectors (4.9). For this purpose in expressions (4.17) one should substitute $f^{rs}(l_1)$, $g_j^{rs}(l_2)$, $h_j^{rs}(l_3)$, u_{js}^t, v_{jr}^t, n'_0, n'_j instead of $f(l_1)$, $g_j(l_2)$, $h_j(l_3)$, u_j^t, v_j^t, n_0, n_j, respectively, where $n'_0 = \min(n_r(\lambda_0), n_s(\lambda_0))$, $n'_j = \min(n_r(\lambda_j), n_s(\lambda_j))$, and u_{jr}^t v_{jr}^t, $t = 0, \dots, n_r(\lambda_j) - 1$ are right and left eigenvectors and associated vectors corresponding to the Jordan block J_j^r of the eigenvalue λ_j. In this case the normalization conditions for the vectors $u_{jr}^{t_1}$, $v_{jr}^{t_2}$ take the form (4.16), while the normalization conditions for the vectors $u_{jr}^{t_1}$, $v_{js}^{t_2}$, $r \neq s$ are

$$v_{jr}^{n_r(\lambda_j)-1} u_{js}^{n_s(\lambda_j)-t} = 0, \quad t = 1, \dots, n'_j, \quad j = 0, \dots, k.$$

Note that Theorems 4.2 and 4.6 are valid in the case of families of complex matrices smoothly depending on a vector of real parameters, if the vectors $g_j(l_1)$, $h_j(l_2)$, and $g_j^{rs}(l_1)$, $h_j^{rs}(l_2)$ are determined separately for each pure imaginary eigenvalue of the matrix A_0 (also for the zero eigenvalue and for each of the complex conjugate eigenvalues, if any) and the expressions containing the vectors $f(l)$ and $f^{rs}(l)$ are excluded from (4.2), (4.10), and (4.11).

If we consider generic families of matrices [1, 2, 3] the condition of linearly independence of the system of vectors in Theorem 4.2 is satisfied. It follows from the fact that, in the generic case, the codimension of a surface in the parameter space $\mathbf{R}^m$ (stratum), on which all the matrices $A(p)$ have the same type of Jordan structure of pure imaginary eigenvalues of the matrix A_0, is equal to $d' = n_0 + 2n_1 + \cdots + 2n_k - k$ (the number of parameters of the miniversal deformation, which correspond to the blocks of pure imaginary eigenvalues minus the number of nonzero pure imaginary eigenvalues, whose magnitudes are arbitrary on the stratum). There is a field of vectors (4.1) determined on this surface (d' vectors at each point). The condition that the system of these vectors is linearly dependent leads to m equations at each point of the surface, where components of one vector are expressed as a linear combination of corresponding components of the other vectors with $(d' - 1)$ arbitrary coefficients. These equations are equivalent to $(m - d' + 1)$ equations which do not contain these coefficients. But, since the dimension of the stratum is equal to $m - d'$ (less than the number of equations), in the generic case these equations have no solution on the stratum. This means that the system of vectors is linearly independent. The analogous statement is correct in the case of Theorem 4.6. Thus, Theorems 4.2 and 4.6 can be applied to all points of a boundary of the stability domain of a generic family of matrices.

5. Conclusion. In this paper the problem of transformation of families of matrices to the Arnold's and Galin's normal forms (miniversal deformations) is solved. A constructive method determining a change of basis and a change of parameters, transforming a family of real or complex matrices to the normal form in the vicinity of the initial value of the parameter vector, is suggested. Finding the miniversal deformation of a family of matrices and describing the transformation to it represent a complete procedure of transformation to the normal form (2.2). The first step in this procedure is transformation of the matrix $A_0 = A(p_0)$ to the Jordan form (2.1) and finding the miniversal deformation $A'(p')$ according to [1, 2, 3]. Then, using Theorems 2.4 or 3.1, the derivatives of the families $C(p)$, $C^{-1}(p)$ and of the mapping $p' = \varphi(p)$, with respect to the parameters at the point $p = p_0$, are determined. At this step a recurrent procedure is used. Note that the procedure consists of only explicit relations containing elementary arithmetic operations. Using the obtained derivatives, the desired functions $p' = \varphi(p)$ and $C(p)$, $C^{-1}(p)$ (describing a change of parameters and a change of basis) are found as Taylor series. Thus, the procedure of transformation of a matrix family to the normal form is complete.

The spectrum of the normal form $A'(\varphi(p))$ coincides with the spectrum of the family $A(p)$, but $A'(p')$ is a family of sparse matrices depending on parameters in a quite simple way. This makes it efficient to apply normal forms (if the mapping $p' = \varphi(p)$ is known) to investigation of different properties of the spectrum of the family $A(p)$. In section 4 the stability domain of a family of matrices in the vicinity of the singular boundary point $p = p_0$ is studied. In the cases when all pure imaginary eigenvalues of the matrix $A(p_0)$ are nonderogatory, tangent cones to the stability domain at the point $p = p_0$ are found. In the case of an arbitrary Jordan structure

of $A(p_0)$, sufficient conditions, which describe directions belong to the tangent cone, are found. These conditions describe subsets of the tangent cones. It turns out that tangent cones are determined by the right and left eigenvectors and associated vectors corresponding to pure imaginary eigenvalues and the first derivatives of the matrix $A(p)$ with respect to parameters at the point $p = p_0$. Expressions determining the tangent cone are explicit and simple. They can be used in different stability problems studying dependence on parameters.

Moreover, by means of using transformation to the normal form, it is possible to investigate local properties of the spectral radius, to solve the problems of finding eigenvalues of perturbed matrices, etc. In this connection it is convenient to investigate the collapse of nonderogatory multiple eigenvalues. In this case the perturbed eigenvalues are determined as roots of a polynomial whose coefficients are obtained from Theorem 2.4.

Acknowledgments. The author thanks Alexander P. Seyranian for his attention to this work and helpful discussions, and Wolfhard Kliem and Michael L. Overton for useful comments.

REFERENCES

[1] V. I. Arnold, *On matrices depending on parameters*, Russian Math. Surveys, 26 (1971), pp. 29–43.
[2] V. I. Arnold, *Lectures on bifurcations and versal families*, Russian Math. Surveys, 27 (1972), pp. 119–184.
[3] V. I. Arnold, *Geometrical Methods in the Theory of Ordinary Differential Equations*, Springer-Verlag, New York, 1983.
[4] J. V. Burke and M. L. Overton, *Stable perturbations of nonsymmetric matrices*, Linear Algebra Appl., 171 (1992), pp. 249–273.
[5] R. Cushman, A. Kelley, and H. Koçak, *Versal normal form at the Lagrangian equilibrium L_4*, J. Differential Equations, 64 (1986), pp. 340–374.
[6] A. Edelman, E. Elmroth, and B. Kågström, *A geometric approach to perturbation theory of matrices and matrix pencils.* I. *Versal deformations*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 653–692.
[7] A. Edelman and H. Murakami, *Polynomial roots from companion matrix eigenvalues*, Math. Comp., 64 (1995), pp. 763–776.
[8] D. M. Galin, *On real matrices depending on parameters*, Uspekhi Mat. Nauk, 27 (1972), pp. 241–242 (in Russian).
[9] F. R. Gantmacher, *The Theory of Matrices*, Vols. 1, 2, Chelsea, New York, 1960.
[10] L. V. Levantovskii, *The boundary of a set of stable matrices*, Russian Math. Surveys, 35 (1980), pp. 249–250.
[11] A. A. Mailybaev and A. P. Seyranian, *On singularities of a boundary of the stability domain*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 106–128.
[12] D. S. Schmidt, *Transformations to versal normal form*, in Computer Aided Proofs in Analysis, K. R. Meyer and D. S. Schmidt, eds., IMA Vol. Math. Appl. 28, Springer, New York, 1989, pp. 235–240.
[13] D. Schmidt, *Versal normal form of the Hamiltonian function of the restricted problem of three bodies near L_4^**, J. Comput. Appl. Math., 52 (1994), pp. 155–176.
[14] L. Stolovitch, *On the computation of a versal family of matrices*, Numer. Algorithms, 4 (1993), pp. 25–56.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 418–430
© 1999 Society for Industrial and Applied Mathematics

CONVERGENCE OF SUBDIVISION SCHEMES ASSOCIATED WITH NONNEGATIVE MASKS*

RONG-QING JIA† AND DING-XUAN ZHOU‡

Abstract. This paper is concerned with refinement equations of the type

$$f = \sum_{\alpha \in \mathbb{Z}^s} a(\alpha) f(M \cdot - \alpha),$$

where f is the unknown function defined on the s-dimensional Euclidean space $\mathbb{R}^s$, a is a finitely supported sequence on $\mathbb{Z}^s$, and M is an $s \times s$ dilation matrix with $m := |\det M|$. The solution of a refinement equation can be obtained by using the subdivision scheme associated with the mask. In this paper we give a characterization for the convergence of the subdivision scheme when the mask is nonnegative. Our method is to relate the problem of convergence to m column-stochastic matrices induced by the mask. In this way, the convergence of the subdivision scheme can be determined in a finite number of steps by checking whether each finite product of those column-stochastic matrices has a positive row. As a consequence of our characterization, we show that the convergence of the subdivision scheme with a nonnegative mask depends only on the location of its positive coefficients. Several examples are provided to demonstrate the power and applicability of our approach.

Key words. subdivision schemes, refinement equations, stochastic matrices, convergent matrix products

AMS subject classifications. 42C15, 41A25, 39A10, 65D17, 15A51

PII. S0895479898342432

1. Introduction.

We are interested in functional equations of the form

$$f = \sum_{\alpha \in \mathbb{Z}^s} a(\alpha) f(M \cdot - \alpha), \tag{1.1}$$

where f is the unknown function defined on the s-dimensional Euclidean space $\mathbb{R}^s$, a is a finitely supported sequence on $\mathbb{Z}^s$, and M is an $s \times s$ integer matrix such that $\lim_{n\to\infty} M^{-n} = 0$. Equation (1.1) is called a *refinement equation*, and the matrix M is called a *dilation matrix*. Correspondingly, the sequence a is called the *refinement mask*. The support of a is defined as $\operatorname{supp} a := \{\alpha \in \mathbb{Z}^s : a(\alpha) \neq 0\}$. Throughout this paper we assume that the mask a satisfies

$$\sum_{\alpha \in \mathbb{Z}^s} a(\alpha) = m := |\det M|. \tag{1.2}$$

In order to solve the refinement equation (1.1), we start with a compactly supported continuous function ϕ and use the iteration scheme $f_n := Q_a^n \phi$, $n = 0, 1, 2, \ldots$,

*Received by the editors July 27, 1998; accepted for publication (in revised form) by R. Brualdi December 5, 1998; published electronically November 2, 1999.
http://www.siam.org/journals/simax/21-2/34243.html

†Department of Mathematical Sciences, University of Alberta, Edmonton, AB, Canada T6G 2G1 (jia@xihu.math.ualberta.ca). The research of this author was supported in part by NSERC Canada grant OGP 121336.

‡Department of Mathematics, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong (mazhou@math.cityu.edu.hk). The research of this author was supported in part by Research Grants Council of Hong Kong grants 9040281 and 9040397 and City University of Hong Kong grant 7000741.

where Q_a is the *cascade operator* on $C(\mathbb{R}^s)$ given by

$$Q_a\phi := \sum_{\alpha\in\mathbb{Z}^s} a(\alpha)\phi(M\cdot - \alpha), \qquad \phi \in C(\mathbb{R}^s). \tag{1.3}$$

This iteration scheme is called a *subdivision scheme*. In particular, we choose ϕ to be the function ϕ_0 given by

$$\phi_0(x) := \prod_{j=1}^{s} \chi(x_j) \qquad \text{for } x = (x_1, \ldots, x_s) \in \mathbb{R}^s, \tag{1.4}$$

where

$$\chi(t) = \max\{1 - |t|, 0\}, \qquad t \in \mathbb{R}.$$

We say that the subdivision scheme associated with mask a *converges* (uniformly) if there is a continuous function f such that

$$\lim_{n\to\infty} \|Q_a^n\phi_0 - f\|_\infty = 0,$$

where ϕ_0 is the function given in (1.4). If the subdivision scheme converges, then the limit function f is a solution of the refinement equation.

In many problems arising from computer-aided geometric design, the mask is nonnegative (see [6] and [12]). The purpose of this paper is to give a characterization for the convergence of subdivision schemes associated with nonnegative masks.

Before proceeding further we introduce some notation. Let $\ell(\mathbb{Z}^s)$ denote the linear space of all real sequences on $\mathbb{Z}^s$, and let $\ell_0(\mathbb{Z}^s)$ denote the linear space of all finitely supported sequences on $\mathbb{Z}^s$. For a finite subset K of $\mathbb{Z}^s$, we denote by $\ell(K)$ the linear subspace of $\ell_0(\mathbb{Z}^s)$ consisting of all sequences supported on K. For $\beta \in \mathbb{Z}^s$ we use δ_β to denote the sequence on $\mathbb{Z}^s$ given by

$$\delta_\beta(\alpha) = \begin{cases} 1 & \text{if } \alpha = \beta, \\ 0 & \text{if } \alpha \in \mathbb{Z}^s \setminus \{\beta\}. \end{cases}$$

In particular, we write δ for δ_0. For a vector $y \in \mathbb{Z}^s$ we use ∇_y to denote the difference operator on $\ell(\mathbb{Z}^s)$ given by

$$\nabla_y u = u - u(\cdot - y), \qquad u \in \ell(\mathbb{Z}^s).$$

Let $e_1, \ldots, e_s$ denote the unit coordinate vectors in $\mathbb{R}^s$. For $j = 1, \ldots, s$, we write ∇_j for ∇_{e_j}.

For $\varepsilon \in \mathbb{Z}$, let A_ε denote the bi-infinite matrix given by

$$A_\varepsilon(\alpha, \beta) := a(\varepsilon + M\alpha - \beta), \qquad \alpha, \beta \in \mathbb{Z}^s. \tag{1.5}$$

We also use A_ε to denote the linear operator on $\ell_0(\mathbb{Z}^s)$ given by

$$A_\varepsilon v(\alpha) = \sum_{\beta\in\mathbb{Z}^s} a(\varepsilon + M\alpha - \beta)v(\beta), \qquad v \in \ell_0(\mathbb{Z}^s),\ \alpha \in \mathbb{Z}^s. \tag{1.6}$$

We observe that the set $\mathbb{Z}^s$ is an Abelian group under addition, and $M\mathbb{Z}^s$ is a subgroup of $\mathbb{Z}^s$. Let E be a complete set of representatives of the distinct cosets of the

quotient group $\mathbb{Z}^s/M\mathbb{Z}^s$. A necessary condition (see [2] and [9]) for the subdivision scheme associated with mask a to converge is that a satisfies the *basic sum rule*:

$$\sum_{\beta\in\mathbb{Z}^s} a(\varepsilon + M\beta) = 1 \qquad \forall\, \varepsilon \in E.$$

A finite subset K of $\mathbb{Z}^s$ is said to be *admissible* for A_ε if $\ell(K)$ is invariant under A_ε. For example, if $s = 1$, $M = (2)$, and a is supported in $[0, N]$, where N is a positive integer, then the set $\{0, 1, \ldots, N-1\}$ is admissible for both A_0 and A_1. Suppose G is a finite subset of $\mathbb{Z}^s$. Let $H := MG \cup (-E + \operatorname{supp} a) \cup \{0\}$, and let

$$K := \left(\sum_{n=1}^{\infty} M^{-n} H\right) \cap \mathbb{Z}^s.$$

In other words, an element $\alpha \in \mathbb{Z}^s$ belongs to K if and only if $\alpha = \sum_{n=1}^{\infty} M^{-n} h_n$ for some sequence of elements $h_n \in H$. Then $K \supseteq G$ and K is admissible $\forall\, A_\varepsilon$, $\varepsilon \in E$ (see [9, Lemma 2.3]).

Suppose K is an admissible set for every A_ε, $\varepsilon \in E$, and $\ell(K)$ contains $\nabla_j \delta$ for $j = 1, \ldots, s$. Let $r := \#K$ and $k := 2^{r^2}$. The following theorem gives the main result of this paper.

THEOREM 1.1. *Let a be a nonnegative mask satisfying* (1.2). *Then the subdivision scheme associated with mask a converges if and only if*

(a) *a satisfies the basic sum rule, and*

(b) *for each k-tuple $(\varepsilon_1, \ldots, \varepsilon_k) \in E^k$, the finite matrix $((A_{\varepsilon_1} \cdots A_{\varepsilon_k})(\alpha, \beta))_{\alpha,\beta\in K}$ has a positive row.*

As a consequence of Theorem 1.1 we have the following useful result.

THEOREM 1.2. *Let a and b be two nonnegative masks satisfying the basic sum rule. Suppose $b(\alpha) > 0$ whenever $a(\alpha) > 0$. If the subdivision scheme associated with mask a converges, then the subdivision scheme associated with mask b also converges.*

Proof. For $\varepsilon \in E$, let B_ε be the linear operator on $\ell_0(\mathbb{Z}^s)$ defined in a way similar to (1.6). Let K be a finite subset of $\mathbb{Z}^s$ such that $\ell(K)$ contains $\nabla_j \delta$ for $j = 1, \ldots, s$, and K is admissible $\forall\, A_\varepsilon$ and B_ε, $\varepsilon \in E$. We also use A_ε and B_ε to denote the matrices $(A_\varepsilon(\alpha, \beta))_{\alpha,\beta\in K}$ and $(B_\varepsilon(\alpha, \beta))_{\alpha,\beta\in K}$, respectively. Since the subdivision scheme associated with mask a converges, for each k-tuple $(\varepsilon_1, \ldots, \varepsilon_k) \in E^k$, the matrix $A_{\varepsilon_1} \cdots A_{\varepsilon_k}$ has a positive row, by Theorem 1.1. By the assumption, $b(\alpha) > 0$ whenever $a(\alpha) > 0$. Hence, for $\alpha, \beta \in K$, $(B_{\varepsilon_1} \cdots B_{\varepsilon_k})(\alpha, \beta) > 0$ whenever $(A_{\varepsilon_1} \cdots A_{\varepsilon_k})(\alpha, \beta) > 0$. Thus, for each k-tuple $(\varepsilon_1, \ldots, \varepsilon_k) \in E^k$, the matrix $B_{\varepsilon_1} \cdots B_{\varepsilon_k}$ has a positive row. By using Theorem 1.1 again, we conclude that the subdivision scheme associated with mask b also converges. □

Here is a brief outline of this paper. In section 2 we review some results on stochastic matrices, which will be needed later. The proof of Theorem 1.1 is given in section 3. In section 4 we provide several examples to illustrate applications of Theorems 1.1 and 1.2.

2. Stochastic matrices. A square matrix $(a_{ij})_{1\le i,j\le r}$ is called a *column-stochastic* matrix if

$$a_{ij} \ge 0 \quad \forall\, i, j = 1, \ldots, r \quad \text{and} \quad \sum_{i=1}^{r} a_{ij} = 1 \quad \forall\, j = 1, \ldots, r.$$

The product of two column-stochastic matrices is column-stochastic.

A column-stochastic matrix A is called *scrambling* if each pair of columns of A have positive entries in some common row. In particular, if A is a column-stochastic matrix and has a positive row, then A is scrambling. Scrambling matrices have the property that the product AB is scrambling if one of A and B is scrambling. Similarly, if A and B are two column-stochastic matrices of the same order, then AB has a positive row provided one of A and B does.

We equip $\mathbb{R}^r$ with the ℓ_1-norm defined by

$$\|u\| := \sum_{j=1}^{r} |u_j| \qquad \text{for } u = (u_1, \ldots, u_r)^T \in \mathbb{R}^r.$$

An $r \times r$ matrix A can be viewed as a linear operator on $\mathbb{R}^r$. The norm of A is induced by the norm in $\mathbb{R}^r$:

$$\|A\| := \sup_{\|u\|=1} \|Au\|.$$

A column-stochastic matrix A always has $\|A\| = 1$. Let V be the subspace of $\mathbb{R}^r$ given by

$$V := \left\{ (v_1, \ldots, v_r)^T \in \mathbb{R}^r \,:\, \sum_{j=1}^{r} v_j = 0 \right\}.$$

The norm in V is inherited from the norm in $\mathbb{R}^r$. Clearly, for $1 \le j, k \le r$, the vector $e_j - e_k$ lies in V.

LEMMA 2.1. *A column-stochastic matrix $A = (a_{ij})_{1\le i,j\le r}$ is scrambling if and only if*

$$\|A|_V\| < 1.$$

Proof. If A is not scrambling, then there exist j and k $(j \ne k)$ such that $a_{ij}a_{ik} = 0$ for $i = 1, \ldots, r$. Let $w := e_j - e_k$. Then $\|w\| = 2$ and

$$Aw = \sum_{i=1}^{r} (a_{ij} - a_{ik}) e_i.$$

Since $a_{ij}a_{ik} = 0$ for $i = 1, \ldots, r$, it follows that

$$\|Aw\| = \sum_{i=1}^{r} |a_{ij} - a_{ik}| = \sum_{i=1}^{r} (a_{ij} + a_{ik}) = 2.$$

Consequently, $\|Aw\|/\|w\| = 1$. This shows $\|A|_V\| = 1$.

Now suppose A is a scrambling matrix. For $1 \le j < k \le r$, let $\lambda_{jk} := \|A(e_j - e_k)\|$. Since A is a scrambling matrix, there exists some i such that $a_{ij}a_{ik} > 0$. For this i we have $|a_{ij} - a_{ik}| < a_{ij} + a_{ik}$. It follows that

$$\lambda_{jk} = \|A(e_j - e_k)\| = \sum_{i=1}^{r} |a_{ij} - a_{ik}| < \sum_{i=1}^{r} (a_{ij} + a_{ik}) = 2.$$

Let $\lambda := \max\{\lambda_{jk} : 1 \le j < k \le r\}$. Then $\lambda < 2$. Pick a nonzero vector $v = (v_1, \ldots, v_r)^T$ from the space V. Let

$$J_+ := \{j \in \{1, \ldots, r\} : v_j > 0\} \quad \text{and} \quad J_- := \{j \in \{1, \ldots, r\} : v_j < 0\}.$$

Observe that $\sum_{j=1}^{r} v_j = 0$. We have

$$\sum_{j\in J_+} v_j = -\sum_{j\in J_-} v_j = (1/2)\sum_{j=1}^{r} |v_j| = \|v\|/2.$$

Since $v = -\frac{2}{\|v\|}\sum_{j\in J_+}\sum_{k\in J_-} v_j v_k (e_j - e_k)$, there exist real numbers c_{jk} $(1 \le j < k \le r)$ such that

$$v = \sum_{1\le j<k\le r} c_{jk}(e_j - e_k) \quad \text{and} \quad \sum_{1\le j<k\le r} |c_{jk}| = \sum_{j\in J_+} v_j = \|v\|/2.$$

It follows that

$$\|Av\| \le \sum_{1\le j<k\le r} |c_{jk}|\,\|A(e_j - e_k)\| = \sum_{1\le j<k\le r} |c_{jk}|\lambda_{jk} \le \lambda \sum_{1\le j<k\le r} |c_{jk}| = (\lambda/2)\|v\|.$$

This shows that $\|A|_V\| \le \lambda/2 < 1$, as desired. □

For a nonnegative matrix $A = (a_{ij})_{1\le i,j\le r}$, its *signum* is the matrix of the same size defined by

$$\sigma(A)_{ij} := \begin{cases} 1 & \text{if } a_{ij} > 0, \\ 0 & \text{if } a_{ij} = 0. \end{cases}$$

For two nonnegative matrices A and B of the same size, we have $\sigma(AB) = \sigma(A)\sigma(B)$. There is a total of 2^{r^2} distinct signum matrices of order r.

A finite multiset of square matrices $\Sigma = \{A_i : i = 1, \ldots, m\}$ is called an *RCP* set (right-convergent product set) if all infinite products with each element drawn from Σ converge.

The following result was established in [14, Theorem 2.1]. Also, see [4, Theorem 6.1] and the references cited there.

LEMMA 2.2. *Let $\Sigma = \{A_i : i = 1, \ldots, m\}$ be an RCP set of column-stochastic matrices. Suppose that, for each sequence $(\varepsilon_n)_{n=1,2,\ldots}$ $(1 \le \varepsilon_n \le m,\ n = 1, 2, \ldots)$, all the columns of the limit matrix*

$$L := \lim_{n\to\infty} A_{\varepsilon_1} \cdots A_{\varepsilon_n}$$

are the same. Then there exists a positive integer $k \le 2^{r^2}$ such that $A_{\varepsilon_1} \cdots A_{\varepsilon_k}$ possesses a positive row $\forall\, \varepsilon_1, \ldots, \varepsilon_k \in \{1, \ldots, m\}$.

Proof. Let $k := 2^{r^2}$. Among the $k+1$ matrices $A_{\varepsilon_1} \cdots A_{\varepsilon_l}$, $l = 1, \ldots, k+1$, there exist two matrices having the same signum. In other words, there exist two integers p and q, $1 \le p < q \le k+1$, such that

$$\sigma(A_{\varepsilon_1} \cdots A_{\varepsilon_p}) = \sigma(A_{\varepsilon_1} \cdots A_{\varepsilon_q}).$$

Let $A := A_{\varepsilon_1} \cdots A_{\varepsilon_p}$ and $B := A_{\varepsilon_{p+1}} \cdots A_{\varepsilon_q}$. Then $\sigma(A) = \sigma(AB)$. For $n = 2, 3, \ldots$, we have

$$\sigma(AB^n) = \sigma((AB)B^{n-1}) = \sigma(AB)\sigma(B^{n-1}) = \sigma(A)\sigma(B^{n-1}) = \sigma(AB^{n-1}).$$

Consequently, $\sigma(AB^n) = \sigma(AB^{n-1}) = \cdots = \sigma(A)$.

Since Σ is an RCP set, $\lim_{n\to\infty} AB^n$ exists. Let $G = (g_{ij})_{1\le i,j\le r}$ be the limit matrix. Then G is a column-stochastic matrix. It follows that $g_{i1} > 0$ for some i. By

our hypothesis, all the columns of G are the same. Hence, $g_{ij} = g_{i1} > 0 \ \ \forall\, j = 1, \ldots, r$. In other words, row i of the matrix G is positive. But $G := \lim_{n\to\infty} AB^n$. So row i of the matrix AB^n is positive for sufficiently large n. On the other hand, $\sigma(A) = \sigma(AB^n)$. It follows that row i of the matrix $A = A_{\varepsilon_1} \cdots A_{\varepsilon_p}$ is positive. Note that $p < q \le k+1$ implies $p \le k$. Consequently, row i of the matrix

$$A_{\varepsilon_1} \cdots A_{\varepsilon_k} = \left(A_{\varepsilon_1} \cdots A_{\varepsilon_p}\right)\left(A_{\varepsilon_{p+1}} \cdots A_{\varepsilon_k}\right)$$

is positive. Since this is true for all possible choices of $(\varepsilon_1, \ldots, \varepsilon_k) \in \{1, \ldots, m\}^k$, the proof of the lemma is complete. □

It was pointed out in [4, Theorem 6.1] that, for a finite set $\Sigma = \{A_1, \ldots, A_m\}$ of $r \times r$ column-stochastic matrices, the following two conditions are equivalent:

(i) There exists a positive integer μ such that $\forall\, n \ge \mu$ all products $A_{\varepsilon_1} \cdots A_{\varepsilon_n}$ $(1 \le \varepsilon_1, \ldots, \varepsilon_n \le m)$ have a positive row.

(ii) There exists a positive integer ν such that $\forall\, n \ge \nu$ all products $A_{\varepsilon_1} \cdots A_{\varepsilon_n}$ $(1 \le \varepsilon_1, \ldots, \varepsilon_n \le m)$ are scrambling.

It was shown in [15, Theorem 4] that if the statement in (ii) is true, then it holds with $\nu \le (3^r - 2^{r+1} + 1)/2$ and this bound is sharp. Thus, condition (b) in Theorem 1.1 can be replaced by the statement in (ii) with $\nu = (3^r - 2^{r+1} + 1)/2$. This reduces the complexity of computation in checking the convergence of the subdivision scheme.

3. Convergence of subdivision schemes. This section is devoted to a proof of Theorem 1.1.

The joint spectral radius of matrices introduced in [16] was used in [3] to investigate refinement equations. Let us recall the definition of the joint spectral radius. Suppose $\mathcal{A}$ is a finite collection of linear operators on a *finite-dimensional* vector space V. A vector norm $\|\cdot\|$ on V induces a norm on the linear operators on V as follows. For a linear operator A on V, define

$$\|A\| := \max_{\|v\|=1} \{\|Av\|\}.$$

For a positive integer n we denote by $\mathcal{A}^n$ the Cartesian power of $\mathcal{A}$:

$$\mathcal{A}^n = \{(A_1, \ldots, A_n) : A_1, \ldots, A_n \in \mathcal{A}\}.$$

When $n = 0$, we interpret $\mathcal{A}^0$ as the set $\{I\}$, where I is the identity mapping on V. Let

$$\|\mathcal{A}^n\|_\infty := \max\{\|A_1 \cdots A_n\| : (A_1, \ldots, A_n) \in \mathcal{A}^n\}.$$

Then the joint spectral radius of $\mathcal{A}$ is defined to be

$$\rho(\mathcal{A}) := \lim_{n\to\infty} \|\mathcal{A}^n\|_\infty^{1/n}.$$

It is easily seen that this limit indeed exists, and

$$\lim_{n\to\infty} \|\mathcal{A}^n\|_\infty^{1/n} = \inf_{n\ge 1} \|\mathcal{A}^n\|_\infty^{1/n}.$$

Now let M be a dilation matrix with $m := |\det M|$, and let a be a refinement mask satisfying (1.2) (a is not necessarily nonnegative). For $\varepsilon \in E$, let A_ε be the bi-infinite matrix given in (1.5). We also view A_ε as the linear operator on $\ell_0(\mathbb{Z}^s)$ defined

in (1.6). Suppose K is an admissible set for every A_ε, $\varepsilon \in E$, and $\ell(K)$ contains $\nabla_j \delta$ for $j = 1, \dots, s$. Let

$$V := \left\{ v \in \ell(K) : \sum_{\alpha \in \mathbb{Z}^s} v(\alpha) = 0 \right\}. \tag{3.1}$$

It was proved in [9, Theorem 3.3] that the subdivision scheme associated with a converges if and only if a satisfies the basic sum rule and $\rho(\{A_\varepsilon|_V : \varepsilon \in E\}) < 1$.

The *subdivision operator* S_a associated with a (see [2]) is the linear operator on $\ell(\mathbb{Z}^s)$ given by

$$S_a u(\alpha) := \sum_{\beta \in \mathbb{Z}^s} a(\alpha - M\beta) u(\beta), \qquad u \in \ell(\mathbb{Z}^s),\ \alpha \in \mathbb{Z}^s.$$

Let Q_a be the cascade operator defined in (1.3). Then for $\phi \in C(\mathbb{R}^s)$ we have

$$Q_a^n \phi = \sum_{\alpha \in \mathbb{Z}^s} S_a^n \delta(\alpha) \phi(M^n \cdot - \alpha). \tag{3.2}$$

This can be easily proved by induction on n.

The following lemma was first proved in [7]. For the reader's convenience, we repeat the proof given in [9].

LEMMA 3.1. *Suppose that α can be written as $\varepsilon_1 + M\varepsilon_2 + \cdots + M^{n-1}\varepsilon_n + M^n\gamma$, where $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n, \gamma \in \mathbb{Z}^s$. Then for any $\beta \in \mathbb{Z}^s$,*

$$S_a^n \delta(\alpha - \beta) = A_{\varepsilon_n} \cdots A_{\varepsilon_1}(\gamma, \beta).$$

Proof. The proof proceeds by induction on n. For $n = 1$ and $\alpha = \varepsilon_1 + M\gamma$, we have

$$S_a \delta(\alpha - \beta) = a(\varepsilon_1 + M\gamma - \beta) = A_{\varepsilon_1}(\gamma, \beta).$$

Suppose $n > 1$ and the lemma has been verified for $n - 1$. For $\alpha = \varepsilon_1 + M\alpha_1$, where $\alpha_1, \varepsilon \in \mathbb{Z}^s$, we have

$$S_a^n \delta(\alpha - \beta) = \sum_{\eta \in \mathbb{Z}^s} a(\alpha - \beta - M\eta) S_a^{n-1} \delta(\eta) = \sum_{\eta \in \mathbb{Z}^s} a(\varepsilon_1 + M\eta - \beta) S_a^{n-1} \delta(\alpha_1 - \eta). \tag{3.3}$$

Suppose $\alpha_1 = \varepsilon_2 + \cdots + M^{n-2}\varepsilon_n + M^{n-1}\gamma$. Then by the induction hypothesis we have

$$S_a^{n-1} \delta(\alpha_1 - \eta) = A_{\varepsilon_n} \cdots A_{\varepsilon_2}(\gamma, \eta).$$

This in connection with (3.3) gives

$$S_a^n \delta(\alpha - \beta) = \sum_{\eta \in \mathbb{Z}^s} A_{\varepsilon_n} \cdots A_{\varepsilon_2}(\gamma, \eta) A_{\varepsilon_1}(\eta, \beta) = A_{\varepsilon_n} \cdots A_{\varepsilon_2} A_{\varepsilon_1}(\gamma, \beta),$$

thereby completing the induction procedure. □

LEMMA 3.2. *Suppose the subdivision scheme associated with mask a converges to a limit function f on $\mathbb{R}^s$. Then for any sequence $(\varepsilon_n)_{n=1,2,\dots}$ in E,*

$$\lim_{n \to \infty} A_{\varepsilon_1} \cdots A_{\varepsilon_n}(\alpha, \beta) = f(x + \alpha) \qquad \forall\, \alpha, \beta \in \mathbb{Z}^s, \tag{3.4}$$

where $x := \sum_{j=1}^{\infty} M^{-j}\varepsilon_j$.

Proof. For $n = 1, 2, \ldots$, let $\phi_n := Q_a^n \phi_0$, where ϕ_0 is the function given in (1.4). By (3.2) we have

$$\phi_n(x) = \sum_{\beta \in \mathbb{Z}^s} (S_a^n \delta)(\beta) \phi_0(M^n x - \beta), \qquad x \in \mathbb{R}^s.$$

Write a_n for $S_a^n \delta$, $n = 1, 2, \ldots$. Setting $x = M^{-n}\alpha$ for $\alpha \in \mathbb{Z}^s$ in the above equation, we obtain

$$\phi_n(M^{-n}\alpha) = a_n(\alpha), \qquad \alpha \in \mathbb{Z}^s. \tag{3.5}$$

Let $(\varepsilon_n)_{n=1,2,\ldots}$ be a sequence in E. By Lemma 3.1 we have

$$A_{\varepsilon_1} \cdots A_{\varepsilon_n}(\alpha, \beta) = a_n\left(\varepsilon_n + M\varepsilon_{n-1} + \cdots + M^{n-1}\varepsilon_1 + M^n\alpha - \beta\right), \qquad \alpha, \beta \in \mathbb{Z}^s.$$

This in connection with (3.5) yields

$$A_{\varepsilon_1} \cdots A_{\varepsilon_n}(\alpha, \beta) = \phi_n\left(\alpha + M^{-1}\varepsilon_1 + \cdots + M^{-n+1}\varepsilon_{n-1} + M^{-n}\varepsilon_n - M^{-n}\beta\right). \tag{3.6}$$

But

$$M^{-1}\varepsilon_1 + \cdots + M^{-n+1}\varepsilon_{n-1} + M^{-n}\varepsilon_n - M^{-n}\beta \to x \quad \text{as } n \to \infty.$$

Moreover, since the subdivision scheme associated with mask a converges to f, we have $\|\phi_n - f\|_\infty \to 0$ as $n \to \infty$. Letting n go to ∞ in (3.6), we obtain (3.4), as desired. □

Now let a be a nonnegative mask satisfying the basic sum rule. Suppose K is a finite subset of $\mathbb{Z}^s$ such that K is admissible $\forall\, A_\varepsilon$, $\varepsilon \in E$. Since a satisfies the basic sum rule, we have

$$\sum_{\alpha \in \mathbb{Z}^s} a(\varepsilon + M\alpha - \beta) = 1 \qquad \forall\, \beta \in K.$$

But $\ell(K)$ is invariant under A_ε. So $\beta \in K$ and $a(\varepsilon + M\alpha - \beta) \neq 0$ implies $\alpha \in K$. Consequently,

$$\sum_{\alpha \in K} a(\varepsilon + M\alpha - \beta) = \sum_{\alpha \in \mathbb{Z}^s} a(\varepsilon + M\alpha - \beta) = 1 \qquad \forall\, \beta \in K.$$

This shows that the matrix $(A_\varepsilon(\alpha, \beta))_{\alpha,\beta \in K}$ is a column-stochastic matrix. Lemma 3.2 tells us that $\{(A_\varepsilon(\alpha, \beta))_{\alpha,\beta \in K} : \varepsilon \in E\}$ is an RCP set.

We are in a position to establish the main result of this paper.

Proof of Theorem 1.1. Suppose K is an admissible set for every A_ε, $\varepsilon \in E$, and $\ell(K)$ contains $\nabla_j \delta$ for $j = 1, \ldots, s$. The norm on $\ell(K)$ is chosen to be the ℓ_1-norm. Let V be the subspace of $\ell(K)$ given in (3.1). Let $r := \#K$ and $k := 2^{r^2}$. We also use A_ε to denote the matrix $(A_\varepsilon(\alpha, \beta))_{\alpha,\beta \in K}$. Suppose that conditions (a) and (b) in Theorem 1.1 are satisfied. In particular, for each k-tuple $(\varepsilon_1, \ldots, \varepsilon_k) \in E^k$, the finite matrix $A_{\varepsilon_1} \cdots A_{\varepsilon_k}$ has a positive row. By Lemma 2.1 we have

$$\|(A_{\varepsilon_1} \cdots A_{\varepsilon_k})\,|_V\| < 1.$$

It follows that

$$\rho(\{A_\varepsilon|_V : \varepsilon \in E\}) < 1.$$

Therefore, the subdivision scheme associated with mask a converges.

Suppose conversely that the subdivision scheme associated with mask a converges. Then $\{A_\varepsilon : \varepsilon \in E\}$ is an RCP set. Moreover, from (3.4) we see that, for each sequence $(\varepsilon_n)_{n=1,2,\ldots}$ in E, all the columns of the limit matrix

$$\lim_{n\to\infty} A_{\varepsilon_1} \cdots A_{\varepsilon_n}$$

are the same. By Lemma 2.2 we conclude that, for each k-tuple $(\varepsilon_1, \ldots, \varepsilon_k) \in E^k$, the finite matrix $A_{\varepsilon_1} \cdots A_{\varepsilon_k}$ has a positive row. □

4. Examples. In this section, we give four examples to illustrate applications of Theorems 1.1 and 1.2.

Our first example is concerned with the refinement equation

$$\phi(x) = \sum_{j=0}^{N} a(j)\phi(2x - j), \qquad x \in \mathbb{R},$$

where N is an integer greater than 1. Suppose that the mask a is nonnegative and satisfies the basic sum rule. In [13, Corollary 5.1], the convergence of the subdivision scheme was established under the condition that $a(j) > 0$ for $j = 0, 1, \ldots, N$. In [8] this condition was relaxed to $a(j) > 0$ for $j = 0, 1, N-1, N$. The following example gives a more general result, which was also obtained in [11, Theorem 8] by using a different method.

EXAMPLE 4.1. *Suppose that the mask a is nonnegative and satisfies the basic sum rule. Then the subdivision scheme associated with mask a converges, provided $0 < a(0) < 1$, $a(1) > 0$, and $0 < a(N) < 1$.*

Proof. Let $K := \{0, 1, \ldots, N-1\}$. Then $\nabla\delta \in \ell(K)$ and K is admissible for both A_0 and A_1. Let

$$A_\varepsilon := (a(\varepsilon + 2i - j))_{i,j=0}^{N-1}, \qquad \varepsilon = 0, 1.$$

By Theorem 1.1, it suffices to show that the matrix $A_{\varepsilon_1} A_{\varepsilon_2} \cdots A_{\varepsilon_N}$ has a positive row for each $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N) \in \{0, 1\}^N$.

We claim that for $2^n \ge N - 1$, $\varepsilon_1 = 1$, and $\varepsilon_2, \ldots, \varepsilon_n \in \{0, 1\}$,

$$A_{\varepsilon_1} A_{\varepsilon_2} \cdots A_{\varepsilon_n}(0, j) > 0, \qquad j = 0, 1, \ldots, N-1.$$

To justify our claim, we set $j_n := j$ and $j_{k-1} := \lfloor (j_k + 1 - \varepsilon_k)/2 \rfloor$ for $k = n, \ldots, 2$, where $\lfloor x \rfloor$ denotes the integer part of a real number x. By our choice of j_k we have $0 \le \varepsilon_k + 2j_{k-1} - j_k \le 1$ for $k = 2, \ldots, n$; hence $A_{\varepsilon_k}(j_{k-1}, j_k) = a(\varepsilon_k + 2j_{k-1} - j_k) > 0$. Moreover, we have

$$2^n j_1 \le 2^{n-1} j_2 + 2^{n-1} \le \cdots \le j_n + 1 + \cdots + 2^{n-1} < j_n + 2^n \le 2^{n+1}.$$

It follows that $0 \le j_1 \le 1$. Consequently, $A_1(0, j_1) = a(1 - j_1) > 0$. Therefore, we obtain

$$A_{\varepsilon_1} A_{\varepsilon_2} \cdots A_{\varepsilon_n}(0, j) \ge A_1(0, j_1) A_{\varepsilon_2}(j_1, j_2) \cdots A_{\varepsilon_n}(j_{n-1}, j_n) > 0.$$

Similarly, if $2^n \ge N - 1$, then $A_0^n(1, j) > 0$ for $j = 1, \ldots, N-1$. Indeed, set $j_n := j$ and $j_{k-1} := \lfloor (j_k + 1)/2 \rfloor$ for $k = n, \ldots, 2$. Then $j_1 = 1$ and

$$A_0^n(1, j) \ge A_0(1, j_1) A_0(j_1, j_2) \cdots A_0(j_{n-1}, j_n) > 0.$$

Since the mask a satisfies the basic sum rule, there exists a positive integer l such that $2l \le N$ and $a(2l) > 0$. It follows that $A_0(l,0) = a(2l) > 0$ and

$$A_0^{n+1}(1,0) \ge A_0^n(1,l)A_0(l,0) > 0.$$

Now consider the product $A_{\varepsilon_1} \cdots A_{\varepsilon_N}$, where $\varepsilon_1, \ldots, \varepsilon_N \in \{0,1\}$. Let $n := \lfloor N/2 \rfloor$. Then $2^n \ge N-1$. If $\varepsilon_1 = \cdots = \varepsilon_n = \varepsilon_{n+1} = 0$, then row 1 of $A_{\varepsilon_1} \cdots A_{\varepsilon_{n+1}} = A_0^{n+1}$ is positive. Otherwise, $\varepsilon_i = 1$ for some i, $1 \le i \le n+1$. Note that the length of the product $A_{\varepsilon_i} \cdots A_{\varepsilon_N}$ is $N - i + 1 \ge N - (n+1) + 1 \ge n$. By what has been proved, row 0 of $A_{\varepsilon_i} \cdots A_{\varepsilon_N}$ is positive. In both cases, $A_{\varepsilon_1} \cdots A_{\varepsilon_N}$ has a positive row. □

EXAMPLE 4.2. *Consider the refinement equation*

$$\phi(x) = \sum_{j \in \mathbb{Z}} a(j)\phi(mx - j), \qquad x \in \mathbb{R},$$

where m is an integer greater than 1. Suppose that the mask a is nonnegative and supported on $[0,N]$, where N is an integer and $N \ge 2m-1$. In addition, we assume that

$$a(j) > 0 \qquad \text{for } j = 0, \ldots, m-1 \quad \text{or} \quad j = N-m+1, \ldots, N,$$

and a satisfies the basic sum rule:

$$\sum_{j \in \mathbb{Z}} a(mj) = \sum_{j \in \mathbb{Z}} a(mj+1) = \cdots = \sum_{j \in \mathbb{Z}} a(mj+m-1) = 1.$$

Then the subdivision scheme associated with mask a converges.

Proof. Let b be the mask given by

$$b(0) = \cdots = b(m-1) = 1/2, \qquad b(N-m+1) = \cdots = b(N) = 1/2,$$

and $b(j) = 0$ for $j \in \mathbb{Z} \setminus \{0, \ldots, m-1, N-m+1, \ldots, N\}$. By Theorem 1.2, in order to prove that the subdivision scheme associated with mask a converges, it suffices to show that the subdivision scheme associated with mask b converges.

The symbol of b is

$$\tilde{b}(z) := \sum_{j=0}^{N} b(j)z^j = \frac{1}{2}(1 + z + \cdots + z^{m-1})(1 + z^{N-m+1}).$$

For $n = 1, 2, \ldots$, let $b_n := S_b^n \delta$, where S_b is the subdivision operator associated with b. It is easily seen that

$$\tilde{b}_n(z) = \tilde{b}(z)\tilde{b}(z^m) \cdots \tilde{b}(z^{m^{n-1}}).$$

Let ∇ denote the difference operator on $\ell(\mathbb{Z})$ given by $\nabla u := u - u(\cdot - 1)$ for $u \in \ell(\mathbb{Z})$. Then the symbol of ∇b_n is

$$\widetilde{\nabla b_n}(z) = (1-z)\tilde{b}_n(z) = \left(\frac{1}{2}\right)^n \left(1 - z^{m^n}\right) \prod_{k=0}^{n-1} \left(1 + z^{(N-m+1)m^k}\right).$$

We observe that $\prod_{k=0}^{n-1}(1 + z^{(N-m+1)m^k})$ is a polynomial of z whose coefficients are either 0 or 1. By using the expression of $\widetilde{\nabla b_n}(z)$ we obtain

$$|\nabla b_n(j)| \le (1/2)^n \cdot 2 = (1/2)^{n-1}.$$

Hence, we have the following estimate for the ℓ_∞-norm of the sequence ∇b_n:

$$(1/2)^n \le \|\nabla b_n\|_\infty \le (1/2)^{n-1}, \qquad n = 1, 2, \ldots.$$

It follows that

$$\lim_{n\to\infty} \|\nabla b_n\|_\infty^{1/n} = 1/2 < 1.$$

By Theorem 3.2 of [9] or [10], we conclude that the subdivision scheme associated with mask b converges. □

EXAMPLE 4.3. *Let X be a multiset of vectors $x^1, \ldots, x^n$ in $\mathbb{R}^s$ whose components are integers. We also use X to denote the $s \times n$ matrix with $x^1, \ldots, x^n$ as its columns. We say that X is* unimodular *if each $s \times s$ submatrix of X has determinant -1, 0, 1. The zonotope generated by the vectors in X is defined by*

$$Z[X] := \left\{ \sum_{j=1}^n t_j x^j \ : \ 0 \le t_1, \ldots, t_n \le 1 \right\}.$$

Let a be a nonnegative sequence on $\mathbb{Z}^s$ such that $a(\alpha) > 0$ if and only if $\alpha \in \mathbb{Z}^s \cap Z[X]$. Suppose a satisfies the basic sum rule. Then the subdivision scheme associated with mask a converges, provided the matrix X is unimodular and $X \setminus x^j$ spans $\mathbb{R}^s$ for every $j = 1, \ldots, n$.

Proof. Let B_X be the box spline associated with X. Precisely, B_X is the distribution on $\mathbb{R}^s$ given by

$$\langle B_X, \varphi \rangle := \int_{[0,1]^n} \varphi\left(u_1 x^1 + \cdots + u_n x^n\right) \, du_1 \cdots du_n, \qquad \varphi \in C(\mathbb{R}^s).$$

By our assumption, $X \setminus x^j$ spans $\mathbb{R}^s$ for every $j = 1, \ldots, n$. Hence, B_X is a compactly supported continuous function. Moreover, B_X is refinable:

$$B_X(x) = \sum_{\alpha\in\mathbb{Z}^s} b(\alpha) B_X(2x - \alpha), \qquad x \in \mathbb{R}^s,$$

where the mask b is nonnegative and $b(\alpha) > 0$ if and only if $\alpha \in \mathbb{Z}^s \cap Z[X]$. Since X is unimodular, the subdivision scheme associated with mask b converges (see, e.g., Theorem 23 in [1, Chapter 7]). In particular, b satisfies the basic sum rule. Since $a(\alpha) > 0$ whenever $b(\alpha) > 0$, by Theorem 1.2 we conclude that the subdivision scheme associated with mask a converges. □

The subdivision scheme associated with a mask supported on a central zonotope was discussed in [2]. Let a be a nonnegative mask on $\mathbb{Z}^s$ satisfying the basic sum rule. Suppose that $a(\alpha) > 0$ if and only if $\alpha \in \mathbb{Z}^s \cap Z[X \cup (-X)]$. It was proved in [2, Theorem 3.4] that the subdivision scheme associated with mask a converges, provided X is unimodular and X spans $\mathbb{R}^s$. Clearly, the conclusion in Example 4.3 extends their result.

EXAMPLE 4.4. *Consider the refinement equation*

$$f(x) = \sum_{\alpha\in\mathbb{Z}^2} a(\alpha) f(Mx - \alpha), \qquad x \in \mathbb{R}^2,$$

where

$$M = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}.$$

Let t be a real number such that $0 < t < 1$. Suppose that $a(0,0) = 1$, $a(1,0) = t$, $a(0,1) = 1 - t$, and $a(\alpha) = 0$ for $\alpha \in \mathbb{Z}^2 \setminus \{(0,0),(1,0),(0,1)\}$. Then the subdivision scheme associated with mask a and dilation matrix M converges.

Proof. By Theorem 1.2 it suffices to deal with the special case $t = 1/2$. Let A_0 and A_1 be the linear operators on $\ell_0(\mathbb{Z}^2)$ given by

$$A_0 v(\alpha) = \sum_{\beta \in \mathbb{Z}^2} a(M\alpha - \beta) v(\beta) \quad \text{and} \quad A_1 v(\alpha) = \sum_{\beta \in \mathbb{Z}^2} a((1,0) + M\alpha - \beta) v(\beta),$$

where $\alpha \in \mathbb{Z}^2$ and $v \in \ell_0(\mathbb{Z}^2)$. Let K be the set

$$([-1,2] \times [-2,2]) \cap \left(\mathbb{Z}^2 \setminus \{(-1,-2),(2,-2),(2,2),(-1,2)\}\right).$$

Then K is admissible for both A_0 and A_1. Clearly, $\ell(K)$ contains $\nabla_1 \delta$ and $\nabla_2 \delta$. We also use A_0 and A_1 to denote the matrices

$$\left(a(M\alpha - \beta)\right)_{\alpha,\beta \in K} \quad \text{and} \quad \left(a((1,0) + M\alpha - \beta)\right)_{\alpha,\beta \in K},$$

respectively. Then A_0 and A_1 are 16×16 matrices. By a computation, we find that, for each $(\varepsilon_1, \ldots, \varepsilon_8) \in \{0,1\}^8$, the matrix $A_{\varepsilon_1} \cdots A_{\varepsilon_8}$ has a positive row. By Theorem 1.1 we conclude that the subdivision scheme associated with mask a and dilation matrix M converges. □

This example was treated in [9, Example 3.6] by using a different method. In [5] a different mask was considered. Their mask is given by

$$b(0,0) = 1, \quad b(1,0) = b(0,1) = b(-1,0) = b(0,-1) = 1/4,$$

and $b(\alpha) = 0$ for $\alpha \in \mathbb{Z}^2 \setminus \{(0,0),(1,0),(0,1),(-1,0),(0,-1)\}$. By Theorem 1.2, the convergence of the subdivision scheme associated with mask b and dilation matrix M can be derived from Example 4.4.

REFERENCES

[1] C. de Boor, K. Höllig, and S. Riemenschneider, *Box Splines*, Springer-Verlag, New York, 1993.

[2] A. S. Cavaretta, W. Dahmen, and C. A. Micchelli, *Stationary Subdivision*, Mem. Amer. Math. Soc. 93, AMS, Providence, RI, 1991.

[3] I. Daubechies and J. C. Lagarias, *Two-scale difference equations.* II. *Local regularity, infinite products of matrices and fractals*, SIAM J. Math. Anal., 23 (1992), pp. 1031–1079.

[4] I. Daubechies and J. C. Lagarias, *Sets of matrices all finite products of which converge*, Linear Algebra Appl., 161 (1992), pp. 227–263.

[5] G. Deslauriers, J. Dubois, and S. Dubuc, *Multidimensional iterative interpolation*, Canadian J. Math., 43 (1991), pp. 297–312.

[6] N. Dyn, *Subdivision schemes in computer-aided geometric design*, in Advances in Numerical Analysis II—Wavelets, Subdivision Algorithms and Radial Functions, W. A. Light, ed., Clarendon Press, Oxford, 1991, pp. 36–104.

[7] T. N. T. Goodman, C. A. Micchelli, and J. D. Ward, *Spectral radius formulas for subdivision operators*, in Recent Advances in Wavelet Analysis, L. L. Schumaker and G. Webb, eds., Academic Press, New York, 1994, pp. 335–360.

[8] D. E. Gonsor, *Subdivision algorithms with nonnegative masks generally converge*, Adv. Comput. Math., 1 (1993), pp. 215–221.

[9] B. Han and R. Q. Jia, *Multivariate refinement equations and convergence of subdivision schemes*, SIAM J. Math. Anal., 29 (1998), pp. 1177–1199.

[10] R. Q. Jia, *Subdivision schemes in L_p spaces*, Adv. Comput. Math., 3 (1995), pp. 309–341.

[11] A. A. Melkman, *Subdivision schemes with non-negative masks converge always—Unless they obviously cannot?*, Ann. Numer. Math., 4 (1997), pp. 451–460.

[12] C. A. Micchelli, *Mathematical Aspects of Geometric Modeling*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 65, SIAM, Philadelphia, 1995.
[13] C. A. Micchelli and H. Prautzsch, *Refinement and subdivision for spaces of integer translates of a compactly supported function*, in Numerical Analysis 1987, D. F. Griffiths and G. A. Watson, eds., Longman, Harlow, UK, 1988, pp. 192–222.
[14] C. A. Micchelli and H. Prautzsch, *Uniform refinement of curves*, Linear Algebra Appl., 114/115 (1989), pp. 841–870.
[15] A. Paz, *Definite and quasidefinite sets of stochastic matrices*, Proc. Amer. Math. Soc., 16 (1965), pp. 634–641.
[16] G.-C. Rota and W. G. Strang, *A note on the joint spectral radius*, Indag. Math., 22 (1960), pp. 379–381.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 431–439
© 1999 Society for Industrial and Applied Mathematics

ANY CIRCULANT-LIKE PRECONDITIONER FOR MULTILEVEL MATRICES IS NOT SUPERLINEAR*

S. SERRA CAPIZZANO† AND E. TYRTYSHNIKOV‡

Abstract. Superlinear preconditioners (those that provide a proper cluster at 1) are very important for the cg-like methods since they make these methods converge superlinearly. As is well known, for Toeplitz matrices generated by a continuous symbol, many circulant and circulant-like (related to different matrix algebras) preconditioners were proved to be superlinear. In contrast, for multilevel Toeplitz matrices there has been no proof of the superlinearity of any multilevel circulants. In this paper we show that such a proof is not possible since any multilevel circulant preconditioner is not superlinear, in the general case of multilevel Toeplitz matrices. Moreover, for matrices not necessarily Toeplitz, we present some general results proving that many popular structured preconditioners cannot be superlinear.

Key words. preconditioner, cluster, conjugate gradient, circulants

AMS subject classifications. 15A12, 15A18, 65F10

PII. S0895479897331941

1. Introduction.

The phenomenon of superlinear convergence for the cg-like methods was scrutinized, first of all, in [2, 27] and has become one of the most topical subjects in papers discussing how to make the cg-like iterations converge superlinearly within a constant number of iterations [5, 8, 10, 14, 15, 16, 17, 18, 19, 22, 25, 28, 29]. The key concept behind these papers is that of a cluster. From some naive approach in the beginning it has eventually developed into a rigorous notion [6, 30, 31, 32, 33] as follows.

Consider a sequence of matrices $A_n \in \mathbb{C}^{n\times n}$ and a set M in the complex plane. Denote by M_ε the ε-extension of M, which is the union of all balls of radius ε centered at points of M. Let $\gamma_n(\varepsilon)$ count those eigenvalues of A_n that do not belong to M_ε. Assume that, for any $\varepsilon > 0$,

$$\gamma_n(\varepsilon) = o(n), \quad n \to \infty.$$

Then M is called a general (some say weak) cluster. If $o(n) = O(1)$, then M is called a proper (some say strong) cluster.

Of course, we are interested in the clusters that are as narrow as possible. In this paper we deal only with one-point clusters and primarily those of the singular values rather than the eigenvalues.

If A_n are Hermitian positive definite matrices and C_n are their Hermitian positive definite preconditioners, then the superlinear convergence for the preconditioned conjugate gradient (PCG) method is guaranteed whenever the eigenvalues of $C_n^{-1}A_n$

*Received by the editors December 30, 1997; accepted for publication (in revised form) by D. Calvetti January 12, 1999; published electronically November 2, 1999. This work was begun when the second author was a visiting professor at Dipartimento di Matematica, Università di Pisa, supported by the CNR, Italy. The work of the second author was supported in part by the Russian Fund of Basic Research and Volkswagen-Stiftung.

http://www.siam.org/journals/simax/21-2/33194.html

†Dipartimento di Energetica, Università Degli Studi di Firenze, Via C. Lombroso 6/17, I-50134 Firenze, Italy, and Dipartimento di Informatica, Università di Pisa Corso Italia 40, 56100 Pisa, Italy (serra@mail.dm.unipi.it).

‡Institute of Numerical Mathematics, Russian Academy of Sciences, Gubkina 8, Moscow 117333, Russia (tee@inm.ras.ru).

have a proper cluster at 1. In the non-Hermitian case, the PCG method can be applied to the symmetrized preconditioned system, and the superlinear convergence can be seen as long as the singular values of $C_n^{-1}A_n$ have a proper cluster at 1 [1, 2].

We would have called a preconditioner C_n for A_n superlinear if the singular values of $C_n^{-1}A_n$ had a proper cluster at 1. However, we will require a little more detail.

A preconditioner C_n for A_n is said to be *sublinear* if the singular values of $C_n^{-1}A_n - I_n$ have a cluster at 0 (I_n is the identity matrix of order n) and *superlinear* if this cluster at 0 is proper.

Useful equivalent definitions also exist in terms of convenient matrix relations. A preconditioner C_n for A_n is sublinear iff, for any $\varepsilon > 0$, there exist matrices $E_{n\,\varepsilon}$ and $R_{n\,\varepsilon}$ such that

$$C_n^{-1}A_n - I_n = E_{n\,\varepsilon} + R_{n\,\varepsilon} \tag{1}$$

with

$$||E_{n\,\varepsilon}||_2 \le c_1(\varepsilon), \qquad \lim_{\varepsilon\to 0} c_1(\varepsilon) = 0; \tag{2}$$

$$\operatorname{rank} R_{n\,\varepsilon} \le c_2(n,\varepsilon), \qquad \lim_{n\to\infty} \frac{c_2(n,\varepsilon)}{n} = 0 \quad \forall\, \varepsilon > 0. \tag{3}$$

A sublinear preconditioner C_n is superlinear iff $c_2(n,\varepsilon)$ is uniformly bounded in n (which is equivalent to the claim that some bound c_2 does not depend on n, and so we may write $c_2 = c_2(\varepsilon)$).

For the singular values of $C_n^{-1}A_n$, it is easy to prove that a sublinear preconditioner provides a general cluster at 1 while a superlinear preconditioner provides a proper cluster, and hence the superlinear convergence. However, beware of claiming the reverse, because a preconditioner providing a proper cluster at 1 for the singular values of $C_n^{-1}A_n$ is not necessarily superlinear according to our definition.

In applications, superlinear preconditioners may not be easily available. Still, some important cases exist for when they are.

If $A_n = A_n(f) = [a_{i-j}]_{ij=1}^n$ are Toeplitz matrices associated with an absolutely convergent Fourier series for a positive symbol

$$f(x) = \sum_{k=-\infty}^{\infty} a_k e^{\mathrm{i}kx},$$

then the circulant preconditioners proposed by Chan and Strang are superlinear [6]. More precisely, this is true for the Strang preconditioner when f belongs to the Dini–Lipschitz class [20], while for the Chan preconditioner this is true when f is merely continuous [7, 20, 21]. Moreover, many other preconditioners related to different matrix algebras [3, 4, 9, 11, 13, 21, 23, 24] can be proved to be optimal or superlinear even in the ill-conditioned case, where the symbol f has zeros [9, 20].

From the practical point of view, the profusion of research devoted to preconditioners for Toeplitz matrices is likely to proceed with the hope of using the same ideas for multilevel Toeplitz matrices (see [29]). In the multilevel case, the constructions similar to those of the unilevel case have led only to sublinear preconditioners [31, 21].

As yet, all attempts to construct a superlinear multilevel circulant or circulant-like preconditioner were unsuccessful. In this paper we show that this construction is impossible.

We thus propose a negative result that can be summarized as follows. If a preconditioner C_n for A_n is sought in the form $C_n = U_n D_n V_n$, with unitary matrices U_n and

V_n (still subject to some assumptions) given beforehand and so independent of A_n, and a diagonal matrix D_n to be chosen, then there are multilevel Toeplitz matrices A_n associated with a multivariate symbol, for which the preconditioner C_n cannot be superlinear.

All the same, much of what we discuss has little to do with the Toeplitz property itself; we use, instead, some general ergodic assumptions (that follow from the Szegő-like theorems in the Toeplitz case). Moreover, our results might apply also to the preconditioners for unilevel matrices.

In section 2, we present the basic theoretical tools allowing us to prove that a preconditioner is not superlinear. Then, in section 3 we recollect the notion, common toolkit, and some ergodic results for the multilevel matrices. In section 4, we present our main negative results concerning superlinear circulant-like preconditioners.

2. Basic statements. All matrices, for the time being, are unilevel of order n (with different n). The following two lemmas are almost evident, so they are stated without proofs.

LEMMA 2.1. *Assume that matrices C_n are nonsingular and uniformly bounded in n, in the spectral norm, together with C_n^{-1}. Then a preconditioner C_n for A_n is sublinear iff the singular values of $A_n - C_n$ have a general cluster at zero, and it is superlinear iff this cluster is proper.*

We will say that A_n and C_n are ε-close up to the rank bound $r = r(n,\varepsilon)$ if, $\forall\, \varepsilon > 0$ and $\forall\, n$, there exist matrices $E_{n\,\varepsilon}$ and $R_{n\,\varepsilon}$ such that

$$A_n - C_n = E_{n\,\varepsilon} + R_{n\,\varepsilon}, \quad ||E_{n\,\varepsilon}||_2 \le \varepsilon, \quad \operatorname{rank} R_{n\,\varepsilon} \le r(n,\varepsilon). \tag{4}$$

LEMMA 2.2. *The singular values of $A_n - C_n$ have a cluster at zero iff the matrices A_n and C_n are ε-close up to the rank bound $r(n,\varepsilon)$ which enjoys the relation*

$$\lim_{n\to\infty} \frac{r(n,\varepsilon)}{n} = 0 \quad \forall\, \varepsilon > 0. \tag{5}$$

The cluster at 0 is proper iff $r(n,\varepsilon)$ does not depend on n.

We call a matrix *equimodular* if all its entries are equal in modulus. It is clear that any unitary equimodular matrix of order n has all its entries equal to $1/\sqrt{n}$ in modulus.

LEMMA 2.3. *Let $C_n = U_n D_n V_n$, where matrices D_n are diagonal, U_n and V_n are unitary, and V_n are equimodular. Assume that, for any $\varepsilon > 0$ and for all sufficiently large n, there exists a column $e_{\varepsilon\,n}$ of I_n such that $||C_n e_{\varepsilon\,n}||_2 \le \varepsilon$. Then the singular values of C_n are clustered at zero.*

Proof. If $D_n = \operatorname{diag}\{d_{1n}, \ldots, d_{nn}\}$, then the singular values of C_n are equal to $|d_{jn}|$, $1 \le j \le n$. From the contrary, assume that zero is not a cluster for them. Then, there are positive δ and c such that

$$\gamma_n(\delta) \ge cn \tag{6}$$

for infinitely many n. In line with the previous notation, $\gamma_n(\delta)$ counts those indices j for which $|d_{jn}| > \delta$. Take up any n for which (6) is valid, and set $v^n_{\varepsilon\,n} = [v_{1,\varepsilon\,n}, \ldots, v_{n,\varepsilon\,n}]^T \equiv V_n\, e_{\varepsilon\,n}$. Then, for those n,

$$||C_n e_{\varepsilon\,n}||_2 \;=\; ||D_n v^n_{\varepsilon\,n}||_2 \;=\; \sqrt{\sum_{j=1}^{n} |d_{jn} v_{j,\varepsilon\,n}|^2} \;\ge\; \sqrt{c}\,\delta,$$

and hence, this cannot be less than ε for an arbitrary $\varepsilon > 0$. □

We will say that A_n are ε-zeroed on $\rho = \rho(n, \varepsilon)$ columns of the identity matrix I_n if, first, they are uniformly bounded in n in the spectral norm, and second, $\forall\ \varepsilon > 0$ and $\forall\ n$, for some $\rho(n, \varepsilon)$ columns $e_{jn\,\varepsilon}$ of I_n it holds that

$$||A_n\, e_{jn\,\varepsilon}||_2 \ \le\ \varepsilon, \quad j = 1, \ldots, \rho(n, \varepsilon).$$

LEMMA 2.4. *Let matrices A_n and C_n be ε-close with the rank bound $r(n, \varepsilon)$, and let A_n be ε-zeroed on $\rho(n, \varepsilon)$ columns of I_n so that*

$$\lim_{n\to\infty} \frac{r(n, \varepsilon)}{\rho(n, \varepsilon)} = 0. \tag{7}$$

Suppose $||C_n||_2$ are uniformly bounded in n. Then, $\forall\ \varepsilon > 0$ and all sufficiently large n, there is a column $e_{n\,\varepsilon}$ of I_n such that

$$||C_n\, e_{n\,\varepsilon}||_2 \ \le\ 3\,\varepsilon.$$

Proof. By contradiction, assume that

$$||C_n\, e_{jn\,\varepsilon}||_2 \ >\ 3\,\varepsilon, \quad j = 1, \ldots, \rho(n, \varepsilon).$$

Since A_n are ε-zeroed on the vectors $e_{jn\,\varepsilon}$, it implies that

$$||(A_n - C_n)\, e_{jn\,\varepsilon}||_2 \ >\ 2\,\varepsilon.$$

Denote by $\sigma_{1n} \ge \cdots \ge \sigma_{nn}$ the singular values of $A_n - C_n$. Then, as is well known,

$$\sum_{j=1}^{\rho} \sigma_{jn}^2 \ \ge\ \sum_{j=1}^{\rho} ||(A_n - C_n)\, e_{jn\,\varepsilon}||_2^2.$$

Since A_n and C_n are ε-close with the rank bound $r(n, \varepsilon)$, it follows that

$$\sigma_{jn} \le \varepsilon \quad \text{for} \quad j > r = r(n, \varepsilon).$$

Using this and taking into account the preceding inequalities, we obtain

$$r\ \sigma_{1n}^2 \ \ge\ 4\varepsilon^2 \cdot \rho - \varepsilon^2\ (\rho - r),$$

and hence

$$\frac{r(n, \varepsilon)}{\rho(n, \varepsilon)}\ ||A_n - C_n||_2^2 \ \ge\ 3\,\varepsilon^2.$$

On the strength of (7), and because the spectral norms of the matrices are bounded, the right-hand side must vanish as $n \to \infty$, which is at odds with the last inequality. □

THEOREM 2.1. *Suppose that $C_n = U_n D_n V_n$, where matrices D_n are diagonal, U_n and V_n are unitary, and V_n are equimodular. Let matrices A_n and C_n be ε-close with the rank bound $r(n, \varepsilon)$, and let A_n be ε-zeroed on $\rho(n, \varepsilon)$ columns of I_n. If the singular values of A_n are not clustered at zero, then* (7) *does not stand; that is,*

$$r(n, \varepsilon) \quad \neq \quad o(\rho(n, \varepsilon)), \qquad n \to \infty. \tag{8}$$

Proof. We do not lose generality if we assume that C_n are bounded in the spectral norm uniformly in n. If it is not the case, we transfer to other matrices $\tilde{C}_n$ satisfying the same premises with $r(n,\varepsilon)$ being possibly replaced by $2\,r(n,\varepsilon)$. Indeed, the number of singular values for C_n that are greater than $||A_n||_2 + \varepsilon$ cannot exceed $r(n,\varepsilon)$, and we may thus get to $\tilde{C}_n$ by cutting off the largest diagonal entries of D_n in the equation $C_n = U_n D_n V_n$.

Assume that the relation (7) is still fulfilled. Then, since all the premises of Lemma 2.4 are fulfilled, we conclude that, for any $\varepsilon > 0$ and for any n, there exists a column $e_{\varepsilon\, n}$ of I_n such that $||C_n e_{\varepsilon\, n}||_2 \le \varepsilon$. From Lemma 2.3, we now deduce that the singular values of C_n have a cluster at 0.

Allowing for (7), we have also $r(n,\varepsilon) = o(n)$, which implies that A_n and C_n have the same one-point clusters [31, 33]. However, the matrices A_n have no singular value cluster at 0, and this cannot be reconciled with the above conclusion that C_n have the cluster at 0. □

THEOREM 2.2. *Under the hypotheses of Theorem* 2.1, *assume that*

$$\lim_{n\to\infty} \rho(n,\varepsilon) = \infty \qquad \forall\, \varepsilon > 0.$$

Then the singular values of $A_n - C_n$ *cannot have a proper cluster at* 0.

Proof. From Lemma 2.2, $r(n,\varepsilon)$ is bounded uniformly in n. Then, obviously, $r(n,\varepsilon) = o(\rho(n,\varepsilon))$, which contradicts (8). □

THEOREM 2.3. *Under the hypotheses of Theorem* 2.2, *consider any matrices* P_n *such that* $P_n + A_n$ *and* $P_n + C_n$ *are nonsingular and uniformly bounded in* n, *in the spectral norm, together with their inverses. Then the matrices* $P_n + C_n$ *cannot be superlinear preconditioners for* $P_n + A_n$.

Proof. According to Lemma 2.1, the preconditioner $P_n + C_n$ for $P_n + A_n$ is superlinear iff the singular values of $A_n - C_n$ have a proper cluster at 0. All now reduces to applying Theorem 2.2. □

It might be useful to consider also *quasi-equimodular* matrices instead of equimodular ones. A sequence of matrices $V_n = [v_{ij}^n]$ of order n is referred to as quasi-equimodular if

$$\frac{c_1}{\sqrt{n}} \;\le\; |v_{ij}^n| \;\le\; \frac{c_2}{\sqrt{n}} \quad \forall\, i,j,$$

for some $0 < c_1 \le c_2$ independent of n. It is clear that the above statements remain valid for a quasi-equimodular sequence of matrices.

However, the same statements also stand in an even more general case of *sub-equimodular* matrices. In addition, this notion applies to a sequence of matrices, not to an individual matrix.

A sequence of matrices $V_n = [v_1^n, \ldots, v_n^n]$ is said to be subequimodular if, for any sequence of indices $j(n)$, $1 \le j(n) \le n$, $n = 1, 2, \ldots$, there exists $\delta > 0$ such that the set $[\delta, +\infty]$ is a cluster for $\{\sqrt{n}|v_{1\,j(n)}^n|, \ldots, \sqrt{n}|v_{n\,j(n)}^n|\}_n$, the absolute values of the entries of the $j(n)$th column times $\sqrt{n}$. It is not difficult to show that the above statements hold true for subequimodular matrices V_n.

3. Multilevel preliminaries. We call a matrix multilevel if it can be viewed as one with a nested block structure: A is a p-level matrix of multiorder $n = (n_1, \ldots, n_p)$ if it is a matrix with $n_1 \times n_1$ blocks, if each of them consists of $n_2 \times n_2$ blocks, and so on. Hence, A is of order $N(n) \equiv n_1 \cdots n_p$. The rows and columns of A can naturally be pointed to through multi-indices $i = (i_1, \ldots, i_p)$ and $j = (j_1, \ldots, j_p)$. Thus,

$$A = [a_{ij}], \quad 1 \le i, j \le n,$$

where $1 = (1, \dots, 1)$, and inequalities between multi-indices are understood in the entrywise sense.

Multilevel matrices may possess some structure: A_n is called a p-level Toeplitz matrix if $A_n = [a_{i-j}]$, and A_n is called a p-level circulant if $A_n = [a_{i-j \,(\mathrm{mod}\, n)}]$, where $i(\mathrm{mod}\, n) \equiv (i_1(\mathrm{mod}\, n_1), \dots, i_p(\mathrm{mod}\, n_p))$.

Given two sequences of p-level matrices A_n and B_n of multiorder n, we call their singular values equally distributed if, for any function F continuous with a bounded support,

$$\frac{1}{N(n)}\Sigma(A_n, F) - \frac{1}{N(n)}\Sigma(B_n, F) = o(1), \quad n \to \infty,$$

where

$$\Sigma(A_n, F) \equiv \sum_{1 \le i \le n} F(\sigma_i(A_n)),$$

and $n \to \infty$ means that every component of the multi-index n tends to infinity.

The singular values of multilevel matrices A_n are called distributed as $f(z)$ for $z \in \Pi$ if, for any F continuous with a bounded support,

$$\frac{1}{N(n)}\Sigma(A_n, F) \to \frac{1}{\mathrm{mes}\,\Pi} \int_\Pi F(f(z))dz.$$

Here, Π is a Lebesgue-measurable set and f is a Lebesgue-measurable function on Π.

If $f(z) = 1$, then the above means that the singular values of A_n have a general cluster at 1. It is not necessarily a proper cluster.

Let $z \in \Pi \equiv [-\pi, \pi]^p$, and consider $f \in L_1(\Pi)$ with the formal Fourier expansion

$$f(z) \sim \sum_k a_k \, e^{\mathrm{i}(k,z)},$$

where

$$k = (k_1, \dots, k_p), \quad (k, z) = k_1 z_1 + \cdots + k_p z_p.$$

Such a function f (we call it a symbol) naturally generates the p-level Toeplitz matrices as follows:

$$A_n(f) = [a_{i-j}], \quad 1 \le i, j \le n.$$

For the unilevel Hermitian–Toeplitz case, the classical distribution results can be found in [12]. The Szegő-like theorem [12] for the singular values of $A_n(f)$ obtained in [34] reads as follows: If a symbol $f \in L_1(\Pi)$ is a complex-valued function of $z \in \Pi = [-\pi, \pi]^p$, then the singular values of $A_n(f)$ are distributed as $|f(z)|$ with $z \in \Pi$.

4. Main negative results. Consider a p-variable symbol $f(x_1, \dots, x_p) = \frac{1}{2}\exp\{\mathrm{i}x_1 + \cdots + \mathrm{i}x_p\}$ and the corresponding p-level Toeplitz matrices $A_n = A_n(f)$, where $n = (n_1, \dots, n_p)$.

THEOREM 4.1. *For $I_n + A_n$, any preconditioner of the form $I_n + C_n$, where C_n is a p-level circulant, is not superlinear.*

Proof. As it is not difficult to verify, the matrices A_n are zeroed (ε-zeroed with $\varepsilon = 0$) on $\rho(n)$ columns of the identity matrix of order $N(n) = n_1 \cdots n_p$, where

$$\rho(n) \;\geq\; c\, N(n) \sum_{k=1}^{p} \frac{1}{n_k} \tag{9}$$

and $c \geq 2^{1-p}$ provided that $n_k \geq 2 \;\forall\, k$.

For example, consider the multilevel Toeplitz matrices A_n with $p = 2$ generated by $f(x_1, x_2) = \frac{1}{2}\, \exp\{\mathrm{i}x_1 + \mathrm{i}x_2\}$. In line with the above,

$$A_n = \frac{1}{2} \begin{bmatrix} 0 & J & & \\ & \ddots & \ddots & \\ & & 0 & J \\ & & & 0 \end{bmatrix},$$

where 0 and J are of order n_2 as $n = (n_1, n_2)$ and J is a Jordan block with $J_{i,i+1} = 1$ for $i = 1, \ldots, n_2 - 1$ and zero otherwise. It is evident that the first n_2 columns of A_n are zero columns. Moreover, each of the $n_1 - 1$ blocks J appearing in A_n has one zero column. Therefore, $\rho(n) = n_2 + n_1 - 1$, which is in accordance with (9).

Clearly, $\rho(n) \to \infty$ as $n \to \infty$ (which means, by definition, that every n_k tends to infinity). Note also that $C_n = V_n^* D_n V_n$, where V_n is a unitary equimodular matrix, for it is the Kronecker product of Fourier matrices. Therefore, we may have recourse to Theorem 2.3, and this completes the proof. □

In effect, the above proof enables us to claim much more. In particular, we can propose a lower estimate on the rank of $R_{n\,\varepsilon}$ in the splitting for $A_n - C_n$ concerning the ε-closeness of A_n and C_n. What is more, it is possible to cover a more general case when C_n is not necessarily a multilevel circulant. We call matrices C_n *circulant-like* if, for any n, it holds that $C_n = V_n^* D_n V_n$, where D_n is a diagonal matrix, and matrices V_n are unitary subequimodular matrices.

THEOREM 4.2. *For $I_n + A_n$, any sublinear preconditioner of the form $I_n + C_n$, where C_n are p-level circulant-like matrices, provides the singular value cluster for which, for some $c(\varepsilon) > 0$ and infinitely many n, it holds that*

$$\gamma_n(\varepsilon) \;\geq\; c(\varepsilon)\, \rho(n),$$

where $\rho(n)$ is defined by (9).

To demonstrate this, we follow the previous proof with the reference, in the end, to Theorem 2.1.

We know that, when using the multilevel T. Chan preconditioner [8], the number of outliers is $O(N(n) \sum_{k=1}^{p} \frac{1}{n_k})$ for any A_n generated by a continuous function f [21]. Therefore, we can say that this preconditioning technique, which is not satisfactory for large p, is the best that we may obtain when quasi-equimodular algebras are considered.

The last theorem apparently pertains to many popular matrix algebra preconditioners. We now have a rigorous proof stating that they cannot be superlinear, at least for the particular multilevel Toeplitz matrix-sequence A_n. However, all is not lost for the multilevel preconditioning in algebra.

Sometimes, it is possible to prove that

(a) multilevel circulant preconditioners cause the singular values to march inside some constant, separated from a zero interval;

(b) moreover, there is a general cluster at 1.

The combination of (a) and (b) can result in a very fast convergence in practice (see [5, 22]). We have as much as (a) and (b) together rather often. For example, if the convex hull of the range of f (complex-valued) is separated from zero, then the singular values of the Toeplitz matrices preconditioned by circulants behave in line with (a) [26]. At the same time, (b) is also valid for this case. Apart from its efficiency in experiments, a rigorous description of the effect of combining (a) and (b) is still to be found.

Note also that the basic results of section 2 allow us to construct many other examples. In particular, there are Hermitian multilevel Toeplitz matrices (generated by a multivariable symbol) for which any Hermitian multilevel circulant-like preconditioner is not superlinear.

To see this, take up the previous $f(x_1, \ldots, x_p)$ and consider the symbols

$$g_1 = \operatorname{Re} f(x_1, \ldots, x_p), \quad g_2 = \operatorname{Im} f(x_1, \ldots, x_p).$$

At least one of the sequences $I + A_n(g_1)$ or $I + A_n(g_2)$ cannot have circulant-like superlinear preconditioners. By contradiction, if both of them have superlinear preconditioners $I + C_{n,1}$ and $I + C_{n,2}$, then the circulant-like preconditioner $I + C_{n,1} + \mathrm{i}C_{n,2}$ is superlinear for $I + A_n(f)$ and this contradicts the main negative results of this section.

Acknowledgment. We are grateful to Dario Bini for encouragement.

REFERENCES

[1] O. Axelsson and V. Barker, *Finite Element Solution of Boundary Value Problems, Theory and Computation*, Academic Press, New York, 1984.

[2] O. Axelsson and G. Lindskög, *The rate of convergence of the preconditioned conjugate gradient method*, Numer. Math., 52 (1986), pp. 499–523.

[3] D. Bini and P. Favati, *On a matrix algebra related to the discrete Hartley transform*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 500–507.

[4] E. Bozzo and C. Di Fiore, *On the use of certain matrix algebras associated with discrete trigonometric transforms in matrix displacement decomposition*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 312–326.

[5] R. H. Chan and M. K. Ng, *Conjugate gradient methods for Toeplitz systems*, SIAM Rev., 38 (1996), pp. 427–482.

[6] R. H. Chan and G. Strang, *Toeplitz equations by conjugate gradients with circulant preconditioner*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 104–119.

[7] R. H. Chan and M. C. Yeung, *Circulant preconditioners for Toeplitz matrices with positive continuous generating functions*, Math. Comp., 58 (1992), pp. 233–240.

[8] T. F. Chan, *An optimal circulant preconditioner for Toeplitz systems*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 766–771.

[9] F. Di Benedetto, *Analysis of preconditioning techniques for ill-conditioned Toeplitz matrices*, SIAM J. Sci. Comput., 16 (1995), pp. 682–697.

[10] F. Di Benedetto, *Preconditioning of block Toeplitz matrices by sine transforms*, SIAM J. Sci. Comput., 18 (1997), pp. 499–515.

[11] F. Di Benedetto and S. Serra Capizzano, *A unifying approach to abstract matrix algebra preconditioning*, Numer. Math., 82 (1999), pp. 57–90.

[12] U. Grenander and G. Szegő, *Toeplitz Forms and Their Applications*, 2nd ed., Chelsea, New York, 1984.

[13] X. A. Jin, *Hartley preconditioners for Toeplitz systems generated by positive continuous functions*, BIT, 34 (1994), pp. 367–371.

[14] T. Kailath and V. Olshevsky, *Displacement structure approach to discrete-trigonometric-transform based preconditioners of G. Strang type and T. Chan type*, Calcolo, 33 (1996), pp. 191–208.

[15] T. Ku and C.-C. Jay Kuo, *On the spectrum of a family of preconditioned block Toeplitz matrices*, SIAM J. Sci. Stat. Comput., 13 (1992), pp. 948–966.

[16] M. Ng, *Band preconditioners for block-Toeplitz–Toeplitz-block systems*, Linear Algebra Appl., 259 (1997), pp. 307–327.
[17] S. Serra, *Preconditioning strategies for asymptotically ill-conditioned block Toeplitz systems*, BIT, 34 (1994), pp. 579–594.
[18] S. Serra, *Preconditioning strategies for Hermitian Toeplitz systems with nondefinite generating functions*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 1007–1019.
[19] S. Serra, *Optimal, quasi-optimal and superlinear band-Toeplitz preconditioners for asymptotically ill-conditioned positive definite Toeplitz systems*, Math. Comp., 66 (1997), pp. 651–665.
[20] S. Serra, *Superlinear PCG methods for symmetric Toeplitz systems*, Math. Comp., 88 (1999), pp. 793–803.
[21] S. Serra, *A Korovkin-type theory for finite Toeplitz operators via matrix algebras*, Numer. Math., 82 (1999), pp. 117–142.
[22] S. Serra, *The rate of convergence of Toeplitz based PCG methods for second order nonlinear boundary value problems*, Numer. Math., 81 (1999), pp. 461–495.
[23] S. Serra Capizzano, *Toeplitz preconditioners constructed from linear approximation processes*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 446–465.
[24] S. Serra Capizzano, *Korovkin theorems and linear positive Gram matrix algebras approximation of Toeplitz matrices*, Linear Algebra Appl., 284 (1998), pp. 307–334.
[25] S. Serra Capizzano and C. Tablino Possio, *Spectral and structural analysis of high precision finite difference matrices for elliptic operators*, Linear Algebra Appl., 293 (1999), pp. 85–131.
[26] S. Serra Capizzano and P. Tilli, *Extreme singular values and eigenvalues of non-Hermitian block Toeplitz matrices*, J. Comput. Appl. Math., 108 (1999), pp. 113–130.
[27] A. van der Sluis and H. A. van der Vorst, *The rate of convergence of conjugate gradients*, Numer. Math., 48 (1986), pp. 543–560.
[28] V. Strela and E. Tyrtyshnikov, *Which circulant preconditioner is better?*, Math. Comp., 65 (1996), pp. 137–150.
[29] E. E. Tyrtyshnikov, *Optimal and superoptimal circulant preconditioners*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 459–473.
[30] E. Tyrtyshnikov, *Circulant preconditioners with unbounded inverses*, Linear Algebra Appl., 216 (1995), pp. 1–23.
[31] E. Tyrtyshnikov, *A unifying approach to some old and new theorems on distribution and clustering*, Linear Algebra Appl., 232 (1996), pp. 1–43.
[32] E. Tyrtyshnikov, A. Yeremin, and N. Zamarashkin, *Clusters, preconditioners, convergence*, Linear Algebra Appl., 263 (1997), pp. 25–48.
[33] E. Tyrtyshnikov, *A Brief Introduction to Numerical Analysis*, Birkhauser, Boston, 1997.
[34] E. Tyrtyshnikov and N. Zamarashkin, *Spectra of multilevel Toeplitz matrices: Advanced theory via simple matrix relationships*, Linear Algebra Appl., 270 (1998), pp. 15–27.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 440–445
© 1999 Society for Industrial and Applied Mathematics

A BOUND ON THE SOLUTION TO A STRUCTURED SYLVESTER EQUATION WITH AN APPLICATION TO RELATIVE PERTURBATION THEORY*

REN-CANG LI†

Abstract. Assuming only that the spectra of A and B are disjoint as opposed to the more restrictive assumption previously used, we obtain a bound in all unitarily invariant norms on the solution to the *structured* Sylvester equation $AX - XB = A^{1/2}EB^{1/2}$. This bound is the first of its kind in all unitarily invariant norms under only the disjointedness assumption. An application of the bound to the relative perturbation theory for scaled Hermitian eigenvalue problems is given.

Key words. structured Sylvester equation, unitarily invariant norm, relative perturbation theory, eigenspace

AMS subject classifications. 15A18, 15A42, 65F15, 65F35, 65G99

PII. S0895479898349586

The operator (matrix) equation $AX - XB = S$, known as the Sylvester equation, appears frequently in the studies of operator theory, stability, and numerical linear algebra; see Bhatia and Rosenthal [6] and references therein for a history of it and its applications. The equation plays a unique role in eigenspace variations as illustrated by Davis and Kahan [7], where it was shown that the singular values of an associated Sylvester equation are the sines of angles between an eigenspace and its perturbation, and consequently, theorems, now known as the Davis–Kahan $\sin\theta$ and $\sin 2\theta$ theorems, were obtained.

Let A and B be Hermitian, and $\lambda(\cdot)$ denote the spectrum of a matrix. It can be proved that for the Frobenius norm $\|X\|_{\mathrm{F}} \stackrel{\mathrm{def}}{=} \sqrt{\sum_{i,j} |(X)_{ij}|^2}$,

$$\|X\|_{\mathrm{F}} \le \|S\|_{\mathrm{F}}/\delta, \quad \delta \stackrel{\mathrm{def}}{=} \min_{\omega\in\lambda(A),\,\gamma\in\lambda(B)} |\omega - \gamma|. \tag{1}$$

This holds under the most mild assumption, i.e., $\lambda(A)\cap\lambda(B) = \emptyset$ as in Figure 1. But to derive bounds for other norms, Davis and Kahan [7] imposed a restriction on how $\lambda(A)$ and $\lambda(B)$ are relatively distributed; see Figure 2. With that restriction Davis and Kahan proved that

$$|||X||| \le |||S|||/\delta \tag{2}$$

for all unitarily invariant norms $|||\cdot|||$ [3, 14], including the spectral (operator) norm $\|\cdot\|_2$. It remained open what kinds of bounds we may have for $|||X|||$ under the situation of Figure 1 until Bhatia, Davis, and McIntosh [4] extended (2) to cover the situation of Figure 1 and proved[1]

$$|||X||| \le (\pi/2)\cdot|||S|||/\delta. \tag{3}$$

(In [4], the constant $\pi/2$ was not given explicitly; see [6] for details.)

*Received by the editors December 11, 1998; accepted for publication (in revised form) by R. Bhatia March 3, 1999; published electronically November 2, 1999. This work was partially supported by the National Science Foundation under grant ACI-9721388.
http://www.siam.org/journals/simax/21-2/34958.html

†Department of Mathematics, University of Kentucky, Lexington, KY 40506 (rcli@ms.uky.edu).

[1] Bhatia, Davis, and McIntosh proved a more general result for the case when A and B are normal.

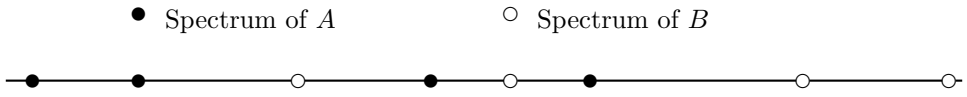


FIG. 1. *The spectrum of A and that of B are disjoint.*

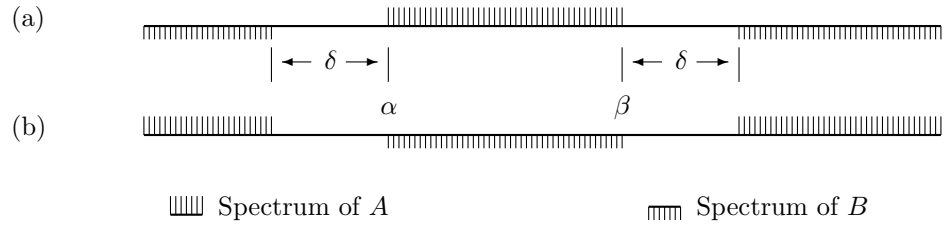


FIG. 2. *The spectrum of A and that of B are separated by two intervals.*

Interestingly, Li [13], in creating a relative perturbation theory for eigenvectors (see also [10]), showed that the sines of angles between an eigenspace and its perturbation are the singular values of an associated *structured* Sylvester equation $AX - XB = S$ in the sense that S takes one of the following forms:

$$AE, FB, AE + FB, \text{ or } A^{1/2}EB^{1/2}, \tag{4}$$

where E and F are some error matrices. Bounds on the norms of X that are inversely proportional to relative gaps between $\lambda(A)$ and $\lambda(B)$ measured by one of[2] $(1 \le p \le \infty)$

$$\frac{|\omega-\gamma|}{|\omega|}, \frac{|\omega-\gamma|}{|\gamma|}, \frac{|\omega-\gamma|}{\sqrt[p]{|\omega|^p+|\gamma|^p}}, \text{ or } \frac{|\omega-\gamma|}{\sqrt{|\omega|\cdot|\gamma|}}, \tag{5}$$

as opposed to the absolute gap δ measured by $|\omega - \gamma|$, are therefore obtained. (Here $A^{1/2}$ can be taken to be any one of A's many square roots. Usually for a positive semidefinite A, it is defined to be the unique square root that is also positive semidefinite.) This was done for the Frobenius norm under the situation of Figure 1, but for all other unitarily invariant norms, a restriction similar to Figure 2 was assumed; see Figure 3. Notice the difference between Figure 2 and Figure 3: in Figure 3, either the spectrum of A or that of B has to scatter around the origin. This is for a good reason: the absolute difference is shift invariant while the relative differences are shift varying; see Li [13] for more discussions. For a sample bound, let $S = A^{1/2}EB^{1/2}$, then under the condition of Figure 3,

$$\|\!|X|\!\| \le \|\!|E|\!\| / \eta, \quad \eta \stackrel{\text{def}}{=} \min_{\omega\in\lambda(A),\,\gamma\in\lambda(B)} \frac{|\omega-\gamma|}{\sqrt{|\omega|\cdot|\gamma|}} \ge \delta/\sqrt{\alpha(\alpha+\delta)}. \tag{6}$$

A natural question arises: under the condition of Figure 1, can we have, e.g., for $AX - XB = A^{1/2}EB^{1/2}$, something like

$$\|\!|X|\!\| \le c\,\|\!|E|\!\| / \eta? \tag{7}$$

[2]It was proved that $\frac{|\omega-\gamma|}{\sqrt[p]{|\omega|^p+|\gamma|^p}}$ is a metric on the set of real numbers in [12] which also conjectured that it might even be a metric on the set of complex numbers. The conjecture was confirmed by Day [8] for $p = \infty$ and recently by Barrlund [2] for all $1 \le p \le \infty$.

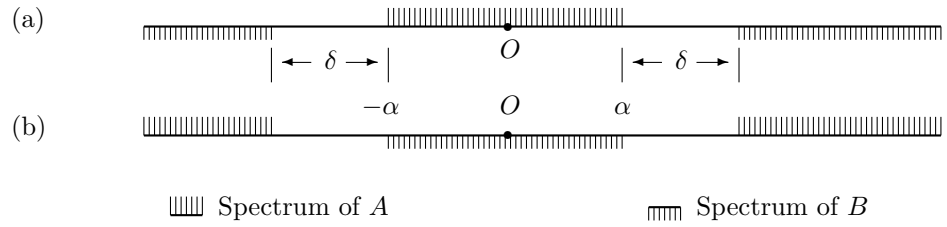


FIG. 3. *The spectrum of A and that of B are separated by two intervals, and one of the spectra scatters around the origin.*

Here $c \geq 1$ is a constant, and $\eta \stackrel{\text{def}}{=} \min_{\omega\in\lambda(A),\,\gamma\in\lambda(B)} \frac{|\omega-\gamma|}{\sqrt{|\omega|\cdot|\gamma|}}$. Similar questions with η defined accordingly for S being of any other forms in (4) can be asked.

While we are not able to give a complete answer to this question, in this note we shall prove something that is very much like it. However, we have no answers for S being of any other forms in (4).

THEOREM 1. *Assume that A and B are positive semidefinite and that $\lambda(A) \cap \lambda(B) = \emptyset$. Then $AX - XB = A^{1/2}EB^{1/2}$ has a unique solution X, and moreover for all unitarily invariant norm $\|\!|\cdot|\!\|$,*

$$|\!|\!| X |\!|\!| \leq (\pi/2) \cdot |\!|\!| E |\!|\!| / \widetilde{\eta}, \quad \widetilde{\eta} \stackrel{\text{def}}{=} \min_{\omega\in\lambda(A),\,\gamma\in\lambda(B)} |\ln(\omega/\gamma)|. \tag{8}$$

It has been proved that all relative measures in (5) are topologically equivalent [12], so they are equally good. The first two in (5) are the classical measures usually used in numerical computations, and the last two were introduced later [1, 9, 12, 13] as circumstances called for. What we see now in (8) is yet another relative measure $|\ln(\omega/\gamma)|$. To see this, let us assume ω and γ are close enough such that $|1-\omega/\gamma| < 1$. (As a matter of fact, $|1-\omega/\gamma| \leq 1/2$ if the most significant binary digit of ω and that of γ are the same; $|1-\omega/\gamma| \leq 1/10$ if the most significant decimal digit of ω and that of γ are the same.) Then

$$\begin{aligned} |\ln(\omega/\gamma)| &= |\ln(1-(1-\omega/\gamma)| \\ &= |(1-\omega/\gamma) + (1-\omega/\gamma)^2/2 + \cdots + (1-\omega/\gamma)^k/k + \cdots|, \end{aligned}$$

and thus the topological equivalence of $|\ln(\omega/\gamma)|$ to any of the measures in (5) is easily established.

Proof of Theorem 1. Without loss of generality, we may assume that both A and B are positive definite; we can always handle the singular case by the often used continuity argument. Our proof relies on the following inequality:

$$|\!|\!| (\ln A)X - X(\ln B) |\!|\!| \leq \left|\!\left|\!\left| A^{1/2}XB^{-1/2} - A^{-1/2}XB^{1/2} \right|\!\right|\!\right|, \tag{9}$$

due to Kosaki [11]; see [5] for another proof. It follows from the equation $AX - XB = A^{1/2}EB^{1/2}$ that $E = A^{1/2}XB^{-1/2} - A^{-1/2}XB^{1/2}$. Thus

$$|\!|\!| E |\!|\!| \geq |\!|\!| (\ln A)X - X(\ln B) |\!|\!| .$$

Now we apply (3), a result of Bhatia, Davis, and McIntosh [4], to complete the proof. ∎

Now let us briefly discuss how Theorem 1 can be applied to the relative perturbation theory. From now on, we shall assume $A = S^*HS$ is an $n \times n$ matrix, perturbed to $\widetilde{A} = S^*\widetilde{H}S$, where S is a scaling matrix and usually diagonal. But this is not necessary to the theorem below. S^* is the complex conjugate transpose of S. The elements of S can vary wildly. H is nonsingular and usually better conditioned than A itself. Set $\Delta H \stackrel{\text{def}}{=} \widetilde{H} - H$, and assume $\|H^{-1}\|_2\|\Delta H\|_2 < 1$. Let A and $\widetilde{A}$ admit eigendecompositions

(10)
$$A = (U_1, U_2)\begin{pmatrix} \Lambda_1 & \\ & \Lambda_2 \end{pmatrix}\begin{pmatrix} U_1^* \\ U_2^* \end{pmatrix} \quad \text{and} \quad \widetilde{A} = (\widetilde{U}_1, \widetilde{U}_2)\begin{pmatrix} \widetilde{\Lambda}_1 & \\ & \widetilde{\Lambda}_2 \end{pmatrix}\begin{pmatrix} \widetilde{U}_1^* \\ \widetilde{U}_2^* \end{pmatrix},$$

where $U = \overset{k}{\begin{pmatrix} U_1 \end{pmatrix}} \overset{n-k}{\begin{pmatrix} U_2 \end{pmatrix}}$, $\widetilde{U} = \overset{k}{\begin{pmatrix} \widetilde{U}_1 \end{pmatrix}} \overset{n-k}{\begin{pmatrix} \widetilde{U}_2 \end{pmatrix}}$ are unitary, and

$$\Lambda_1 = \operatorname{diag}(\lambda_1, \ldots, \lambda_k), \quad \Lambda_2 = \operatorname{diag}(\lambda_{k+1}, \ldots, \lambda_n), \tag{11}$$

$$\widetilde{\Lambda}_1 = \operatorname{diag}(\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_k), \quad \widetilde{\Lambda}_2 = \operatorname{diag}(\widetilde{\lambda}_{k+1}, \ldots, \widetilde{\lambda}_n). \tag{12}$$

Notice that

$$\begin{aligned} A &= S^*HS = (H^{1/2}S)^*\, H^{1/2}S, \\ \widetilde{A} &= S^*H^{1/2}\big(I + H^{-1/2}(\Delta H)H^{-1/2}\big)H^{1/2}S \\ &= \Big(\big(I + H^{-1/2}(\Delta H)H^{-1/2}\big)^{1/2}H^{1/2}S\Big)^*\, \big(I + H^{-1/2}(\Delta H)H^{-1/2}\big)^{1/2}H^{1/2}S. \end{aligned}$$

Set $B = S^*H^{1/2}$ and $\widetilde{B} = S^*H^{1/2}\big(I + H^{-1/2}(\Delta H)H^{-1/2}\big)^{1/2} \stackrel{\text{def}}{=} BD$, where $D = \big(I + H^{-1/2}(\Delta H)H^{-1/2}\big)^{1/2}$. Then $A = BB^*$ and $\widetilde{A} = \widetilde{B}\widetilde{B}^*$. Given the eigendecompositions of A and $\widetilde{A}$ as in (10)–(12), it can be seen that B and $\widetilde{B}$ admit the following singular value decompositions:

$$B = (U_1, U_2)\begin{pmatrix} \Lambda_1^{1/2} & \\ & \Lambda_2^{1/2} \end{pmatrix}\begin{pmatrix} V_1^* \\ V_2^* \end{pmatrix} \text{ and } \widetilde{B} = (\widetilde{U}_1, \widetilde{U}_2)\begin{pmatrix} \widetilde{\Lambda}_1^{1/2} & \\ & \widetilde{\Lambda}_2^{1/2} \end{pmatrix}\begin{pmatrix} \widetilde{V}_1^* \\ \widetilde{V}_2^* \end{pmatrix},$$

where U_i, $\widetilde{U}_i$ are the same as in (10), $\overset{k}{\begin{pmatrix} V_1 \end{pmatrix}} \overset{n-k}{\begin{pmatrix} V_2 \end{pmatrix}}$ and $\overset{k}{\begin{pmatrix} \widetilde{V}_1 \end{pmatrix}} \overset{n-k}{\begin{pmatrix} \widetilde{V}_2 \end{pmatrix}}$ are unitary. We have

$$\widetilde{A} - A = \widetilde{B}\widetilde{B}^* - BB^* = \widetilde{B}D^*B^* - \widetilde{B}D^{-1}B^* = \widetilde{B}(D^* - D^{-1})B^*.$$

Pre- and post-multiply the equations by $\widetilde{U}_2^*$ and U_1, respectively, to get

$$\widetilde{\Lambda}_2\widetilde{U}_2^*U_1 - \widetilde{U}_2^*U_1\Lambda_1 = \widetilde{\Lambda}_2^{1/2}\widetilde{V}_2^*(D^* - D^{-1})V_1\Lambda_1^{1/2}, \tag{13}$$

a *structured* Sylvester equation for $\widetilde{U}_2^*U_1$. Notice that for any unitarily invariant norm

$$\begin{aligned}
\left|\!\left|\!\left| \widetilde{V}_2^*(D^* - D^{-1})V_1 \right|\!\right|\!\right| &\le \left|\!\left|\!\left| D^* - D^{-1} \right|\!\right|\!\right| \\
&= \left|\!\left|\!\left| \left(I + H^{-1/2}(\Delta H)H^{-1/2}\right)^{1/2} - \left(I + H^{-1/2}(\Delta H)H^{-1/2}\right)^{-1/2} \right|\!\right|\!\right| \\
&\le \left\| \left(I + H^{-1/2}(\Delta H)H^{-1/2}\right)^{-1/2} \right\|_2 \left|\!\left|\!\left| H^{-1/2}(\Delta H)H^{-1/2} \right|\!\right|\!\right| \\
&\le \frac{\|H^{-1}\|_2 \left|\!\left|\!\left| \Delta H \right|\!\right|\!\right|}{\sqrt{1 - \|H^{-1}\|_2 \|\Delta H\|_2}}.
\end{aligned}$$

It is known that $|||\sin\Theta(U_1, \widetilde{U}_1)||| \equiv |||\widetilde{U}_2^* U_1|||$ for all unitarily invariant norms [14]. The following theorem is now a consequence of (13) and Theorem 1.

THEOREM 2. *Let $A = S^*HS$ and $\widetilde{A} = S^*\widetilde{H}S$ be two $n \times n$ Hermitian matrices with eigendecompositions* (10)–(12), *where H is positive definite and $\|H^{-1}\|_2\|\Delta H\|_2 < 1$. If*

$$\widetilde{\eta} \stackrel{\text{def}}{=} \min_{\mu \in \lambda(\Lambda_1),\, \widetilde{\mu} \in \lambda(\widetilde{\Lambda}_2)} |\ln(\mu/\widetilde{\mu})| > 0,$$

then for any unitarily invariant norm $|||\cdot|||$

$$\left|\!\left|\!\left| \sin\Theta(U_1, \widetilde{U}_1) \right|\!\right|\!\right| \le (\pi/2) \cdot \frac{\left|\!\left|\!\left| D - D^{-1} \right|\!\right|\!\right|}{\widetilde{\eta}}, \tag{14}$$

$$\le (\pi/2) \cdot \frac{\|H^{-1}\|_2}{\sqrt{1 - \|H^{-1}\|_2\|\Delta H\|_2}} \frac{\left|\!\left|\!\left| \Delta H \right|\!\right|\!\right|}{\widetilde{\eta}}, \tag{15}$$

where $D = (I + H^{-1/2}(\Delta H)H^{-1/2})^{1/2} = D^*$.

An analogous theorem for all unitarily invariant norms except those under the condition of Figure 3 is given in [13].

We have obtained a bound in all unitarily invariant norms on a *structured* Sylvester equation $AX - XB = A^{1/2}EB^{1/2}$. Though its proof is rather simple once we realize the two crucial results due to Kosaki [11] and to Bhatia, Davis, and McIntosh [4], respectively, it is significant in that for *structured* Sylvester equations it is the first of its kind in all unitarily invariant norms under the weakest assumption on A's and B's spectrum distributions. It remains to be seen whether similar results can be proved for other kinds of *structured* Sylvester equations.

Acknowledgment. I would like to thank Professor R. Bhatia for bringing the inequality (9) to my attention and for his constructive suggestions which improved the presentation considerably.

REFERENCES

[1] J. BARLOW AND J. DEMMEL, *Computing accurate eigensystems of scaled diagonally dominant matrices*, SIAM J. Numer. Anal., 27 (1990), pp. 762–791.
[2] A. BARRLUND, *The p-relative distance is a metric*, SIAM J. Matrix Anal. Appl., to appear.
[3] R. BHATIA, *Matrix Analysis*, Grad. Texts in Math. 169, Springer, New York, 1996.
[4] R. BHATIA, C. DAVIS, AND A. MCINTOSH, *Perturbation of spectral subspaces and solution of linear operator equations*, Linear Algebra Appl., 52–53 (1983), pp. 45–67.
[5] R. BHATIA AND K. R. PARTHASARATHY, *Positive Definite Functions and Operator Inequalities*, unpublished manuscript, 1998.
[6] R. BHATIA AND P. ROSENTHAL, *How and why to solve the operator equation $AX - XB = Y$*, Bull. London Math. Soc., 29 (1997), pp. 1–21.

[7] C. Davis and W. Kahan, *The rotation of eigenvectors by a perturbation.* III, SIAM J. Numer. Anal., 7 (1970), pp. 1–46.
[8] D. Day, *A New Metric on the Complex Numbers*, Technical Report Sand 98-1754, Sandia National Laboratories, Albuquerque, NM, 1998.
[9] J. Demmel and K. Veselić, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1204–1245.
[10] S. C. Eisenstat and I. C. F. Ipsen, *Relative perturbation techniques for singular value problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1972–1988.
[11] H. Kosaki, *Arithmetic-Geometric Mean and Related Inequalities for Operators*, Preprint Kyushu-MPS-1997-29, Kyushu University, Fukuoka, 1997.
[12] R.-C. Li, *Relative perturbation theory:* I. *Eigenvalue and singular value variations*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 956–982.
[13] R.-C. Li, *Relative perturbation theory:* II. *Eigenspace and singular subspace variations*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 471–492.
[14] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 446–453
© 1999 Society for Industrial and Applied Mathematics

THE SOLUTION TO A STRUCTURED MATRIX APPROXIMATION PROBLEM USING GRASSMAN COORDINATES*

CRAIG S. MACINNES†

Abstract. A method for finding the best approximation of a matrix A by a full rank Hankel matrix is given. The initial problem of best approximation of one matrix by another is transformed to a problem involving best approximation of a given vector by a second vector whose elements are constrained so that its inverse image is a Hankel matrix. The map from a matrix to a vector is the invertible map between a subspace represented as the row space of the matrix A and the Grassman vector representing that subspace. The relation between the principle angles associated with a pair of subspaces and the angle between the Grassman vectors associated with the subspaces is established.

Key words. Grassman vector, principal angle, symmetric polynomial, Hankel matrix

AMS subject classifications. 15A15, 15A21, 15A57

PII. S0895479898335453

1. Introduction. The structured matrix approximation problem is of particular interest in signal processing [3], [10], where replacing a data matrix corresponding to a signal plus noise with a matrix known to have the correct underlying structure is used as a means of increasing the relative signal strength. This paper deals with the approximation of an arbitrary matrix, A, by a full rank Hankel matrix of the same dimensions. The approximation is accomplished by first finding the vector v whose entries are the Grassman coordinates for the subspace represented by the rows of A. Then, a Grassman vector g is fitted to v, where g is structured so that its inverse image is a matrix H', which is then premultiplied by a certain Hankel matrix, giving the entire matrix a Hankel structure.

Grassman coordinates are introduced in section 2. A relation between a vector subspace in $\mathbf{C}^N$ and the corresponding Grassman vector representing that subspace can be made whenever certain constraints are placed on the elements of the vectors. In general these constraints are the Plücker relations [6]. For the case where the matrix is Hankel, more restrictive relations are satisfied by the elements of the corresponding Grassman vectors. In section 3 it is shown that these relations are equivalent to requiring that the elements of the Grassman vectors be specific functions of the elementary symmetric polynomials. In section 4, a relation between the principal angles between two subspaces and the inner product of the corresponding Grassman vectors is established. Conclusions are given in section 5.

2. Grassman coordinates. Let U be a K-dimensional subspace of an N-dimensional vector space V, and let U be expressed as a $K \times N$ matrix A of *row vectors* [6]. Then, the Grassman vector $\varphi(U)$ associated with the subspace U is the Kth compound matrix of A [7]. As an example, consider the following matrix of row

*Received by the editors March 11, 1998; accepted for publication (in revised form) by R. Brualdi December 28, 1998; published electronically November 2, 1999. This work was performed by an employee of the U.S. Government or under U.S. Government contract. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/simax/21-2/33545.html

†Department of Physics, Informatics and Mathematics, Universidad Peruana Cayetano Heredia, Av. Honorio Delgado 430, Lima 31, Peru (ximefe@tsi.com.pe).

vectors representing a two-dimensional subspace in $\mathbf{R}^4$:

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{pmatrix}. \tag{2.1}$$

Then, the Grassman (Plücker) coordinates of A are

$$\varphi(A) = \left(\begin{vmatrix} 1 & 2 \\ 5 & 6 \end{vmatrix}, \begin{vmatrix} 1 & 3 \\ 5 & 7 \end{vmatrix}, \begin{vmatrix} 1 & 4 \\ 5 & 8 \end{vmatrix}, \begin{vmatrix} 2 & 3 \\ 6 & 7 \end{vmatrix}, \begin{vmatrix} 2 & 4 \\ 6 & 8 \end{vmatrix}, \begin{vmatrix} 3 & 4 \\ 7 & 8 \end{vmatrix} \right).$$

The components of $\varphi(A)$ are ordered lexicographically:

$$\varphi(A) = (p_{12}, p_{13}, p_{14}, p_{23}, p_{24}, p_{34}).$$

The Grassman vectors corresponding to any two sets of row vectors spanning the same space differ only by a scalar multiple. There is an inverse map φ^{-1} taking a Grassman vector to a corresponding subspace. In general the matrix representing the subspace can be formed by choosing any element $p_{i_1,i_2,\ldots,i_N}$ of the Grassman vector. The jth row of the matrix representing the subspace is then given by the elements

$$p_{i_1,\ldots,i_{j-1},1,i_{j+1},\ldots,i_N}, p_{i_1,\ldots,i_{j-1},2,i_{j+1},\ldots,i_N}, \ldots, p_{i_1,\ldots,i_{j-1},N,i_{j+1},\ldots,i_N}.$$

The subspace corresponding to the Grassman vector $\varphi(A)$ above can be found by choosing any nonzero element, say, p_{12}, and forming the following row matrix:

$$\widehat{A} \overset{\text{def}}{=} \begin{pmatrix} \cdots & p_{i2} & \cdots \\ \cdots & p_{1j} & \cdots \end{pmatrix} \equiv \begin{pmatrix} p_{12} & p_{22} & p_{32} & p_{42} \\ p_{11} & p_{12} & p_{13} & p_{14} \end{pmatrix} = \begin{pmatrix} -4 & 0 & 4 & 8 \\ 0 & -4 & -8 & -12 \end{pmatrix}, \tag{2.2}$$

where the row spaces of A and $\hat{A}$ coincide. It should be noted that in general, not all vectors are Grassman vectors, i.e., represent subspaces.

3. The power sum relations. Consider a $K \times N$ Hankel matrix, H. The $K + N - 1$ distinct elements of H,

$$\{h_{11},\ h_{21}, \ldots, h_{K1},\ h_{K2},\ \ldots,\ h_{KN}\},$$

occur in the first column and last row. The following theorem states that these elements may be considered to be weighted sums of exponentials; see Grenander [5].

THEOREM 3.1 (due to Carathéodory). *Let $c_1, c_2, \ldots, c_n$ be given complex constants not all zero, with $n > 1$. There exist an integer m, $1 \le m \le n$, and certain constants $\rho_k, e_k,\ k = 1, 2, \ldots, m$, such that $\rho_k > 0,\ \|e_k\| = 1,\ e_k \ne e_l, k \ne l$, and such that*

$$c_\nu = \sum_{k=1}^{m} \rho_k e_k^\nu, \quad \nu = 1, 2, \ldots, n.$$

The integer m and the constants $\rho_k,\ e_k$ are uniquely determined.

Lanczos [8] called this type of term a *weighted power sum* and this terminology will be used here.

The relationship between a Hankel (or Toeplitz) matrix and power sums is of practical interest because Hankel matrices with elements which are power sums occur in many signal processing applications. One of these applications involves the so-called

frequencies of sinusoids problem [12], as follows. Consider a sequence $\{b_n\}$ with each term consisting of the weighted sum of K complex exponentials, with frequencies f_j, magnitudes a_j, and phases φ_j, for $j = 1, 2, \ldots, K$ and $n = 1, 2, \ldots$:

$$b_n \stackrel{\text{def}}{=} a_1 \exp(i(2\pi f_1 n + \varphi_1)) + \cdots + a_K \exp(i(2\pi f_K n + \varphi_K)) = \sum_{j=1}^{K} \alpha_j z_j^n, \tag{3.1}$$

where

$$\alpha_j = a_j \exp(i\varphi_j) \text{ and } z_j = \exp(i2\pi f_j), \quad j = 1, 2, \ldots, K \quad \text{with } i \stackrel{\text{def}}{=} \sqrt{-1}.$$

Given the sequence $\{b_n\}$, there are various methods for finding the parameters $\{\alpha_j,\ z_j\}$. The linear prediction method begins by arranging the elements of the sequence in Hankel or Toeplitz matrix form [11].

When a Hankel matrix A representing a subspace has entries as in (3.1) which are weighted power sums, any Grassman vector which represents the corresponding subspace satisfies conditions in addition to the Plücker relations. These conditions will be called the *power sum relations* and will now be derived.

Let $s_i,\ i = 1, \ldots, K$, denote the ith elementary symmetric polynomial; see Lang [9] where it is shown that any symmetric polynomial is a polynomial function of the elementary symmetric polynomials.

Let A be a $K \times N$, $(K \leq N)$, Hankel matrix whose entries are power sums, where $M = K$:

$$A \stackrel{\text{def}}{=} \begin{pmatrix} \sum_{i=1}^{M} \alpha_i z_i^1 & \sum_{i=1}^{M} \alpha_i z_i^2 & \cdots & \sum_{i=1}^{M} \alpha_i z_i^N \\ \sum_{i=1}^{M} \alpha_i z_i^2 & \sum_{i=1}^{M} \alpha_i z_i^3 & \cdots & \sum_{i=1}^{M} \alpha_i z_i^{N+1} \\ \vdots & \vdots & & \vdots \\ \sum_{i=1}^{M} \alpha_i z_i^K & \sum_{i=1}^{M} \alpha_i z_i^{K+1} & \cdots & \sum_{i=1}^{M} \alpha_i z_i^{N+K-1} \end{pmatrix}. \tag{3.2}$$

Consider the matrix consisting of the first K columns of A: $P = [a_1\ a_2, \ldots, a_K]$. Then, $\det P$ is the first element $p_{1\ 2,\ldots,K}$ in the Grassman vector $\varphi(A)$.

LEMMA 3.2. *The determinant of the power sum matrix P is*

$$\det P = \left(\prod_{i=1}^{K} \alpha_i \right) \left(\prod_{i=1}^{K} z_i \right) (\det V)^2, \tag{3.3}$$

where V is the Vandermonde matrix

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ z_1 & z_2 & \cdots & z_K \\ \vdots & \vdots & & \vdots \\ z_1^{K-1} & z_2^{K-1} & \cdots & z_K^{K-1}, \end{pmatrix}.$$

Proof. We have that $\det V = \prod_{\substack{i,j=1 \\ i>j}}^{K} (z_i - z_j)$. Note that the determinant is a multilinear function of its columns. By viewing each column of $\det P$ as a sum of K

columns, we deduce that $\det P$ is a sum of K^K terms, $K!$ of which are nonvanishing. These terms occur whenever $z_i \neq z_j,\ i \neq j$, from the expression for the determinant above.

The entire determinant $\det P$ is

$$\det P = \left(\prod_{i=1}^{K} \alpha_i\right) \left\{ \sum_{\sigma \in S_K} \left(\prod_{i=1}^{K} z_{\sigma(i)}^{i}\right) (-1)^{\sigma} \right\} \det V. \tag{3.4}$$

Here, the subscript $\sigma(i)$ of $z_{\sigma(i)}^{i}$ is an index and the superscript, i, is an exponent. S_K is the group of permutations on K letters.

The second factor in (3.4) is simply the Vandermonde determinant again times the product of the $z_i's$:

$$\left\{ \sum_{\sigma \in S_K} \left(\prod_{i=1}^{K} z_{\sigma(i)}^{i}\right) (-1)^{\sigma} \right\} = \left(\prod_{i=1}^{K} z_i\right) \det V. \qquad \square$$

Consider a general power sum matrix $GP = [a_{q_1}, a_{q_2}, \ldots, a_{q_K}]$ where the K columns $\{q_1, q_2, \ldots, q_K\}$ are from the original Hankel matrix, A, given in equation (3.2). Assume without loss of generality that $1 \leq q_1 < q_2 < \cdots < q_K \leq N$:

$$GP \stackrel{\text{def}}{=} \begin{pmatrix} \sum_{i=1}^{M} \alpha_i z_i^{q_1} & \sum_{i=1}^{M} \alpha_i z_i^{q_2} & \cdots & \sum_{i=1}^{M} \alpha_i z_i^{q_K} \\ \sum_{i=1}^{M} \alpha_i z_i^{q_1+1} & \sum_{i=1}^{M} \alpha_i z_i^{q_2+1} & \cdots & \sum_{i=1}^{M} \alpha_i z_i^{q_K+1} \\ \vdots & \vdots & & \vdots \\ \sum_{i=1}^{M} \alpha_i z_i^{q_1+K-1} & \sum_{i=1}^{M} \alpha_i z_i^{q_2+K-1} & \cdots & \sum_{i=1}^{M} \alpha_i z_i^{q_K+K-1} \end{pmatrix}.$$

Here, $\det GP$ corresponds to the Grassman vector element $p_{q_1 \ \cdots \ q_K}$. Following the same line of argument as above, the factorization can be shown to be

$$\det GP = \left(\prod_{i=1}^{N} \alpha_i\right) \left\{ \sum_{\sigma \in S_K} \left(\prod_{i=1}^{K} z_{\sigma(i)}^{q_i}\right) (-1)^{\sigma} \right\} \det V.$$

The second term involves the determinant of what will be called the *general Vandermonde matrix*, GV:

$$\left\{ \sum_{\sigma \in S_K} \left(\prod_{i=1}^{K} z_{\sigma(i)}^{q_i}\right) (-1)^{\sigma} \right\} = \left(\prod_{i=1}^{K} z_i^{q_1}\right) \det \begin{pmatrix} 1 & 1 & \cdots & 1 \\ z_1^{r_2} & z_2^{r_2} & \cdots & z_K^{r_2} \\ \vdots & \vdots & & \vdots \\ z_1^{r_K} & z_2^{r_K} & \cdots & z_K^{r_K} \end{pmatrix}.$$

Here, r_i is defined as $(q_i - q_1),\ i = 2, \ldots, K$. The entire factorization of the determinant of the general power sum matrix GP is then

$$\det GP = \left(\prod_{i=1}^{N} \alpha_i\right) \left(\prod_{i=1}^{K} z_i^{q_1}\right) \det V \det GV. \tag{3.5}$$

We now have the following lemma.

LEMMA 3.3. *Let a Grassman vector, g, be derived from a Hankel matrix. Then the elements of g normalized by the first element are symmetric polynomials.*

Proof. Divide the factorization for $\det GP$ in equation (3.5) by the factorization of $\det P$ in (3.3):

$$\frac{p_{q_1\ q_2\ \cdots\ q_K}}{p_{1\ 2\ \cdots\ K}} = \frac{\det GP}{\det P} = \frac{\left(\prod_{i=1}^K \alpha_i\right)\left(\prod_{i=1}^K z_i^{q_1}\right)(\det V)(\det GV)}{\left(\prod_{i=1}^K \alpha_i\right)\left(\prod_{i=1}^K z_i\right)(\det V)^2} = \left(\prod_{i=1}^K z_i^{q_1-1}\right)\frac{\det GV}{\det V}. \tag{3.6}$$

If $z_i = z_j$, $i \neq j$, then $\det GV = 0$. Therefore, the polynomial $\det GV$ is divisible by $z_i - z_j$ and it follows that $\frac{\det GV}{\det V}$ is a polynomial. Interchanging z_i and z_j, $i \neq j$, the numerator and denominator in (3.6) can be seen to be skew-symmetric. ∎

Using the result of Theorem 3.1, it follows that the elements of a normalized Grassman vector corresponding to any full rank Hankel matrix are symmetric polynomials. Consider the normalized Grassman vector for a 2×4 Hankel matrix A_1 with elements $(A_1)_{ij} = x^n + y^n,\ n = i + j - 1;\ i = 1, 2;\ j = 1, \ldots, 4$. The normalization of the vector $\varphi(A_1)$ can be written in terms of the elementary symmetric polynomials $s_1 = x + y$ and $s_2 = xy$ as

$$\begin{aligned}\varphi(A_1)(1/p_{12}) &= (1/p_{12})(p_{12}, p_{13}, p_{14}, p_{23}, p_{24}, p_{34}) \\ &= (1,\ s_1,\ s_1^2 - s_2,\ s_2,\ s_1 s_2,\ s_2^2).\end{aligned} \tag{3.7}$$

The normalized Grassman vector corresponding to a subspace represented by a 2×5 Hankel matrix A_2 can be written as

$$\begin{aligned}\varphi(A_2)/(1/p_{12}) &= (1/p_{12})(p_{12}, p_{13}, p_{14}, p_{15}, p_{23}, p_{24}, p_{25}, p_{34}, p_{35}, p_{45}) \\ &= (1, s_1, s_1^2 - s_2, s_1(s_1^2 - 2s_2), s_2, s_1 s_2, s_2(s_1^2 - s_2), s_2^2, s_1 s_2^2, s_2^3).\end{aligned}$$

4. Formulation of an alternate problem. Consider a $K \times N$ complex Hankel matrix, A, and corresponding Grassman vector

$$\varphi(A) = \tilde{v} = (p_{1\ 2\ \cdots\ K}, \ldots, p_{N-K+1\ N-K+2\ \cdots\ N}).$$

Let $v = \tilde{v}/p_{12\cdots K}$.

The *distance* between two subspaces A and B of equal dimension can be defined as

$$\operatorname{dist}(A, B) \stackrel{\text{def}}{=} \|P_A - P_B\|_2,$$

where P_A and P_B are the unique orthogonal projections onto A and B, respectively.

What is desired is a matrix, C, with Hankel structure (denote this by $C \in \mathcal{H}$) such that

$$\|P_A - P_C\|_2 = \min_{B \in \mathcal{H}} \|P_A - P_B\|_2. \tag{4.1}$$

The problem in (4.1) will be transformed to the maximization of the inner product of two vectors. In order to do this the notion of *principal angles* [1], [4] will now be introduced.

Let U and W be subspaces with

$$p = \dim(U) \geq \dim(W) = q \geq 1.$$

The *principal angles* $\theta_1, \ldots, \theta_q \in [0, \pi/2]$ between U and W are defined recursively by

$$\cos(\theta_k) = \max_{u \in U} \max_{w \in W} |u^H w|,$$

subject to

$$\begin{aligned} \|u\|_2 = \|w\|_2 &= 1, \\ u^H u_i &= 0, \qquad i = 1, \ldots, k-1, \\ w^H w_i &= 0, \qquad i = 1, \ldots, k-1. \end{aligned}$$

Note that $0 \leq \theta_1 \leq \cdots \leq \theta_q \leq \pi/2$. The vectors $\{u_1, \ldots, u_p\}$ and $\{w_1, \ldots, w_q\}$ are called the *principal vectors* of the subspace pair (U, W). Although the principal vectors aren't unique, the principal angles are. The largest principal angle is related to the distance between equidimensional subspaces given above. If $p = q$, then

$$\text{dist } (U, W) = \sqrt{1 - \cos(\theta_q)^2} = \sin \theta_q.$$

The norm in (4.1) is therefore $\sin \theta_k$, where θ_k is the largest principal angle between the subspaces represented by the row spaces of A and B.

Consider a second subspace distance involving the product of the cosines of all the principal angles:

$$\text{dist}_x(U, W) \stackrel{\text{def}}{=} 1 - \left| \prod_{i=1}^{q} \cos \theta_i \right|. \tag{4.2}$$

We now show that this distance arises naturally when the principal angles associated with the subspaces are related to the angle between the Grassman vectors corresponding to the subspaces. In the following let A^H denote the conjugate transpose of the matrix, A. Let the singular value decomposition of the $N \times M$ matrix A be $A = Y \Sigma Z^H$ where Y and Z are unitary and where the $N \times M$ diagonal matrix, Σ, consists of the singular values in order of decreasing magnitude: $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p$, where $p = \min\{N, M\}$. A generalization of the Courant–Fischer minimax theorem for symmetric matrices is the following [4].

LEMMA 4.1. *If $A \in \mathbf{C}^{N \times M}$, then for $k = 1$ to $\min\{M, N\}$,*

$$\sigma_k(A) = \max_{\substack{\dim(S)=k \\ \dim(T)=k}} \min_{\substack{x \in S \\ y \in T}} \frac{|y^H A x|}{\|x\|_2 \|y\|_2} = \max_{\dim S = k} \min_{x \in S} \frac{\|Ax\|_2}{\|x\|_2}.$$

We now have the following theorem [4].

THEOREM 4.2. *Let $Q_U \in \mathbf{C}^{\mathbf{m} \times \mathbf{p}}$ and $Q_W \in \mathbf{C}^{\mathbf{m} \times \mathbf{q}}$ be orthogonal bases for the subspaces U and W, respectively. Let the principal vectors of U and W be $[u_1 \ \cdots \ u_p]$ and $[w_1 \ \cdots \ w_q]$, respectively. From the minimax characterization of singular values above, if $Y \Sigma Z^H$ is the singular value decomposition of $Q_U^H Q_W$, then the following are true:*

$$\begin{aligned} [u_1 \ \cdots \ u_p] &= Q_U Y, \\ [w_1 \ \cdots \ w_q] &= Q_W Z, \\ \cos(\theta_k) &= \sigma_k, \ k = 1, \ldots, q, \end{aligned}$$

where $\Sigma = \text{diag}\,(\sigma_1, \ldots, \sigma_q)$ *and where* $\theta_1, \ldots, \theta_k$ *are the principal angles associated with the subspaces* U *and* W.

The following result is related to a theorem by Brualdi, Friedland, and Pothem [2] concerning the angle between two Grassman vectors.

THEOREM 4.3. *Let* A *and* B *be matrices of equal dimensions. Then the cosine of the angle between the Grassman vectors* $\varphi(A)$ *and* $\varphi(B)$ *corresponding to the row spaces of* A *and* B *is equal to the product of the cosines of the principal angles between* A *and* B.

Proof. Consider $K \times N$ full rank matrices A and B, $K \leq N$. Then, A can be written as $A = R_A\, Q_A$ where Q_A has orthonormal rows. Note that $\|\varphi(Q_A)\|_2 = 1$ and so $\|\varphi(A)\|_2 = |\det(R_A)|$. In a similar manner, $B = R_B Q_B$ and $\|\varphi(B)\|_2 = |\det(R_B)|$. Then, the cosine of the angle between $\varphi(A)$ and $\varphi(B)$ is

$$\frac{\left|\varphi(A)\varphi(B)^H\right|}{\|\varphi(A)\|_2\|\varphi(B)\|_2} = \frac{\left|\det(R_A)\ \varphi(Q_A)\ \det(R_B)^H\ \varphi(Q_B)^H\right|}{\|\varphi(R_A)\|_2\ \|\varphi(R_B)\|_2}$$

$$= |\varphi(Q_A Q_B^H)| = |\det(Q_A Q_B^H)|.$$

Now, use Theorem 4.2. ∎

Let the principal angles between the subspaces represented by the rows of A and B be $0 \leq \theta_1 \leq \cdots \leq \theta_q \leq \pi/2$, where $p = \dim(R(A^T)) \geq \dim(R(B^T)) = q \geq 1$. The number r, $1 \leq r \leq q$, of the cosines of the principal angles that are equal to one is the dimension of the subspace $A \cap B$. If the number s of cosines of principle angles that are 0 is greater than 0, then there is a subspace $C \subset A$ of dimension $p - q + s$ such that $C \bigcap B = \emptyset$.

We now return to the formulation of the alternate optimization problem. Consider A from (2.1):

$$v \stackrel{\text{def}}{=} \varphi(A) = (-4, -8, -12, -4, -8, -4).$$

The normalized Grassman vector, g, is, from (3.7),

$$g = (1,\ s_1,\ s_1^2 - s_2,\ s_2,\ s_1 s_2,\ s_2^2).$$

The alternate problem is then

$$\max_{s_1, s_2} \frac{|v\ g(s_1, s_2)^H|}{\|v\|_2 \|g(s_1, s_2)\|_2}.$$

Once g is found, the Hankel matrix H is obtained by substituting s_1 and s_2 in $H = \varphi^{-1}(p_{12})\ \varphi^{-1}(g)$, where

$$\varphi^{-1}(p_{12}) = \begin{pmatrix} s_1 & s_1^2 - 2s_2 \\ s_1^2 - 2s_2 & s_1^3 - 3s_1 s_2 \end{pmatrix}$$

and $\varphi^{-1}(g)$ is found using (2.2).

5. Conclusion. The set of constraints which a Grassman vector must satisfy in order for its inverse image to be a Hankel matrix have been developed. A new subspace distance has been introduced and its use in transforming the original matrix approximation problem into a simpler vector approximation problem has been demonstrated.

Acknowledgments. The author would like to thank G. Martin for advice and guidance during the work on this paper and the anonymous reviewers for their helpful suggestions.

REFERENCES

[1] A. Bjorck and G. H. Golub, *Numerical methods for computing angles between linear subspaces*, Math. Comp., 27 (1973), pp. 579–594.
[2] R. A. Brualdi, S. Friedland, and A. Pothen, *The sparse basis problem and multilinear algebra*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1–20.
[3] D. R. Fuhrmann, *Application of Toeplitz covariance estimation to adaptive beamforming and detection*, IEEE Trans. Signal Process., 39 (1991), pp. 2194–2198.
[4] G. Golub and C. V. Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[5] U. Grenander and G. Szego, *Toeplitz Forms and Their Applications*, Chelsea Publishing Company, New York, 1984.
[6] W. Hodge and D. Pedoe, *Methods of Algebraic Geometry,* I, Cambridge University Press, Cambridge, 1968.
[7] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1990.
[8] C. Lanczos, *Applied Analysis*, Dover Books, New York, 1988.
[9] S. Lang, *Algebra*, Addison–Wesley, Reading, MA, 1984.
[10] S. Morgera, *The role of abstract algebra in structured estimation theory*, IEEE Trans. Inform. Theory, 38 (1992), pp. 1053–1065.
[11] L. L. Scharf, *Statistical Signal Processing*, Addison–Wesley, Reading, MA, 1990.
[12] D. Tufts and R. Kumaresan, *Estimation of frequencies of multiple sinusoids: Making linear prediction perform like maximum likelihood*, Proc. IEEE, 70 (1982), pp. 975–989.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 454–476
© 1999 Society for Industrial and Applied Mathematics

CONDENSED FORMS FOR SKEW-HAMILTONIAN/HAMILTONIAN PENCILS*

CHRISTIAN MEHL†

Abstract. In this paper we consider real or complex skew-Hamiltonian/Hamiltonian pencils $\lambda S - H$, i.e., pencils where S is a skew-Hamiltonian and H is a Hamiltonian matrix. These pencils occur, for example, in the theory of continuous time, linear quadratic optimal control problems. We reduce these pencils to canonical and Schur-type forms under structure-preserving transformations, i.e., J-congruence-transformations $(\lambda S-H) \mapsto -JP^*J(\lambda S-H)P$, where P is nonsingular or unitary.

Key words. Hamiltonian matrices, skew-Hamiltonian matrices, matrix pencils, canonical forms, Schur forms, Riccati equations, descriptor systems, optimal control

AMS subject classifications. 15A21, 15A22, 15A57, 93B25

PII. S0895479898336513

1. Introduction.

The motivation for the study of pencils $\lambda S - H$, where S is a skew-Hamiltonian and H is a Hamiltonian matrix, comes mainly from the linear quadratic optimal control problem; see [10], [13], [14], and the references therein. This is the problem of minimizing the cost functional

$$\frac{1}{2}\int_{t_0}^{\infty} \Big(y(t)^*Qy(t) + u(t)^*Ru(t)\Big)dt$$

subject to the dynamics

$$E\dot{x}(t) = Ax(t) + Bu(t), \tag{1}$$

$$x(t_0) = x_0, \tag{2}$$

$$y(t) = Cx(t), \tag{3}$$

where $A, E \in \mathcal{C}^{n\times n}, Q \in \mathcal{C}^{p\times p}, B \in \mathcal{C}^{n\times m}, C \in \mathcal{C}^{p\times n}, R \in \mathcal{C}^{m\times m}$, B with full column rank, C with full row rank, Q Hermitian, and R Hermitian positive definite. Solutions of (1) can be obtained via the solution of the *generalized continuous algebraic Riccati equation*

$$C^*QC + A^*XE + E^*XA - (B^*XE)^*R^{-1}(B^*XE) = 0, \tag{4}$$

or directly via the computation of deflating subspaces of the pencil

$$\lambda S_0 - H_0 = \lambda \begin{bmatrix} E & 0 \\ 0 & E^* \end{bmatrix} - \begin{bmatrix} A & BR^{-1}B^* \\ C^*QC & -A^* \end{bmatrix}; \tag{5}$$

see again [10], [13], [14], and the references therein.

From (5) we obtain immediately that S_0 is a skew-Hamiltonian and H_0 is a Hamiltonian matrix.

DEFINITION 1.1. *Let* $J_n := \left[\begin{smallmatrix} 0 & I_n \\ -I_n & 0 \end{smallmatrix}\right]$. *(We drop the index* n *if the dimension is clear from the context.)*

*Received by the editors March 30, 1998; accepted for publication (in revised form) by V. Mehrmann December 15, 1998; published electronically November 10, 1999.

http://www.siam.org/journals/simax/21-2/33651.html

†Fakultät für Mathematik, Technische Universität Chemnitz, D-09107 Chemnitz, Germany (mehl@mathematik.tu-chemnitz.de).

1. *A matrix* $H \in \mathcal{C}^{2n\times 2n}$ *is called* Hamiltonian *if*

$$HJ + JH^* = 0.$$

The set of Hamiltonian matrices is denoted by $\mathcal{H}_{2n}$.

2. *A matrix* $S \in \mathcal{C}^{2n\times 2n}$ *is called* skew-Hamiltonian *if*

$$SJ - JS^* = 0.$$

The set of skew-Hamiltonian matrices is denoted by $\mathcal{SH}_{2n}$.

3. *A pencil* $\lambda S - H \in \mathcal{C}^{2n\times 2n}$ *is called* skew-Hamiltonian/Hamiltonian *if* S *is skew-Hamiltonian and* H *is Hamiltonian.*

The set of Hamiltonian matrices and the set of skew-Hamiltonian matrices have a special algebraic structure. The first is a Lie algebra (see, e.g., [21]), and the latter is a Jordan algebra; see [1]. If the matrices under consideration are complex, there exists a vector space isomorphism between these two algebras.

LEMMA 1.2. *The map* $A \mapsto iA$ *is a vector space isomorphism between* $\mathcal{H}_{2n}$ *and* $\mathcal{SH}_{2n}$.

Proof. The proof follows directly from the definition. ∎

If E in (5) is the identity matrix, then the study of the pencil $\lambda S_0 - H_0$ reduces to the study of a Hamiltonian matrix. Condensed forms for Hamiltonian matrices have been extensively studied in recent years; see, e.g., [2], [3], [4], [11], [17], or [5] for a more general approach concerning elements from classical Lie or Jordan algebras. There also have been extensive studies concerning symplectic matrices and symplectic pencils that occur in the context of the discrete algebraic Riccati equation corresponding to (5); see, e.g., [6], [11], [22].

If E in (5) is nonsingular, then it is well known that

$$S_0^* \begin{bmatrix} 0 & E^{-*} \\ -E^{-1} & 0 \end{bmatrix} H_0 = -H_0^* \begin{bmatrix} 0 & E^{-*} \\ -E^{-1} & 0 \end{bmatrix} S_0; \tag{6}$$

i.e., $\lambda(iS_0) - H_0$ is an $i\left[\begin{smallmatrix} 0 & E^{-*} \\ -E^{-1} & 0 \end{smallmatrix}\right]$-self-adjoint pencil or in the real case $\lambda S_0 - H_0$ is a $\left[\begin{smallmatrix} 0 & E^{-*} \\ -E^{-1} & 0 \end{smallmatrix}\right]$-skew-symmetric pencil in the sense of [10], where canonical forms for these kinds of pencils were discussed. Obviously this terminology cannot be used if E in (5) is singular. But in recent years, there has been great interest in the singular case; see [13] and the references therein. Therefore we use in this paper the concept of skew-Hamiltonian/Hamiltonian pencils. The case that E is singular is included here.

In the following we will look for condensed forms, i.e., forms from which the eigenvalues of the pencil can be read off in a simple way. For an arbitrary pencil these are the Kronecker canonical form (see [8]) that yields the eigenvalues and eigenvectors and the (generalized) Schur form (see [9]) that yields the eigenvalues and a nested set of invariant subspaces and that is obtained by a transformation $(\lambda A - B) \mapsto U(\lambda A - B)V$, where U and V are unitary matrices. Following the notation in [13] we use the term "Kronecker canonical form" also in the case of regular pencils, but we note that in this case the canonical form is due to Weierstraß; see [8].

The following lemma is helpful in finding out what kind of structure-preserving transformations we may use to obtain corresponding forms for skew-Hamiltonian/ Hamiltonian pencils.

LEMMA 1.3. *The map* $H \mapsto JH$ *is a bijection between* $\mathcal{H}_{2n}$ *and the set of* $2n \times 2n$ *Hermitian matrices.*

Proof. The proof follows directly from the definitions. ∎

The general structure-preserving equivalence transformations for Hermitian matrices are the congruence transformations $A \mapsto P^*AP$, where P is nonsingular. Lemma 1.3 implies that the transformations $H \mapsto -JP^*JHP$, where P is nonsingular, preserve the Hamiltonian structure. We will call these transformations J-congruence transformations by analogy to the congruence transformations. More general are the transformations $H \mapsto \mu JP^*JHP$, where P is nonsingular and $\mu \in \mathcal{R}$. These transformations are the analogue of the μ-congruence transformations $A \mapsto \mu P^*AP$ (see also [16] for related μ-symplectic transformations).

DEFINITION 1.4.

1. *Two matrices $A, B \in \mathcal{C}^{2n\times 2n}$ are called* J-congruent *if there exists a nonsingular matrix $P \in \mathcal{C}^{n\times n}$ such that*

$$-JP^*JAP = B. \tag{7}$$

2. *Two pencils $\lambda A - B, \lambda C - D \in \mathcal{C}^{2n\times 2n}$ are called* J-congruent *if there exists a nonsingular matrix $P \in \mathcal{C}^{n\times n}$ such that*

$$-JP^*J(\lambda A - B)P = \lambda C - D. \tag{8}$$

Analogously we define μ-J-congruence. We easily obtain the following lemma.

LEMMA 1.5. *Let $S \in \mathcal{C}^{2n\times 2n}$ be a skew-Hamiltonian, $H \in \mathcal{C}^{2n\times 2n}$ a Hamiltonian, and $P \in \mathcal{C}^{2n\times 2n}$ an arbitrary matrix and let $\mu \in \mathcal{R}$. Then*

1. *μJP^*JSP is skew-Hamiltonian,*
2. *μJP^*JHP is Hamiltonian,*
3. *$\mu JP^*J(\lambda S - H)P$ is skew-Hamiltonian/Hamiltonian.*

Proof. The proof follows directly from the definitions. ∎

If $\lambda S - H$ is a skew-Hamiltonian/Hamiltonian pencil, then $\lambda(iJS) - (JH)$ is a Hermitian pencil in the sense of [19], where a canonical form for Hermitian pencils is given. In the theory of Hermitian pencils this form is the analogue of the Kronecker canonical form in the theory of general matrix pencils. The numerical computation of the Kronecker canonical form is in general ill-conditioned (see, e.g., [9]), in contrast to the computation of the Schur form; see again [9]. But for a Hermitian matrix the Schur form and the Jordan canonical from coincide; i.e., a triangular form of a Hermitian pencil would be diagonal. Thus, the problem of finding a triangular form for a Hermitian pencil under unitary congruence transformations is equivalent to the problem of the simultaneous diagonalization of two Hermitian matrices. This is possible if and only if the matrices commute; see [18].

In contrast to this, the concept of skew-Hamiltonian/Hamiltonian pencils enables us to obtain both a canonical form and a Schur-type form under J-congruence. After stating some preliminary results in section 2 we discuss this canonical form in section 3. In regard to our interest in finding eigenvalues we restrict our examination to regular pencils. In section 4 we discuss under which conditions it is possible to obtain a structured Schur form under unitary J-congruence transformations, i.e., J-congruence transformations, where the matrix P in (8) is unitary. We will call this form a J-Schur form. Since in general not every skew-Hamiltonian/Hamiltonian pencil has a J-Schur form, we present in section 5 a closely related form that is almost a Schur form and that is obtained by using also some nonunitary J-congruence transformations.

Throughout the paper we use the following notation and expressions:

1. By the *direct sum* of two square matrices A, B, we mean the matrix $\left[\begin{smallmatrix} A & 0 \\ 0 & B \end{smallmatrix}\right]$. Analogously we define the direct sum of square pencils.

2. By $\mathrm{Span}(x_1, \ldots, x_m)$ we denote the subspace of $\mathcal{C}^k$ spanned by the vectors $x_1, \ldots, x_m \in \mathcal{C}^k$.
3. By e_i we denote the ith unit vector.
4. By $\sim_c$ we denote the equivalence relation by congruence.
5. Let $\lambda A - B \in \mathcal{C}^{n\times n}$ be a regular pencil. Introducing homogeneous parameters, $\alpha A - \beta B$ (see [8]), the eigenvalues of $\alpha A - \beta B$ can be represented by pairs $(\alpha, \beta) \in \mathcal{C}^2 \backslash \{0\}$, such that

$$\alpha A x - \beta B x = 0 \quad \text{for an } x \in \mathcal{C}^n \backslash \{0\}.$$

Obviously $(t\alpha, t\beta)$ represents the same eigenvalue for all $t \in \mathcal{C}\backslash\{0\}$; thus, we denote them by $\lambda = \frac{\alpha}{\beta}$ if $\beta \neq 0$. Pairs $(\alpha, 0), \alpha \neq 0$ represent the eigenvalue infinity of $\alpha A - \beta B$ that we will denote by ∞. We note that we always include the eigenvalue ∞ when we are talking about real or about purely imaginary eigenvalues.

2. Preliminaries. We first review some properties of the pencils under consideration.

LEMMA 2.1. *Let $\lambda S - H$ be a skew-Hamiltonian/Hamiltonian pencil. If λ_0 is an eigenvalue of $\lambda S - H$, then so is $-\lambda_0^*$ with the same partial multiplicities.*

Proof. Since S is skew-Hamiltonian and H is Hamiltonian we find that

$$J(\lambda S - H)J = -\lambda S^* - H^*;$$

i.e., $\lambda S - H$ and $-\lambda S^* - H^*$ are equivalent. Thus, these two pencils have the same Kronecker canonical form. ∎

To every block associated with an eigenvalue λ_0 with nonzero real part in the Kronecker canonical form, we find by Lemma 2.1 a paired block of the same size associated with the eigenvalue $-\lambda_0^*$. In general we have no pairing of blocks for the purely imaginary eigenvalues.

In the following we will use the concept of principle vectors of matrix pencils (see [7], [8], and [13]) that is analogous to the concept of principle vectors of matrices or the concept of Jordan chains. Since the chains of principle vectors lead in the pencil case to the Kronecker canonical form, we call them Kronecker chains.

DEFINITION 2.2. *Let $\lambda A - B \in \mathcal{C}^{n\times n}$ be a regular pencil.*

1. *A* Kronecker chain *associated with a finite eigenvalue λ_0 of $\lambda A - B$ is a tuple $(x_1, \ldots, x_m)$ of vectors from $\mathcal{C}^n\backslash\{0\}$ such that*

$$\lambda_0 A x_1 = B x_1 \tag{9}$$

$$\text{and} \quad \lambda_0 A x_i - B x_i = -A x_{i-1} \quad \text{for } i = 2, \ldots, m. \tag{10}$$

2. *A* Kronecker chain *associated with the eigenvalue ∞ of $\lambda A - B$ is a tuple $(y_1, \ldots, y_k)$ of vectors from $\mathcal{C}^n\backslash\{0\}$ such that*

$$A y_1 = 0 \tag{11}$$

$$\text{and} \quad A y_i = B y_{i-1} \quad \text{for } i = 2, \ldots, k. \tag{12}$$

It is clear that if the Kronecker chains in Definition 2.2 are of maximal length, we find a block $\lambda_0 I_m - J_m(\lambda)$ in the Kronecker canonical form of $\lambda A - B$, whose associated deflating subspace is spanned by $x_1, \ldots, x_m$ and a block $\lambda J_k(0) - I_k$, whose associated deflating subspace is spanned by $y_1, \ldots, y_k$. Here $J_l(\mu)$ denotes an $l \times l$ Jordan block

with eigenvalue μ:

$$J_l(\mu) = \begin{bmatrix} \mu & 1 & 0 & \dots & 0 \\ 0 & \mu & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \dots & \dots & 0 & \mu \end{bmatrix}.$$

Furthermore, we have the following property.

LEMMA 2.3. *Let $\lambda A - B \in \mathcal{C}^{n\times n}$ be a regular pencil. If $(x_1, \dots, x_m)$ is a Kronecker chain associated with the eigenvalue 0 of $\lambda A - B$, then $(x_1, \dots, x_m)$ is a Kronecker chain associated with the eigenvalue ∞ of $\lambda B - A$.*

Proof. The proof follows directly from Definition 2.2. ∎

The structure of skew-Hamiltonian/Hamiltonian pencils leads to special properties of the Kronecker chains.

LEMMA 2.4. *Let $\lambda S - H \in \mathcal{C}^{2n\times 2n}$ be a regular skew-Hamiltonian/Hamiltonian pencil. Furthermore let $(x_1, \dots, x_m)$ be a Kronecker chain associated with the eigenvalue $\lambda_0 \in \mathcal{C} \cup \{\infty\}$ and $(y_1, \dots, y_l)$ a Kronecker chain associated with the eigenvalue $\mu_0 \in \mathcal{C} \cup \{\infty\}$, where $\mu_0 \neq -\lambda_0^*$. Then for all i, j, where $1 \le i \le l$ and $1 \le j \le m$, we have*

$$y_i^* JSx_j = y_i^* JHx_j = 0. \tag{13}$$

Proof. The proof proceeds via induction on $k = i + j$.

1. $k = 2$, i.e., $i = j = 1$: Since $\mu \neq -\lambda^*$, at most one of these eigenvalues is infinite. Thus, we may assume without loss of generality (w.l.o.g.) that $\lambda \neq \infty$.

(a) In the case $\mu = \infty$, we have $Sy_1 = 0$. Since S is skew-Hamiltonian, we have

$$y_1^* JSx_1 = y_1^* S^* Jx_1 = 0.$$

Furthermore, we obtain from $\lambda Sx_1 = Hx_1$ that

$$y_1^* JHx_1 = \lambda y_1^* JSx_1 = 0.$$

(b) In the case $\mu \neq \infty$, we have

$$\mu Sy_1 = Hy_1.$$

Noting that S is skew-Hamiltonian and H is Hamiltonian, we see that this implies

$$\begin{aligned} \mu^* y_1^* JSx_1 &= (\mu x_1^* JSy_1)^* \\ &= (x_1^* JHy_1)^* \\ &= -y_1^* JHx_1. \end{aligned}$$

Since we also have $\lambda y_1^* JSx_1 = y_1^* JHx_1 = 0$, we obtain

$$(\lambda + \mu^*) y_1^* JSx_1 = 0.$$

Thus, we obtain $y_1^* JSx_1 = 0$, since $\mu \neq -\lambda^*$, and therefore also $y_1^* JHx_1 = 0$.

2. $k \Rightarrow k + 1$: Suppose that we have already proved (13) for all i, j, where $i \le l$, $j \le m$, and $i + j = k$. Obviously it is sufficient to show that

$$y_i^* JSx_{j+1} = y_i^* JHx_{j+1} = 0$$

for all i, j, where $i \leq l$, $j < m$, and $i + j = k$. Now assume again w.l.o.g. that $\lambda \neq \infty$ and let $y_0 := 0$.

(a) In the case $\mu = \infty$ we have $Sy_i = Hy_{i-1}$. Note that this holds also in the case $i = 1$, since $y_0 = 0$. By this or in the case $i > 1$ by induction, we obtain

$$\begin{aligned} y_i^* JSx_{j+1} &= y_i^* S^* Jx_{j+1} \\ &= y_{i-1}^* H^* Jx_{j+1} \\ &= 0. \end{aligned}$$

From $\lambda Sx_{j+1} - Hx_{j+1} = -Sx_j$ we also obtain

$$y_i^* JHx_{j+1} = \lambda y_i^* JSx_{j+1} + y_i^* JSx_j = 0.$$

(b) In the case $\mu \neq \infty$ we have $\mu Sy_i - Hy_i = -Sy_{i-1}$. Since $x_{j+1}^* JSy_{i-1} = 0$, this yields

$$\begin{aligned} \mu^* y_i^* JSx_{j+1} &= (\mu x_{j+1}^* JSy_i)^* \\ &= (x_{j+1}^* JHy_i)^* \\ &= -y_i^* JHx_{j+1}, \end{aligned}$$

once more using the fact that S is skew-Hamiltonian and that H is Hamiltonian. Noting that $y_i^* JSx_j = 0$, we also have

$$\lambda y_i^* JSx_{j+1} = y_i^* JHx_{j+1}.$$

Therefore, we obtain

$$(\lambda + \mu^*) y_i^* JSx_{j+1},$$

and finally

$$y_i^* JSx_{j+1} = y_i^* JHx_{j+1} = 0. \qquad \square$$

LEMMA 2.5. *Let $\lambda S - H \in \mathcal{C}^{2n \times 2n}$ be a regular skew-Hamiltonian/Hamiltonian pencil. Furthermore let $(x_1, \ldots, x_m)$ be a Kronecker chain associated with the eigenvalue $\lambda \in \mathcal{C} \cup \{\infty\}$ and $(y_1, \ldots, y_l)$ a Kronecker chain associated with the eigenvalue μ, where $\mu = -\lambda^*$. Then for all i, j where $1 \leq i \leq l$, $1 \leq j \leq m$ and $i + j \leq \max(l, m)$, we have*

$$y_i^* JSx_j = y_i^* JHx_j = 0.$$

Proof. If $\lambda = \infty$, then Lemma 2.3 implies that $(x_1, \ldots, x_m)$ and $(y_1, \ldots, y_l)$ are Kronecker chains of $\lambda H - S$ associated with the eigenvalue zero. This also holds for the skew-Hamiltonian/Hamiltonian pencil $\lambda(iH) - (iS)$. Thus, it remains to prove the assertion for the case that $\lambda \neq \infty$.

1. First of all we prove that for all i, j, where $1 \leq i \leq l$ and $0 \leq j < m$, we have

$$y_i^* JSx_j = -y_{i-1}^* JSx_{j+1} \quad \text{and} \quad y_i^* JHx_j = -y_{i-1}^* JHx_{j+1}, \tag{14}$$

where $x_0 := 0$ and $y_0 := 0$. Since $(x_1, \ldots, x_m)$ and $(y_1, \ldots, y_l)$ are Kronecker chains we have

$$\lambda Sx_{j+1} - Hx_{j+1} = -Sx_j \quad \text{and} \quad -\lambda^* Sy_i - Hy_i = -Sy_{i-1}. \tag{15}$$

Note that this is also true for $i = 1$ and $j = 0$. From (15) and since S is skew-Hamiltonian and H is Hamiltonian, we obtain

$$\begin{aligned} -y_{i-1}^* JSx_{j+1} &= -y_{i-1}^* S^* Jx_{j+1} \\ &= -\lambda y_i^* S^* Jx_{j+1} - y_i^* H^* Jx_{j+1} \\ &= -\lambda y_i^* JSx_{j+1} + y_i^* JHx_{j+1} \\ &= y_i^* JSx_j \end{aligned}$$

and analogously

$$\begin{aligned} -y_{i-1}^* JHx_{j+1} &= -\lambda y_{i-1}^* JSx_{j+1} - y_{i-1}^* JSx_j \\ &= \lambda y_i^* JSx_j - y_{i-1}^* JSx_j \\ &= (\lambda y_i^* S^* - y_{i-1}^* S^*)Jx_j \\ &= -y_i^* H^* Jx_j \\ &= y_i^* JHx_j. \end{aligned}$$

2. Now assume w.l.o.g. that $l = \max(l, m)$. Since $i + j \leq l$ and $x_0 = 0$ we obtain by (14) that

$$y_i^* JSx_j = (-1)^j y_{i+j}^* JSx_0 = 0$$

and

$$y_i^* JHx_j = (-1)^j y_{i+j}^* JHx_0 = 0. \qquad \square$$

3. A canonical form under *J*-congruence. In this section we present a canonical form for skew-Hamiltonian/Hamiltonian pencils under J-congruence. In the study of skew-Hamiltonian/Hamiltonian pencils this form is the analogue of the Kronecker canonical form in the study of general matrix pencils.

THEOREM 3.1. *Let $\lambda S - H \in \mathcal{C}^{2n\times 2n}$ be a regular skew-Hamiltonian/Hamiltonian pencil. Then there exists a nonsingular matrix $P \in \mathcal{C}^{2n\times 2n}$ such that*

$$\begin{aligned} &-JP^*J(\lambda S - H)P \\ &\quad = \lambda \begin{bmatrix} S_{11} & 0 & 0 & 0 \\ 0 & S_{22} & 0 & S_{24} \\ 0 & 0 & S_{11}^* & 0 \\ 0 & S_{42} & 0 & S_{22}^* \end{bmatrix} - \begin{bmatrix} H_{11} & 0 & 0 & 0 \\ 0 & H_{22} & 0 & H_{24} \\ 0 & 0 & -H_{11}^* & 0 \\ 0 & H_{42} & 0 & -H_{22}^* \end{bmatrix}, \end{aligned} \tag{16}$$

where

1. *$\lambda S_{11} - H_{11}$ is a pencil in Kronecker canonical form and has only eigenvalues with positive real part.*
2. *$\lambda\hat{S} - \hat{H} := \lambda\left[\begin{smallmatrix} S_{22} & S_{24} \\ S_{42} & S_{22}^* \end{smallmatrix}\right] - \left[\begin{smallmatrix} H_{22} & H_{24} \\ H_{42} & -H_{22}^* \end{smallmatrix}\right]$ is a skew-Hamiltonian/Hamiltonian pencil such that $J(\lambda\hat{S} - \hat{H})$ is the direct sum of blocks of the form*

$$\lambda\varepsilon \begin{bmatrix} 0 & & & i \\ & & \cdot & \\ & \cdot & & \\ i & & & 0 \end{bmatrix} - \varepsilon \begin{bmatrix} 0 & & & \mu \\ & & \cdot & 1 \\ & \cdot & \cdot & \\ \mu & 1 & & 0 \end{bmatrix} \in \mathcal{C}^{k\times k} \tag{17}$$

or

$$\lambda\varepsilon \begin{bmatrix} 0 & & & 0 \\ & & \cdot & i \\ & \cdot & \cdot & \\ 0 & i & & 0 \end{bmatrix} - \varepsilon \begin{bmatrix} 0 & & & 1 \\ & & \cdot & \\ & \cdot & & \\ 1 & & & 0 \end{bmatrix} \in \mathcal{C}^{k\times k}, \tag{18}$$

where μ is real and $\varepsilon = \pm 1$.

Proof.

1. We first prove that $\lambda S - H$ is, up to the permutation of rows and columns, J-congruent to the direct sum of two skew-Hamiltonian/Hamiltonian pencils such that the eigenvalues of one pencil have all nonzero real part and the eigenvalues of the other pencil are all purely imaginary.

Let us consider a basis $(x_1, \ldots, x_m, y_1, \ldots, y_m, v_1, \ldots, v_{2(n-m)})$ of $\mathcal{C}^{2n}$ consisting of Kronecker chains of $\lambda S - H$ such that $(x_1, \ldots, x_m)$ is a basis of the deflating subspace associated with the eigenvalues with positive real part, $(y_1, \ldots, y_m)$ is a basis of the deflating subspace associated with the eigenvalues with negative real part, and $(v_1, \ldots, v_{2(n-m)})$ is a basis of the deflating subspace associated with the purely imaginary eigenvalues. Setting

$$Q = [x_1, \ldots, x_m, v_1, \ldots, v_{n-m}, y_1, \ldots, y_m, v_{n-m+1}, \ldots, v_{2(n-m)}],$$

we find by Lemma 2.4 that

$$\begin{aligned} &-JQ^*J(\lambda S - H)Q \\ &\quad = \lambda \begin{bmatrix} \tilde{S}_{11} & 0 & 0 & 0 \\ 0 & \tilde{S}_{22} & 0 & \tilde{S}_{24} \\ 0 & 0 & \tilde{S}_{11}^* & 0 \\ 0 & \tilde{S}_{42} & 0 & \tilde{S}_{22}^* \end{bmatrix} - \begin{bmatrix} \tilde{H}_{11} & 0 & 0 & 0 \\ 0 & \tilde{H}_{22} & 0 & \tilde{H}_{24} \\ 0 & 0 & -\tilde{H}_{11}^* & 0 \\ 0 & \tilde{H}_{42} & 0 & -\tilde{H}_{22}^* \end{bmatrix}, \end{aligned} \tag{19}$$

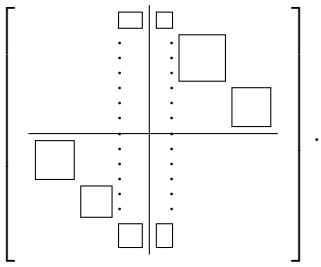
where $\lambda\tilde{S}_{11} - \tilde{H}_{11}$ has only eigenvalues with positive real part and the skew-Hamiltonian/Hamiltonian pencil $\lambda\tilde{S} - \tilde{H} := \lambda\left[\begin{smallmatrix} \tilde{S}_{22} & \tilde{S}_{24} \\ \tilde{S}_{42} & \tilde{S}_{22}^* \end{smallmatrix}\right] - \left[\begin{smallmatrix} \tilde{H}_{22} & \tilde{H}_{24} \\ \tilde{H}_{42} & -\tilde{H}_{22}^* \end{smallmatrix}\right]$ has only purely imaginary eigenvalues.

2. For the proof of (16) we note that $J(\lambda(i\tilde{S}) - \tilde{H})$ is a Hermitian pencil that has only purely imaginary eigenvalues. Thus, applying Lemma 2 from [19], we find that there exists a nonsingular matrix $X = \left[\begin{smallmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{smallmatrix}\right]$, where $X_{jk} \in \mathcal{C}^{(n-m)\times(n-m)}$, such that $X^*J(\lambda\tilde{S} - \tilde{H})X$ is a direct sum of blocks of the form (17) or (18). Furthermore, we find nonsingular matrices R, T such that $R(\lambda\tilde{S}_{11} - \tilde{H}_{11})T$ is in Kronecker canonical form. Hence we obtain (16) by setting

$$P = Q \cdot \begin{bmatrix} R & 0 & 0 & 0 \\ 0 & X_{11} & 0 & X_{12} \\ 0 & 0 & -T^* & 0 \\ 0 & X_{21} & 0 & X_{22} \end{bmatrix}. \qquad \square$$

Note that the blocks of the submatrices $\hat{S}$ and $\hat{H}$ in (16) in general occur in

patterns such as



In general it is not possible to reduce the subpencil $\lambda\hat{S} - \hat{H}$ to a direct sum of skew-Hamiltonian/Hamiltonian pencils that have only one eigenvalue. As an example, consider the pencil

$$\lambda \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Further reduction is impossible, since the pencil has purely imaginary eigenvalues with odd algebraic multiplicity.

LEMMA 3.2. *Let $\lambda S - H \in \mathcal{C}^{2n\times 2n}$ be a regular skew-Hamiltonian/Hamiltonian pencil with pairwise distinct eigenvalues $\lambda_1, \ldots, \lambda_k$ that are all purely imaginary. Then the following statements are equivalent:*

1. *The eigenvalues $\lambda_1, \ldots, \lambda_k$ all have even algebraic multiplicity.*
2. *There exists a nonsingular matrix P such that*

$$-JP^*JSP = \left[\begin{array}{ccc|ccc} S_{11} & & 0 & S_{1,k+1} & & 0 \\ & \ddots & & & \ddots & \\ 0 & & S_{kk} & 0 & & S_{k,2k} \\ \hline S_{k+1,1} & & 0 & S_{k+1,k+1} & & 0 \\ & \ddots & & & \ddots & \\ 0 & & S_{2k,k} & 0 & & S_{2k,2k} \end{array}\right]$$

and

$$-JP^*JHP = \left[\begin{array}{ccc|ccc} H_{11} & & 0 & H_{1,k+1} & & 0 \\ & \ddots & & & \ddots & \\ 0 & & H_{kk} & 0 & & H_{k,2k} \\ \hline H_{k+1,1} & & 0 & H_{k+1,k+1} & & 0 \\ & \ddots & & & \ddots & \\ 0 & & H_{2k,k} & 0 & & H_{2k,2k} \end{array}\right],$$

where $\lambda\left[\begin{smallmatrix} S_{ii} & S_{i,k+i} \\ S_{k+i,i} & S_{k+i,k+i} \end{smallmatrix}\right] - \left[\begin{smallmatrix} H_{ii} & H_{i,k+i} \\ H_{k+i,i} & H_{k+i,k+i} \end{smallmatrix}\right], i = 1, \ldots, k$, is a skew-Hamiltonian/Hamiltonian pencil having only one eigenvalue.

Proof. $1 \Rightarrow 2$: For the deflating subspace associated with λ_i, we find a basis $(x_1, \ldots, x_{m_i})$ consisting of Kronecker chains. Now set

$$P_{i1} = [x_1, \ldots, x_{\frac{m_i}{2}}] \quad \text{and} \quad P_{i2} = [x_{\frac{m_i}{2}+1}, \ldots, x_{m_i}].$$

This is possible, since every eigenvalue has even algebraic multiplicity. Setting

$$P = [P_{11}, \ldots, P_{k1}, P_{12}, \ldots, P_{k2}],$$

Lemma 2.4 implies the required result.

$2 \Rightarrow 1$: is clear, since the subpencils

$$\lambda \begin{bmatrix} S_{ii} & S_{i,k+i} \\ S_{k+i,i} & S_{k+i,k+i} \end{bmatrix} - \begin{bmatrix} H_{ii} & H_{i,k+i} \\ H_{k+i,i} & H_{k+i,k+i} \end{bmatrix}$$

have even sizes. □

There exists a real analogue to Theorem 3.1.

THEOREM 3.3. *Let $\lambda S - H \in \mathcal{R}^{2n\times 2n}$ be a regular skew-Hamiltonian/Hamiltonian pencil. Then there exists a nonsingular matrix $P \in \mathcal{R}^{2n\times 2n}$ such that*

$$\begin{aligned} &-JP^*J(\lambda S - H)P \\ &\quad = \lambda \begin{bmatrix} S_{11} & 0 & 0 & 0 \\ 0 & S_{22} & 0 & S_{24} \\ 0 & 0 & S_{11}^* & 0 \\ 0 & S_{42} & 0 & S_{22}^* \end{bmatrix} - \begin{bmatrix} H_{11} & 0 & 0 & 0 \\ 0 & H_{22} & 0 & H_{24} \\ 0 & 0 & -H_{11}^* & 0 \\ 0 & H_{42} & 0 & -H_{22}^* \end{bmatrix}, \end{aligned} \tag{20}$$

where

1. *$\lambda S_{11} - H_{11}$ is a pencil in real Kronecker canonical form and has only eigenvalues with positive real part.*
2. *$\lambda \hat{S} - \hat{H} := \lambda \left[\begin{smallmatrix} S_{22} & S_{24} \\ S_{42} & S_{22}^* \end{smallmatrix}\right] - \left[\begin{smallmatrix} H_{22} & H_{24} \\ H_{42} & -H_{22}^* \end{smallmatrix}\right]$ is a skew-Hamiltonian/Hamiltonian pencil such that $J(\lambda \hat{S} - \hat{H})$ is the direct sum of blocks either of the form*

$$\lambda\varepsilon \begin{bmatrix} & & R \\ & R & \\ R & & \end{bmatrix} - \varepsilon \begin{bmatrix} & & \mu I_2 \\ & \mu I_2 & I_2 \\ \mu I_2 & I_2 & \end{bmatrix} \in \mathcal{C}^{2k\times 2k}, \tag{21}$$

where $\mu > 0$, $R = \left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$ and $I_2 = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$, or of one of the forms

$$\lambda\varepsilon \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \Delta_k \\ 0 & -\Delta_k & 0 \end{bmatrix} - \varepsilon\Delta_{2k+1},\ k \in \mathcal{N}, \tag{22}$$

$$\lambda \begin{bmatrix} 0 & \Lambda_k \\ -\Lambda_k & 0 \end{bmatrix} - \begin{bmatrix} 0 & \Delta_k \\ \Delta_k & 0 \end{bmatrix},\ k \text{ even}, \tag{23}$$

$$\lambda\varepsilon \begin{bmatrix} 0 & \Delta_k \\ -\Delta_k & 0 \end{bmatrix} - \varepsilon\Lambda_{2k},\ k \in \mathcal{N}, \tag{24}$$

or

$$\lambda \begin{bmatrix} 0 & \Delta_k \\ -\Delta_k & 0 \end{bmatrix} - \begin{bmatrix} 0 & \Lambda_k \\ \Lambda_k & 0 \end{bmatrix},\ k \text{ odd}, \tag{25}$$

where $\varepsilon = \pm 1$ and furthermore

$$\Delta_k := \begin{bmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & 0 \end{bmatrix} \quad \text{and} \quad \Lambda_k := \begin{bmatrix} 0 & & & 0 \\ & & \iddots & 1 \\ & \iddots & \iddots & \\ 0 & 1 & & 0 \end{bmatrix}.$$

Proof. The proof is similar to the proof of Theorem 3.1 using results on pencils $\lambda A - B$, where A is skew symmetric and B is symmetric; see, e.g., [20]. ∎

COROLLARY 3.4. *Let $\lambda S - H$ be a real skew-Hamiltonian/Hamiltonian pencil. Then the following conditions apply:*

1. *Every block of odd size belonging to the eigenvalue zero in the (general) Kronecker canonical form occurs with even multiplicity.*
2. *Every block of even size belonging to the eigenvalue ∞ in the (general) Kronecker canonical form occurs with even multiplicity.*

Proof. This follows directly from the structures of the blocks (23) and (25) in the structured canonical form (20). ∎

4. The existence of *J*-Schur forms. As shown in [17] not every Hamiltonian matrix can be reduced by unitary symplectic similarity transformations to a matrix in a Schur-type form that is called Schur–Hamiltonian form in [17] or Hamiltonian Schur form in [13]. This is a Hamiltonian matrix

$$\begin{bmatrix} T & R \\ 0 & -T^* \end{bmatrix}, \tag{26}$$

where $T \in \mathcal{C}^{n\times n}$ is upper triangular and $R \in \mathcal{C}^{n\times n}$ is Hermitian. The analogous Schur-type form for a skew-Hamiltonian matrix is

$$\begin{bmatrix} T & R \\ 0 & T^* \end{bmatrix}, \tag{27}$$

where $T \in \mathcal{C}^{n\times n}$ is upper triangular and $R \in \mathcal{C}^{n\times n}$ is skew-Hermitian. We call matrices of the form (26) or (27) matrices in J-Schur form. In [12] and [15] a necessary and sufficient condition for the existence of the Hamiltonian Schur form, resp., J-Schur form, is presented.

THEOREM 4.1. *Let $H \in \mathcal{C}^{2n\times 2n}$ be a Hamiltonian matrix and let $\lambda_1, \dots, \lambda_m$ be the pairwise distinct purely imaginary eigenvalues of H with algebraic multiplicities $p_1, \dots, p_m$. Then the following statements are equivalent:*

1. *There exists a unitary symplectic matrix $P \in \mathcal{C}^{2n\times 2n}$ such that PHP^{-1} is in J-Schur form* (26).
2. *If the columns of $U_k \in \mathcal{C}^{2n\times p_k}$, $k = 1, \dots, m$, form a basis of the invariant subspace associated with λ_k, then*

$$U_k^* J U_k \sim_c J_{p_k/2}. \tag{28}$$

Note that (28) implies that the multiplicities p_k are even.

Proof. See [12] or [15].

An analogous result holds in the case that H is skew-Hamiltonian.

THEOREM 4.2. *Let $S \in \mathcal{C}^{2n\times 2n}$ be a skew-Hamiltonian matrix and let $\lambda_1, \dots, \lambda_m$ be the pairwise distinct real eigenvalues of S with algebraic multiplicities $p_1, \dots, p_m$. Then the following statements are equivalent:*

1. *There exists a unitary symplectic matrix $P \in \mathcal{C}^{2n\times 2n}$ such that PSP^{-1} is in J-Schur form* (27).
2. *If the columns of $U_k \in \mathcal{C}^{2n\times p_k}$, $k = 1, \dots, m$, form a basis of the invariant subspace associated with λ_k, then*

$$U_k^* J U_k \sim_c J_{p_k/2}. \tag{29}$$

Proof. The proof follows directly from Lemma 1.2 and Theorem 4.1. ∎

We also obtain a similar result for the skew-Hamiltonian/Hamiltonian pencils. For the proof of this result, we need the following lemma.

LEMMA 4.3. *Let $P \in \mathcal{C}^{2n\times 2n}$. Then there exists a factorization*

$$P = Q \cdot \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22}^* \end{bmatrix}, \tag{30}$$

where $Q \in \mathcal{C}^{2n\times 2n}$ is unitary and $R_{11}, R_{22} \in \mathcal{C}^{n\times n}$ are upper triangular.

Proof. After performing n steps of the QR Householder algorithm (see, e.g., [9]), we find a unitary matrix $Q_1 \in \mathcal{C}^{n\times n}$ such that

$$P = Q_1 \cdot \begin{bmatrix} R_{11} & \tilde{P}_{12} \\ 0 & \tilde{P}_{22} \end{bmatrix},$$

where $R_{11} \in \mathcal{C}^{n\times n}$ is upper triangular. Then we compute the QL factorization

$$\tilde{P}_{22} = Q_2 \cdot R_{22}^*,$$

where $Q_2 \in \mathcal{C}^{n\times n}$ is unitary and $R_{22}^* \in \mathcal{C}^{n\times n}$ is lower triangular; see again [9]. Thus, setting $Q = Q_1 \cdot \left[\begin{smallmatrix} I_n & 0 \\ 0 & Q_2 \end{smallmatrix}\right]$ implies (30). ∎

THEOREM 4.4. *Let $\lambda S - H \in \mathcal{C}^{2n\times 2n}$ be a skew-Hamiltonian/Hamiltonian pencil. Let $\lambda_1, \ldots, \lambda_m$ be the pairwise distinct, finite, purely imaginary eigenvalues of $\lambda S - H$ with algebraic multiplicities $p_1, \ldots, p_m$ and let p_∞ be the algebraic multiplicity of the eigenvalue ∞. Then the following statements are equivalent:*

1. *There exists a nonsingular matrix $P \in \mathcal{C}^{2n\times 2n}$ such that*

$$-JP^*J(\lambda S - H)P = \lambda \begin{bmatrix} S_{11} & S_{12} \\ 0 & S_{11}^* \end{bmatrix} - \begin{bmatrix} H_{11} & H_{12} \\ 0 & -H_{11}^* \end{bmatrix}, \tag{31}$$

   where S_{11} and H_{11} are upper triangular.

2. *There is a unitary matrix $Q \in \mathcal{C}^{2n\times 2n}$ such that*

$$-JQ^*J(\lambda S - H)Q = \lambda \begin{bmatrix} \hat{S}_{11} & \hat{S}_{12} \\ 0 & \hat{S}_{11}^* \end{bmatrix} - \begin{bmatrix} \hat{H}_{11} & \hat{H}_{12} \\ 0 & -\hat{H}_{11}^* \end{bmatrix}, \tag{32}$$

   where $\hat{S}_{11}$ and $\hat{H}_{11}$ are upper triangular.

3. *If the columns of $U_k \in \mathcal{C}^{2n\times p_k}$, $k = 1, \ldots, m$, form a basis of the deflating subspace associated with λ_k, then*

$$U_k^* J S U_k \sim_c J_{p_k/2}. \tag{33}$$

   Furthermore, if $p_\infty \neq 0$ and if the columns of $U_\infty \in \mathcal{C}^{2n\times p_\infty}$ form a basis of the deflating subspace associated with the eigenvalue ∞, then

$$U_\infty^* J H U_\infty \sim_c i J_{p_\infty/2}. \tag{34}$$

Proof. 1 ⇒ 2: If $P \in \mathcal{C}^{2n\times 2n}$ is nonsingular such that (31) holds, Lemma 4.3 implies that there exists a factorization (30). Using this we obtain from (31), that

$$\begin{aligned} &-JQ^*J(\lambda S - H)Q \\ &\quad = J \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22}^* \end{bmatrix}^{-*} J \left(\lambda \begin{bmatrix} S_{11} & S_{12} \\ 0 & S_{11}^* \end{bmatrix} - \begin{bmatrix} H_{11} & H_{12} \\ 0 & -H_{11}^* \end{bmatrix} \right) \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22}^* \end{bmatrix}^{-1} \\ &\quad = \lambda \begin{bmatrix} -R_{22}^{-1} S_{11} R_{11}^{-1} & * \\ 0 & -R_{11}^{-*} S_{11}^* R_{22}^{-*} \end{bmatrix} - \begin{bmatrix} -R_{22}^{-1} H_{11} R_{11}^{-1} & * \\ 0 & R_{11}^{-*} H_{11}^* R_{22}^{-*} \end{bmatrix}, \end{aligned}$$

where the upper left subpencil $-R_{22}^{-1}(\lambda S_{11} - H_{11})R_{11}^{-1}$ is still upper triangular.

$2 \Rightarrow 3$: Let $\lambda S - H$ be in the form (32). First we show that the pencil is (up to the permutation of rows and columns) J-congruent to a direct sum of skew-Hamiltonian/Hamiltonian pencils in the form (32) that have only one eigenvalue. Therefore, consider a 4×4 subpencil

$$\lambda \check{S} - \check{H} = \lambda \begin{bmatrix} s_{11} & s_{12} & s_{13} & s_{14} \\ 0 & s_{22} & -s_{14}^* & s_{24} \\ 0 & 0 & s_{11}^* & 0 \\ 0 & 0 & s_{12}^* & s_{22}^* \end{bmatrix} - \begin{bmatrix} h_{11} & h_{12} & h_{13} & h_{14} \\ 0 & h_{22} & h_{14}^* & h_{24} \\ 0 & 0 & -h_{11}^* & 0 \\ 0 & 0 & -h_{12}^* & -h_{22}^* \end{bmatrix}$$

of the pencil in (32) such that $\lambda s_{11} - h_{11}$ has the eigenvalue λ_0 and such that $\lambda s_{22} - h_{22}$ has the eigenvalue μ_0. Furthermore, let $\lambda_0 \neq \mu_0, -\mu_0^*$. Next we show that the (1,2)-elements of $\check{S}$ and $\check{H}$, respectively, can be simultaneously set to zero. Therefore, let

$$P = \begin{bmatrix} 1 & x & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & y & 1 \end{bmatrix},$$

where $x, y \in \mathcal{C}$, and consider $-JP^*J\check{S}P$ and $-JP^*J\check{H}P$. If we want to set the (1,2)-elements of both matrices to zero, we have to solve the equations

$$\begin{aligned} s_{12} + s_{22}y^* + s_{11}x &= 0, \\ \text{and} \quad h_{12} + h_{22}y^* + h_{11}x &= 0. \end{aligned}$$

This implies $(h_{22}s_{11} - h_{11}s_{22})x = s_{22}h_{12} - h_{22}s_{12}$. Since $\lambda_0 \neq \mu_0$, it follows that $h_{22}s_{11} - h_{11}s_{22} \neq 0$. Therefore, x (and y) can be chosen such that the (1,2)-elements of both $-JP^*J\hat{S}P$ and $-JP^*J\hat{H}P$ are zero. In an analogous way we show that using a J-congruence transformation with the matrix

$$P = \begin{bmatrix} 1 & 0 & 0 & x \\ 0 & 1 & y & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

the (1,4)-elements of $\check{S}$ and $\check{H}$ can also be simultaneously set to zero. Here, we need $\lambda_0 \neq -\mu^*$. (We note that this transformation has no effect on the 2×2 diagonal blocks of $\lambda\check{S} - \check{H}$.) Now it is clear that a sequence of J-congruence transformations can be used to obtain a pencil that is (up to the permutation of rows and columns) a direct sum of skew-Hamiltonian/Hamiltonian pencils in the form (32) such that each pencil has only one eigenvalue, and we may consider the blocks separately.

Furthermore, we may assume w.l.o.g. that the bases U_k and U_∞ are canonical, for if (33) holds for a special basis U_k, it holds for every basis of the deflating subspace associated with λ_k, because a change of basis means the transition from U_k to $U_k X$, where $X \in \mathcal{C}^{p_k \times p_k}$ is nonsingular. (The same argument holds for U_∞.) Thus, if

$$\lambda\tilde{S} - \tilde{H} = \lambda \begin{bmatrix} \tilde{S}_{11} & \tilde{S}_{12} \\ 0 & \tilde{S}_{11}^* \end{bmatrix} - \begin{bmatrix} \tilde{H}_{11} & \tilde{H}_{12} \\ 0 & -\tilde{H}_{11}^* \end{bmatrix}$$

is a skew-Hamiltonian/Hamiltonian pencil satisfying (31) and that has only one eigenvalue that is finite and purely imaginary, it remains to show that $J\tilde{S}$ is congruent to

J. This also proves (34), for we can reduce the case of the eigenvalue ∞ to the case of the eigenvalue zero by considering $\lambda(i\tilde{H}) - (i\tilde{S})$ according to Lemma 2.3. Now we obtain

$$\begin{aligned}
&\begin{bmatrix} \tilde{S}_{11}^{-1} & -\frac{1}{2}\tilde{S}_{11}^{-1}\tilde{S}_{12} \\ 0 & I \end{bmatrix}^* J\tilde{S} \begin{bmatrix} \tilde{S}_{11}^{-1} & -\frac{1}{2}\tilde{S}_{11}^{-1}\tilde{S}_{12} \\ 0 & I \end{bmatrix} \\
&\quad = \begin{bmatrix} \tilde{S}_{11}^{-*} & 0 \\ -\frac{1}{2}\tilde{S}_{12}^{*}\tilde{S}_{11}^{-*} & I \end{bmatrix} \begin{bmatrix} 0 & \tilde{S}_{11}^{*} \\ -\tilde{S}_{11} & -\tilde{S}_{12} \end{bmatrix} \begin{bmatrix} \tilde{S}_{11}^{-1} & -\frac{1}{2}\tilde{S}_{11}^{-1}\tilde{S}_{12} \\ 0 & I \end{bmatrix} \\
&\quad = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} = J,
\end{aligned}$$

i.e., $J\tilde{S}$ is congruent to J.

$3 \Rightarrow 1$: Since (33) and (34) imply that all the algebraic multiplicities of the purely imaginary eigenvalues are even, it is by Theorem 3.1 and Lemmas 3.2 and 2.3 again sufficient to consider the case that $\lambda S - H$ has only one eigenvalue that is finite and purely imaginary. Since JS is congruent to J we have that S is J-congruent to the identity matrix. Thus there exists $P_1 \in \mathcal{C}^{2n\times 2n}$ nonsingular such that

$$-JP_1^*J(\lambda S - H)P_1 = \lambda I - \hat{H},$$

where $\hat{H}$ is a Hamiltonian matrix having only one eigenvalue, which is purely imaginary. Thus, $\hat{H}$ trivially satisfies (28) and by Theorem 4.1 we find a symplectic matrix $P_2 \in \mathcal{C}^{2n\times 2n}$, i.e., $P_2^{-1} = -JP_2^*J$, such that $P_2^{-1}\hat{H}P_2$ is in J-Schur form (26). We obtain (31) by setting $P = P_1P_2$. ∎

We call a pencil of the form (31) a pencil in J-Schur form.

5. An almost-Schur form under *J*-congruence. We have seen in section 4 that not every skew-Hamiltonian/Hamiltonian pencil can be reduced to a J-Schur form by unitary J-congruence transformation. Thus, if we want to obtain a condensed form from which the eigenvalues can be read off in a simple way, we have to use also nonunitary J-congruence transformations. In this section we will present a condensed form that we will call almost-Schur form and that can be obtained by applying J-congruence transformations, where P in (8) is a product of unitary or *almost-unitary* matrices.

DEFINITION 5.1. *Let $P = [p_1, \ldots, p_n] \in \mathcal{C}^n$ be nonsingular. Then P is called* almost unitary *if there exists an index k such that $(p_1, \ldots, p_{k-1}, p_{k+1}, \ldots, p_n)$ is an orthonormal tuple of vectors.*

THEOREM 5.2. *Let $\lambda S - H \in \mathcal{C}^{2n\times 2n}$ be a regular skew-Hamiltonian/Hamiltonian pencil. Then there exists a nonsingular matrix $P \in \mathcal{C}^{2n\times 2n}$ and $k \in \mathcal{N}$ such that*

$$\begin{aligned}
&-JP^*J(\lambda S - H)P \\
&\quad = \lambda \begin{bmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ 0 & 0 & -S_{14}^* & S_{24} \\ 0 & 0 & S_{11}^* & 0 \\ 0 & S_{42} & S_{12}^* & 0 \end{bmatrix} - \begin{bmatrix} H_{11} & H_{12} & H_{13} & H_{14} \\ 0 & 0 & H_{14}^* & H_{24} \\ 0 & 0 & -H_{11}^* & 0 \\ 0 & H_{42} & -H_{12}^* & 0 \end{bmatrix},
\end{aligned} \tag{35}$$

where the following conditions hold:

1. *$\lambda S_{11} - H_{11} \in \mathcal{C}^{k\times k}$ is an upper triangular pencil having only eigenvalues with nonnegative real part.*

2. $S_{24}, S_{42}, H_{24}, H_{42} \in \mathcal{C}^{(n-k)\times(n-k)}$ *are diagonal and all the eigenvalues of the skew-Hamiltonian/Hamiltonian pencil*

$$\lambda \begin{bmatrix} 0 & S_{24} \\ S_{42} & 0 \end{bmatrix} - \begin{bmatrix} 0 & H_{24} \\ H_{42} & 0 \end{bmatrix}$$

are purely imaginary.

3. *The spectrum of* $\lambda S - H$ *is equal to the union of the spectra of*

$$\lambda \begin{bmatrix} S_{11} & S_{13} \\ 0 & S_{11}^* \end{bmatrix} - \begin{bmatrix} H_{11} & H_{13} \\ 0 & -H_{11}^* \end{bmatrix} \quad \text{and} \quad \lambda \begin{bmatrix} 0 & S_{24} \\ S_{42} & 0 \end{bmatrix} - \begin{bmatrix} 0 & H_{24} \\ H_{42} & 0 \end{bmatrix}.$$

4. P *is a product of unitary and almost unitary matrices.*

Proof. We consider two cases:

(a) Let us assume that there exists an eigenvector $x \neq 0$ associated with an eigenvalue λ_0 of $\lambda S - H$ such that

$$x^* JSx = 0 \ \text{ and } \ x^* JHx = 0. \tag{36}$$

Since $\lambda S - H$ is regular, we have $JSx \neq 0$ or $JHx \neq 0$. Assume that $JSx \neq 0$; i.e., $\lambda_0 \neq \infty$. (Otherwise the argument proceeds analogously by exchanging JSx and JHx.) Then (36) implies in particular that x and JSx are linearly independent. Let $(q_2, \ldots, q_n, q_{n+2}, \ldots, q_{2n})$ be an orthonormal basis of $\mathrm{Span}(x, JSx)^\perp$. We assume that $\mathrm{Re}(\lambda_0) \geq 0$ and set

$$Q := [x, q_2, \ldots, q_n, JSx, q_{n+2}, \ldots, q_{2n}]. \tag{37}$$

(In the case $\mathrm{Re}(\lambda_0) < 0$ we set $Q = [JSx, q_2, \ldots, q_n, x, q_{n+2}, \ldots, q_{2n}]$, and the argument proceeds analogously.)

Note that (after having normalized x and JSx) the matrix Q is unitary by (36) and by definition of the q_i. We obtain

$$-JQ^* JSQe_1 = \begin{bmatrix} -(JSx)^* JSx \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} -x^* S^* Sx \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

and

$$-JQ^* JHQe_1 = \begin{bmatrix} -(JSx)^* JHx \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} -x^* S^* Hx \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Since $-JQ^* J(\lambda S - H)Q$ is still skew-Hamiltonian/Hamiltonian, we obtain

$$\begin{aligned} &-JQ^* J(\lambda S - H)Q \\ &\quad = \lambda \begin{bmatrix} s_{11} & * & * & * \\ 0 & S_{22} & * & S_{24} \\ 0 & 0 & s_{11}^* & 0 \\ 0 & S_{42} & * & S_{22}^* \end{bmatrix} - \begin{bmatrix} h_{11} & * & * & * \\ 0 & H_{22} & * & H_{24} \\ 0 & 0 & -h_{11}^* & 0 \\ 0 & H_{42} & * & -H_{22}^* \end{bmatrix}, \end{aligned}$$

where $S_{22}, S_{24}, S_{42}, H_{22}, H_{24}, H_{42} \in \mathcal{C}^{(n-1)\times(n-1)}$, and furthermore $s_{11} = -x^*S^*Sx$ and $h_{11} = -x^*S^*Hx$. Obviously the pencil $\lambda s_{11} - h_{11}$ has the eigenvalue λ_0 and by permutation of rows and columns we see that the spectrum of $\lambda S - H$ is equal to the union of the spectra of

$$\lambda \begin{bmatrix} s_{11} & * \\ 0 & s_{11}^* \end{bmatrix} - \begin{bmatrix} h_{11} & * \\ 0 & -h_{11}^* \end{bmatrix} \quad \text{and} \quad \lambda \begin{bmatrix} S_{22} & S_{24} \\ S_{42} & S_{22}^* \end{bmatrix} - \begin{bmatrix} H_{22} & H_{24} \\ H_{42} & -H_{22}^* \end{bmatrix}.$$

(b) If the assumption in (a) does not hold, then for every eigenvector $x \neq 0$ associated with an eigenvalue of $\lambda S - H$ we have

$$x^*JSx \neq 0 \quad \text{or} \quad x^*JHx \neq 0. \tag{38}$$

In this case we see by Lemma 2.4 that all the eigenvalues of $\lambda S - H$ are purely imaginary and by Lemma 2.5 they all have partial multiplicities equal to one. Once again we assume $x^*JSx \neq 0$. (Otherwise the argument proceeds analogously by exchanging JSx and JHx.) Let $(q_1, \ldots, q_{n-1}, q_{n+1}, \ldots, q_{2n})$ be an orthonormal basis of $\mathrm{Span}(JSx)^\perp$. Then

$$P_1 := [q_1, \ldots, q_{n-1}, x, q_{n+1}, \ldots, q_{2n}] \tag{39}$$

is invertible. This follows, since if $\alpha_1, \ldots, \alpha_{2n} \in \mathcal{C}$ are such that

$$\alpha_1 q_1 + \cdots + \alpha_{n-1} q_{n-1} + \alpha_n x + \alpha_{n+1} q_{n+1} + \cdots + \alpha_{2n} q_{2n} = 0, \tag{40}$$

then multiplying (40) from the left by x^*S^*J we obtain $\alpha_{2n} = 0$, since $x^*JSx \neq 0$, and therefore $\alpha_i = 0$ for all $i = 1, \ldots, 2n$, since the q_i form a basis. Hence we obtain

$$-JP_1^*JSx = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ s_{42} \end{bmatrix} \quad \text{and} \quad -JP_1^*JHx = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ h_{42} \end{bmatrix},$$

where $s_{42} = x^*JSx$ and $h_{42} = x^*JHx$. This implies that

$$\begin{aligned} &-JP_1^*J(\lambda S - H)P_1 \\ &\quad = \lambda \begin{bmatrix} S_{11} & 0 & S_{13} & S_{14} \\ S_{21} & 0 & -S_{14}^* & S_{24} \\ S_{31} & 0 & S_{11}^* & S_{21}^* \\ 0 & s_{42} & 0 & 0 \end{bmatrix} - \begin{bmatrix} H_{11} & 0 & H_{13} & H_{14} \\ H_{21} & 0 & H_{14}^* & H_{24} \\ H_{31} & 0 & -H_{11}^* & -H_{21}^* \\ 0 & h_{42} & 0 & 0 \end{bmatrix}, \end{aligned} \tag{41}$$

where $S_{11}, H_{11} \in \mathcal{C}^{(n-1)\times(n-1)}$ and the other blocks have corresponding sizes. Now set

$$\lambda\tilde{S} - \tilde{H} := \lambda \begin{bmatrix} S_{11} & S_{13} & S_{14} \\ S_{21} & -S_{14}^* & S_{24} \\ S_{31} & S_{11}^* & S_{21}^* \end{bmatrix} - \begin{bmatrix} H_{11} & H_{13} & H_{14} \\ H_{21} & H_{14}^* & H_{24} \\ H_{31} & -H_{11}^* & -H_{21}^* \end{bmatrix} \tag{42}$$

and let y be an eigenvalue of $\lambda\tilde{S} - \tilde{H}$ associated with the eigenvalue μ_0. By (38) we find

$$y^*J\tilde{S}y \neq 0 \quad \text{or} \quad y^*J\tilde{H}y \neq 0,$$

which implies, in particular, that also μ_0 is purely imaginary with partial multiplicity equal to one. Again we assume $y^*J\tilde{S}y \neq 0$ and choose an orthonormal basis $(p_1, \dots, p_{2n-2})$ of $\mathrm{Span}(J\tilde{S}x)^\perp$. Analogous to the proof of the nonsingularity of (39), we can show that the vectors y and $(p_1, \dots, p_{2n-2})$ are linearly independent. Embedding these vectors canonically in $\mathcal{C}^n$ and then setting

$$P_2 := [p_1, \dots, p_{n-1}, e_n, p_n, \dots, p_{2n-2}, y],$$

we obtain

$$\begin{aligned} &-JP_2^*P_1^*J(\lambda S - H)P_1P_2 \\ &\qquad = \lambda \begin{bmatrix} \tilde{S}_{11} & 0 & \tilde{S}_{13} & 0 \\ 0 & 0 & 0 & s_{24} \\ \tilde{S}_{31} & 0 & \tilde{S}_{11}^* & 0 \\ 0 & s_{42} & 0 & 0 \end{bmatrix} - \begin{bmatrix} \tilde{H}_{11} & 0 & \tilde{H}_{13} & 0 \\ 0 & 0 & 0 & h_{24} \\ \tilde{H}_{31} & 0 & -\tilde{H}_{11}^* & 0 \\ 0 & h_{42} & 0 & 0 \end{bmatrix}, \end{aligned} \tag{43}$$

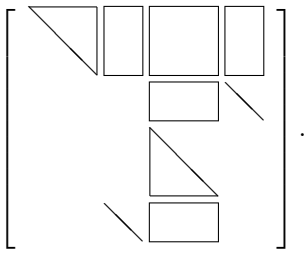
where $\tilde{S}_{11}, \tilde{S}_{13}, \tilde{S}_{31}, \tilde{H}_{11}, \tilde{H}_{13}, \tilde{H}_{31} \in \mathcal{C}^{(n-1)\times(n-1)}$, and furthermore $s_{24} = y^*J\tilde{S}y$ and $h_{24} = y^*J\tilde{H}y$. Thus, the spectrum of $\lambda S - H$ is equal to the union of the spectra of

$$\lambda \begin{bmatrix} \tilde{S}_{11} & \tilde{S}_{13} \\ \tilde{S}_{31} & \tilde{S}_{11}^* \end{bmatrix} - \begin{bmatrix} \tilde{H}_{11} & \tilde{H}_{13} \\ \tilde{H}_{31} & -\tilde{H}_{11}^* \end{bmatrix} \quad \text{and} \quad \lambda \begin{bmatrix} 0 & s_{24} \\ s_{42} & 0 \end{bmatrix} - \begin{bmatrix} 0 & h_{24} \\ h_{42} & 0 \end{bmatrix},$$

and the eigenvalues of $\lambda\left[\begin{smallmatrix} 0 & s_{24} \\ s_{42} & 0 \end{smallmatrix}\right] - \left[\begin{smallmatrix} 0 & h_{24} \\ h_{42} & 0 \end{smallmatrix}\right]$ are λ_0 and μ_0.

In both cases we have reduced the problem to a problem of smaller dimension and the proof follows by induction. □

REMARK 5.3. The matrices $-JP^*JSP$ and $-JP^*JHP$ in (35) have the pattern



.

REMARK 5.4. The proof of Theorem 5.2 also suggests a procedure to compute the almost-Schur form (35). But during the computation the following difficulties may occur:

1. In the first step we may compute only one eigenvalue and then reduce the pencil as indicated in part (a) or (b) of the proof of Theorem 5.2. This means that if we compute an eigenvector satisfying (38), we do not know whether (38) holds for all the eigenvalues or if there exists an eigenvalue such that (36) holds. In particular we might fail in finding an eigenvector of the subpencil (42) that satisfies (38). Consider, for example, the pencil

$$\lambda \begin{bmatrix} 0 & 0 & i & i \\ i & 0 & i & 0 \\ 0 & 0 & 0 & -i \\ 0 & -i & 0 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 1 & 2 \\ 1 & 0 & 2 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

The subpencil

$$\lambda \begin{bmatrix} 0 & i & i \\ i & i & 0 \\ 0 & 0 & -i \end{bmatrix} - \begin{bmatrix} 0 & 1 & 2 \\ 1 & 2 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

has the eigenvalue i with partial multiplicity 3. Thus, Lemma 2.5 implies that we do not find an eigenvector satisfying (38). Therefore, the following observation may be helpful.

Given a skew-Hamiltonian/Hamiltonian pencil in the form (41) and an eigenvector x satisfying (36), we can perform a step as indicated in part (a) of the proof of Theorem 5.2 by choosing the basis of $\mathrm{Span}(x, JSx)^{\perp}$ (resp., $\mathrm{Span}(x, JHx)^{\perp}$) such that q_n in (37) is equal to e_n. This is always possible, since e_n is an eigenvector. Then we obtain by Lemmas 2.4 or 2.5 that

$$e_n^* JSx = 0$$

and by the special structure of the pencil (41) also

$$e_n^* x = 0.$$

Hence we can proceed reducing the matrix as indicated in (a) until we find another eigenvector y satisfying (38).

2. If we compute an eigenvector satisfying (38) during the reduction of $\lambda S - H$, we do not achieve the J-Schur form by proceeding as indicated in part (b) of the proof of Theorem 5.2. But this does in general not mean that there is no J-Schur form. Consider as an example the skew-Hamiltonian matrix

$$S := \begin{bmatrix} -1 - 0.5i & 0 & 0.5i & 0 \\ 0 & -1 - 0.5i & 0 & 0.5i \\ -0.5i & 0 & -1 + 0.5i & 0 \\ 0 & -0.5i & 0 & -1 + 0.5i \end{bmatrix}$$

and the skew-Hamiltonian/Hamiltonian pencil $\lambda S - (iS)$. We note that e_1 and e_3 are eigenvectors associated with the eigenvalue i. Now we have

$$e_1^* JSe_1 \neq 0 \quad \text{and} \quad e_3^* JSe_3 \neq 0,$$

but

$$(e_1 + e_3)^* JS(e_1 + e_3) = 0 = (e_1 + e_3)^* JH(e_1 + e_3).$$

Thus, if we choose the eigenvector e_1 instead of $(e_1 + e_3)$ for the reduction of the pencil, we would create the almost-Schur form, although there exists a J-Schur form. Setting

$$U := \frac{1}{\sqrt{2}} J \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$$

we obtain

$$-JU^* J(\lambda S - iS)U = -\lambda \begin{bmatrix} 1 & 0 & -i & 0 \\ 0 & 1 & 0 & -i \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} + \begin{bmatrix} i & 0 & 1 & 0 \\ 0 & i & 0 & 1 \\ 0 & 0 & i & 0 \\ 0 & 0 & 0 & i \end{bmatrix}.$$

Since we do not know how to compute an eigenvector satisfying (36), if it exists, we currently see no method to avoid this problem.

In the real case we obtain a result similar to Theorem 5.2.

DEFINITION 5.5. *Let $P = [p_1, \dots, p_n] \in \mathcal{R}^n$ be nonsingular. Then P is called* almost orthogonal *if there exists indices k, l such that*

$$(p_1, \dots, p_{k-1}, p_{k+1}, \dots, p_{l-1}, p_{l+1}, \dots, p_n)$$

is an orthonormal tuple of vectors.

THEOREM 5.6. *Let $\lambda S - H \in \mathcal{R}^{2n\times 2n}$ be a regular skew-Hamiltonian/Hamiltonian pencil. Then there exists a nonsingular matrix $P \in \mathcal{R}^{2n\times 2n}$ and $k \in \mathcal{N}$ such that*

$$\begin{aligned} &-JP^*J(\lambda S - H)P \\ &\quad = \lambda \begin{bmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ 0 & 0 & -S_{14}^* & S_{24} \\ 0 & 0 & S_{11}^* & 0 \\ 0 & S_{42} & S_{12}^* & 0 \end{bmatrix} - \begin{bmatrix} H_{11} & H_{12} & H_{13} & H_{14} \\ 0 & 0 & H_{14}^* & H_{24} \\ 0 & 0 & -H_{11}^* & 0 \\ 0 & H_{42} & -H_{12}^* & 0 \end{bmatrix}, \end{aligned} \tag{44}$$

where the following conditions hold:

1. *$\lambda S_{11} - H_{11} \in \mathcal{R}^{k\times k}$ is a quasi upper triangular pencil having only eigenvalues with nonnegative real part.*
2. *The pencils $\lambda S_{24} - H_{24}, \lambda S_{42} - H_{42} \in \mathcal{R}^{(n-k)\times(n-k)}$ are block diagonal and the diagonal blocks are either of the form*

$$\lambda \begin{bmatrix} 0 & & \\ & \ddots & \\ & & 0 \end{bmatrix} - \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix} \in \mathcal{R}^{m\times m} \tag{45}$$

for $m \in \mathcal{N}$ or of the form

$$\lambda \begin{bmatrix} 0 & s \\ -s & 0 \end{bmatrix} - \begin{bmatrix} h & 0 \\ 0 & h \end{bmatrix}, \tag{46}$$

where $s, h \in \mathcal{R}\backslash\{0\}$. In particular, all the eigenvalues of the skew-Hamiltonian/Hamiltonian pencil

$$\lambda \begin{bmatrix} 0 & S_{24} \\ S_{42} & 0 \end{bmatrix} - \begin{bmatrix} 0 & H_{24} \\ H_{42} & 0 \end{bmatrix} \tag{47}$$

are nonzero and purely imaginary.

3. *The spectrum of $\lambda S - H$ is equal to the union of the spectra of*

$$\lambda \begin{bmatrix} S_{11} & S_{13} \\ 0 & S_{11}^* \end{bmatrix} - \begin{bmatrix} H_{11} & H_{13} \\ 0 & -H_{11}^* \end{bmatrix} \quad \text{and} \quad \lambda \begin{bmatrix} 0 & S_{24} \\ S_{42} & 0 \end{bmatrix} - \begin{bmatrix} 0 & H_{24} \\ H_{42} & 0 \end{bmatrix}.$$

4. *P is a product of orthogonal and almost orthogonal matrices.*

Proof. Analogous to the proof of Theorem 5.2 we distinguish two cases.

(a) Let us assume that there exists an eigenvector $x \in \mathcal{C}^n\backslash\{0\}$ associated with an eigenvalue $\lambda_0 \in \mathcal{C}$ of $\lambda S - H$ such that

$$x^*JSx = 0 \ \text{ and } \ x^*JHx = 0. \tag{48}$$

If λ_0 and x are real we can proceed as in the proof of Theorem 5.2. So let us assume that λ_0 has nonzero real part and that $x = y + iz$, where $y, z \in \mathcal{R}^n$. Furthermore we assume once again w.l.o.g. that $JSx \neq 0$. (Otherwise the argument proceeds analogously by exchanging JSx and JHx.) This means in particular that both JSy and JSz do not vanish. Otherwise y or z would be an eigenvector associated with the eigenvalue ∞. It is well known that y, z form a basis of a two-dimensional subspace; i.e., in particular y and z are linearly independent. Furthermore, we have by (48) that

$$\begin{aligned} y^*JSy + z^*JSz + i(y^*JSz - z^*JSy) \\ = (y + iz)^*JS(y + iz) = 0. \end{aligned} \tag{49}$$

Since $\lambda_0^* \neq -\lambda_0^*$ and $y - iz$ is an eigenvalue associated with λ_0^*, we also have by Lemma 2.4 that

$$\begin{aligned} y^*JSy - z^*JSz + i(y^*JSz - z^*JSy) \\ = (y - iz)^*JS(y + iz) = 0. \end{aligned} \tag{50}$$

Hence by adding, resp., subtracting (49) and (50) we obtain

$$y^*JSy = z^*JSz = z^*JSy = y^*JSz = 0.$$

If we choose a real orthonormal basis (r_1, r_2) of $\mathrm{Span}(y, z)$, we still have

$$r_1, r_2 \perp JSr_1, JSr_2.$$

Now let $(q_1, \ldots, q_{2n-4})$ be a real orthonormal basis of $\mathrm{Span}(r_1, r_2, JSr_1, JSr_2)$. Then (assuming again w.l.o.g. that λ_0 has nonnegative real part and after having normalized r_1, r_2, JSr_1, and JSr_2) the matrix

$$Q := [r_1, r_2, q_1, \ldots, q_{n-2}, JSr_1, JSr_2, q_{n-1}, \ldots, q_{2n-4}]$$

is orthogonal and we obtain

$$\begin{aligned} &-JQ^*J(\lambda S - H)Q \\ &\quad = \lambda \begin{bmatrix} S_{11} & * & * & * \\ 0 & S_{22} & * & S_{24} \\ 0 & 0 & S_{11}^* & 0 \\ 0 & S_{42} & * & S_{22}^* \end{bmatrix} - \begin{bmatrix} H_{11} & * & * & * \\ 0 & H_{22} & * & H_{24} \\ 0 & 0 & -H_{11}^* & 0 \\ 0 & H_{42} & * & -H_{22}^* \end{bmatrix}, \end{aligned}$$

where $S_{11}, H_{11} \in \mathcal{R}^{2\times 2}$ and $S_{22}, S_{24}, S_{42}, H_{22}, H_{24}, H_{42} \in \mathcal{R}^{(n-2)\times(n-2)}$. Obviously the pencil $\lambda S_{11} - H_{11}$ has the eigenvalues λ_0 and λ_0^* and the spectrum of $\lambda S - H$ is equal to the union of the spectra of

$$\lambda \begin{bmatrix} S_{11} & * \\ 0 & S_{11}^* \end{bmatrix} - \begin{bmatrix} H_{11} & * \\ 0 & -H_{11}^* \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} S_{22} & S_{24} \\ S_{42} & S_{22}^* \end{bmatrix} - \begin{bmatrix} H_{22} & H_{24} \\ H_{42} & -H_{22}^* \end{bmatrix}.$$

(b) If the assumption in (a) does not hold, then for every eigenvector $x \in \mathcal{C}^n \backslash \{0\}$ associated with an eigenvalue of $\lambda S - H$ we have

$$x^*JSx \neq 0 \quad \text{or} \quad x^*JHx \neq 0. \tag{51}$$

This means again that all the eigenvalues of $\lambda S - H$ are purely imaginary and all their partial multiplicities are one. In addition we have that $\lambda S - H$ cannot have the

eigenvalue zero, for if $w \in \mathcal{R}^n$ is an eigenvector associated with zero, i.e., $Hw = 0$, since S is skew-Hamiltonian, we obtain

$$w^*JSw = (w^*JSw)^* = w^*S^*Jw = -w^*JSw;$$

i.e., $w^*JSw = 0 = w^*JHw$. In the case of the eigenvalue ∞, we can find a real eigenvector and so we proceed as in the proof of Theorem 5.2. Thus it remains to consider the case that x is an eigenvector associated with the eigenvalue $i\mu$ for a $\mu \in \mathcal{R}\backslash\{0\}$ and that $x = y + iz$, where $y, z \in \mathcal{R}^n$. Again y and z are linearly independent and we have both $JSy \neq 0$ and $JSz \neq 0$. Now let $(q_1, \ldots, q_{n-2}, q_{n+1}, \ldots, q_{2n})$ be a real orthonormal basis of $\mathrm{Span}(JSy, JSz)^\perp$. Analogous to the proof of the nonsingularity of (39) we can show that

$$\tilde{P} := [q_1, \ldots, q_{n-2}, y, z, q_{n+1}, \ldots, q_{2n}]$$

is nonsingular. We obtain

$$\begin{aligned} &-J\tilde{P}^*J(\lambda S - H)\tilde{P} \\ &\quad = \lambda \begin{bmatrix} S_{11} & 0 & S_{13} & S_{14} \\ S_{21} & 0 & -S_{14}^* & S_{24} \\ S_{31} & 0 & S_{11}^* & S_{21}^* \\ 0 & S_{42} & 0 & 0 \end{bmatrix} - \begin{bmatrix} H_{11} & 0 & H_{13} & H_{14} \\ H_{21} & 0 & H_{14}^* & H_{24} \\ H_{31} & 0 & -H_{11}^* & -H_{21}^* \\ 0 & H_{42} & 0 & 0 \end{bmatrix}, \end{aligned} \tag{52}$$

where

$$S_{42} = \begin{bmatrix} y^*JSy & y^*JSz \\ z^*JSy & z^*JSz \end{bmatrix} \quad \text{and} \quad H_{42} = \begin{bmatrix} y^*JHy & y^*JHz \\ z^*JHy & z^*JHz \end{bmatrix},$$

and $S_{11}, H_{11} \in \mathcal{R}^{(n-2)\times(n-2)}$. The other blocks have corresponding sizes. Since S_{42} is skew symmetric and $y^*JSz = -z^*JSy$ we obtain

$$S_{42} = \begin{bmatrix} 0 & s \\ -s & 0 \end{bmatrix},$$

where $s := -y^*JSz$. This yields in particular that $y^*JSy = 0$. On the other hand, we have $i\mu S(y + iz) = H(y + iz)$; i.e.,

$$-\mu Sz = Hy \quad \text{and} \quad \mu Sy = Hz.$$

Therefore, we obtain that

$$y^*JHz = \mu y^*JSy = 0$$

and

$$y^*JHy = -\mu y^*JSz = \mu z^*S^*Jy = \mu z^*JSy = z^*JHz.$$

Thus,

$$H_{42} = \begin{bmatrix} h & 0 \\ 0 & h \end{bmatrix},$$

where $h := -y^*JHy$. The rest of the proof of (b) follows analogous to the proof of Theorem 5.2, reducing the subpencil

$$\lambda \begin{bmatrix} S_{11} & S_{13} & S_{14} \\ S_{21} & -S_{14}^* & S_{24} \\ S_{31} & S_{11}^* & S_{21}^* \end{bmatrix} - \begin{bmatrix} H_{11} & H_{13} & H_{14} \\ H_{21} & H_{14}^* & H_{24} \\ H_{31} & -H_{11}^* & -H_{21}^* \end{bmatrix}$$

by choosing a nonreal eigenvector associated with an eigenvalue $i\tilde{\mu}, \tilde{\mu} \in \mathcal{R}$, or two real eigenvectors associated with the eigenvalue ∞. (Note that the eigenvalue ∞ must have even algebraic multiplicity, since all the other eigenvalues of the pencil occur in pairs and the size of the pencil is even.)

In both cases we have reduced the problem to a problem of smaller dimension and the proof follows by induction. □

LEMMA 5.7. *Let $\lambda S - H$ be a real skew-Hamiltonian/Hamiltonian pencil. Then the following conditions hold:*

1. *Every block of odd size belonging to the eigenvalue zero in the Kronecker canonical form occurs with even multiplicity.*
2. *Every block of even size belonging to the eigenvalue ∞ in the Kronecker canonical form occurs with even multiplicity.*

Proof. See the proof in [20] of the analogous result for pencils $\lambda A - B$, where A is symmetric and B is skew symmetric. □

6. Conclusions. We have discussed canonical forms and almost-Schur forms for skew-Hamiltonian/Hamiltonian pencils under J-congruence. Also a necessary and sufficient condition for the existence of a J-Schur form has been given.

Acknowledgment. We thank V. Mehrmann for helpful discussions and for initiating this research.

REFERENCES

[1] H. BRAUN AND M. KOECHER, *Jordan-Algebren*, Springer-Verlag, Berlin, Heidelberg, New York, 1966.
[2] N. BURGOYNE AND R. CUSHMAN, *Normal forms for real Hamiltonian matrices*, in The 1976 Ames Research Center Conference on Geometric Control Theory, C. Martin and R. Hermann, eds., Mathematical Sciences Press, Brookline, MA, 1976, pp. 483–529.
[3] A. CIAMPI, *Classical Hamiltonian Linear Systems*, Queen's Papers in Pure and Appl. Math. 31, Queen's University, Kingston, ON, Canada, 1972.
[4] A. CIAMPI, *Classification of Hamiltonian linear systems*, Indiana Univ. Math. J., 23 (1973), pp. 513–526.
[5] D.-Ž. DJOKOVIĆ, J. PATERA, P. WINTERWITZ, AND H. ZASSENHAUS, *Normal forms of elements of classical real and complex Lie and Jordan algebras*, J. Math. Phys., 24 (1983), pp. 1363–1374.
[6] A. FERRANTE AND B. C. LEVY, *Canonical form of symplectic matrix pencils*, Linear Algebra Appl., 274 (1998), pp. 259–300.
[7] F. GANTMACHER, *Theory of Matrices*, Vol. 1, Chelsea, New York, 1959.
[8] F. GANTMACHER, *Theory of Matrices*, Vol. 2, Chelsea, New York, 1959.
[9] G. GOLUB AND C. VAN LOAN, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1983.
[10] P. LANCASTER AND L. RODMAN, *Algebraic Riccati Equations*, Clarendon Press, Oxford, 1995.
[11] A. J. LAUB AND K. MEYER, *Canonical forms for symplectic and Hamiltonian matrices*, Celestial Mech., 9 (1974), pp. 213–238.
[12] W. W. LIN AND T. C. HO, *On Schur Type Decompositions for Hamiltonian and Symplectic Pencils*, Technical report, Institute of Applied Mathematics, National Tsing Hua University, Taiwan, 1990.
[13] V. MEHRMANN, *The Autonomous Linear Quadratic Control Problem, Theory and Numerical Solution*, Lecture Notes in Control and Information Sciences 163, Springer-Verlag, Heidelberg, 1991.
[14] V. MEHRMANN, *A step towards a unified treatment of continuous and discrete time control problems*, Linear Algebra Appl., 241–243 (1996), pp. 749–779.
[15] V. MEHRMANN AND H. XU, *Canonical Forms for Hamiltonian and Symplectic Matrices and Pencils*, Preprint SFB 393/98-07, Sonderforschungsbereich 393 Numerische Simulation auf massiv parallelen Rechnern, Fakultät für Mathematik, Technische Universität Chemnitz, D-09107 Chemnitz, FRG, 1998.

[16] K. Meyer and G. Hall, *Introduction to Hamiltonian Dynamical Systems and the N-Body Problem*, Springer, New York, 1992.
[17] C. Paige and C. Van Loan, *A Schur decomposition for Hamiltonian matrices*, Linear Algebra Appl., 14 (1981), pp. 11–32.
[18] I. Smirnov, *Linear Algebra and Group Theory*, McGraw–Hill, New York, Toronto, London, 1961.
[19] R. Thompson, *The characteristic polynomial of a principal subpencil of a hermitian matrix pencil*, Linear Algebra Appl., 14 (1976), pp. 135–177.
[20] R. Thompson, *Pencils of complex and real symmetric and skew matrices*, Linear Algebra Appl., 147 (1991), pp. 323–371.
[21] H. Weyl, *The Classical Groups*, Princeton University Press, Princeton, NJ, 1946.
[22] H. K. Wimmer, *Normal forms of symplectic pencils and discrete-time algebraic Riccati equations*, Algebra Appl., 147 (1991), pp. 411–440.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 477–480
© 1999 Society for Industrial and Applied Mathematics

ON OPTIMAL BANDED PRECONDITIONERS FOR THE FIVE-POINT LAPLACIAN*

ZHANYE TONG† AND AHMED SAMEH†

Abstract. We prove the following theorem, which includes a conjecture by Greenbaum and Rodrigue [*BIT*, 29 (1989), pp. 610–634] as a special case:

Let A_h be the five-point Laplace matrix on a unit square for grid size h, and let M_h be any symmetric positive definite preconditioner for A_h with half bandwidth $k = O(1)$. Then the condition number $\kappa(M_h^{-\frac{1}{2}} A_h M_h^{-\frac{1}{2}})$ satisfies $\kappa(M_h^{-\frac{1}{2}} A_h M_h^{-\frac{1}{2}}) = O(h^{-2})$.

Key words. optimal, preconditioner, tridiagonal

AMS subject classification. 65F10

PII. S0895479898344558

1. Introduction.

In Greenbaum and Rodrigue [1], optimal preconditioners of a given sparsity pattern are discussed. By using the Overton optimization code [3], optimal tridiagonal preconditioners are computed for the five-point Laplacian on a unit square

$$A_h = \begin{bmatrix} T & -I & & \\ -I & T & \ddots & \\ & \ddots & \ddots & -I \\ & & -I & T \end{bmatrix}_{n^2 \times n^2}, \quad T = \begin{bmatrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{bmatrix}_{n \times n}, \tag{1.1}$$

where I is the identity matrix of order n and $h = 1/(n+1)$ is the mesh size. It turns out that the optimal tridiagonal preconditioner for A_h is slightly better than the block diagonal of A_h which is the optimal block diagonal preconditioner. Based on the numerical results, they made the following conjecture.

Conjecture. Let A_h be the five-point Laplace matrix of (1.1) for grid size h, and let M_h be any symmetric positive definite tridiagonal preconditioner for A_h. Then the condition number of the preconditioned matrix $M_h^{-\frac{1}{2}} A_h M_h^{-\frac{1}{2}}$ satisfies

$$\kappa(M_h^{-\frac{1}{2}} A_h M_h^{-\frac{1}{2}}) = O(h^{-2}). \tag{1.2}$$

In this short note, we show that the conjecture is indeed true. Actually, we shall prove the following theorem.

THEOREM 1.1. *Let A_h be the five-point Laplace matrix of* (1.1) *for grid size h, and let M_h be any symmetric positive definite preconditioner for A_h with half bandwidth $k = O(1)$. Then the condition number of the preconditioned matrix $M_h^{-\frac{1}{2}} A_h M_h^{-\frac{1}{2}}$ satisfies*

$$\kappa(M_h^{-\frac{1}{2}} A_h M_h^{-\frac{1}{2}}) = O(h^{-2}). \tag{1.3}$$

*Received by the editors September 9, 1998; accepted for publication by A. Greenbaum May 2, 1999; published electronically November 10, 1999.
http://www.siam.org/journals/simax/21-2/34455.html
†Department of Computer Sciences, Purdue University, West Lafayette, IN 47907 (tong@cs.purdue.edu, sameh@purdue.edu).

Moreover, for $k = 1$, we have

$$\kappa(M_h^{-\frac{1}{2}} A_h M_h^{-\frac{1}{2}}) \geq \frac{5}{13}\kappa(A_h).$$

2. The proof of Theorem 1.1. We first introduce two lemmas.

LEMMA 2.1. *Let n be a positive integer, $h = \frac{1}{n+1}$. Then for any $k, 1 \leq k \leq n-1$, the matrix*

$$\begin{bmatrix} \sin[2(n-k+1)h\pi] & \sin[3(n-k+1)h\pi] & \cdots & \sin[(k+1)(n-k+1)h\pi] \\ \sin[2(n-k+2)h\pi] & \sin[3(n-k+2)h\pi] & \cdots & \sin[(k+1)(n-k+2)h\pi] \\ \vdots & \vdots & & \vdots \\ \sin[2nh\pi] & \sin[3nh\pi] & \cdots & \sin[(k+1)nh\pi] \end{bmatrix} \tag{2.1}$$

is invertible.

Proof. Let $\theta = h\pi$. It suffices to prove that the matrix

$$C = \begin{bmatrix} \sin[2\theta] & \sin[3\theta] & \cdots & \sin[(k+1)\theta] \\ \sin[4\theta] & \sin[6\theta] & \cdots & \sin[2(k+1)\theta] \\ \vdots & \vdots & & \vdots \\ \sin[2k\theta] & \sin[3k\theta] & \cdots & \sin[k(k+1)\theta] \end{bmatrix} \tag{2.2}$$

is invertible. It is well known that for any integer $s \geq 1$

$$\sin(s\phi) = \sin(\phi)\left(c_s \cos^{s-1}(\phi) + p_{s-2}(\cos(\phi))\right),$$

in which $c_s \neq 0$ is a constant independent of ϕ and p_{s-2} is a polynomial of order $s-2$. It follows that

$$\det(C) = c\prod_{i=2}^{k+1} \sin(i\theta) \times \det \begin{vmatrix} 1 & 1 & \cdots & 1 \\ \cos[2\theta] & \cos[3\theta] & \cdots & \cos[(k+1)\theta] \\ \vdots & \vdots & & \vdots \\ \cos^{k-1}[2\theta] & \cos^{k-1}[3\theta] & \cdots & \cos^{k-1}[(k+1)\theta] \end{vmatrix} \neq 0. \quad \square$$

LEMMA 2.2. *Let n be a positive integer, $h = \frac{1}{n+1}$. Then we have*

$$\frac{\left(\sin^2(\frac{nh\pi}{2}) + \sin^2(\frac{h\pi}{2})\right)\sin^2(2h\pi) + \left(\sin^2(\frac{nh\pi}{2}) + \sin^2(\frac{2h\pi}{2})\right)\sin^2(h\pi)}{\left(\sin^2(\frac{h\pi}{2}) + \sin^2(\frac{h\pi}{2})\right)\sin^2(2h\pi) + \left(\sin^2(\frac{h\pi}{2}) + \sin^2(\frac{2h\pi}{2})\right)\sin^2(h\pi)} \geq \frac{5}{13} \times \frac{\cos^2(\frac{h\pi}{2})}{\sin^2(\frac{h\pi}{2})}.$$

Proof. The proof is straightforward and hence omitted. $\square$

Now, let us prove Theorem 1.1. By the definition of the condition number, we have

$$\begin{aligned} \kappa(M_h^{-\frac{1}{2}} A_h M_h^{-\frac{1}{2}}) &= \frac{\lambda_{\max}(M_h^{-1}A_h)}{\lambda_{\min}(M_h^{-1}A_h)} \\ &= \frac{\max_{x\neq 0} \frac{x^T A_h x}{x^T M_h x}}{\min_{y\neq 0} \frac{y^T A_h y}{y^T M_h y}} \\ &\geq \frac{x^T A_h x}{y^T A_h y} \times \frac{y^T M_h y}{x^T M_h x} \text{ for any } x, y \neq 0. \end{aligned}$$

For any matrix M_h with half bandwidth $k = O(1)$, we are going to construct two vectors, x_0 and y_0, such that

$$\frac{x_0^T A_h x_0}{y_0^T A_h y_0} = O(h^{-2}) \quad \text{and} \quad x_0^T M_h x_0 = y_0^T M_h y_0. \tag{2.3}$$

The eigenvalues and the corresponding eigenvectors of A_h are [2]

$$\lambda^{(i,j)} = 4\left(\sin^2\left(\frac{ih\pi}{2}\right) + \sin^2\left(\frac{jh\pi}{2}\right)\right), \quad 1 \le i, j \le n,$$

and

$$v^{(i,j)} = \begin{bmatrix} \sin(ih\pi)V_j \\ \sin(2ih\pi)V_j \\ \vdots \\ \sin(nih\pi)V_j \end{bmatrix}, \quad V_j = \begin{bmatrix} \sin(jh\pi) \\ \sin(2jh\pi) \\ \vdots \\ \sin(njh\pi) \end{bmatrix}, \quad 1 \le i, j \le n.$$

Particularly, for any $j, 1 \le j \le n$, we have

$$v^{(1,j)} = \begin{bmatrix} \sin(h\pi)V_j \\ \sin(2h\pi)V_j \\ \vdots \\ \sin(nh\pi)V_j \end{bmatrix}, \quad v^{(n,j)} = \begin{bmatrix} \sin(h\pi)V_j \\ -\sin(2h\pi)V_j \\ \vdots \\ (-1)^{n+1}\sin(nh\pi)V_j \end{bmatrix}.$$

Let us consider two vectors

$$x_0 = v^{(n,1)} + \sum_{i=1}^{k} a_i v^{(n,i+1)} = [X_1^T, X_2^T, \ldots, X_n^T]^T, \tag{2.4}$$

$$y_0 = v^{(1,1)} + \sum_{i=1}^{k} a_i v^{(1,i+1)} = [Y_1^T, Y_2^T, \ldots, Y_n^T]^T, \tag{2.5}$$

which are partitioned conformally with A_h. It is easy to see that x_0 and y_0 have the following properties:

(1) For any a_i, $1 \le i \le k$, we have $X_i = (-1)^{i+1} Y_i$, $1 \le i \le n$.
(2) For $k = O(1)$, each eigenvector component in x_0 corresponds to an eigenvalue of order $O(1)$, and each eigenvector component in y_0 corresponds to an eigenvalue of order $O(h^2)$.
(3) There exists a set of $a_i, 1 \le i \le k$, such that the last k components in X_i and Y_i are zeros for $1 \le i \le n$. This is because the linear system

$$\begin{bmatrix} \sin[2(n-k+1)h\pi] & \cdots & \sin[(k+1)(n-k+1)h\pi] \\ \vdots & & \vdots \\ \sin[2nh\pi] & \cdots & \sin[(k+1)nh\pi] \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_k \end{bmatrix} = -\begin{bmatrix} \sin[(n-k+1)h\pi] \\ \cdots \\ \sin[nh\pi] \end{bmatrix},$$

has a unique nonzero solution for any k, $1 \le k \le n-1$.

Suppose the constants $a_i's$ in (2.4) and (2.5) are chosen such that the last k components in X_i and Y_i are zeros for $1 \le i \le n$. Then, it is easy to verify that

$$x_0^T M_h x_0 = y_0^T M_h y_0$$

for any symmetric matrix M_h with half bandwidth k.

On the other hand,

$$\begin{aligned}\frac{x_0^T A_h x_0}{y_0^T A_h y_0} &= \frac{\lambda^{(n,1)} + \sum_{i=1}^{k} \lambda^{(n,i+1)} a_i^2}{\lambda^{(1,1)} + \sum_{i=1}^{k} \lambda^{(1,i+1)} a_i^2} \\ &\ge \frac{\left(\min_{i=1}^{k+1} \lambda^{(n,i)}\right)(1 + a_1^2 + \cdots + a_k^2)}{\left(\max_{i=1}^{k+1} \lambda^{(1,i)}\right)(1 + a_1^2 + \cdots + a_k^2)} \\ &= O(h^{-2}) \qquad \text{for } k = O(1).\end{aligned}$$

In the case of $k = 1$, we have

$$a_1 = -\frac{\sin(h\pi)}{\sin(2h\pi)}.$$

Direct computation shows that

$$\begin{aligned}\frac{x_0^T A_h x_0}{y_0^T A_h y_0} &= \frac{\lambda^{(n,1)} + \lambda^{(n,2)} a_1^2}{\lambda^{(1,1)} + \lambda^{(1,2)} a_1^2} \\ &= \frac{\left(\sin^2(\frac{nh\pi}{2}) + \sin^2(\frac{h\pi}{2})\right)\sin^2(2h\pi) + \left(\sin^2(\frac{nh\pi}{2}) + \sin^2(\frac{2h\pi}{2})\right)\sin^2(h\pi)}{\left(\sin^2(\frac{h\pi}{2}) + \sin^2(\frac{h\pi}{2})\right)\sin^2(2h\pi) + \left(\sin^2(\frac{h\pi}{2}) + \sin^2(\frac{2h\pi}{2})\right)\sin^2(h\pi)} \\ &\ge \frac{5}{13} \times \frac{\cos^2(\frac{h\pi}{2})}{\sin^2(\frac{h\pi}{2})} \\ &= \frac{5}{13}\kappa(A_h).\end{aligned}$$

This completes the proof of Theorem 1.1.

The bound for $k = 1$ is not sharp, but it is very tight for $h = \frac{1}{5}, \frac{1}{6}, \frac{1}{7}$ according to the calculated bounds in [1].

REFERENCES

[1] A. Greenbaum and G. H. Rodrigue, *Optimal preconditioner of a given sparsity pattern*, BIT, 29 (1989), pp. 610–634.

[2] W. Hackbusch, *Iterative Solution of Large Sparse Systems of Equations*, Springer-Verlag, New York, 1994.

[3] M. L. Overton, *On minimizing the maximum eigenvalue of a symmetric matrix*, SIAM J. Matrix. Anal. Appl., 9 (1988), pp. 256–268.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 481–507
© 1999 Society for Industrial and Applied Mathematics

CLASSIFICATION OF LINEAR PERIODIC DIFFERENCE EQUATIONS UNDER PERIODIC OR KINEMATIC SIMILARITY*

I. GOHBERG†, M. A. KAASHOEK‡, AND J. KOS§

Abstract. In this paper linear periodic systems of difference equations are classified with respect to periodic similarity and kinematic similarity. Complete sets of invariants of periodic difference equations relative to such similarity transformations are given, and corresponding canonical forms are described. Also the irreducible periodic difference equations, i.e., those that cannot be reduced by such similarities to a nontrivial direct sum, are identified.

Key words. linear periodic difference equations, periodic similarity, kinematic similarity, invariants, canonical forms

AMS subject classifications. 39A10, 15A21

PII. S0895479898334307

Introduction and main results. This paper deals with the problem of classifying linear N-periodic systems of difference equations. Such equations have the form

$$x_{n+1} = A_n x_n, \quad n \in \mathbb{Z} \quad (A_n : \mathbb{C}^{r_n} \to \mathbb{C}^{r_{n+1}},\ n \in \mathbb{Z}), \tag{0.1}$$

with $A_n = A_{n+N}$ for $n \in \mathbb{Z}$. The coefficient A_n in (0.1) is assumed to be a linear transformation and, by N-periodicity, $r_{n+N} = r_n$ for each n.

Our aim is to classify N-periodic difference equations under two types of time-variant basis transformations. The first is that of kinematic similarity. The system (0.1) is said to be *kinematically similar* to the system

$$y_{n+1} = B_n y_n, \quad n \in \mathbb{Z} \quad (B_n : \mathbb{C}^{r_n} \to \mathbb{C}^{r_{n+1}},\ n \in \mathbb{Z}), \tag{0.2}$$

if there exists a time-variant change of basis $y_n = F_n^{-1} x_n$, where F_n is an invertible linear transformation on $\mathbb{C}^{r_n}$, which transforms (0.1) into (0.2), i.e., $B_n = F_{n+1}^{-1} A_n F_n$ for each n, and which satisfies an additional boundedness condition, namely,

$$\sup_{n \in \mathbb{Z}} \{\|F_n\|, \|F_n^{-1}\|\} < \infty. \tag{0.3}$$

The second type of basis transformation appears if, in place of (0.3), we require that $F_n = F_{n+N}$ for $n \in \mathbb{Z}$; in this case (0.1) and (0.2) are called *N-periodically similar*. Obviously, any periodic similarity is kinematic.

In this paper we give necessary and sufficient conditions for two N-periodic difference equations to be kinematically similar or N-periodically similar. Complete sets

*Received by the editors February 17, 1998; accepted for publication (in revised form) by P. Van Dooren February 15, 1999; published electronically November 10, 1999.
http://www.siam.org/journals/simax/21-2/33430.html

†School of Mathematical Sciences, Tel-Aviv University, Ramat-Aviv, 69978 Tel-Aviv, Israel (gohberg@math.tau.ac.il).

‡Faculteit der Exacte Wetenschappen, Vrije Universiteit, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands (kaash@cs.vu.nl).

§Mathematical Models and Methods Department, Informatics Division, National Aerospace Laboratory, P.O. Box 90502, 1006 BM Amsterdam, The Netherlands (kos@nlr.nl). The research for this paper was carried out mainly during this author's doctorate work at the Faculteit Wiskunde en Informatica, Vrije Universiteit, Amsterdam.

of invariants for N-periodic difference equations relative to both types of similarities are given and corresponding canonical forms are described. We also give necessary and sufficient conditions so that a periodic difference equation is periodically or kinematically similar to a periodic difference equation with a different period.

In connection with the construction of canonical forms, the question of whether the N-periodic difference equation (0.1) is kinematically similar to a nontrivial direct sum

$$x_{n+1} = \begin{pmatrix} A_{1,n} & 0 \\ 0 & A_{2,n} \end{pmatrix} x_n, \quad n \in \mathbb{Z},$$

plays an important role. Here nontrivial means that the space $\mathbb{C}^{r_{1,n}}$ on which $A_{1,n}$ acts is nonzero for at least one n and different from $\mathbb{C}^{r_n}$ for at least one (possibly different) n. We say that the N-periodic difference equation (0.1) is *irreducible* (or *N-periodically irreducible*) *in the class of N-periodic difference equations* if (0.1) is not kinematically similar (or not N-periodically similar) to a nontrivial direct sum of N-periodic difference equations. A complete description of these irreducible and N-periodically irreducible difference equations will be given.

For square N-periodic difference equations, i.e., when all the coefficients are square matrices, we have a special interest in (N-periodic) irreducibility in the class of square N-periodic difference equations. A square N-periodic difference equation is called *irreducible* (or *N-periodically irreducible*) *in the class of square N-periodic difference equations* if it is not kinematically similar (or not N-periodically similar) to a nontrivial direct sum of square N-periodic difference equations. These squarely irreducible systems will be described too.

Let us now review our results in more detail. For this purpose we need the linear transformation $A_N \cdots A_1$, which will be referred to as the *monodromy operator* of the N-periodic equation (0.1). We associate with the monodromy operator a Jordan matrix, denoted by $[J_{A_N \cdots A_1}]$, which has nonnegative eigenvalues only, and which is obtained by replacing the diagonal entries in the Jordan normal form $J_{A_N \cdots A_1}$ of $A_N \cdots A_1$ by their absolute values. We are now ready to state our first two theorems.

THEOREM 0.1. *The N-periodic difference equations* (0.1) *and* (0.2) *are kinematically similar if and only if, for each $n \in \{1, 2, \ldots, N\}$,*

$$\operatorname{rank}(A_{j+n-1}A_{j+n-2}\cdots A_n) = \operatorname{rank}(B_{j+n-1}B_{j+n-2}\cdots B_n),\ 1 \le j \le N(\max_{1\le n\le N} r_n), \tag{0.4}$$

and the Jordan matrices $[J_{A_N \cdots A_1}]$ and $[J_{B_N \cdots B_1}]$ are similar.

THEOREM 0.2. *The N-periodic difference equations* (0.1) *and* (0.2) *are N-periodically similar if and only if, for each $n \in \{1, 2, \ldots, N\}$, the rank conditions* (0.4) *are satisfied and the monodromy matrices $A_N A_{N-1} \cdots A_1$ and $B_N B_{N-1} \cdots B_1$ are similar.*

To describe the irreducible equations, we need some additional notation. Given the N-periodic difference equation (0.1), we define the nonnegative integers

$$s_{n,j} = \begin{cases} \operatorname{rank}(A_{n+j-1}\cdots A_{n+1}) - \operatorname{rank}(A_{n+j-1}\cdots A_{n+1}A_n), & j \ge 2,\ n = 1, 2, \ldots, N, \\ r_{n+1} - \operatorname{rank} A_n, & j = 1,\ n = 1, 2, \ldots, N. \end{cases} \tag{0.5}$$

Notice that these numbers do not change under kinematic similarity.

THEOREM 0.3. *The N-periodic difference equation* (0.1) *is irreducible in the class of N-periodic difference equations (either by N-periodic similarity or by kinematic*

similarity) if and only if $A_N A_{N-1} \cdots A_1$ *is an invertible matrix of which the Jordan form has one block only or* $A_N A_{N-1} \cdots A_1$ *is nilpotent and there exist integers* m *and* q *with* $1 \le m \le N$ *and* $q \ge 1$ *such that*

$$s_{n,j} = \begin{cases} 1 & \text{if } 1 \le j \le q \text{ and } n = m, \\ 0 & \text{otherwise}, \end{cases} \tag{0.6}$$

where the numbers $s_{n,j}$ *are defined by* (0.5).

In particular, an N-periodic difference equation is N-periodically irreducible in the class of N-periodic difference equations if and only if it is kinematically irreducible in this class.

Next we shall define three types of N-periodic difference equations. We shall say that the N-periodic difference equation (0.1) is of *type* I if A_n does not depend on n and is equal to a single Jordan block with a nonzero eigenvalue λ satisfying $0 \le \arg \lambda < \frac{2\pi}{N}$. Similarly, we say that (0.1) is a N-periodic difference equation of *type* I_+ if A_n does not depend on n and is equal to a single Jordan block with a nonzero eigenvalue λ satisfying $\lambda > 0$. By Theorem 0.3 the N-periodic difference equations of types I and I_+ are irreducible.

We proceed with the second type. We say that the N-periodic difference equation (0.1) is of type II if there exist nonnegative integers m, p, and r with $1 \le m \le N$ and $m < p < m + N$ such that

$$A_n = \begin{cases} \begin{pmatrix} I_r \\ 0 \end{pmatrix} : \mathbb{C}^r \to \mathbb{C}^{r+1} & \text{if } n = m, \\ I_{r+1} : \mathbb{C}^{r+1} \to \mathbb{C}^{r+1} & \text{if } n = m+1, \dots, p-1, \\ (\,0 \quad I_r\,) : \mathbb{C}^{r+1} \to \mathbb{C}^r & \text{if } n = p, \\ I_r : \mathbb{C}^r \to \mathbb{C}^r & \text{if } n = p+1, \dots, m+N-1. \end{cases} \tag{0.7}$$

Here I_r stands for the identity operator on $\mathbb{C}^r$. In particular, if (0.1) is an N-periodic difference equation of type II, then the N coefficients $A_1, A_2, \dots, A_N$ are all identity operators except for two of them. The latter two coefficients are the canonical projection on the last r coordinates and the canonical embedding on the first r coordinates. The order of these two coefficients in $A_1, A_2, \dots, A_N$ depends on whether $p > N$. Notice that for a N-periodic difference equation of type II the monodromy operator is nilpotent, and one computes that if for the N-periodic system (0.1) the coefficients are given by (0.7), then the equalities (0.6) hold true with $q = p - m + rN$. It follows from Theorem 0.3 that systems of type II are irreducible.

Finally, we say that the N-periodic difference equation (0.1) is of type III if there exist $1 \le m \le N$ and $r \ge 1$ such that

$$A_n = \begin{cases} \begin{pmatrix} 0 & I_{r-1} \\ 0 & 0 \end{pmatrix} : \mathbb{C}^r \to \mathbb{C}^r & \text{if } n = m, \\ I_r : \mathbb{C}^r \to \mathbb{C}^r & \text{if } n = m+1, \dots, m+N-1. \end{cases} \tag{0.8}$$

Again in this case the monodromy operator is nilpotent and one computes that equalities (0.6) holds true with $q = rN$. Hence, by Theorem 0.3, systems of type III are also irreducible.

THEOREM 0.4. *Systems of type* I_+, II, *or* III *are irreducible in the class of* N*-periodic difference equations under kinematic similarity. Any* N*-periodic difference equation is kinematically similar to a unique (up to reordering) direct sum of difference equations of types* I_+, II, *and* III.

THEOREM 0.5. *Systems of type* I, II, *or* III *are irreducible under N-periodic similarity. Any N-periodic difference equation is N-periodically similar to a unique (up to reordering) direct sum of difference equations of types* I, II, *and* III.

We shall also describe precisely how many times each of the irreducible difference equations I, I_+, II, and III appears in the direct sums referred to in Theorems 0.4 and 0.5. Analogues of Theorems 0.4 and 0.5 for square periodic difference equations will also be given.

The present paper consists of seven sections not counting this introduction. Section 1 presents a canonical form for N-periodic difference equations under N-periodic similarity and contains the proof of Theorem 0.2. In section 2 a similar program is carried out, but with kinematic similarity in place of N-periodic similarity. In particular, in this section Theorem 0.1 is proved. Furthermore, an open question in section 8 of [4] is answered for periodic difference equations. The third section concerns the notion of reducibility in the class of N-periodic difference equations. Here we prove Theorems 0.3–0.5. Results on reducibility by kinematic similarity in the general class of time-varying systems are also given. Reducibility in the class of square periodic systems is the main topic of section 4. In section 5 similarity of periodic difference equations with different periods is analyzed. In section 6 the results of section 1 are restated in terms of a certain class of finite block matrices. The final section presents an example illustrating how the canonical forms referred to in Theorems 0.4 and 0.5 can be constructed and how one can compute the corresponding kinematic and N-periodic similarities.

Problems that are equivalent to classification problems for periodic difference equations have been studied in different settings. The problem of classifying 2-periodic difference equations under 2-periodic similarity may be restated as the problem of classifying pairs of matrices under the following equivalence relation:

$$(A_1, A_2) \simeq (B_1, B_2) \Leftrightarrow \exists X, Y : Y^{-1}A_1X = B_1,\ X^{-1}A_2Y = B_2. \tag{0.9}$$

In this form, the problem has been fully treated by Dobrovol'skaja and Ponomarev in [1], which contains a description of a complete set of invariants, a decomposition into irreducible pairs of matrices, and canonical forms. The classification problem related with (0.9) (or, equivalently, the classification problem for 2-periodic difference equations) can also been seen as a special case of the problem of classifying quadruples of subspaces (see Gel'fand and Ponomarev [3]). In more recent papers by Lenard [8] and Horn and Merino [7] the classification problem related with (0.9) is studied independently. Chapter 2 of [7] gives a canonical form and invariants, and [8] presents the decomposition into irreducible pairs; these results are similar to those obtained in [1].

The problems of classifying N-periodic difference equations under N-periodic similarity may be seen as classification problems for certain block matrices under block diagonal similarity (see Proposition 6.1). The latter classification problems may be restated as problems of classifying representations of a special type of quivers. These problems for quiver representations are studied in Nazarova [9]. (A quiver is a directed graph, possibly with loops and multiple edges.) In terms of N-periodic difference equations, the results of [9] mean that the number of N-periodically irreducible equations is infinite and that it is possible to describe the latter equations. Our results include, additionally, explicit descriptions of a complete set of invariants, canonical forms, the irreducible equations, and geometric methods to obtain the corresponding similarities. Problems related to the classification of N-periodic difference equations under

N-periodic similarity are also studied in Gel'fand and Ponomarev [2] in connection with the representations of the Lorentz group.

Periodic difference equations have also been studied in Van Dooren and Sreedhar [10], where necessary and sufficient conditions are given so that a periodic difference equation is periodically similar to a time-invariant one.

The problem of classifying general (not necessarily periodic) systems of difference equations of the form $x_{n+1} = A_n x_n$ $(n \in \mathbb{Z})$ under kinematic similarity has been studied by the authors in [4]. That paper covers in detail the scalar case, the time invariant case (when the coefficients A_n do not depend on n), the case of finitely supported difference equations (when the number of nonzero coefficients A_n is finite), and some other examples. The present paper and [4] are complementary.

1. Invariants and a canonical form under periodic similarity. In this section a canonical form for N-periodic difference equations under N-periodic similarity will be presented. Furthermore, Theorem 0.2 will be proved. To introduce the canonical form, we need the following lemma.

LEMMA 1.1. *Let $(A_n)_{n=-\infty}^{\infty}$ be an N-periodic sequence of linear operators A_n from $\mathbb{C}^{r_n}$ to $\mathbb{C}^{r_{n+1}}$. Put $s = N \max_{1\le n\le N} r_n$. For each $1 \le n \le N$ the space $\mathbb{C}^{r_n}$ admits a direct sum decomposition*

$$\mathbb{C}^{r_n} = \mathcal{N}_{n,1} \oplus \mathcal{N}_{n,2} \oplus \cdots \oplus \mathcal{N}_{n,s+1} \tag{1.1}$$

such that relative to the decomposition (1.1) *the operator A_n has the block matrix representation*

$$A_n = \left(\begin{array}{cccc|c} 0 & A_{n,1,2} & \dots & A_{n,1,s} & 0 \\ \vdots & 0 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & A_{n,s-1,s} & 0 \\ 0 & 0 & \dots & 0 & 0 \\ \hline 0 & 0 & \dots & 0 & A_{n,s+1} \end{array}\right) : \oplus_{j=1}^{s+1}\mathcal{N}_{n,j} \to \oplus_{j=1}^{s+1}\mathcal{N}_{n+1,j}, \tag{1.2}$$

where $A_{n,j,j+1}$ is injective, $j = 1, 2, \dots, s-1$, $A_{n,s+1}$ is invertible, and $\mathcal{N}_{N+p,j} = \mathcal{N}_{p,j}$, for all $p \ge 1$ and $j = 1, 2, \dots, s+1$.

Proof. The proof is divided into three steps.

Step 1. In the first step of the proof, the subspaces $\mathcal{N}_{n,j}$ will be defined. Put

$$\mathcal{M}_{n,j} = \operatorname{Ker}(A_{n+j-1} \cdots A_{n+1} A_n),\ j \ge 0,\ n \in \mathbb{Z}. \tag{1.3}$$

By definition, $\mathcal{M}_{n,0} = \{0\}$. Fix $1 \le n \le N$. Notice that $\mathcal{M}_{n,j-1} \subset \mathcal{M}_{n,j}$, $j \ge 1$. For $1 \le j \le s$, let $\mathcal{N}_{n,j}$ be a complement of $\mathcal{M}_{n,j-1}$ into $\mathcal{M}_{n,j}$, i.e.,

$$\mathcal{M}_{n,j-1} \oplus \mathcal{N}_{n,j} = \mathcal{M}_{n,j}. \tag{1.4}$$

To define $\mathcal{N}_{n,s+1}$ for $1 \le n \le N$, let $\mathcal{N}_{1,s+1}$ be the maximum invariant subspace of the monodromy operator $A_N \cdots A_1$ on which $A_N \cdots A_1$ is invertible. Put $\mathcal{N}_{n,s+1} = A_{n-1} \cdots A_1 \mathcal{N}_{1,s+1},\ n = 2, 3, \dots, N$.

Step 2. In this step of the proof it will be shown that

$$\mathcal{M}_{n,s} \oplus \mathcal{N}_{n,s+1} = \mathbb{C}^{r_n},\ n = 1, 2, \dots, N. \tag{1.5}$$

To see (1.5), fix $1 \le n \le N$. Notice that $\mathcal{M}_{n,N(r_n+j)} = \mathcal{M}_{n,N(r_n+j+1)}$, $j = 0, 1, \dots$. In particular, $\mathcal{M}_{n,s} = \mathcal{M}_{n,Nr_n} = \operatorname{Ker}(A_{n+N-1} \cdots A_n)^{r_n}$. Thus it suffices to show that

$\mathcal{N}_{n,s+1}$ is the maximum $(A_{n+N-1}\cdots A_n)$-invariant subspace on which $A_{n+N-1}\cdots A_n$ is invertible. Indeed, if this property has been proved, then

$$\mathcal{N}_{n,s+1} = \operatorname{Im}(A_{n+N-1}\cdots A_n)^{r_n}$$

and (1.5) follows trivially. Notice that the subspace $\mathcal{N}_{n,s+1}$ is $(A_{n+N-1}\cdots A_n)$-invariant, since

$$A_{n+N-1}\cdots A_n\mathcal{N}_{n,s+1} = A_{n-1}\cdots A_1A_N\cdots A_1\mathcal{N}_{1,s+1} = A_{n-1}\cdots A_1\mathcal{N}_{1,s+1} = \mathcal{N}_{n,s+1}.$$

Assume that $\mathcal{N}$ is an $(A_{n+N-1}\cdots A_n)$-invariant subspace on which $A_{n+N-1}\cdots A_n$ is invertible. It follows that $A_N\cdots A_n\mathcal{N}$ is an $(A_N\cdots A_1)$-invariant subspace and hence $\dim A_N\cdots A_n\mathcal{N} \le \dim \mathcal{N}_{1,s+1}$. The left-hand side of this equality is equal to $\dim \mathcal{N}$; the right-hand side is equal to $\dim \mathcal{N}_{n,s+1}$. So $\dim \mathcal{N} \le \dim \mathcal{N}_{n,s+1}$, and hence $\mathcal{N}_{n,s+1}$ is the maximum $(A_{n+N-1}\cdots A_n)$-invariant subspace on which $A_{n+N-1}\cdots A_n$ is invertible.

Step 3. In this step the proof will be finished. Since $\mathcal{M}_{n,0} = \{0\}$, it follows from (1.4) that

$$\mathcal{N}_{n,1} \oplus \mathcal{N}_{n,2} \oplus \cdots \oplus \mathcal{N}_{n,j} = \mathcal{M}_{n,j},\ 1 \le j \le s. \tag{1.6}$$

In particular, (1.5) and $j = s$ in (1.6) give the partitioning (1.1). Next set $\mathcal{N}_{N+1,j} = \mathcal{N}_{1,j}$, $j = 1, 2, \ldots, s$. It can be easily seen that the following identities are satisfied:

$$A_n\mathcal{N}_{n,s+1} = \mathcal{N}_{n+1,s+1},\ n = 1, 2, \ldots, N-1. \tag{1.7}$$

Furthermore, from (1.6) and $A_n\mathcal{N}_{n,j} \subset A_n\mathcal{M}_{n,j} \subset \mathcal{M}_{n+1,j-1}$ it follows that

$$A_n\mathcal{N}_{n,1} = \{0\}, \quad A_n\mathcal{N}_{n,j} \subset \mathcal{N}_{n+1,1} \oplus \cdots \oplus \mathcal{N}_{n+1,j-1},\ j = 1, 2, \ldots, s,\ n = 1, 2, \ldots, N. \tag{1.8}$$

Finally, the maps

$$\pi_{n+1,j-1}A_n \mid \mathcal{N}_{n,j} : \mathcal{N}_{n,j} \to \mathcal{N}_{n+1,j-1},\ j = 1, 2, \ldots, s,\ n = 1, 2, \ldots, N, \tag{1.9}$$

are injective, where $\pi_{n+1,j-1} : \mathbb{C}^{r_{n+1}} \to \mathcal{N}_{n+1,j-1}$ is the canonical projection with respect to the decomposition (1.1). To see this, assume that $x \in \mathcal{N}_{n,j}$ with $\pi_{n+1,j-1}A_nx = 0$. It follows from (1.8) that $A_nx \in \mathcal{M}_{n+1,j-2}$, and hence $x \in \mathcal{M}_{n,j-1}$. Thus $x = 0$ and we conclude that the map (1.9) is injective.

The partitioning (1.2) now follows from (1.7), (1.8), the injectivity of the map (1.9), and the invertibility of the monodromy operator $A_N\cdots A_1 \mid \mathcal{N}_{1,s+1} : \mathcal{N}_{1,s+1} \to \mathcal{N}_{1,s+1}$. ∎

For Lemma 1.1 to hold, it is not necessary to take $s = N\max_{1\le n\le N} r_n$; other choices with smaller s are also good. In fact, from the proof of Lemma 1.1 it can be seen that Lemma 1.1 holds for any $s \ge s'$, where s' is the smallest integer such that $\mathcal{M}_{n,j-1} = \mathcal{M}_{n,j}$ for $j \ge s'$ and $n = 1, 2, \ldots, N$. Here $\mathcal{M}_{n,j}$ is as in (1.3).

From the partitioning induced by the lines in the block matrix in (1.2), the following theorem is immediately obtained.

THEOREM 1.2. *Each N-periodic system is N-periodically similar to the direct sum of an N-periodic system with invertible coefficients and an N-periodic system with a nilpotent monodromy operator.*

By Theorem 1.2, to find canonical forms for N-periodic systems of difference equations it suffices to consider two special cases, namely, the case of N-periodic

systems with invertible coefficients and the case of N-periodic systems with nilpotent monodromy operators. The following two theorems present canonical forms for these two cases.

THEOREM 1.3. *An N-periodic system with invertible coefficients is N-periodically similar to a unique (up to reordering) direct sum of N-periodic difference equations of type* I. *This direct sum is the time-invariant system*

$$x_{n+1} = Jx_n,\ n \in \mathbb{Z}, \tag{1.10}$$

where J is the Jordan matrix which one obtains if in the Jordan normal form of the monodromy matrix the diagonal entries are replaced by their Nth roots in the sector $\{z \in \mathbb{C} | 0 \le \arg z < 2\pi/N\}$.

For the second theorem we need to introduce the following notation. For nonnegative integers m, p, and r with $1 \le m \le N$, $m < p < m + N$, and $r \ge 0$, we denote by $\Sigma_{\mathrm{II}}(m, p, r)$ the N-periodic difference equation (0.1) with the coefficients A_n being given by (0.7). Similarly, for nonnegative integers m and r with $1 \le m \le N$ and $r \ge 0$, we denote by $\Sigma_{\mathrm{III}}(m, r)$ the N-periodic difference equation (0.1) with the coefficients A_n being given by (0.8). With the N-periodic difference equation (0.1) we associate the numbers $\eta_{j,k}((A_n))$ given by

$$\eta_{j,k}((A_n)) = \begin{cases} \operatorname{rank}(A_{j+1}A_j) - \operatorname{rank} A_j - (\operatorname{rank} A_{j+1} - r_{j+1}), & j = k, \\ \operatorname{rank}(A_{j+1} \cdots A_k) - \operatorname{rank}(A_j \cdots A_k) & \\ \quad - (\operatorname{rank}(A_{j+1} \cdots A_{k+1}) - \operatorname{rank}(A_j \cdots A_{k+1})), & j \ge k+1. \end{cases} \tag{1.11}$$

THEOREM 1.4. *An N-periodic system $x_{n+1} = A_n x_n$, $n \in \mathbb{Z}$, with a nilpotent monodromy operator is N-periodically similar to a unique (up to reordering) direct sum of N-periodic systems of the forms $\Sigma_{\mathrm{II}}(m, p, r)$ and $\Sigma_{\mathrm{III}}(m, r)$. The number of times the system $\Sigma_{\mathrm{II}}(m, p, r)$ appears in the direct sum is equal to $\eta_{p-1+rN,m}((A_n))$; the number of times the system $\Sigma_{\mathrm{III}}(m, r)$ appears is equal to $\eta_{m-1+rN,m}((A_n))$, where the numbers $\eta_{j,k}((A_n))$ are given by* (1.11).

Theorem 1.4 will be proved in section 3. In the remaining part of this section, we prove Theorems 1.3 and 0.2.

Proof of Theorem 1.3. We split the proof into two parts. The first part concerns the uniqueness statement.

Part 1. Let $\tilde{J}$ be a Jordan matrix with nonzero eigenvalues λ in the sector $0 \le \arg \lambda < \frac{2\pi}{N}$, and assume that the system $y_{n+1} = \tilde{J} y_n$, $n \in \mathbb{Z}$, is N-periodically similar to the system (1.10). We have to prove that $\tilde{J}$ is equal to J up to a reordering of the blocks. By N-periodic similarity, $\tilde{J} = E_{n+1}^{-1} J E_n$, $n \in \mathbb{Z}$, where E_n is invertible and $E_n = E_{n+N}$ for each n. It follows that

$$\tilde{J}^N = E_{N+1}^{-1} J^N E_1 = E_1^{-1} J^N E_1.$$

Hence $\tilde{J}^N$ is equal to J^N up to a reordering of the blocks. Notice that the Jordan normal form of the Nth power of a Jordan matrix with nonzero eigenvalues is obtained by replacing each diagonal entry of the Jordan matrix by its Nth power (which can be proved by reduction to the case of a single Jordan block with a nonzero eigenvalue). Since the eigenvalues λ of $\tilde{J}$ and J are in the sector $0 \le \arg \lambda < \frac{2\pi}{N}$, we conclude that $\tilde{J}$ is equal to J up to a reordering of the blocks.

Part 2. In this part we show that the N-periodic equation (0.1) with A_n invertible for each n is N-periodically similar to the time-invariant difference equation (1.10). Let E_1 be an invertible matrix such that $E_1^{-1} A_N \cdots A_1 E_1$ is a Jordan matrix $J_{A_N \cdots A_1}$

of $A_N \cdots A_1$. Define E_n to be the invertible matrix $E_n = A_{n-1} \cdots A_1 E_1$ for $1 \leq n \leq N$. Extend the sequence $(E_n)_{n=1}^N$ to a N-periodic sequence $(E_n)_{n=-\infty}^\infty$. The N-periodic similarity $(E_n)_{n=-\infty}^\infty$ transforms $x_{n+1} = A_n x_n$, $n \in \mathbb{Z}$, into the N-periodic difference equation

$$\hat{x}_{n+1} = \hat{A}_n \hat{x}_n,\ n \in \mathbb{Z}, \tag{1.12}$$

where $\hat{A}_n = I$ for $1 \leq n \leq N-1$ and $\hat{A}_N = J_{A_N \cdots A_1}$.

Finally, we define an N-periodic similarity $(F_n)_{n=-\infty}^\infty$ which transforms (1.12) into (1.10). Notice that $\hat{A}_N$ is similar to J^N. Let F_1 be an invertible matrix satisfying $F_1^{-1} \hat{A}_N F_1 = J^N$, and put

$$F_n = F_1 J^{-n+1},\ n = 1, 2, \ldots, N.$$

Recall that $\hat{A}_n = I$ for $1 \leq n \leq N-1$. So we obtain the identities

$$F_{n+1}^{-1} \hat{A}_n F_n = J,\ n = 1, 2, \ldots, N,$$

where $F_{N+1} = F_1$. Extend the sequence $(F_n)_{n=1}^N$ to an N-periodic sequence $(F_n)_{n=-\infty}^\infty$. We conclude that the difference equations (1.12) and (1.10) are N-periodically similar, and hence (0.1) and (1.10) are N-periodically similar. □

Proof of Theorem 0.2. We split the proof into two parts. In the first part we prove the necessity of the conditions.

Part 1. Assume that $B_n = E_{n+1}^{-1} A_n E_n$, $n \in \mathbb{Z}$, where $(E_n)_{n=-\infty}^\infty$ is an N-periodic similarity. Iterating these equations, we obtain that

$$E_{n+j}^{-1} A_{n+j-1} \cdots A_n E_n = B_{n+j-1} \cdots B_n, \quad j \geq 1,\ 1 \leq n \leq N. \tag{1.13}$$

In particular, taking ranks in (1.13), we get the equalities (0.4). Substitution of $j = N$ and $n = 1$ in (1.13) yields the identity $E_{N+1}^{-1} A_N \cdots A_1 E_1 = B_N \cdots B_1$. But $E_{N+1} = E_1$, since $(E_n)_{n=-\infty}^\infty$ is N-periodic. Hence $A_N \cdots A_1$ is similar to $B_N \cdots B_1$.

Part 2. Assume that the monodromy matrices of (0.1) and (0.2) are similar and that the rank conditions (0.4) are satisfied. We have to show that the systems (0.1) and (0.2) are N-periodically similar. By Lemma 1.1 and Theorem 1.3 we may assume, without loss of generality, that the monodromy matrices are nilpotent. It suffices to show that the N-periodic difference equation (0.1) with nilpotent monodromy matrix is N-periodically similar to a N-periodic difference equation with coefficients of the form (1.2) with, for each $n = 1, 2, \ldots, N$, the additional properties

$$\mathcal{N}_{n,s+1} = \{0\},\ A_{n,j,k} = 0,\ j < k-1, \tag{1.14}$$

$$A_{n,j,j+1} = \begin{pmatrix} I \\ 0 \end{pmatrix} : \mathcal{N}_{n,j+1} \to \mathcal{N}_{n,j+1} \oplus \mathcal{K}_{n,j+1} = \mathcal{N}_{n+1,j},\ j = 1, 2, \ldots, s-1 \tag{1.15}$$

for some subspaces $\mathcal{K}_{n,j}$, $j = 1, 2, \ldots, s-1$. Indeed, the form (1.2) with the properties (1.14), (1.15) for $n = 1, 2, \ldots, N$ is completely determined by the ranks in the left-hand side of (0.4). Similarly, the N-periodic difference equation (0.2) with nilpotent monodromy matrix is N-periodically similar to a N-periodic difference equation which is completely determined by the ranks in the right-hand side of (0.4). Thus (0.1) and (0.2) are N-periodically similar.

To show that (0.1) with nilpotent monodromy matrix is N-periodically similar to a N-periodic difference equation with coefficients of the form (1.2) with, for each

$n = 1, 2, \ldots, N$, the properties (1.14), (1.15), we construct complements $\mathcal{N}_{n,j}$ as in (1.4) in a special way. The construction is as follows.

Fix $1 \le n \le N$. Let $\mathcal{N}_{n,s}$ be a complement of $\mathcal{M}_{n,s-1}$ into $\mathcal{M}_{n,s}$, i.e., $\mathcal{M}_{n,s-1} \oplus \mathcal{N}_{n,s} = \mathcal{M}_{n,s}$. We proceed recursively. Assume that $\mathcal{N}_{m,j+1}$ is a complement of $\mathcal{M}_{m,j}$ into $\mathcal{M}_{m,j+1}$ for some nonnegative integers m and j with the properties $1 \le m \le N$, $1 \le j \le s-1$, $m + j + 1 = n + s$ modulo N. Notice that $\mathcal{N}_{m,j+1}$ has the following properties:

$$A_m \mathcal{N}_{m,j+1} \subset A_m \mathcal{M}_{m,j+1} \subset \mathcal{M}_{m+1,j}, \tag{1.16}$$

$$A_m \mathcal{N}_{m,j+1} \cap \mathcal{M}_{m+1,j-1} \subset A_m[\mathcal{M}_{m,j} \cap \mathcal{N}_{m,j+1}] = \{0\}. \tag{1.17}$$

Hence there exists a subspace $\mathcal{K}_{m+1,j}$ such that

$$\mathcal{M}_{m+1,j-1} \oplus A_m \mathcal{N}_{m,j+1} \oplus \mathcal{K}_{m+1,j} = \mathcal{M}_{m+1,j}.$$

Now we define $\mathcal{N}_{m+1,j}$ to be equal to $A_m \mathcal{N}_{m,j+1} \oplus \mathcal{K}_{m+1,j}$. If in this process $m = N$, and hence $\mathcal{N}_{m+1,j} = \mathcal{N}_{N+1,j}$, we put $\mathcal{N}_{1,j} = \mathcal{N}_{N+1,j}$ and continue the recursive procedure with $\mathcal{N}_{1,j}$.

Carry out the above recursive procedure for $1 \le n \le N$. We obtain subspaces $\mathcal{N}_{n,j}$ such that

$$\mathcal{M}_{n,j-1} \oplus \mathcal{N}_{n,j} = \mathcal{M}_{n,j},\ j = 1, 2, \ldots, s,\ n = 1, 2, \ldots, N. \tag{1.18}$$

Thus the subspaces $\mathcal{N}_{n,j}$ decompose $\mathbb{C}^{r_n}$ as in (1.1) with $\mathcal{N}_{n,s+1} = \{0\}$, $n = 1, 2, \ldots, N$. Furthermore, for each $1 \le n \le N$, the subspaces $\mathcal{N}_{n,j}$ have the properties

$$A_n \mathcal{N}_{n,j+1} \subset \mathcal{N}_{n+1,j},\ j = 1, 2, \ldots, s-1, \tag{1.19}$$

$$\pi_{n+1,j} A_n | \mathcal{N}_{n,j+1} = \begin{pmatrix} I \\ 0 \end{pmatrix} : \mathcal{N}_{n,j+1} \to A_n \mathcal{N}_{n,j+1} \oplus \mathcal{K}_{n+1,j} = \mathcal{N}_{n+1,j},\ j = 1, 2, \ldots, s-1, \tag{1.20}$$

where $\pi_{n+1,j}$ is the projection of $\mathbb{C}^{r_{n+1}}$ onto $\mathcal{N}_{n+1,j}$ according to the decomposition (1.1), $\mathcal{N}_{N+1,j} = \mathcal{N}_{1,j}$, $j = 1, 2, \ldots, s$, and $\pi_{N+1,j} = \pi_{1,j}$, $j = 1, 2, \ldots, s$. By Lemma 1.1 and (1.18), (1.19), (1.20) the N-periodic difference equation (0.1) with nilpotent monodromy matrix is N-periodically similar to the N-periodic difference equation with coefficients given by (1.2) and (1.14), (1.15). ∎

2. Invariants and a canonical form under kinematic similarity. In this section a canonical form for N-periodic difference equations under kinematic similarity will be presented. Furthermore, Theorem 0.1 will be proven. Finally, we shall answer an open question in [4] for periodic systems. First, we state the analogues of Theorems 1.2–1.4 for N-periodic similarity replaced by kinematic similarity.

THEOREM 2.1. *Each N-periodic system is kinematically similar to the direct sum of an N-periodic system with invertible coefficients and an N-periodic system with a nilpotent monodromy operator.*

THEOREM 2.2. *An N-periodic system with invertible coefficients is kinematically similar to a unique (up to reordering) direct sum of N-periodic difference equations of type* I_+. *This direct sum is the time-invariant system*

$$x_{n+1} = Jx_n,\ n \in \mathbb{Z}, \tag{2.1}$$

where J is the Jordan matrix which one obtains if in the Jordan normal form of the monodromy matrix the diagonal entries are replaced by the positive Nth roots of their absolute values.

THEOREM 2.3. *An N-periodic system $x_{n+1} = A_n x_n$, $n \in \mathbb{Z}$, with a nilpotent monodromy operator is kinematically similar to a unique (up to reordering) direct sum of N-periodic systems of the forms $\Sigma_{\mathrm{II}}(m,p,r)$ and $\Sigma_{\mathrm{III}}(m,r)$. The number of times the system $\Sigma_{\mathrm{II}}(m,p,r)$ appears in the direct sum is equal to $\eta_{p+rN,m}((A_n))$; the number of times the system $\Sigma_{\mathrm{III}}(m,r)$ appears is equal to $\eta_{m+rN,m}((A_n))$, where the numbers $\eta_{j,k}((A_n))$ are given by* (1.11).

From Theorems 1.4 and 2.3 the following corollary is immediate.

COROLLARY 2.4. *Two N-periodic difference equations with nilpotent monodromy operators are kinematically similar if and only if they are N-periodically similar.*

Theorem 2.3 will be proved in section 3. Since an N-periodic similarity is a kinematic similarity, Theorem 2.1 is an immediate consequence of Theorem 1.2.

Proof of Theorem 2.2. To prove Theorem 2.2 we shall use Theorem 4.1 in [4]. Indeed, according to Theorem 1.3 an N-periodic system is kinematically similar to a time-invariant system $x_{n+1} = \tilde{J}x_n$, $n \in \mathbb{Z}$, where $\tilde{J}$ is the Jordan matrix which one obtains if in the Jordan normal form of the monodromy matrix the diagonal entries are replaced by their Nth roots in the sector $\{z \in \mathbb{C} \mid 0 \le \arg z < \frac{2\pi}{N}\}$. By Theorem 4.1 in [4] the system $x_{n+1} = \tilde{J}x_n$, $n \in \mathbb{Z}$, is kinematically similar to the time-invariant system $x_{n+1} = Jx_n$, $n \in \mathbb{Z}$, where J is obtained from $\tilde{J}$ by replacing the diagonal entries in $\tilde{J}$ by their absolute values. This yields Theorem 2.2, except for the statement about uniqueness. The latter, however, also follows directly from Theorem 4.1 in [4]. □

Proof of Theorem 0.1. We split the proof into two parts.

Part 1. In this part we prove the necessity of the condition in Theorem 0.1. Assume that $B_n = E_{n+1}^{-1} A_n E_n$, $n \in \mathbb{Z}$, where $(E_n)_{n=-\infty}^{\infty}$ is a kinematic similarity. Iterating the last equations, we obtain that

$$E_{n+j}^{-1} A_{n+j-1} \cdots A_n E_n = B_{n+j-1} \cdots B_n,\ j \ge 1,\ 1 \le n \le N. \tag{2.2}$$

In particular, taking ranks in (2.2), we get the equalities (0.4). Substitution of $n = Nm + 1$ and $j = N$ into (2.2) yields the identities

$$E_{N(m+1)+1}^{-1} A_N \cdots A_1 E_{Nm+1} = B_N \cdots B_1,\ m \in \mathbb{Z}. \tag{2.3}$$

Thus the time-invariant difference equations $x_{n+1} = A_N \cdots x_n$, $n \in \mathbb{Z}$, and $y_{n+1} = B_N \cdots B_1 y_n$, $n \in \mathbb{Z}$, are kinematically similar. From Theorem 4.1 in [4] it follows that the Jordan matrices $[J_{A_N \cdots A_1}]$ and $[J_{B_N \cdots B_1}]$ are similar.

Part 2. Assume that the rank conditions (0.4) are satisfied and that the Jordan matrices $[J_{A_N \cdots A_1}]$ and $[J_{B_N \cdots B_1}]$ are similar. We shall show that the systems (0.1) and (0.2) are kinematically similar. By Theorem 2.1 the system (0.1) is kinematically similar to the direct sum of the two N-periodic difference equations

$$x_{n+1} = A_n^{(1)} x_n,\ n \in \mathbb{Z}, \tag{2.4}$$

$$x_{n+1} = A_n^{(2)} x_n,\ n \in \mathbb{Z}, \tag{2.5}$$

with $A_n^{(1)}$ being invertible for each n and the monodromy matrix $A_N^{(2)} \cdots A_1^{(2)}$ being nilpotent. Similarly, (0.2) is kinematically similar to the direct sum of the two N-periodic difference equations

$$y_{n+1} = B_n^{(1)} y_n,\ n \in \mathbb{Z}, \tag{2.6}$$

$$y_{n+1} = B_n^{(2)} y_n,\ n \in \mathbb{Z}, \tag{2.7}$$

with $B_n^{(1)}$ being invertible for each n and the monodromy matrix $B_N^{(2)} \cdots B_1^{(2)}$ being nilpotent. Notice that the Jordan matrices $[J_{A_N^{(1)} \cdots A_1^{(1)}}]$ and $[J_{B_N^{(1)} \cdots B_1^{(1)}}]$ are similar. Thus it follows from Theorem 2.2 that the systems (2.4) and (2.6) are kinematically similar. In addition notice that the matrices $A_n^{(2)}$ and $B_n^{(2)}$ are of the same sizes for each n and that for each $n \in \{1, 2, \ldots, N\}$

$$\operatorname{rank}(A_{j+n-1}^{(2)} A_{j+n-2}^{(2)} \cdots A_n^{(2)}) = \operatorname{rank}(B_{j+n-1}^{(2)} B_{j+n-2}^{(2)} \cdots B_n^{(2)}),\ 1 \le j \le N(\max_{1 \le n \le N} r_n).$$

From Theorem 0.2 it follows that the systems (2.5) and (2.7) are N-periodically similar and, consequently, kinematically similar. We conclude that the direct sum of (2.4) and (2.5) is kinematically similar to the direct sum of (2.6) and (2.7). Hence the systems (0.1) and (0.2) are kinematically similar. □

Theorem 0.1 allows us to answer an open question in section 8 of [4] for periodic difference equations. To state the result, we recall that two families $(a_n)_{n \in \tau}$ and $(b_n)_{n \in \tau}$ of nonnegative real numbers are called *strongly equivalent* if there exists $M \ge 1$ such that $b_n/M \le a_n \le M b_n$, $n \in \tau$. Furthermore, we denote by $s_k(A)$ the kth singular value of the matrix A.

THEOREM 2.5. *Let $(A_n)_{n=-\infty}^{\infty}$ and $(B_n)_{n=-\infty}^{\infty}$ be N-periodic sequences of linear operators $A_n, B_n : \mathbb{C}^{r_n} \to \mathbb{C}^{r_{n+1}}$. Assume that the families*

$$(s_k(A_j A_{j-1} \cdots A_i))_{i,j \in \mathbb{Z},\, i \le j}, \quad (s_k(B_j B_{j-1} \cdots B_i))_{i,j \in \mathbb{Z},\, i \le j} \tag{2.8}$$

are strongly equivalent for each k. Then the N-periodic difference equations $x_{n+1} = A_n x_n$, $n \in \mathbb{Z}$, and $y_{n+1} = B_n y_n$, $n \in \mathbb{Z}$, are kinematically similar.

Proof. Assume that the families (2.8) are strongly equivalent. In particular, by N-periodicity, the families

$$(s_k(A_N \cdots A_1)^j)_{j=0}^{\infty}, \quad (s_k(B_N \cdots B_1)^j)_{j=0}^{\infty}$$

are strongly equivalent for each k. According to Theorem 4.1 in [4], the Jordan matrices $[J_{A_N \cdots A_1}]$ and $[J_{B_N \cdots B_1}]$ are similar. The strong equivalence of the families (2.8) also yields that

$$s_k(A_j \cdots A_i) = 0 \Leftrightarrow s_k(B_j \cdots B_i) = 0$$

for all $i \le j$ and all k. Hence $\operatorname{rank}(A_j \cdots A_i) = \operatorname{rank}(B_j \cdots B_i)$ for all $i \le j$. From Theorem 0.1 we conclude that the systems $x_{n+1} = A_n x_n$, $n \in \mathbb{Z}$, and $y_{n+1} = B_n y_n$, $n \in \mathbb{Z}$, are kinematically similar. □

The converse of Theorem 2.5 is also true; see section 8 of [4].

3. Reducibility. In this section we shall prove Theorems 0.3, 0.4, 0.5, 1.4, and 2.3, which are all related to reducibility in the class of N-periodic systems. In the final paragraph reducibility by kinematic similarity in the general class of time-varying systems is analyzed.

The key lemma in the proofs follows.

LEMMA 3.1. *Let $x_{n+1} = A_n x_n$, $n \in \mathbb{Z}$, be an N-periodic system with a nilpotent monodromy matrix $A_N A_{N-1} \cdots A_1$. Let the numbers $s_{n,j}$ be given by (0.5). The system $x_{n+1} = A_n x_n$, $n \in \mathbb{Z}$, is reducible in the class of N-periodic systems by N-periodic (or kinematic) similarity if and only if there exists an integer m, $1 \le m \le N$, such that the numbers $s_{m,j}$ may be decomposed as*

$$s_{m,j} = s_{m,j}^{(1)} + s_{m,j}^{(2)},\ j \ge 1, \tag{3.1}$$

where for $i = 1, 2$ the nonnegative integer $s_{m,1}^{(i)}$ is nonzero and where the nonnegative integers $s_{m,j}^{(i)}$ are ordered in a decreasing way:

$$s_{m,j}^{(i)} \geq s_{m,j+1}^{(i)},\ j \geq 1. \tag{3.2}$$

We introduce the following terminology. An array $(s_{n,j})_{n=1,j=1}^{N,\ \infty}$ will be called *admissible* if, for $n = 1, 2, \ldots, N$,

$$s_{n,j} \geq 0,\ j \geq 1, \tag{3.3}$$

$$s_{n,j} = 0,\ j \geq J \text{ for some nonnegative integer } J, \tag{3.4}$$

$$s_{n,j} \geq s_{n,j+1},\ j \geq 1. \tag{3.5}$$

Notice that (3.3) follows from (3.4) and (3.5).

LEMMA 3.2. *For an N-periodic system of the form* (0.1) *the array $(s_{n,j})_{n=1,j=1}^{N\ \infty}$ defined by* (0.5) *is admissible.*

Proof. The properties (3.3) and (3.4) can be easily shown for all n. The number J in (3.4) can be taken equal to $N \max_{1 \leq n \leq N} r_n$. To see that (3.5) is satisfied, we use Lemma 1.1 and its proof. The operators $A_{n,j,j+1}$ in Lemma 1.1 are injective and hence $\dim \mathcal{N}_{n,j} \geq \dim \mathcal{N}_{n+1,j-1}$ for all j and n. From the direct sum decomposition (1.4) we obtain the identity $\dim \mathcal{M}_{n,j} - \dim \mathcal{M}_{n,j-1} \geq \dim \mathcal{M}_{n+1,j-1} - \dim \mathcal{M}_{n+1,j-2}$. This identity may be rewritten as

$$\dim \mathcal{M}_{n,j} - \dim \mathcal{M}_{n+1,j-1} \geq \dim \mathcal{M}_{n,j-1} - \dim \mathcal{M}_{n+1,j-2},\ j \geq 2. \tag{3.6}$$

The left-hand side of (3.6) is equal to $r_n - r_{n+1} - s_{n,j}$; the right-hand side of (3.6) is equal to $r_n - r_{n+1} - s_{n,j-1}$. We conclude that $s_{n,j-1} \geq s_{n,j}$ for all n and $j \geq 2$. □

An admissible array will be called *reducible* if it can be written as the sum of two nonzero admissible arrays. Here the sum of two arrays is the componentwise sum, and an array is called *nonzero* if at least one of its entries is nonzero. Notice that an admissible array $(s_{n,j})_{n=1,j=1}^{N,\ \infty}$ is reducible if and only if there exists an integer m, $1 \leq m \leq N$, for which the numbers $s_{m,j}$ may be decomposed as in (3.1) such that for $i = 1, 2$ the nonnegative integer $s_{m,1}^{(i)}$ is nonzero, and the nonnegative integers $s_{m,j}^{(i)}$ are ordered in a decreasing way as in (3.2). Thus Lemma 3.1 may be restated: "The N-periodic difference equation (0.1) with nilpotent monodromy matrix is reducible in the class of N-periodic systems by N-periodic (or kinematic) similarity if and only if the admissible array $(s_{n,j})_{n=1,j=1}^{N\ \infty}$ is reducible, where the numbers $s_{n,j}$ are defined by (0.5)." It can be easily seen that an admissible array $(s_{n,j})_{n=1,j=1}^{N,\ \infty}$ is irreducible if and only if there exist nonnegative integers m and q, $1 \leq m \leq N$ and $q \geq 1$, such that

$$s_{n,j} = \begin{cases} 1 & \text{if } n = m, 1 \leq j \leq q, \\ 0 & \text{otherwise.} \end{cases}$$

Proof of Lemma 3.1. The proof consists of two parts.

Part 1. Assume that the N-periodic difference equation (0.1) with nilpotent monodromy matrix is kinematically similar to the nontrivial direct sum of the N-periodic difference equations

$$x_{n+1} = A_n^{(1)} x_n,\ n \in \mathbb{Z}, \quad (A_n^{(1)} : \mathbb{C}^{r_{n,1}} \to \mathbb{C}^{r_{n+1,1}}), \tag{3.7}$$

$$x_{n+1} = A_n^{(2)} x_n,\ n \in \mathbb{Z}, \quad (A_n^{(2)} : \mathbb{C}^{r_{n,2}} \to \mathbb{C}^{r_{n+1,2}}). \tag{3.8}$$

In particular, the monodromy matrices of the N-periodic difference equations (3.7), (3.8) are nilpotent. For $i = 1, 2$ let $s_{n,j}^{(i)}$ be given by the right-hand side of (0.5) with A_m replaced by $A_m^{(i)}$ for each $m \in \mathbb{Z}$. It is evident that $s_{n,j} = s_{n,j}^{(1)} + s_{n,j}^{(2)}$ for all n and j. Moreover, $s_{n,j}^{(i)} \geq s_{n,j+1}^{(i)}$ for all n and j by Lemma 3.2. Thus it suffices to show that the arrays $(s_{n,j}^{(i)})_{n=1,j=1}^{N\ \infty}, i = 1, 2$, are nonzero. Suppose that $s_{n,j}^{(1)} = 0$ for all j and n. Thus rank $(A_{n+j-1}^{(1)} \cdots A_n^{(1)})$ does not depend on $j \geq 0$. In particular $r_{n,1} = \text{rank}\,(A_{n+N-1}^{(1)} \cdots A_n^{(1)})^{r_{n,1}}$ for all n. The monodromy matrix of (3.7) is nilpotent and hence the latter ranks are zero for all n. But then $r_{n,1} = 0$ for all n and the difference equation (3.7) is trivial. We conclude that the array $(s_{n,j}^{(1)})_{n=1,j=1}^{N\ \infty}$ is nonzero. In a similar way it can be shown that the array $(s_{n,j}^{(2)})_{n=1,j=1}^{N\ \infty}$ is nonzero.

Part 2. Suppose that the admissible array $(s_{n,j})_{n=1,j=1}^{N\ \infty}$ is reducible. Decompose it as a sum of nonzero irreducible admissible arrays

$$s_{n,j} = s_{n,j}^{(1)} + \cdots + s_{n,j}^{(M)}. \tag{3.9}$$

So for each $i = 1, 2, \ldots, M$ there exists nonnegative integers m_i and q_i, $1 \leq m_i \leq N$, $q_i \geq 1$, such that

$$s_{n,j}^{(i)} = \begin{cases} 1 & \text{if } n = m_i, 1 \leq j \leq q_i, \\ 0 & \text{otherwise.} \end{cases} \tag{3.10}$$

For $i = 1, 2, \ldots, M$, write $m_i + q_i$ as $m_i + q_i = p_i + r_i N$ for nonnegative integers p_i and r_i with $m_i \leq p_i < m_i + N$. Notice that $m_i = p_i$ if and only if q_i is divisible by N. Let $x_{n+1} = A_n^{(0)} x_n$, $n \in \mathbb{Z}$, be the direct sum of the N-periodic difference equations $\Sigma_{\text{II}}(m_i, p_i, r_i) : x_{n+1} = A_n^{(i)} x_n$, $n \in \mathbb{Z}$, where i runs over all integers $1, 2, \ldots, M$ for which q_i not divisible by N and of the N-periodic difference equations $\Sigma_{\text{III}}(m_i, r_i) : x_{n+1} = A_n^{(i)} x_n$, $n \in \mathbb{Z}$, where i runs over all integers $1, 2, \ldots, M$ for which q_i divisible by N. Put for $i = 0, 1, \ldots, M$ and $n = 1, 2, \ldots, N$

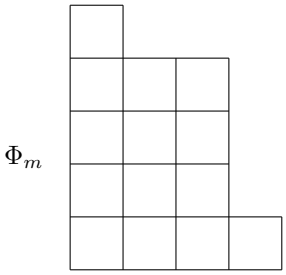
$$t_{n,j}^{(i)} = \begin{cases} \text{rank}\,(A_{n+j-1}^{(i)} \cdots A_{n+1}^{(i)}) - \text{rank}\,(A_{n+j-1}^{(i)} \cdots A_{n+1}^{(i)} A_n^{(i)}), & j \geq 2, \\ r_{n+1}^{(i)} - \text{rank}\, A_n^{(i)}, & j = 1, \end{cases}$$

where $r_{n+1}^{(i)}$ is the dimension of the space on which $A_{n+1}^{(i)}$ acts. Notice that $t_{n,j}^{(0)} = t_{n,j}^{(1)} + \cdots + t_{n,j}^{(M)}$. Furthermore, from the definition of $t_{n,j}^{(i)}$ and the definitions of the systems $x_{n+1} = A_n^{(i)} x_n$, $n \in \mathbb{Z}$, for $i = 1, 2, \ldots, M$, it can easily be computed that $t_{n,j}^{(i)}$ is equal to the right-hand side of (3.10) for $n = 1, 2, \ldots, N$, $j = 1, 2, \ldots$, and $i = 1, 2, \ldots, M$. Thus, according to the decomposition (3.9), $t_{n,j}^{(0)} = s_{n,j}$ for all n and j. The matrix $A_N \cdots A_1$ is nilpotent and hence the numbers $s_{n,j}$ are in a one-to-one relation with the numbers r_n, $n = 1, 2, \ldots, N$, and the ranks in the left-hand side of (0.4). Thus it follows from Theorem 0.2 that the N-periodic difference equations (0.1) and $x_{n+1} = A_n^{(0)} x_n$, $n \in \mathbb{Z}$, are N-periodically similar. The latter N-periodic difference equation is reducible and hence (0.1) is reducible as well. □

Part 2 of the proof of Lemma 3.1 shows that any N-periodic difference equation of the form (0.1) with nilpotent monodromy matrix is N-periodically similar to the

direct sum of N-periodic difference equations of types II and III. To find these N-periodic difference equations, one may decompose the admissible array $(s_{n,j})_{n=1,j=1}^{N\ \infty}$ into irreducible admissible arrays, as in (3.9). Such a decomposition can be made as follows. For each $1 \leq n \leq N$ make a Young tableau of the sequence $(s_{n,j})_{j=1}^{\infty}$ by representing each nonzero number $s_{n,j}$ as a rectangle with height $s_{n,j}$ and width 1 and putting these rectangles one next to another (cf. [5, p. 131]). The result is a ladder-like figure Φ_n. For instance, if $n = m$ and $s_{m,1} = 5$, $s_{m,2} = s_{m,3} = 4$, $s_{m,4} = 1$, and $s_{m,j} = 0$ for $j > 4$, then Φ_m is the following figure:

(3.11)



Now view Φ_m as a set of rectangles of height 1, which are put one above another. The widths of these rectangles determine a decomposition

$$s_{m,j} = \tilde{s}_{m,j}^{(1)} + \cdots + \tilde{s}_{m,j}^{M_m},\ j \geq 1. \tag{3.12}$$

For instance, if Φ_m is as in (3.11), then the decomposition (3.12) is given by $M_m = 5$ and

$$\tilde{s}_{m,j}^{(1)} = \begin{cases} 1, & 1 \leq j \leq 4, \\ 0, & j > 4, \end{cases} \qquad \tilde{s}_{m,j}^{(i)} = \begin{cases} 1, & 1 \leq j \leq 3, \\ 0, & j > 3, \end{cases} \quad i = 2, 3, 4,$$

$$\tilde{s}_{m,j}^{(5)} = \begin{cases} 1, & j = 1, \\ 0, & j > 1. \end{cases}$$

In this way we find integers $\tilde{s}_{m,j}^{(i)}$ for $m = 1, 2, \ldots, N$ and for $i = 1, 2, \ldots, M_m$, $j \geq 1$. Next put $S_m = \sum_{n=1}^{m} M_n$ for $M = 1, 2, \ldots, N$ and put $M = S_N$. For integers i between $1 + S_m$ and S_{m+1}, $1 + S_m \leq i \leq S_{m+1}$, we define numbers $s_{n,j}^{(i)}$, $n = 1, 2, \ldots, N$, $j \geq 1$, by

$$s_{n,j}^{(i)} = \begin{cases} \tilde{s}_{m,j}^{(i-S_m)} & \text{if } n = m \text{ and } j \geq 1, \\ 0 & \text{if } n \neq m \text{ and } j \geq 1. \end{cases}$$

The numbers $s_{n,j}^{(i)}$ satisfy the decomposition (3.9) and the array $(s_{n,j}^{(i)})_{n=1,j=1}^{N\ \infty}$ is irreducible for all i.

Proof of Theorem 1.4. The proof is divided into two steps.

Step 1. In the first step we make a little computation. Let the N-periodic difference equation (0.1) be of type II or of type III, i.e., (0.1) is equal to $\Sigma_{\mathrm{II}}(m, p, r)$ or (0.1) is equal to $\Sigma_{\mathrm{III}}(m, r)$. In the latter case, put $p = m$. In both cases the numbers $\eta_{j,k}((A_n))$ in (1.11) are given by

$$\eta_{j,k,m,p,r} := \eta_{j,k}((A_n)) = \begin{cases} 1 & \text{if } k = m \text{ and } j = p - 1 + rN, \\ 0 & \text{otherwise.} \end{cases} \tag{3.13}$$

Step 2. Assume that the N-periodic difference equation (0.1) has a nilpotent monodromy matrix. By the paragraph preceding the present proof, (0.1) is N-periodically similar to the direct sum Σ of N-periodic difference equations of types II and III, say,

$$\Sigma = \oplus_{i=1}^{t} \alpha_i \Sigma_{\mathrm{II}}(m_i, p_i, r_i) \oplus \oplus_{i=t+1}^{t'} \alpha_i \Sigma_{\mathrm{III}}(m_i, r_i)$$

with $(m_i, p_i, r_i) \neq (m_j, p_j, r_j)$ for all $i \neq j$, $1 \leq i, j \leq t$ and $(m_i, r_i) \neq (m_j, r_j)$ for $i \neq j$, $t+1 \leq i, j \leq t'$. Let B_n be the nth coefficient of Σ. According to the direct sum decomposition of Σ,

$$\eta_{j,k}((B_n)) = \sum_{i=1}^{t'} \alpha_i \eta_{j,k,m_i,p_i,r_i} \tag{3.14}$$

for all j and k. By N-periodic similarity the number $\eta_{j,k}((A_n))$ is also equal to the right-hand side of (3.14) for each j and k. It follows from the computation in Step 1 that the numbers α_i, $i = 1, \ldots, t'$, are uniquely determined and given by $\alpha_i = \eta_{p_i - 1 + r_i N, m_i}((A_n))$ for $i = 1, 2, \ldots t'$. □

Proof of Theorem 2.3. Theorem 2.3 may be proved in a similar way as Theorem 1.4 by replacing N-periodic similarity by kinematic similarity. □

Proofs of Theorems 0.4 *and* 0.5. Equations of types II and III are irreducible under N-periodic or kinematic similarity by Lemma 3.1. Equations of type I are kinematically irreducible by Theorem 4.3 in [4]. Hence they are also irreducible under N-periodic similarity, and equations of type I_+ are kinematically irreducible.

From Theorems 1.2–1.4 it follows that any N-periodic difference equation is N-periodically similar to a direct sum of equations of types I, II, and III. According to Theorems 2.1–2.3, any N-periodic difference equation is kinematically similar to a direct sum of equations of types I_+, II, and III.

It remains to show the uniqueness of the direct sums in Theorems 0.4 and 0.5. We shall show the uniqueness in the former theorem; the uniqueness in the latter theorem can be shown similarly. So assume that (0.1) and (0.2) are N-periodically similar and that both of them are direct sums of systems of types I, II, and III. System (0.1) is the direct sum of the N-periodic difference equations $\Sigma_{1,1}$ and $\Sigma_{1,2}$, where $\Sigma_{1,1}$ is the sum of all systems of type I in (0.1) and $\Sigma_{1,2}$ is the sum of all systems of types II and III in (0.1). In a similar way system (0.2) is the direct sum of the N-periodic difference equations $\Sigma_{2,1}$ and $\Sigma_{2,2}$, where $\Sigma_{2,1}$ is the direct sum of all systems of type I in (0.2) and $\Sigma_{2,2}$ is the direct sum of all systems of types II and III in (0.2). From Theorem 0.2 we obtain that the systems $\Sigma_{1,1}$ and $\Sigma_{2,1}$ are N-periodically similar. According to Theorem 1.3, the direct sums $\Sigma_{1,1}$ and $\Sigma_{2,1}$ are the same up to reordering. Theorem 0.2 also yields that the systems $\Sigma_{1,2}$ and $\Sigma_{2,2}$ are N-periodically similar. By Theorem 1.4 the direct sums $\Sigma_{1,2}$ and $\Sigma_{2,2}$ are the same up to reordering. We conclude that the direct sums (0.1) and (0.2) are the same up to reordering. □

Proof of Theorem 0.3. By Theorem 0.5 the system (0.1) is irreducible under N-periodic similarity if and only if it is either N-periodically similar to a system of type I or N-periodically similar to a system of type II or III. According to Theorem 0.2, the system (0.1) is N-periodically similar to a system of type I if and only if the monodromy matrix of (0.1) is invertible and its Jordan form has one block only. The system (0.1) is N-periodically similar to a system of type II or III if and only if (0.6) holds for some integers m and q with $1 \leq m \leq N$ and $q \geq 1$. □

Finally, we come to reducibility of N-periodic difference equations by kinematic similarity in the general class of time-varying (possibly non-N-periodic) difference equations.

THEOREM 3.3. *The N-periodic difference equation* (0.1) *is irreducible by kinematic similarity in the class of time-varying difference equations if and only if the monodromy matrix $A_N \cdots A_1$ is invertible and the Jordan form of $A_N \cdots A_1$ consists of one block only.*

Proof. The proof is split into two parts.

Part 1. Assume that the monodromy matrix of the system (0.1) is invertible and has one Jordan block only. By Theorem 0.1 the system (0.1) is kinematically similar to a time-invariant system $x_{n+1} = Jx_n$, $n \in \mathbb{Z}$, where J is the Jordan matrix of the monodromy matrix of (0.1). The latter system is kinematically irreducible in the class of time-varying systems by Theorem 4.3 in [4]. Hence (0.1) is kinematically irreducible in the class of time-varying systems.

Part 2. Assume that the system (0.1) is kinematically irreducible in the class of time-varying systems. By Theorem 0.4 the system (0.1) is either kinematically similar to a system of type I_+ or kinematically similar to a system of type II or type III. If (0.1) is kinematically similar to a system of type I_+, then, according to Theorem 0.2, the monodromy matrix of (0.1) is invertible and its Jordan form has one block only. So it suffices to show that any system of type II or type III is the direct sum of time-varying (non-N-periodic) systems. Let (0.1) be equal to $\Sigma_{\mathrm{II}}(m,p,r)$ or to $\Sigma_{\mathrm{III}}(m,r)$. In the latter case, put $p = m$. Let $\Sigma^{(1)} : x_{n+1} = A_n^{(1)} x_n$, $n \in \mathbb{Z}$, and $\Sigma^{(2)} : x_{n+1} = A_n^{(2)} x_n$, $n \in \mathbb{Z}$, be the time-varying systems defined by

$$A_n^{(1)} = \begin{cases} A_n : \mathbb{C}^{r_n} \to \mathbb{C}^{r_{n+1}} & \text{if } n \le p, \\ I : \mathbb{C}^{r-k} \to \mathbb{C}^{r-k} & \text{if } p + kN < n < p + (k+1)N \text{ for some} \\ & k \in \{0, 1, \dots, r-1\}, \\ \begin{pmatrix} 0 & I \end{pmatrix} : \mathbb{C}^{r-k} \to \mathbb{C}^{r-k-1} & \text{if } n = p + (k+1)N \text{ for some} \\ & k \in \{0, 1, \dots, r-1\}, \\ 0 : \{0\} \to \{0\} & \text{if } n > p + rN, \end{cases}$$

$$A_n^{(2)} = \begin{cases} 0 : \{0\} \to \{0\} & \text{if } n < m + N, \\ I : \mathbb{C}^{\ell+1} \to \mathbb{C}^{\ell+1} & \text{if } m + (\ell+1)N < n < m + (\ell+2)N \text{ for some} \\ & \ell \in \{0, 1, \dots, r-1\}, \\ \begin{pmatrix} I \\ 0 \end{pmatrix} : \mathbb{C}^{\ell} \to \mathbb{C}^{\ell+1} & \text{if } n = m + (\ell+1)N \text{ for some } \ell \in \{0, 1, \dots, r-1\}, \\ A_n & \text{if } n \ge m + (r+1)N. \end{cases}$$

The N-periodic difference equation (0.1) is the direct sum of the time-varying difference equations $\Sigma^{(1)}$ and $\Sigma^{(2)}$. ∎

4. Reducibility in the class of square periodic systems. In this section we consider square reducibility in the class of N-periodic systems.

In the first two theorems we shall describe the squarely irreducible systems. For this purpose we call the square N-periodic system $x_{n+1} = A_n x_n$, $n \in \mathbb{Z}$, a *system of type* IV if precisely one of the coefficients A_n is a singular single Jordan block and each of the other coefficients A_n is either the identity matrix or an identity matrix in which precisely one of the 1's is replaced by a 0. Recall that the system (0.1) is called a system of type I if it is time-invariant and if $A_n = A$ is an invertible single Jordan block with a nonzero eigenvalue λ satisfying $0 \le \arg \lambda < \frac{2\pi}{N}$, and that (0.1) is called a system of type I_+ if, additionally, $A_n = A$ has positive eigenvalues only.

THEOREM 4.1. *Square N-periodic systems of type* I_+ *or* IV *are irreducible in the class of square N-periodic systems under kinematic similarity. Any square N-periodic*

system is kinematically similar to a unique (up to reordering) direct sum of systems of types I_+ *and* IV.

THEOREM 4.2. *Square* N*-periodic systems of type* I *or* IV *are irreducible in the class of square* N*-periodic systems under* N*-periodic similarity. Any square* N*-periodic system is* N*-periodically similar to a unique (up to reordering) direct sum of systems of types* I *and* IV.

In the proofs of Theorems 4.1 and 4.2 we shall use some terminology from graph theory (see, e.g., Chapter 1 in [6]). Let $\mathcal{G} = (V, E)$ be a directed graph with loops and multiple arcs. A sequence $v_1, v_2, \dots, v_n$ of vertices in V is called a *circuit* if $(v_j, v_{j+1}) \in E$ for $1 \le j \le n-1$, and $v_n = v_1$. A circuit $v_1, v_2, \dots, v_n$ is called *elementary* if $v_i \ne v_j$ for $n \ne i \ne j \ne n$. Let $E' \subset E$ be a subset of arcs of $\mathcal{G}$. The *subgraph spanned by* E' is defined to be the graph $\mathcal{G}' = (V', E')$ with V' the set of vertices which are initial point or terminal point of the arcs in E'.

Proofs of Theorems 4.1 *and* 4.2. According to Theorem 0.5, any square N-periodic system is N-periodically similar to a unique direct sum of systems of types I, II, and III. A similar result holds under kinematic similarity (Theorem 0.4). Notice that each system of type III is also of type IV. Hence, in order to prove Theorems 4.1 and 4.2, it suffices to show that systems of type IV are kinematically irreducible in the class of square N-periodic systems and that any direct sum of systems of type II that is a square system may be seen as a unique direct sum of systems of type IV by changing the order in the direct sum. We shall first show the latter statement.

Let $\Sigma(q_i, t_i)$, $i = 1, 2, \dots, M$, be a collection of systems $x_{n+1} = C_n^{(2,i)} x_n$, $n \in \mathbb{Z}$, of type II. Here, for $i = 1, 2, \dots, M$, the numbers q_i, t_i denote the unique indices between 1 and N of the coefficients $C_q^{(2,i)}$ and $C_t^{(2,i)}$, respectively, which are of the form

$$C_q^{(2,i)} = (\,0 \quad I\,) : \mathbb{C}^{r+1} \to \mathbb{C}^r, \quad C_t^{(2,i)} = \begin{pmatrix} I \\ 0 \end{pmatrix} : \mathbb{C}^r \to \mathbb{C}^{r+1}$$

for some r. Notice that $t_i \ne q_i$ for systems of type II. With this collection $\Sigma(q_i, t_i)$, $i = 1, 2, \dots, M$, of systems of type II we associate a directed graph $\mathcal{G} = (V, E)$, which may have loops and multiple arcs. Namely, let

$$V = \{q_i\}_{i=1}^M \cup \{t_i\}_{i=1}^M, \ E = \{(q_i, t_i) \mid i = 1, 2, \dots, M\},$$

where equal elements of the vertex set V are taken only once and equal elements of the edge set E are taken as often as they appear when i runs from 1 to M. Thus a subset E' of E corresponds with a direct sum of the systems $\Sigma(q_i', t_i')$, where (q_i', t_i') runs through E'. Notice that a direct sum of systems of the form $\Sigma(q_i, t_i)$ is square if and only if the corresponding set of arcs spans a circuit in $\mathcal{G}$.

Assume that the direct sum of all systems $\Sigma(q_i, t_i)$, $i = 1, 2, \dots, M$, is a square system. Hence the graph $\mathcal{G}$ is a circuit. It may be easily seen that the arc set E (i.e., the set E consisting of all arcs of $\mathcal{G}$) allows a unique decomposition

$$E = E_1 \cup E_2 \cup \cdots \cup E_P \tag{4.1}$$

into mutually disjoint arc sets E_i, $i = 1, 2, \dots, P$, such that for each $1 \le i \le P$ the subgraph of $\mathcal{G}$ spanned by E_i is an elementary circuit. We claim that the direct sum that corresponds to an elementary circuit is a system of type IV. To see this, let the graph spanned by $\{(q_i, t_i) \mid i = 1, 2, \dots, M'\}$ be an elementary circuit. The direct sum

$\oplus_{i=1}^{M'}\Sigma(q_i,t_i)$ of the systems $\Sigma(q_i,t_i)$ is the N-periodic system $x_{n+1}=C_nx_n$, $n\in\mathbb{Z}$, with

$$C_n=\begin{cases} J_{0,r} & \text{if } n=q_1,\\ \tilde{I}_r & \text{if } n=q_j,\ j\neq 1,\\ I_r & \text{if } n\neq q_j,\ j=1,2,\dots,M', \end{cases} \qquad 1\le n\le N,$$

for some r, where $\tilde{I}_r$ denotes a matrix that is obtained by replacing a 1 by a 0 in the $r\times r$ identity matrix. So the direct sum $\oplus_{i=1}^{M'}\Sigma(q_i,t_i)$ is a system of type IV. We conclude from the decomposition (4.1) that the direct sum of all systems $\Sigma(q_i,t_i)$, $i=1,2,\dots,M$, is a unique direct sum of P systems of type IV.

To see that any system of type IV is kinematically irreducible in the class of square N-periodic systems, notice that any system of type IV is (kinematically similar) to a direct sum of systems of types II and III and that the circuit in $\mathcal{G}$ that corresponds to this direct sum is elementary. Thus any system of type IV is kinematically irreducible in the class of square N-periodic systems. ∎

5. Similarity of difference equation with different periods. In this section we will study periodic difference equations that are kinematically (or periodically) similar to periodic systems with a different period. The main result is as follows.

THEOREM 5.1. *Let $x_{n+1}=A_nx_n$, $n\in\mathbb{Z}$, be an N-periodic difference equation with $A_n:\mathbb{C}^{r_n}\to\mathbb{C}^{r_{n+1}}$ for $n\in\mathbb{Z}$. Let M be a positive integer. The following statements are equivalent:*

(i) *The difference equation $x_{n+1}=A_nx_n$, $n\in\mathbb{Z}$, is kinematically similar to an M-periodic difference equation.*

(ii) *The difference equation $x_{n+1}=A_nx_n$, $n\in\mathbb{Z}$, is N-periodically similar to an M-periodic difference equation.*

(iii) *The numbers r_n are M-periodic in n, and for each $1\le j\le M(\max_{1\le n\le M} r_n)$ the ranks*

$$\operatorname{rank}\left(A_{n+j-1}A_{n+j-2}\cdots A_n\right),\ n\in\mathbb{Z}, \tag{5.1}$$

are M-periodic in n.

(iv) *The numbers r_n are M-periodic in n, and for each $1\le j\le N(\max_{1\le n\le N} r_n)$ the ranks (5.1) are M-periodic in n.*

Proof. (ii) ⇒ (i). The proof is evident.

(i) ⇒ (iii). Assume that $x_{n+1}=A_nx_n$, $n\in\mathbb{Z}$, is kinematically similar to the M-periodic difference equation $\tilde{x}_{n+1}=B_n\tilde{x}_n$, $n\in\mathbb{Z}$. It is evident that the numbers r_n are M-periodic in n. Let $(E_n)_{n=-\infty}^{\infty}$ be a sequence of invertible matrices which satisfy $E_{n+1}^{-1}A_nE_n=B_n$, $n\in\mathbb{Z}$. Iterating these equations, we obtain the identities

$$E_{n+j}^{-1}A_{n+j-1}\cdots A_nE_n=B_{n+j-1}\cdots B_n,\ n=1,2,\dots,\ j=1,2,\dots. \tag{5.2}$$

The product in the right-hand side of (5.2) is M-periodic in n. Thus, taking ranks in (5.2), we obtain that the ranks in (5.1) are M-periodic in n.

(iv) ⇒ (ii). Assume that the numbers r_n and the ranks (5.1) are M-periodic in n. By Theorem 0.5 the system (0.1) is N-periodically similar to a direct sum Σ of systems of types I, II, and III. Let P be the greatest common divisor of M and N. The ranks (5.1) are also P-periodic in n, and hence

$$\eta_{j+\ell P,k+\ell P}((A_n))=\eta_{j,k}((A_n)),\ \ell\in\mathbb{Z},$$

for all j and k. Hence, according to Theorem 1.4, the number of systems $\Sigma_{\mathrm{II}}(m,p,r)$ in the direct sum Σ is equal to the number of systems $\Sigma_{\mathrm{II}}(m+\ell P, p+\ell P, r)$ in the direct sum Σ for all ℓ. Thus the coefficients 1 to N of the system $\Sigma_{\mathrm{II}}(m,p,r)$ in Σ are P-periodic. We conclude that the system $\Sigma_{\mathrm{II}}(m,p,r)$ in the direct sum Σ is P-periodic for all triples (m,p,r). In a similar way the system $\Sigma_{\mathrm{III}}(m,r)$ in the direct sum Σ is P-periodic for all pairs (m,r). Since systems of type I are time-invariant, we conclude that, in particular, Σ is M-periodic.

(iii) $\Rightarrow$ (iv). Put $t_{n,k} = \sum_{j=1}^{k} s_{n-j+1,j}$ for $n \in \mathbb{Z}$ and $k = 1, 2, \ldots$, where $s_{n,j}$ is defined by (0.5). Notice that, by definition, $t_{n,k} = r_{n+1} - \operatorname{rank}(A_n A_{n-1} \cdots A_{n-j+1}) \leq r_{n+1}$ for each n and k.

Assume that (iii) holds. It suffices to show that $s_{n,j} = 0$ for all n and for all $j > M \max_{1\leq n\leq M} r_n$. Indeed, in this case the ranks (5.1) are independent of j for $j \geq M \max_{1\leq n\leq M} r_n$, and hence (iv) follows immediately. To see that $s_{n,j} = 0$ for all n and for all $j > M \max_{1\leq n\leq M} r_n$, assume that $s_{m,i} \neq 0$ for some $m \in \mathbb{Z}$ and $i > M \max_{1\leq n\leq M} r_n$. Let ℓ be the entier of $\frac{i-1}{M}$. Notice that $\ell \geq \max_{1\leq n\leq M} r_n$. The array $(s_{n,j})_{n=1,j=1}^{N\ \infty}$ is admissible (Lemma 3.2), and hence (3.3) and (3.5) are satisfied. Notice that

$$t_{m+i-1,i} \geq \sum_{k=0}^{\ell} s_{m+kM,i-kM}, \tag{5.3}$$

because of (3.3). By M-periodicity of the numbers r_n and of the ranks (5.1), the right-hand side of (5.3) is equal to $\sum_{k=0}^{\ell} s_{m,i-kM}$. According to (3.5) the latter sum is larger than or equal to $(\ell+1)s_{m,i}$. We conclude that

$$t_{m+i-1,i} \geq (\ell+1)s_{m,i} \geq \ell + 1 \geq 1 + \max_{1\leq n\leq M} r_n \geq 1 + r_{m+i-1}.$$

However, in the first lines of the proof of (iii) $\Rightarrow$ (iv) we have shown that $t_{m+i-1,i} \leq r_{m+i-1}$. We conclude that $s_{n,j} = 0$ for all n and all $j > M \max_{1\leq n\leq M} r_n$. □

For $M = 1$ the equivalence of (ii) and (iii) in Theorem 5.1 appeared recently in Van Dooren–Sreedhar [10].

6. Special similarities for a class of finite block matrices. The following proposition shows that the problem of N-periodic similarity may be restated as a problem for finite-dimensional matrices.

PROPOSITION 6.1. *Two N-periodic difference equations $x_{n+1} = A_n x_n$, $n \in \mathbb{Z}$, and $y_{n+1} = B_n y_n$, $n \in \mathbb{Z}$, are N-periodically similar if and only if the $N \times N$ block matrices*

$$A = \begin{pmatrix} 0 & \ldots & \ldots & \ldots & 0 & A_N \\ A_1 & 0 & \ldots & \ldots & \ldots & 0 \\ 0 & A_2 & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & \ldots & 0 & A_{N-1} & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & \ldots & \ldots & \ldots & 0 & B_N \\ B_1 & 0 & \ldots & \ldots & \ldots & 0 \\ 0 & B_2 & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & \ldots & 0 & B_{N-1} & 0 \end{pmatrix} \tag{6.1}$$

are block-diagonally similar.

Proof. Let $(E_n)_{n=-\infty}^{\infty}$ be an N-periodic similarity that transforms the difference equation $x_{n+1} = A_n x_n$, $n \in \mathbb{Z}$, into $\tilde{x}_{n+1} = B_n \tilde{x}_n$, $n \in \mathbb{Z}$. So, we have the

identities $E_{n+1}^{-1}A_nE_n = B_n$, $n = 1, 2, \ldots, N-1$, and $E_1^{-1}A_NE_N = B_N$. Hence $E := \operatorname{diag}(E_n)_{n=1}^N$ satisfies $E^{-1}AE = B$.

Conversely, assume that the block matrices A and B are block-diagonally similar. Let $E = \operatorname{diag}(E_n)_{n=1}^N$ satisfy $E^{-1}AE = B$. Extend the sequence $(E_n)_{n=1}^N$ to an N-periodic sequence $(E_n)_{n=-\infty}^{\infty}$. From $E^{-1}AE = B$ we obtain the identities $E_{n+1}^{-1}A_nE_n = B_n$, $n = 1, 2, \ldots, N$. We conclude that $(E_n)_{n=-\infty}^{\infty}$ is an N-periodic similarity that transforms $x_{n+1} = A_nx_n$, $n \in \mathbb{Z}$, into $\tilde{x}_{n+1} = B_n\tilde{x}_n$, $n \in \mathbb{Z}$. ∎

Notice that block-diagonal similarity of A and B in (6.1) is not the same as similarity of A and B. To see this, let

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix},$$

where $N = 3$, $r_1 = r_2 = r_3 = 1$. For this example, clearly A and B are similar, but they are not diagonally similar. Necessary and sufficient conditions both for similarity and for block-diagonal similarity of the block matrices in (6.1) will be given in Theorem 6.2 below.

In what follows, we use the following notation and definition. Given a Jordan matrix J of order r, we denote by J^{inv} an invertible Jordan matrix of which the Jordan blocks are the nonsingular Jordan blocks of J. The nonsingular Jordan matrix J^{inv} is referred to as the *nonsingular* (or *invertible*) *part* of J.

From now on we fix a positive integer N and nonnegative integers r_n, $1 \le n \le N$. Put $s = N \max_{1\le n\le N} r_n$. Other choices for s are possible (cf. the discussion that follows the proof of Lemma 1.1).

THEOREM 6.2. *Let A and B be the $N \times N$ block matrices*

$$A = \begin{pmatrix} 0 & \ldots & \ldots & \ldots & 0 & A_N \\ A_1 & 0 & \ldots & \ldots & \ldots & 0 \\ 0 & A_2 & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & \ldots & 0 & A_{N-1} & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & \ldots & \ldots & \ldots & 0 & B_N \\ B_1 & 0 & \ldots & \ldots & \ldots & 0 \\ 0 & B_2 & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & \ldots & 0 & B_{N-1} & 0 \end{pmatrix}, \tag{6.2}$$

where, for $1 \le i \le N$, the zero in the ith block-diagonal entry is a square matrix of order r_i. The matrices A and B are block-diagonally similar if and only if $A_NA_{N-1}\cdots A_1$ and $B_NB_{N-1}\cdots B_1$ are similar, and for each $1 \le n \le N$

$$\operatorname{rank}(A_{j+n-1}A_{j+n-2}\cdots A_n) = \operatorname{rank}(B_{j+n-1}B_{j+n-2}\cdots B_n), \ j = 1, 2, \ldots, s, \tag{6.3}$$

where $A_{n+N} = A_n$, $n = 1, 2, \ldots$, and $B_{n+N} = B_n$, $n = 1, 2, \ldots$. Furthermore, let $J^{\mathrm{inv}}_{A_NA_{N-1}\cdots A_1}$ (resp., $J^{\mathrm{inv}}_{B_NB_{N-1}\cdots B_1}$) be the nonsingular part of a Jordan matrix of $A_NA_{N-1}\cdots A_1$ (resp., $B_NB_{N-1}\cdots B_1$). The matrices A and B are similar if and only if $J^{\mathrm{inv}}_{A_NA_{N-1}\cdots A_1}$ is similar to $J^{\mathrm{inv}}_{B_NB_{N-1}\cdots B_1}$, and for each $1 \le j \le s$

$$\sum_{n=1}^{N} \operatorname{rank}(A_{n+j-1}A_{n+j-2}\cdots A_n) = \sum_{n=1}^{N} \operatorname{rank}(B_{n+j-1}B_{n+j-2}\cdots B_n), \tag{6.4}$$

where $A_{n+N} = A_n$, $n = 1, 2, \ldots$, and $B_{n+N} = B_n$, $n = 1, 2, \ldots$.

From Proposition 6.1 and Theorem 1.4 we may obtain a canonical form for block matrices of the form (6.2) under block-diagonal similarity.

Proof of Theorem 6.2. By Proposition 6.1 the block matrices A and B in (6.2) are block-diagonally similar if and only if the N-periodic difference equations $x_{n+1} = A_n x_n$, $n \in \mathbb{Z}$, and $\tilde{x}_{n+1} = B_n \tilde{x}_n$, $n \in \mathbb{Z}$, are N-periodically similar. So from Theorem 0.2 we conclude that A and B are block-diagonally similar if and only if $A_N \cdots A_1$ is similar to $B_N \cdots B_1$, and for each $1 \le n \le N$ the rank conditions (6.3) are satisfied.

Let us consider the block matrix A in the left-hand side of (6.2). By Proposition 6.1 and Theorems 1.2 and 0.2 it is (block-diagonally) similar to a block matrix C_A of the form

$$C_A = \begin{pmatrix} 0 & \dots & \dots & \dots & 0 & C_N \\ C_1 & 0 & \dots & \dots & \dots & 0 \\ 0 & C_2 & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & C_{N-1} & 0 \end{pmatrix},$$

where $C_n = \begin{pmatrix} C_n^{(1)} & 0 \\ 0 & C_n^{(2)} \end{pmatrix}$ with

$$C_n^{(1)} = \begin{cases} I & 1 \le n \le N-1, \\ J^{\mathrm{inv}}_{A_N \cdots A_1}, & n = N, \end{cases}$$

and with the product $C_N^{(2)} \cdots C_1^{(2)}$ being nilpotent. Put

$$C_{A,1} = \begin{pmatrix} 0 & \dots & \dots & 0 & J^{\mathrm{inv}}_{A_N \cdots A_1} \\ I & \ddots & & \ddots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & I & 0 \end{pmatrix}, \quad C_{A,2} = \begin{pmatrix} 0 & \dots & \dots & 0 & C^{(2)}_{A,N} \\ C^{(2)}_{A,1} & \ddots & & \ddots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & C^{(2)}_{A,N-1} & 0 \end{pmatrix}.$$

Let us redecompose C_A as a $2N \times 2N$ block matrix using the 2×2 decompositions of the C_n's. With respect to this decomposition of C_A, the block matrix $C_{A,1}$ is the submatrix in the odd block rows and columns, and $C_{A,2}$ is the submatrix in the even block rows and columns. Hence C_A is similar to $\tilde{C}_A = \begin{pmatrix} C_{A,1} & 0 \\ 0 & C_{A,2} \end{pmatrix}$. Notice that $C_{A,2}$ is nilpotent.

Following the preceding paragraph, the block matrix B in the right-hand side of (6.2) is similar to $\tilde{C}_B = \begin{pmatrix} C_{B,1} & 0 \\ 0 & C_{B,2} \end{pmatrix}$ with $C_{B,1}$ an invertible $N \times N$ block matrix and $C_{B,2}$ a nilpotent matrix of the forms

$$C_{B,1} = \begin{pmatrix} 0 & \dots & \dots & 0 & J^{\mathrm{inv}}_{B_N \cdots B_1} \\ I & \ddots & & \ddots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & I & 0 \end{pmatrix}, \quad C_{B,2} = \begin{pmatrix} 0 & \dots & \dots & 0 & C^{(2)}_{B,N} \\ C^{(2)}_{B,1} & \ddots & & \ddots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & C^{(2)}_{B,N-1} & 0 \end{pmatrix}.$$

Now the matrices A and B are similar if and only if the 2×2 block matrices $\tilde{C}_A$ and $\tilde{C}_B$ are similar. Both $C_{A,1}$ and $C_{B,1}$ are invertible, and both $C_{A,2}$ and $C_{B,2}$ are nilpotent. Hence A and B are similar if and only if $C_{A,1}$ is similar to $C_{B,1}$ and $C_{A,2}$ is similar to $C_{B,2}$. It suffices to show that

(i) the matrices $C_{A,1}$ and $C_{B,1}$ are similar if and only if the Jordan matrices $J^{\mathrm{inv}}_{A_N \cdots A_1}$ and $J^{\mathrm{inv}}_{B_N \cdots B_1}$ are similar, and

(ii) if $C_{A,1}$ and $C_{B,1}$ are of the same order, then the matrices $C_{A,2}$ and $C_{B,2}$ are similar if and only if the rank conditions (6.4) are satisfied.

To see (i), assume that $C_{A,1}$ is similar to $C_{B,1}$. It is evident that $C^N_{A,1}$ and $C^N_{B,1}$ are similar. From the form of these two block-diagonal matrices we conclude that $J^{\mathrm{inv}}_{A_N \cdots A_1}$ is similar to $J^{\mathrm{inv}}_{B_N \cdots B_1}$. Conversely, if the latter two Jordan matrices are similar, the block matrices $C_{A,1}$ and $C_{B,1}$ are (block-diagonally) similar.

To see (ii), assume that $C_{A,1}$ and $C_{B,1}$ are of the same order. Hence $C_{A,2}$ and $C_{B,2}$ are of the same order. Both of them are nilpotent. Thus they are similar if and only if $\operatorname{rank} C^j_{A,2} = \operatorname{rank} C^j_{B,2}$ for $j = 1, 2, \ldots, s$. These identities may be rewritten as

$$\sum_{n=1}^{N} \operatorname{rank} (C^{(2)}_{A,n+j-1} \cdots C^{(2)}_{A,n}) = \sum_{n=1}^{N} \operatorname{rank} (C^{(2)}_{B,n+j-1} \cdots C^{(2)}_{B,n}), \ j = 1, 2, \ldots, s. \tag{6.5}$$

Since both $C_{A,1}$ and $C_{B,1}$ are invertible the rank conditions (6.5) hold if and only if the rank conditions (6.4) hold. □

With $A_N = B_N = 0$, the matrices A and B in (6.2) are finite block-weighted shift matrices as in section 5 of [4]. Hence all results for kinematic similarity of finitely supported difference equations in section 5 of [4] may be derived from the present results for N-periodic similarity of N-periodic difference equations.

7. An example. We conclude the paper with an example to illustrate how canonical forms of N-periodic systems are computed (cf. the results in Theorems 0.4, 0.5, 1.2–1.4, and 2.1–2.3) and how the corresponding N-periodic and kinematic similarities can be constructed.

Let (0.1) be the 2-periodic system given by

$$A_1 = \begin{pmatrix} 0 & 1 & -1 & 0 \\ 0 & -2 & 2 & 0 \\ 0 & 3 & -2 & 1 \\ 0 & -2 & 1 & 0 \end{pmatrix}, \ A_2 = \begin{pmatrix} -4 & -2 & -2 & -3 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 \end{pmatrix}. \tag{7.1}$$

Both r_1 and r_2 are 4, and $s = 8$. By direct computation we obtain that

$$A_2A_1 = \begin{pmatrix} 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \ A_1A_2A_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix},$$

$$A_1A_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & -1 & -1 & -1 \end{pmatrix}.$$

TABLE 7.1
rank $(A_j \cdots A_k)$.

$k \setminus j$	0	1	2	3	4	5	6	7	8
1	4	3	2	1	1	1	1	1	1
2	-	4	2	1	1	1	1	1	1
3	-	-	4	3	2	1	1	1	1

TABLE 7.2
rank $(A_{j+1} \cdots A_k)$ − rank $(A_j \cdots A_k)$.

$k \setminus j$	0	1	2	3	4	5	6	7	8
1	-1	-1	-1	0	0	0	0	0	0
2	-	-2	-1	0	0	0	0	0	0
3	-	-	-1	-1	-1	0	0	0	0

First, we show how the canonical form under N-periodic similarity in Theorem 0.5 can be computed. The system (0.1) is 2-periodically similar to a unique direct sum of 2-periodic systems of types I, II, and III. This direct sum will be denoted by

$$x_{n+1} = C_n x_n, \quad n \in \mathbb{Z}. \tag{7.2}$$

To compute the systems of type I in the direct sum (7.2), notice that -1 is the only nonzero eigenvalue of the monodromy operator A_2A_1 and that both the algebraic and geometric multiplicity of -1 are 1. It follows from Theorems 1.2 and 1.3 that

$$x_{n+1} = ix_n, \quad n \in \mathbb{Z}, \tag{7.3}$$

is the only system of type I in the direct sum (7.2). Next we compute the systems of types II and III in the direct sum (7.2). Tables 7.1 and 7.2 can be computed easily, where the second table has been derived from the first one.

By Table 7.2 the triples (m, p, r) of positive integers which satisfy $1 \le m \le 2$, $m < p < m + 2$, and $\eta_{p-1+2r,m}((A_n)) > 0$ are $(m, p, r) = (2, 3, 1)$ and $(m, p, r) = (1, 2, 0)$. Thus, according to Theorem 1.4, the direct sum (7.2) contains the following two 2-periodic systems of types II:

$$x_{n+1} = C_n^{(\mathrm{II}),1} x_n, \quad n \in \mathbb{Z}, \qquad x_{n+1} = C_n^{(\mathrm{II}),2} x_n, \quad n \in \mathbb{Z},$$

where the coefficients are given by

$$C_n^{(\mathrm{II}),1} = \begin{cases} \begin{pmatrix} 0 & 1 \end{pmatrix} : \mathbb{C}^2 \to \mathbb{C}, & n = 1, \\ \begin{pmatrix} 1 \\ 0 \end{pmatrix} : \mathbb{C} \to \mathbb{C}^2, & n = 2, \end{cases} \qquad C_n^{(\mathrm{II}),2} = \begin{cases} 0 : \{0\} \to \mathbb{C}, & n = 1, \\ 0 : \mathbb{C} \to \{0\}, & n = 2. \end{cases}$$

By Table 7.2 the only pair (m, r) of positive integers which satisfies $1 \le m \le 2$ and $\eta_{m-1+2r,m} > 0$ is $(m, r) = (2, 1)$. Thus, according to Theorem 1.4, the direct sum (7.2) contains the following 2-periodic system of type III:

$$x_{n+1} = C_n^{(\mathrm{III})} x_n, \quad n \in \mathbb{Z},$$

where the coefficients $C_n^{(\mathrm{III})}$ are given by

$$C_n^{(\mathrm{III})} = \begin{cases} 1, & n = 1, \\ 0, & n = 2. \end{cases} \tag{7.4}$$

We conclude that the canonical system (7.2) is given by

$$C_1 = \left(\begin{array}{c|cc||c} i & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right), \quad C_2 = \left(\begin{array}{c|c|c|c} i & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ \hline\hline 0 & 0 & 0 & 0 \end{array}\right). \tag{7.5}$$

Next we show how to compute the 2-periodic similarity that brings (0.1) with coefficients being given by (7.1) into the canonical form (7.2) with coefficients given by (7.5). In the computation two 2-periodic similarities are constructed. The first 2-periodic similarity $(E_n)_{n=-\infty}^{\infty}$ is obtained by applying the constructions in the proof of Theorem 1.3 and in Part 2 of the proof of Theorem 0.2 to the system (0.1) with coefficients being given by (7.1). Following the proof of Theorem 1.3, the first column $e_{1,1}$ of E_1 is defined to be an eigenvector of A_2A_1 with eigenvalue -1. So we take

$$e_{1,1} = (\,3 \quad -1 \quad -1 \quad 1\,)^T.$$

Furthermore, we define the first column $e_{2,1}$ of E_2 to be the vector

$$e_{2,1} = A_1e_{1,1} = (\,0 \quad 0 \quad 0 \quad 1\,)^T.$$

Let the kernels $\mathcal{M}_{n,j}$ be defined as in (1.3). So

$$\mathcal{M}_{1,3} = \cdots = \mathcal{M}_{1,8} = \operatorname{span}\{(\,1 \quad 0 \quad 0 \quad 0\,)^T, (\,0 \quad 1 \quad 0 \quad 0\,)^T, (\,0 \quad 0 \quad 1 \quad 0\,)^T\}, \tag{7.6}$$

$$\mathcal{M}_{2,2} = \cdots = \mathcal{M}_{2,8} = \operatorname{span}\{(\,1 \quad -1 \quad 0 \quad 0\,)^T, (\,0 \quad 1 \quad -1 \quad 0\,)^T, (\,0 \quad 0 \quad 1 \quad -1\,)^T\}.$$

Now we follow the construction of the complements $\mathcal{N}_{n,j}$ in Part 2 of the proof of Theorem 0.2. We define

$$\mathcal{N}_{1,8} = \mathcal{K}_{1,8} = \mathcal{N}_{2,7} = \mathcal{K}_{2,7} = \mathcal{N}_{1,6} = \mathcal{K}_{1,6} = \mathcal{N}_{2,5}$$

$$= \mathcal{K}_{2,5} = \mathcal{N}_{1,4} = \mathcal{K}_{1,4} = \mathcal{N}_{2,3} = \mathcal{K}_{2,3} = \{0\}.$$

The kernels $\mathcal{M}_{1,1}$ and $\mathcal{M}_{1,2}$ are given by

$$\mathcal{M}_{1,1} = \operatorname{span}\{(\,1 \quad 0 \quad 0 \quad 0\,)^T\}, \quad \mathcal{M}_{1,2} = \operatorname{span}\{(\,1 \quad 0 \quad 0 \quad 0\,)^T, (\,0 \quad 1 \quad 0 \quad 0\,)^T\}, \tag{7.7}$$

and thus we set

$$\mathcal{N}_{1,2} = \mathcal{K}_{1,2} = \operatorname{span}\{(\,0 \quad 1 \quad 0 \quad 0\,)^T\}.$$

Notice that $A_1\mathcal{N}_{1,2} = \operatorname{span}\{(\,1 \quad -2 \quad 3 \quad -2\,)^T\}$ and

$$\mathcal{M}_{2,0} \oplus A_1\mathcal{N}_{1,2} \oplus \mathcal{K}_{2,1} = \mathcal{M}_{2,1}$$

with $\mathcal{K}_{2,1} = \operatorname{span}\{(\,0 \quad 1 \quad -1 \quad 0\,)^T\}$. Thus we put

$$\mathcal{N}_{2,1} = A_1\mathcal{N}_{1,3} \oplus \mathcal{K}_{2,1} = \operatorname{span}\{(\,1 \quad -2 \quad 3 \quad -2\,)^T, (\,0 \quad 1 \quad -1 \quad 0\,)^T\}.$$

Next we define

$$\mathcal{N}_{2,8} = \mathcal{K}_{2,8} = \mathcal{N}_{1,7} = \mathcal{K}_{1,7} = \mathcal{N}_{2,6} = \mathcal{K}_{2,6} = \mathcal{N}_{1,5} = \mathcal{K}_{1,5} = \mathcal{N}_{2,4} = \mathcal{K}_{2,4} = \{0\}.$$

Equations (7.6) for $\mathcal{M}_{1,2}$ and (7.7) for $\mathcal{M}_{1,3}$ suggest that we define $\mathcal{N}_{1,3}$ and $\mathcal{K}_{1,3}$ by

$$\mathcal{N}_{1,3} = \mathcal{K}_{1,3} = \operatorname{span}\{(\,0 \quad 0 \quad 1 \quad 0\,)^T\}.$$

We compute that $A_1\mathcal{N}_{1,3} = \operatorname{span}\{(\,-1 \quad 2 \quad -2 \quad 1\,)^T\}$ and that

$$\mathcal{M}_{2,1} = \operatorname{span}\{(\,1 \quad 1 \quad 0 \quad -2\,)^T, (\,0 \quad 1 \quad -1 \quad 0\,)^T\}.$$

Hence we obtain the identity

$$\mathcal{K}_{2,2} = \{0\}, \quad \mathcal{N}_{2,2} = \operatorname{span}\{(-1, 2, -2, 1)^T\}.$$

Finally, notice that $A_2\mathcal{N}_{2,2} = \operatorname{span}\{(1, 0, 0, 0)^T\} = \operatorname{span}\mathcal{M}_{1,1}$. So we get that

$$\mathcal{N}_{1,1} = \operatorname{span}\{(1, 0, 0, 0)^T\}.$$

Now we collect the above basis vectors of the complements $\mathcal{N}_{1,j}$, $j = 1, 2, \ldots, 8$, in the second, third, and fourth columns of E_1 and we collect the above basis vectors of the complements $\mathcal{N}_{2,j}$, $j = 1, 2, \ldots 8$, in the second, third, and fourth columns of E_2, i.e.,

$$E_1 = \begin{pmatrix} 3 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \; E_2 = \begin{pmatrix} 0 & 1 & 0 & -1 \\ 0 & -2 & 1 & 2 \\ 0 & 3 & -1 & -2 \\ 1 & -2 & 0 & 1 \end{pmatrix}.$$

Put $B_1 = E_2^{-1}A_1E_1$ and $B_2 = E_1^{-1}A_2E_2$. According to the proofs of Theorems 1.3 and 0.2, we get the identities

$$B_1 = \left(\begin{array}{c|c|c|c} 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right), \quad B_2 = \left(\begin{array}{c|c|c|c} -1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \end{array}\right).$$

The second 2-periodic similarity $(F_n)_{n=-\infty}^{\infty}$ is obtained by applying the constructions in the proofs of Theorems 1.2 and 0.2 to the canonical 2-periodic system (7.2) with the coefficients being given by (7.5). Following these constructions we obtain, after making computations as in the previous paragraph, the matrices

$$F_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad F_2 = \begin{pmatrix} i & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

and the identities $F_2^{-1}C_1F_1 = B_1$ and $F_1^{-1}C_2F_2 = B_2$. Let $(G_n)_{n=-\infty}^{\infty}$ be the 2-periodic sequence given by $G_n = E_nF_n^{-1}$, $n = 1, 2$, i.e.,

$$G_1 = \begin{pmatrix} 3 & 1 & 0 & 0 \\ -1 & 0 & 0 & 1 \\ -1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \quad G_2 = \begin{pmatrix} 0 & -1 & 1 & 0 \\ 0 & 2 & -2 & 1 \\ 0 & -2 & 3 & -1 \\ -i & -1 & -2 & 0 \end{pmatrix}.$$

By construction the 2-periodic sequence $(G_n)_{n=-\infty}^{\infty}$ transforms the system (0.1) with the coefficients being given by (7.1) into the system (7.2) with the coefficients being given by (7.5).

The canonical form of the system (0.1) with the coefficients being given by (7.1) under kinematic similarity in Theorem 0.4 is obtained in the following way: In the systems of type I appearing in the canonical form under N-periodic similarity, one replaces the diagonal entries by their absolute values. Thus the canonical form in Theorem 0.4 is the 2-periodic system $x_{n+1} = \tilde{C}_n x_n$, $n \in \mathbb{Z}$, where

$$\tilde{C}_1 = \left(\begin{array}{c|cc||c} 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 \end{array}\right), \quad \tilde{C}_2 = \left(\begin{array}{c|c|c|c} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ \hline\hline 0 & 0 & 0 & 0 \end{array}\right).$$

The corresponding kinematic similarity $(\tilde{G}_n)_{n=-\infty}^{\infty}$ is obtained from $(G_n)_{n=-\infty}^{\infty}$ by multiplying G_n from the right by the diagonal matrix diag $(i^{n-1}, 1, 1, 1)$ (cf. the proof of Theorem 4.1 in [4]). Hence $(\tilde{G}_n)_{n=-\infty}^{\infty}$ is the 4-periodic sequence given by

$$\tilde{G}_1 = \begin{pmatrix} 3 & 1 & 0 & 0 \\ -1 & 0 & 0 & 1 \\ -1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \quad \tilde{G}_2 = \begin{pmatrix} 0 & -1 & 1 & 0 \\ 0 & 2 & -2 & 1 \\ 0 & -2 & 3 & -1 \\ 1 & -1 & -2 & 0 \end{pmatrix},$$

$$\tilde{G}_3 = \begin{pmatrix} -3 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix}, \quad \tilde{G}_4 = \begin{pmatrix} 0 & -1 & 1 & 0 \\ 0 & 2 & -2 & 1 \\ 0 & -2 & 3 & -1 \\ -1 & -1 & -2 & 0 \end{pmatrix}.$$

Since the system considered in this section is square, we may also apply the results of section 4. For instance, in the class of square periodic systems the canonical form for the present example is a direct sum of three square periodic systems, namely, of the time-invariant system (7.3), the 2-periodic system $x_{n+1} = C_n^{(\mathrm{III})} x_n$, $n \in \mathbb{Z}$, where $C_n^{(\mathrm{III})}$ is given by (7.4), and the 2-periodic system $x_{n+1} = C_n^{(\mathrm{II})} x_n$, $n \in \mathbb{Z}$, where

$$C_1^{(\mathrm{II})} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad C_2^{(\mathrm{II})} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

Let us also remark that from Theorem 5.1 it follows that our example is not kinematically similar to one that is time-invariant.

REFERENCES

[1] N.M. Dobrovol'skaja and V.A Ponomarev, *Pairs of opposite operators*, Uspekhi Math. Nauk, 20 (1965), pp. 81–86 (in Russian).

[2] I.M. Gel'fand and V.A. Ponomarev, *Non-decomposable representations of the Lorentz group*, Russian Math. Surveys, 23 (1968), pp. 1–58.

[3] I.M. Gel'fand and V.A. Ponomarev, *Problems of linear algebra and classification of quadruples of subspaces in a finite-dimensional vector space*, in Hilbert Space Operators and Operator Algebras, Béla Sz.-Nagy, ed., Colloq. Math. Soc. János Bolyai 5, Hilbert Space Operators, Tihany (Hungary), North-Holland, Amsterdam, 1970, pp. 163–237.

[4] I. Gohberg, M.A. Kaashoek, and J. Kos, *Classification of linear time-varying difference equations under kinematic similarity*, Integral Equations Operator Theory, 25 (1996), pp. 445–480.

[5] I. Gohberg, P. Lancaster, and L. Rodman, *Invariant Subspaces of Matrices with Applications*, Canad. Math. Soc. Ser. Monogr. Adv. Texts, Wiley, New York, 1986.
[6] M. Gondran and M. Minoux, *Graphs and Algorithms*, Wiley-Intersci. Ser. Discrete Math. Optim., Wiley, New York, 1984.
[7] R.A. Horn and D.I. Merino, *Contragredient equivalence: A canonical form and some applications*, Linear Algebra Appl., 214 (1995), pp. 43–92.
[8] A. Lenard, *Structure theory of reciprocal pairs of linear transformations*, Linear Algebra Appl., 207 (1994), pp. 21–35.
[9] L.A. Nazarova, *Representations of quivers of infinite type*, Math. USSR-Izv., 7 (1973), pp. 749–763.
[10] P. Van Dooren and J. Sreedhar, *When is a periodic discrete-time system equivalent to a time invariant one?*, Linear Algebra Appl., 212/213 (1994), pp. 131–151.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 508–521
© 1999 Society for Industrial and Applied Mathematics

ANALYSIS OF ITERATIVE LINE SPLINE COLLOCATION METHODS FOR ELLIPTIC PARTIAL DIFFERENTIAL EQUATIONS*

A. HADJIDIMOS†‡, E. N. HOUSTIS†, J. R. RICE†, AND E. VAVALIS†‡

Abstract. In this paper we present the convergence analysis of iterative schemes for solving linear systems resulting from discretizing multidimensional linear second-order elliptic partial differential equations (PDEs) defined in a hyperparallelepiped Ω and subject to Dirichlet boundary conditions on some faces of Ω and Neumann on the others, using line cubic spline collocation (LCSC) methods. Specifically, we derive analytic expressions or obtain sharp bounds for the spectral radius of the corresponding Jacobi iteration matrix and from this we determine the convergence ranges and compute the optimal parameters for the extrapolated Jacobi and successive overrelaxation (SOR) methods. Experimental results are also presented.

Key words. SOR iterative method, elliptic partial differential equations, collocation methods

AMS subject classifications. 65N35, 65N05, 65F10

PII. S0895479896309827

1. Introduction.

We consider the following second-order linear elliptic PDE,

$$Lu \equiv \sum_{i=1}^{k} \alpha_i \frac{\partial^2 u}{\partial x_i^2} + \sum_{i=1}^{k} \beta_i \frac{\partial u}{\partial x_i} + \gamma u = f \quad \text{in} \quad \Omega, \tag{1.1a}$$

subject to Dirichlet and/or Neumann boundary conditions

$$Bu = g \text{ on } \partial\Omega \equiv \text{boundary of } \Omega, \tag{1.1b}$$

where Bu is u or $\frac{\partial u}{\partial x_i}$, $\Omega \equiv \prod_{i=1}^{k} \otimes [a_i, b_i]$ is a rectangular domain in R^k (the space of k real variables), and $\alpha_i (< 0)$, β_i, $\gamma (\geq 0)$, f, and g are functions of k variables.

If line cubic spline collocation (LCSC) methods are used to solve (1.1a), (1.1b), then the differential operator is discretized along lines in each direction independently and then the line discretization stencils are combined into one large linear system. Although not extensively analyzed yet, collocation methods combine certain desirable algorithmic and computational characteristics of both finite element and finite difference discretization schemes and compete well with them in accuracy, efficiency, and applicability. In particular the LCSC methods exploit the general idea of being alternating direction implicit (ADI), without being real ADI schemes, in a convenient and natural way. In contrast to ADI methods LCSC can be applied to any PDE operator that can be formulated in a tensor product form. Relative to other methods, their efficiency increases with the dimensionality of the PDE problem, and their implementation is relatively easy. Certain additional desirable properties (inherent

*Received by the authors September 29, 1996; accepted for publication by M. Eiermann April 9, 1998; published electronically November 23, 1999. The work was supported in part by NSF awards CCR 9202536 and CDA 9123502, AFOSR award F 49620–J–0069, and ARPA–ARO award DAAH04–94–6-0010.

http://www.siam.org/journals/simax/21-2/30982.html

†Purdue University, Computer Science Department, West Lafayette, IN 47907 (enh@cs.purdue.edu, jrr@cs.purdue.edu, mav@cs.purdue.edu).

‡Authors' permanent address: University of Crete, Mathematics Department, 714 09 Heraklion, Greece, and Institute of Applied and Computational Mathematics, FORTH, 711 10 Heraklion, Greece (hadjidim@math.uch.gr).

parallelism, methodology for choosing the proper directions of the lines, etc.) of LCSC already have been exploited [5]. In section 2 we briefly describe LCSC discretization procedures. We also present and discuss the resulting coefficient linear system of collocation equations which is nonsymmetric and lacks many of the properties found in Ritz–Galerkin-type finite element methods.

In [5], we were able to formulate and analyze iterative schemes for solving the LCSC linear systems in the case of general elliptic second-order PDE with Dirichlet boundary conditions and constant coefficients; that is,

$$Lu \equiv \sum_{i=1}^{k} \alpha_i D_{x_i}^2 u + \gamma u = f \qquad \text{in } \Omega, \tag{1.2a}$$

$$u = g \qquad \text{on } \partial\Omega, \tag{1.2b}$$

where $D_x^2 u$ denotes $\partial^2 u/\partial x^2$. Unfortunately the convergence analysis presented in [5], as it stands, cannot be applied in the case of the presence of Neumann boundary conditions on some of the faces of Ω. In section 3 we present a convergence analysis of block Jacobi, extrapolated Jacobi (EJ), and successive overrelaxation (SOR) schemes for the iterative solution of the collocation equations that arise from discretizing elliptic problems (1.2a) subject to Neumann (N) boundary conditions on one or more (but not on all if $\gamma = 0$) faces of Ω and Dirichlet (D) ones on the others. More specifically, analytic expressions or sharp bounds for the spectral radius of the corresponding block Jacobi iteration matrix are derived and from these we determine the convergence ranges and compute the optimal parameters for the extrapolated Jacobi and SOR methods. Furthermore, based on our analysis, certain conclusions of significant practical importance are drawn.

Finally, in section 4 we present the results of various numerical experiments designed for the verification of the theoretical behavior of the iterative LCSC solvers. The experiments show very good agreement with the theoretical predictions regarding the convergence of the iterative method used. Although the theoretical results presented here hold for the model problem (1.2a), (1.2b), our experimental results indicate that the behavior of the iterative LCSC solvers on the general problem (1.1a), (1.1b) is similar.

2. The LCSC method. In this section we briefly describe the LCSC discretization method and introduce some notation to be used later. We start by introducing one extra point beyond each end of the intervals $[a_i,b_i]$. Each of these enlarged intervals is then discretized uniformly with step size h_i by

$$\Delta_i \equiv \left\{ \tau_\ell^i = a_i + \ell h_i \ ; \ \ell = -1, \ldots, N_i + 1 \text{ with } h_i = \frac{b_i - a_i}{N_i} \right\}.$$

The tensor product of these discretized lines, $\Delta \equiv \prod_{i=1}^{k} \otimes \Delta_i$, provides an extended uniform partition of Ω. A collocation approximation u_Δ of u in the space $S_{3,\Delta}$ of cubic splines in k dimensions is defined by requiring that it satisfies (1.1a) at all the mesh points of Δ and (1.1b) on the boundary mesh points. In the sequel the interior mesh points are denoted by $(\tau_{n_1}^1, \tau_{n_2}^2, \ldots, \tau_{n_k}^k)$ for all $\mathbf{n} \in \prod_{i=1}^{k} \otimes I_i$, where $I_i \equiv \{n_i \ : \ 1 \le n_i \le N_i - 1\}$.

In the line cubic spline collocation methods we consider in this paper the collocation approximation is made on each set $\mathcal{L}^i$ of lines parallel to the x_i axis. More

specifically, this collocation approximation $u^i_\Delta \in S_{3,\Delta}$ on each line in $\mathcal{L}^i$ is represented as follows:

$$u^i_\Delta(\mathbf{x}) = \sum_{\ell=-1}^{N_i+1} U^i_\ell(\mathbf{x}^i) B^i_\ell(x_i), \tag{2.1}$$

where $B^i_\ell(x_i)$ are the B-spline bases defined on Δ_i, $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ and $\mathbf{x}^i = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$. Observe that (2.1) is the sum of one-dimensional splines in the x_i variable whose coefficients $U^i(\mathbf{x}^i)$ are functions of the other $k-1$ variables. Furthermore, this approximation is *redundant* in that there are k choices for representing $u^i_\Delta(\mathbf{x})$, one for each coordinate direction.

2.1. The second-order LCSC method. From the assumed representation (2.1) of u^i_Δ and the nature of the B-spline basis functions we conclude easily (see also [9] and [5]) that

$$D^2_{x_i} u^i_\Delta|_{\mathbf{P}^i_\ell} = [U^i_{\ell-1}(\mathbf{n}_i) - 2U^i_\ell(\mathbf{n}_i) + U^i_{\ell+1}(\mathbf{n}_i)]/h_i^2, \tag{2.2}$$

$$D_{x_i} u^i_\Delta|_{\mathbf{P}^i_\ell} = [U^i_{\ell-1}(\mathbf{n}_i) - U^i_{\ell+1}(\mathbf{n}_i)]/2h_i, \tag{2.3}$$

and

$$u^i_\Delta|_{\mathbf{P}^i_\ell} = [U^i_{\ell-1}(\mathbf{n}_i) + 4U^i_\ell(\mathbf{n}_i) + U^i_{\ell+1}(\mathbf{n}_i)]/6 \tag{2.4}$$

at the mesh (collocation) points $\mathbf{P}^i_\ell$ for $\ell = 1, 2, \ldots, N_i - 1$ on each line of $\mathcal{L}^i$. These points are represented by the vectors

$$\mathbf{P}^i_\ell \equiv (\tau^1_{n_1}, \ldots, \tau^{i-1}_{n_{i-1}}, \tau^i_\ell, \tau^{i+1}_{n_{i+1}}, \ldots, \tau^k_{n_k}), \quad \ell = 0, \ldots, N_i,$$

where $\mathbf{n}_i \equiv (n_1, \ldots, n_{i-1}, n_{i+1}, \ldots, n_k)$ indexes the $\mathcal{L}^i$ lines.

First, we observe that the collocation equations obtained from the boundary conditions and the differential equation at the end points of each line can be explicitly determined as follows.

Dirichlet (D) boundary conditions on $\mathbf{P}^i_0$:

$$U^i_0 = \frac{h_i^2 f(\mathbf{P}^i_0)}{6\alpha_i}, \qquad U^i_{-1} = -U^i_1 + \frac{2h_i^2 f(\mathbf{P}^i_0)}{3\alpha_i}; \tag{2.5}$$

Neumann (N) boundary conditions on $\mathbf{P}^i_0$:

$$U^i_0 = U^i_1 + \frac{h_i^2 f(\mathbf{P}^i_0)}{2\alpha_i}, \qquad U^i_{-1} = U^i_1; \tag{2.6}$$

Dirichlet (D) boundary conditions on $\mathbf{P}^i_{N_i}$:

$$U^i_{N_i} = \frac{h_i^2 f(\mathbf{P}^i_{N_i})}{6\alpha_i}, \qquad U^i_{N_i+1} = -U^i_{N_i-1} + \frac{2h_i^2 f(\mathbf{P}^i_{N_i})}{3\alpha_i}; \tag{2.7}$$

Neumann (N) boundary conditions on $\mathbf{P}^i_{N_i}$:

$$U^i_{N_i} = U^i_{N_i-1} + \frac{h_i^2 f(\mathbf{P}^i_{N_i})}{2\alpha_i}, \quad U^i_{N_i+1} = U^i_{N_i-1}, \tag{2.8}$$

where for simplicity we have assumed that all boundary conditions are homogeneous. The rest of the unknowns are determined by solving the so-called *interior collocation equations*, that is, on those lines of $\mathcal{L}^i$ not in $\partial\Omega$,

$$\left(\sum_{i=1}^{k} \alpha_i \, D^2_{x_i} \, u^i_\Delta + \gamma u^1_\Delta\right)\Bigg|_{\mathbf{P}^i_\ell} = f\,|_{\mathbf{P}^i_\ell} \quad \text{for} \quad \ell = 1, 2, \ldots, N_i - 1. \tag{2.9}$$

Equations (2.9) *away* ($2 \le \ell \le N_i - 2$) from the boundary are given by

$$\begin{aligned} &\sum_{i=1}^{k} \frac{\alpha_i}{h_i^2}\left[U^i_{\ell-1}(\mathbf{n}_i) - 2U^i_\ell(\mathbf{n}_i) + U^i_{\ell+1}(\mathbf{n}_i)\right] \\ &+ \frac{\gamma}{6}[U^1_{\ell-1}(\mathbf{n}_1) + 4U^1_\ell(\mathbf{n}_1) + U^1_{\ell+1}(\mathbf{n}_1)] = f|_{\mathbf{P}^i_\ell}, \end{aligned} \tag{2.10}$$

while for lines next to the boundary they have similar form with appropriately modified right sides [5]. The full matrix form of these equations is given in section 3. After eliminating the predetermined (from equations (2.5)–(2.8)) boundary unknowns, each collocation approximation $u^i_\Delta(\mathbf{x})$ involves $\mathcal{K} = \Pi_i(N_i - 1)$ unknown coefficients U^i_ℓ. This redundancy is handled by requiring that all these approximations agree on the mesh points; that is,

$$u^1_\Delta = u^2_\Delta = \cdots = u^k_\Delta \quad \text{on the mesh} \quad \Delta. \tag{2.11}$$

It has been shown [5, 7, 9] that the above-described method leads to a second-order collocation approximation of u.

2.2. The fourth-order LCSC method. To derive the fourth-order LCSC method we use high-order approximations of the derivatives $D^j_{x_i} u$, $j = 1, 2$, defined as appropriate linear combinations of the spline interpolant and its derivatives at the mesh points [7]. Specifically, we approximate the second derivatives in the PDE operator by the difference scheme

$$\begin{aligned} &D^2_{x_i} u^i_\Delta \Big|_{\mathbf{P}^i_\ell} \\ &= \left(U^i_{\ell-2}(\mathbf{n}_i) + 8U^i_{\ell-1}(\mathbf{n}_i) - 18U^i_\ell(\mathbf{n}_i) + 8U^i_{\ell+1}(\mathbf{n}_i) + U^i_{\ell+1}(\mathbf{n}_i)\right)/(12h_i^2). \end{aligned} \tag{2.12}$$

The collocation equations corresponding to the mesh points on a line $\mathcal{L}^i$ *away* ($2 \le \ell \le N_i - 2$) from the boundary are written at the point $\mathbf{P}^i_\ell$ as

$$\begin{aligned} &\sum_{i=1}^{k} \frac{\alpha_i}{12h_i^2}\left[U^i_{\ell-2}(\mathbf{n}_i) + 8U^i_{\ell-1}(\mathbf{n}_i) - 18U^i_\ell(\mathbf{n}_i) + 8U^i_{\ell+1}(\mathbf{n}_i) + U^i_{\ell+2}(\mathbf{n}_i)\right] \\ &+ \frac{\gamma}{6}[U^1_{\ell-1}(\mathbf{n}_i) + 4U^1_\ell(\mathbf{n}_i) + U^1_{\ell+1}(\mathbf{n}_1)] = f|_{\mathbf{P}^i_\ell}. \end{aligned} \tag{2.13}$$

The redundancy on the coefficients is handled in the same way as in the second-order case, with the only basic difference being that the stencils now have five, rather than three, points along each coordinate direction.

3. Iterative solution of the interior LCSC equations. The LCSC equations can be written in the form

$$\sum_{i=1}^{k} A_i \mathbf{U}^i(\mathbf{n}) = \mathbf{F}^1, \tag{3.1}$$

where the coefficient matrices are defined by

$$A_i \equiv \left(\prod_{j=1}^{k-i} \otimes I\right) \otimes \mathcal{E}_i \otimes \left(\prod_{j=k-i+2}^{k} \otimes I\right), \ i = 1, \ldots, k, \tag{3.2}$$

where I denotes a unit matrix of appropriate order and

$$\mathcal{E}_1 = \frac{\alpha_1}{h_1^2}\mathcal{H}_1 + \frac{\gamma}{6}\mathbf{T}_4^1 \quad \text{and} \quad \mathcal{E}_i = \frac{\alpha_i}{h_i^2}\mathcal{H}_i \ , \ i = 2, \ldots, k. \tag{3.3}$$

The matrix $\mathcal{H}_i$ depends both on the discretization scheme applied and the nature (Dirichlet (D) or Neumann (N)) of the boundary conditions on the *left* and *right* end points of the line direction with which the matrix $\mathcal{H}_i, i = 1, \ldots, k$, is associated. Specifically, $\mathcal{H}_i = T^i_{\alpha,\beta}$ for the second-order discretization and $\mathcal{H}_i = \frac{1}{12}T^i_{10}T^i_{\alpha,\beta}$ for the fourth-order one. The matrices $T^i_{\alpha,\beta}$, T^i_4, and T^i_{10} are given by

$$T^i_{\alpha,\beta} := \begin{pmatrix} \alpha & 1 & & & & \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & & \ddots & & \\ & & & 1 & -2 & 1 \\ & & & & 1 & \beta \end{pmatrix}_{(N_i-1)\times(N_i-1)}, \tag{3.4}$$

$$T^i_4 = 6I + T^i_{\alpha,\beta} \text{ and } T^i_{10} = 12I + T^i_{-2,-2},$$

where

$$(\alpha, \beta) = \begin{cases} (-2,-2) & \text{in case of (D) conditions on both ends,} \\ (-1,-2) & \text{in case of (N) conditions on the } \textit{left} \text{ and} \\ & \text{(D) conditions on the } \textit{right} \text{ end,} \\ (-2,-1) & \text{in case of (D) conditions on the } \textit{left} \text{ and} \\ & \text{(N) conditions on the } \textit{right} \text{ end,} \\ (-1,-1) & \text{in case of (N) conditions on both ends.} \end{cases} \tag{3.5}$$

Note. Since in the analysis that will follow the existence of A_1^{-1} is a necessary requirement, we assume, without loss of generality, that on at least one of the faces of Ω perpendicular to the x_1-direction (D) boundary conditions have been imposed so that the invertibility of A_1 is guaranteed even if $\gamma = 0$.

Equation (3.1), coupled with the $k-1$ interpolation conditions (2.11) leads us to the order-kK system of linear equations which, if we order according to the ordering of unknowns $\mathbf{U}^1(\mathbf{n})$, is given by

$$\begin{pmatrix} A_1 & A_2 & A_3 & \ldots & A_k \\ -B_2 & D_2 & & & \\ -B_3 & & D_3 & & \\ \vdots & & & \ddots & \\ -B_k & & & & D_k \end{pmatrix} \begin{pmatrix} \mathbf{U}^1 \\ \mathbf{U}^2 \\ \mathbf{U}^3 \\ \vdots \\ \mathbf{U}^k \end{pmatrix} = \begin{pmatrix} \mathbf{F}^1 \\ \mathbf{F}^2 \\ \mathbf{F}^3 \\ \vdots \\ \mathbf{F}^k \end{pmatrix}, \tag{3.6}$$

where for $i = 2, \ldots, k$ we have that

$$B_i = \left(\prod_{j=1}^{k-1} \otimes I\right) \otimes T_4^1 \ , \ D_i = \left(\prod_{j=1}^{k-i} \otimes I\right) \otimes T_4^i \otimes \left(\prod_{j=k-i+2}^{k} \otimes I\right). \tag{3.7}$$

The matrices in (3.3) and (3.7) may be linear or quadratic functions of the matrix $T^i_{-2,-2}$ if *all* the boundary conditions are Dirichlet (D). With some Neumann (N) conditions present, this is no longer the case. Besides, in the case of the fourth-order discretization scheme, the product $T^i_{10}T^i_{\alpha,\beta}$ is not a symmetric matrix in the case where $(\alpha,\beta) \neq (-2,-2)$. So, the analysis presented in section 3 of [5] does not always apply to the present case. We present a unified analysis for all cases based on the following three lemmas. The first one is a well-known result. For the proofs of the second and the third the reader is referred to [11] and [6, Theorem 4.2.12].

LEMMA 3.1. *If A and B are Hermitian positive definite matrices of the same order, then* (i) *AB (and BA) possesses a complete set of linearly independent eigenvectors.* (ii) *Further, let $X \in \mathcal{C}^{n,n}$ and $\sigma(X) \subset \mathcal{R}$, where $\sigma(.)$ denotes the spectrum of a matrix. Let also λ_X, $\lambda_{X,m}$, and $\lambda_{X,M}$ denote any, the smallest, and the largest eigenvalue of X, respectively. Then for the eigenvalues of the matrix AB there holds*

$$0 < \lambda_{A,m}\lambda_{B,m} \leq \lambda_{AB,m} \leq \lambda_{AB,M} \leq \lambda_{A,M}\lambda_{B,M}. \tag{3.8}$$

Note. The assertions of Lemma 3.1 hold even if one of A or B is nonnegative definite. The only difference is that the two leftmost inequalities in (3.8) become equalities; that is, $0 = \lambda_{A,m}\lambda_{B,m} = \lambda_{AB,m}$, since $\det(AB) = \det(A)\det(B) = 0$.

LEMMA 3.2. *Consider the real symmetric tridiagonal matrix $T_{\alpha,\beta}$ of order m given by*

$$T_{\alpha,\beta} := \begin{pmatrix} \alpha & 1 & & & & & \\ 1 & -2 & 1 & & & & \\ & 1 & -2 & 1 & & & \\ & & & \ddots & & & \\ & & & 1 & -2 & 1 \\ & & & & 1 & \beta \end{pmatrix}, \tag{3.9}$$

where $(\alpha,\beta) = (-2,-2), (-2,-1), (-1,-2)$, or $(-1,-1)$. Its eigenvalues $\lambda_{T_{\alpha,\beta}}$ are given for $\ell = 0,\ldots,m-1$ by the expressions

$$\begin{aligned} &\text{(i) } \textit{If } (\alpha,\beta) = (-2,-2), \textit{ then } \lambda_{T_{\alpha,\beta}} = -4\sin^2\left(\frac{(\ell+1)\pi}{2(m+1)}\right), \\ &\text{(ii) } \textit{If } (\alpha,\beta) = (-2,-1) \textit{ or } (-1,-2), \textit{ then } \lambda_{T_{\alpha,\beta}} = -4\sin^2\left(\frac{(2\ell+1)\pi}{2(2m+1)}\right), \\ &\text{(iii) } \textit{If } (\alpha,\beta) = (-1,-1), \textit{ then } \lambda_{T_{\alpha,\beta}} = -4\sin^2\left(\frac{\ell\pi}{2m}\right). \end{aligned} \tag{3.10}$$

LEMMA 3.3. *Let the $n \times n$ matrices A_i, $i = 1,\ldots,k$, possess complete sets of linearly independent eigenvectors $\mathbf{y}^{(i,j)}$, with corresponding eigenvalues $\lambda^{(j)}_{A_i}$ ($j = 1,\ldots,k$, $i = 1,\ldots,k$). Then the matrix $A \equiv A_1 \otimes A_2 \otimes \cdots \otimes A_k$ possesses the n^k linearly independent eigenvectors $\mathbf{y}^{(\underline{j})} \equiv \mathbf{y}^{(1,j_1)} \otimes \mathbf{y}^{(2,j_2)} \otimes \cdots \otimes \mathbf{y}^{(k,j_k)}$, where $\underline{j} \equiv (j_1,\ldots,j_k)$ and $j_i = 1,\ldots,n$, $i = 1,\ldots,k$, with corresponding eigenvalues $\lambda^{(\underline{j})} = \lambda^{(j_1)}_{A_i}\lambda^{(j_2)}_{A_2}\ldots\lambda^{(j_k)}_{A_k}$.*

Having developed the background material required we are able to go on with the analysis of the three methods associated with the linear system (3.6): block Jacobi (J),

$$K\mathbf{U}^{(s+1)} = (L+M)\mathbf{U}^{(s)} + \mathbf{F}\ ,\ s = 0,1,2,\ldots;$$

block EJ,

$$K\mathbf{U}^{(s+1)} = (1-\omega)K\mathbf{U}^{(s)} + \omega(L+M)\mathbf{U}^{(s)} + \omega\mathbf{F}\ ,\ s = 0, 1, 2, \ldots;$$

and block SOR,

$$K\mathbf{U}^{(s+1)} = (1-\omega)K\mathbf{U}^{(s)} + \omega L\mathbf{U}^{(s+1)} + \omega M\mathbf{U}^{(s)} + \omega\mathbf{F}\ ,\ s = 0, 1, 2, \ldots,$$

where

$$K \equiv \mathrm{diag}(A_1, D_2, D_3, \ldots, D_k),$$

and $-L$ and $-M$ are the strictly lower and the strictly upper triangular parts of the matrix coefficient in (3.6).

The block Jacobi iteration matrix $J \equiv K^{-1}(L+M) = \{J_{i,j}\}$, associated with the matrix coefficient in (3.6), can be described as

$$\begin{aligned} J_{1,1} = 0,\ \{\{J_{i,j}\}_{j=2}^{k}\}_{i=2}^{k} = 0,\ \{J_{i,j} := H_j = -A_1^{-1}A_j\}_{j=2}^{k}, \\ \{J_{i,1} := G_i = D_i^{-1}B_i\}_{i=2}^{k}. \end{aligned} \tag{3.11}$$

Let H and G denote the matrices

$$H := [H_2, H_3, \ldots, H_k]\,,\ G := \left[G_2^T, G_3^T, \ldots, G_k^T\right]^T \tag{3.12}$$

of dimensions $\mathcal{K} \times (k-1)\mathcal{K}$ and $(k-1)\mathcal{K} \times \mathcal{K}$, respectively. Consider then the matrix J^2 which, apparently, is a block diagonal matrix with diagonal blocks HG and GH of orders $\mathcal{K}$ and $(k-1)\mathcal{K}$, respectively. Therefore, we have (see [12, Theorem 1.12, p. 16]) that

$$\sigma(J^2) \setminus \{0\} = \sigma(HG) \setminus \{0\} = \sigma(GH) \setminus \{0\}. \tag{3.13}$$

However, we have

$$HG = \sum_{j=2}^{k} H_j G_j \tag{3.14}$$

and each term in the above sum can be found by using (3.11), (3.2), (3.7) and simple tensor product properties to be

$$H_jG_j = -\Pi_{\ell=1}^{k-j}(\otimes I) \otimes (\mathcal{H}_j D_j^{-1}) \otimes (\Pi_{\ell=k-j+2}^{k-1} \otimes I) \otimes (A_1^{-1}T_4^1),\ j = 2, \ldots, k. \tag{3.15}$$

For the second-order LCSC, we introduce the matrices S_j and find their value to be

$$S_j := \mathcal{H}_j D_j^{-1} = \frac{\alpha_j}{h_j^2} T_{\alpha,\beta}^j \left(6I + T_{\alpha,\beta}^j\right)^{-1},\ j = 2, \ldots, k, \tag{3.16}$$

$$S_1 := A_1^{-1}T_4^1 = \left[\frac{\alpha_1}{h_1^2} T_{\alpha,\beta}^1 \left(6I + T_{\alpha,\beta}^1\right)^{-1} + \frac{\gamma}{6} I\right]^{-1}. \tag{3.17}$$

For the fourth-order LCSC, we have similarly

$$S_j := \mathcal{H}_j D_j^{-1} = \frac{1}{12}\left(12I + T_{-2,-2}^j\right) \frac{\alpha_j}{h_j^2} T_{\alpha,\beta}^j \left(6I + T_{\alpha,\beta}^j\right)^{-1},\ j = 2, \ldots, k, \tag{3.18}$$

$$S_1 := A_1^{-1}T_4^1 = \left[\left(6I + T^1_{\alpha,\beta}\right)^{-1} \frac{1}{12}\left(12I + T^1_{-2,-2}\right) \frac{\alpha_1}{h_1^2} T^1_{\alpha,\beta} + \frac{\gamma}{6} I\right]^{-1}. \tag{3.19}$$

Due to the presence of the unit matrix factors in the tensor product form (3.15), all H_jG_j's possess a linearly independent set of $\mathcal{K}$ common eigenvectors if and only if each $\mathcal{H}_j D_j^{-1}$, $j = 2, \ldots, k$, and $A_1^{-1}T_4^1$ possess complete sets of linearly independent eigenvectors. For the matrices in (3.17) this is obvious because $T^j_{\alpha,\beta}$, $j = 2, \ldots, k$, is real symmetric negative (or nonpositive) definite, $T^1_{\alpha,\beta}$ is real symmetric negative definite, and, in addition, $\sigma(T^j_{\alpha,\beta}) \subset (-4, 0]$, $j = 1, \ldots, k$ (see Lemma 3.2), $\alpha_j < 0$, $j = 1, \ldots, k$, and $\gamma \geq 0$. As one can readily see, each of the matrices S_j in (3.18) is the product of the two real symmetric positive definite matrices $\frac{1}{12}(12I + T^j_{-2,-2})$ and $\frac{\alpha_j}{h_j^2} T^j_{\alpha,\beta}(6I + T_{\alpha,\beta})^{-1}$. (The second matrix factor might also be nonnegative definite.) Therefore Lemma 3.1 (or its note) applies. For the matrix $A_1^{-1}T_4^1$ in (3.19) first observe that the first matrix term in the brackets is similar to

$$\frac{1}{12}(12I + T^1_{-2,-2}) \frac{\alpha_1}{h_1^2} T^1_{\alpha,\beta}(6I + T^1_{\alpha,\beta})^{-1},$$

which, in turn, is of exactly the same form as the matrices S_j in (3.18), except that the second matrix factor considered previously is now always positive definite. Consequently, in all possible cases of the LCSC matrices, all terms $H_j\ G_j$ in (3.15) possess a linearly independent set of $\mathcal{K}$ common eigenvectors. By virtue of this result and in view of Lemma 3.3, it is implied from (3.15) that

$$\lambda_{HG} = -\lambda_{S_1} \sum_{j=2}^{k} \lambda_{S_j}, \tag{3.20}$$

where λ_X is used to denote any eigenvalue of the matrix X. However, from the previous discussion it follows that $\lambda_{S_i} \geq 0$, $j = 2, \ldots, k$, while $\lambda_{S_1} > 0$. Therefore $\lambda_{HG} \leq 0$. So, from (3.13) it is implied that J^2 possesses nonpositive eigenvalues and hence the block Jacobi matrix J has a purely imaginary spectrum. From the analysis so far it becomes clear that the eigenvalues of J^2 and therefore those of J can be given analytically in the following cases:

(1) In all the cases of the second-order LCSC we are dealing with when Dirichlet *and/or* Neumann boundary conditions are imposed on the faces of $\partial\Omega$.
(2) In the fourth-order one when *only* Dirichlet boundary conditions are imposed on the faces of $\partial\Omega$.

This is an immediate consequence of the fact that each matrix S_j, $j = 1, \ldots, k$, in (3.16)–(3.19) is a simple *real rational* matrix function of the matrix $T^j_{\alpha,\beta}$, $j = 1, \ldots, k$, in case (1) and of the matrix $T^j_{\alpha,\beta}$, $j = 2, \ldots, k$, and $T^1_{-2,-2}$ in case (2), respectively.

Having in mind the various conclusions we have arrived at in the analysis so far, one can state the following theorem which gives analytic expressions for the eigenvalues of the block Jacobi matrix J in (3.11).

THEOREM 3.4. *The eigenvalues μ of the block Jacobi iteration matrix $K^{-1}(L+M)$ defined in* (3.11) *are as follows. We have $\mu = 0$ of multiplicity $(k-2)\mathcal{K}$. For the second-order collocation scheme the others are pure imaginary:*

$$\mu = \pm i \left\{ \frac{\sum_{j=2}^{k} \frac{\alpha_j}{h_j^2} \lambda^j_{\alpha,\beta}(6 + \lambda^j_{\alpha,\beta})^{-1}}{\frac{\alpha_1}{h_1^2} \lambda^1_{\alpha,\beta}(6 + \lambda^1_{\alpha,\beta})^{-1} + \frac{\gamma}{6}} \right\}^{1/2}. \tag{3.21}$$

For the fourth-order scheme, where (2.1) *is assumed to be subjected to Dirichlet boundary conditions only, the others are pure imaginary:*

$$\mu = \pm i \left\{ \frac{\sum_{j=2}^{k} \frac{1}{12}(12+\lambda_{-2,-2}^{j})\frac{\alpha_j}{h_j^2}\lambda_{-2,-2}^{j}(6+\lambda_{-2,-2}^{j})^{-1}}{\frac{1}{12}(12+\lambda_{-2,-2}^{1})\frac{\alpha_1}{h_1^2}\lambda_{-2,-2}^{1}(6+\lambda_{-2,-2}^{1})^{-1}+\frac{\gamma}{6}} \right\}^{-1/2}. \tag{3.22}$$

In (3.21) *and* (3.22) $\lambda_{\alpha,\beta}^{j}$, $j=1,\ldots,k$, *with* $(\alpha,\beta)=(-2,-2)$, $(-2,-1)$, $(-1,-2)$, $(-1,-1)$, *are the* $n=N_j-1$ *eigenvalues of the matrix* $T_{\alpha,\beta}^{j}$, $j=1,\ldots,k$, *as obtained by the application of Lemma* 3.2. □

Note. It should be pointed out that analytic expressions for the eigenvalues of J *cannot* be derived, in terms of the matrices $T_{\alpha,\beta}^{j}$ involved, in the case of the fourth-order LCSC where at least on one face of $\partial\Omega$ the Neumann (N) boundary conditions are imposed. This is due to the noncommutativity of the matrices $T_{-2,-2}^{j}$ and $T_{\alpha,\beta}^{j}$ for $(\alpha,\beta)\neq(-2,-2)$. However, one can trivially give analytic expressions in terms of the eigenvalues of the matrices $(12I+T_{2,-2}^{j})\,T_{\alpha,\beta}^{j}\,(6I+T_{\alpha,\beta}^{j})^{-1}$, $j=1,\ldots,k$, and also, by virtue of Lemma 3.1, lower and upper bounds can be given for the eigenvalues of H_j G_j, $j=1,\ldots,k$, in terms of the extreme eigenvalues of $T_{-2,-2}^{j}$ and $T_{\alpha,\beta}^{j}$, $j=1,\ldots,k$, of Lemma 3.2.

To derive the spectral radii of the Jacobi matrices of Theorem 3.4 and also upper bounds in the case of the previous note, we introduce some notation first. Let, then, $\lambda_{\alpha,\beta}^{j}$ denote an eigenvalue of the matrices $T_{\alpha,\beta}^{j}$, $j=1,\ldots,k$, as in Theorem 3.4. To simplify the notation we omit using another index on the generic $\lambda_{\alpha,\beta}^{j}$ and we omit the index j from the pair of subscripts (α,β). Let also

$$c_j := c_j(\alpha,\beta) := \min \lambda_{\alpha,\beta}^{j}, \tag{3.23}$$

$$s_j := s_j(\alpha,\beta) := \max \lambda_{\alpha,\beta}^{j} \tag{3.24}$$

denote the smallest and the largest eigenvalues in $\lambda_{\alpha,\beta}^{j}$, $j=1,\ldots,k$, respectively. Finally, define the functions

$$y_j := y_j(\lambda_{\alpha,\beta}^{j}) := \frac{\alpha_j}{h_j^2}\lambda_{\alpha,\beta}^{j}(6+\lambda_{\alpha,\beta}^{j})^{-1}, \tag{3.25}$$

$$z_j := z_j(\lambda_{-2,-2}^{j},\lambda_{\alpha,\beta}^{j}) := \frac{1}{12}(12+\lambda_{-2,-2}^{j})y_j \tag{3.26}$$

in terms of the eigenvalues $\lambda_{\alpha,\beta}^{j}$ of the matrices $T_{\alpha,\beta}^{j}$, $j=1,\ldots,k$. Then we have Theorem 3.5

THEOREM 3.5. (i) *The squares of the spectral radii of the block Jacobi matrices corresponding to the collocation schemes considered in Theorem* 3.4 *are given by the following expressions.*

Second order:

$$\rho^2(J) \equiv \sum_{j=2}^{k} y_j(c_j) \Big/ \left(y_1(s_1)+\frac{\gamma}{6}\right). \tag{3.27}$$

Fourth order:

$$\rho^2(J) \equiv \frac{\sum_{j=2}^{k} z_j\,(c_j(-2,-2),c_j(\alpha,\beta))}{z_1\,(s_1(-2,-2),s_1(\alpha,\beta))+\frac{\gamma}{6}}. \tag{3.28}$$

(ii) *Moreover the expression* (3.28) *is also a strict upper bound for the square of the spectral radius of the block Jacobi matrix in the case of the fourth-order scheme corresponding to an elliptic problem* (1.2a) *where at least on one of the faces of* $\partial\Omega$ *the Neumann* (N) *boundary conditions are imposed.*

Proof. We notice that in (3.21) and (3.22) $\alpha_j < 0$, $\lambda^j_{\alpha,\beta} \in [c_j, s_j]$, $j = 1, \ldots, k$, and $\gamma \geq 0$. Hence the functions $y_j := y_j(\lambda^j_{\alpha,\beta}) := y_j(\lambda)$ in (3.23) and (3.24) are nonnegative and independent of each other, and so are the expressions $z_j := z_j(\lambda^j_{-2,-2}, \lambda^j_{-2,-2}) := z(\lambda, \lambda)$ in (3.26). To determine the spectral radius of the block Jacobi matrix of the second-order collocation scheme we examine the extreme values of $y_j(\lambda)$, $\lambda \in [c_j, s_j]$, $j = 1, \ldots, k$. By differentiation we have

$$\frac{\partial y_j}{\partial \lambda} = \frac{6}{(6+\lambda)^2} \frac{\alpha_j}{h_j^2} < 0$$

and, consequently,

$$\max y_j = y_j(c_j), \quad \min y_j = y_j(s_j). \tag{3.29}$$

For the corresponding quantity of the fourth-order scheme of Theorem 3.4, working in a similar way, we obtain

$$\frac{\partial z_j}{\partial \lambda} = \frac{1}{12(6+\lambda)^2} \frac{\alpha_j}{h_j^2} [(6+\lambda)^2 + 36] < 0,$$

which implies that

$$\begin{aligned} \max z_j &= z_j \left(c_j(-2,-2), c_j(-2,-2)\right), \\ \min z_j &= z_j \left(s_j(-2,2), s_j(-2,-2)\right). \end{aligned} \tag{3.30}$$

It is obvious now using (3.21)–(3.26) and (3.29), (3.30) that the results (3.27) and (3.28) follow. This concludes the proof for part (i) of the theorem.

For part (ii) we use the analysis preceding the statement of Theorem 3.4 and refer to the matrices in (3.18) and (3.19), especially for those indices $j = 1, \ldots, k$ for which $(\alpha, \beta) \neq (-2, -2)$, and also recall the note immediately following Theorem 3.4. It follows from this and Lemma 3.1, its note, and Lemma 3.2 that the lower and upper bounds for the eigenvalues of the matrices in (3.18) and (3.19) depend directly on the extreme eigenvalues of the two matrices

$$\frac{1}{12}(12I + T^j_{-2,-2}) \quad \text{and} \quad \frac{\alpha_j}{h_j^2} T^j_{\alpha,\beta} (6I + T^j_{\alpha,\beta})^{-1}.$$

These are readily seen to be

$$\frac{1}{12}(12 + c_j(-2,2)) \quad \text{and} \quad \frac{1}{12}(12 + s_j(-2,-2))$$

for the minimum and the maximum eigenvalues for the former matrix and $y_j(c_j)$ and $y_j(s_j)$ for the latter one, where c_j, s_j, and y_j are the expressions in (3.23) and (3.24). From Lemma 3.1, its note, and using the expressions (3.26), the bound for the spectral radius of the corresponding block Jacobi matrix is readily shown to be the expression in the right-hand side of (3.28). □

Remark. The analysis so far has been made on the assumption that the x_1-direction is somehow predetermined. However, since there are k possible choices for

the x_1-direction in any particular case, the choice should be made in such a way as to give the smallest possible values in (3.27) and (3.28). Consider then the quantities

$$\min_i \left(\sum_{\substack{j=1 \\ j\neq i}}^{k} y_j(c_j) / \left(y_i(s_i) + \frac{\gamma}{6} \right) \right) \tag{3.31}$$

taken over all $i = 1, \ldots, k$ for which they are well defined and also the quantities

$$\min_i \left(\sum_{\substack{j=1 \\ j\neq i}}^{\kappa} \frac{z_j\left(c_j(-2,2), c_j(\alpha,\beta)\right)}{z_i\left(s_i(-2,-2), s_i(\alpha,\beta)\right) + \frac{\gamma}{6}} \right) \tag{3.32}$$

considered in the same way. The indices i in (3.31) and (3.32) for which the corresponding minima take place should be interchanged with the index 1 and therefore the x_i-direction should be taken as the x_1-direction. If, in either case (3.31) or (3.32), more than one index i gives the same minimum value, then any such i will do. It should be noted that after having chosen the x_1-direction this way, the expressions in (3.27) and (3.28) give then the smallest possible values for the spectral radius or for an upper bound on it, as was explained, for the particular problem at hand. □

From this point on, the analysis is almost identical with that in [5], and the interested reader is referred to it. For completeness, we simply mention some of the theoretical results obtained in [5], which are based on the corresponding theory developed in [12, 13] and in [1, 2, 3, 4, 10].

(i) The block EJ method and the block SOR method corresponding to the block Jacobi method of this paper converge for values of their parameters varying in some open intervals whose left end is 0 and the right one is a function of $\rho(J)$. However, the optimum SOR is always superior to the optimum EJ, and the corresponding optimal parameters for the SOR method are given by

$$\omega_{opt} = \frac{2}{1 + \sqrt{1 + \rho^2(J)}}, \ \rho(\mathcal{L}_{\omega_{opt}}) = 1 - \omega_{opt}. \tag{3.33}$$

(ii) For the efficiency of both the serial and the parallel iterative solution of the linear system (3.6), a cyclic natural ordering of the unknowns U^i, $i = 2, \ldots, k$, is adopted according to which

$$\mathcal{U}^i \equiv \left\{ \ldots \left\{ \left\{ \ldots \left\{ U^i_{j_1,\ldots,j_{i-1},j_i,\ldots,j_k} \right\}_{j_i=1}^{N_i-1} \cdots \right\}_{j_k=1}^{N_k-1} \right\}_{j_1=1}^{N_1-1} \cdots \right\}_{j_{i-1}=1}^{N_{i-1}-1}. \tag{3.34}$$

The equations in (3.1) are reordered according to the ordering of $\mathcal{U}^1$, and each block of the auxiliary conditions is reordered according to the ordering of the unknowns $\mathcal{U}^i$. It is then proved that the new coefficient matrix $\mathcal{A}$ is obtained by a permutation similarity transformation of the matrix coefficient A in (3.6) having the same $k \times k$ block structure, and, therefore, the associated block Jacobi matrix $\mathcal{J}$ is similar to the previous one J. Consequently, the convergence results are identically the same so that all the formulas in connection with eigenvalues, spectral radii, etc., of the Jacobi, the EJ, and the SOR method studied in this section remain unchanged when these reorderings are made.

(iii) The new structure of the collocation coefficient matrix $\mathcal{A}$, for the second-order and fourth-order scheme in two dimensions, is given in [5].

4. Numerical experiments. In this section we summarize the results of some numerical experiments that verify the convergence properties of the iterative solution methods analyzed in section 3. We mention that although we present numerical data for only the $O(h^2)$, 3-dimensional case, these are very representative of problems with different dimensionality or with $O(h^4)$ discretization schemes. For experimental data on the convergence properties of the LCSC method and the iterative solvers in the case of only Dirichlet boundary conditions the reader is referred to [5] and [7]. The parallel implementation details and the performance of the iterative LCSC schemes, for two-dimensional problems, on several multiprocessing systems can be found in [8].

We have applied the LCSC discretization techniques with uniform (NGRID by NGRID by NGRID) meshes to approximate the known solution of the following set of PDEs:

PDE 1. $D_x^2u + D_y^2u + D_z^2u = -f$.

PDE 2. $4D_x^2u + 2D_y^2u + D_z^2u - 10u = -f$.

PDE 3. $D_x^2u + \frac{.1+x-y+z}{8}D_y^2u + \frac{x+z}{9}D_z^2u + e^{x+y+z}u = -f$.

Each is defined on the unit cube and subject to one of the following types of boundary conditions:

BC 1. Dirichlet conditions on all faces of Ω.

BC 2. Neumann conditions on the faces $y = 1$ and $z = 0$ of Ω and Dirichlet ones on the rest.

BC 3. Neumann conditions on the face $x = 1$ of Ω and Dirichlet on the rest.

These PDEs and boundary conditions are combined to give nine problems; our theory is applicable only to six of these (those excluding PDE 3). The right-hand side f is selected so that the true solution is always

$$u(x, y, z) = e^{x+y+z}(x^2 - x)(y^2 - y)(z^2 - z).$$

The linear systems from the LCSC discretization were solved by the proposed SOR iteration method with the termination criterion being that $||\mathbf{U}^{(s+1)} - \mathbf{U}^{(s)}||_\infty$ is within the interval $(0, 10^{-7})$. All experiments were performed, in double precision, on a SUN4-110 workstation.

In Figure 4.1 we present the theoretically estimated (using the material developed in section 3 when applicable) and the experimentally determined (by systematic search) values of the optimum SOR relaxation parameter ω_{opt}. Specifically, the points we plot are the experimentally observed optimal values of ω for various values of $NGRID$. The lines we plot show the relation (determined using (3.33) and (3.28) of Theorem 3.5) between ω_{opt} and the discretization parameter $NGRID$. Since our theory cannot be directly used to determine such relation in the case of PDE 3 we plot only the experimentally determined ω_{opt}.

Our first observation is that there is a good agreement between our theory and the experiments in the six cases where it applies. The theoretical values for ω_{opt} are close to (although always larger than) the measured ones and exhibit the same dependence on the discretization and PDE problem parameters. Besides confirming our theory, these experiments also show that for PDEs where our analysis is not applicable the proposed SOR scheme still converges. Furthermore, relaxing with ω_{opt} determined by our theory leads us to comparable rates of convergence.

Another interesting observation is that ω_{opt} seems to go fast and asymptotically to a number in the interval $[.03, .08]$. It is therefore expected that the rate of convergence will not decrease rapidly as $NGRID$ increases beyond 30. This is confirmed

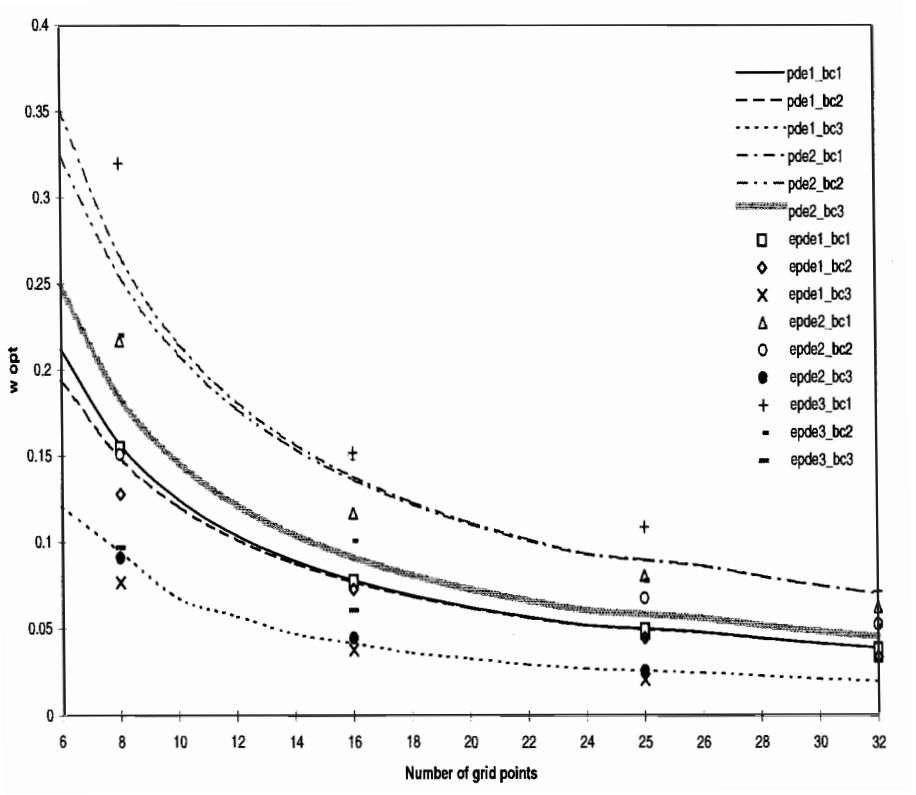


FIG. 4.1. *The graph of the ω_{opt} vs. the number of grid points for the second-order LCSC method applied to the three three-dimensional PDE operators, with different types of boundary conditions on faces of $\partial\Omega$. The text pde1_bc2, etc., indicates the combination of the PDE operator and the boundary condition. Experimental values have "e" in front in the legend; the others are theoretically established values. Experimentally determined ω_{opt} values are given by the "symbol" points alone. Theoretical ω_{opt} values are given by lines.*

TABLE 4.1
The required number of SOR iterations to solve the LCSC equations for PDE 2 and various boundary conditions.

NGRID	BC 1	BC 2	BC 3
8	42	55	85
16	63	98	172
24	73	125	255
32	105	169	320

in Table 4.1, where we observe that increasing $NGRID$ from 24 to 32 increases the number of iterations only by at most 30%.

Table 4.1 presents the SOR iterations required to solve the discretized equations using the optimal value for the relaxation parameter ω and a 10^{-7} stopping criterion for the problems defined by PDE 2 and various boundary conditions. We also note here that the measured discretization errors for all the experiments confirm the expected second order of convergence of the collocation discretization scheme. Specifically, the measured order in all cases is in the interval $[1.8, 2.1]$.

REFERENCES

[1] J. de Pillis and M. Neumann, *Iterative methods with k-part splittings*, IMA J. Numer. Anal., 1 (1981), pp. 65–79.

[2] A. Hadjidimos, *The optimal solution of the extrapolation problem of a first order scheme*, Internat. J. Comput. Math., 13 (1983), pp. 153–168.

[3] A. Hadjidimos, *On the extrapolation technique for the solution of linear systems*, Calcolo, 23 (1986), pp. 35–43.

[4] A. Hadjidimos, *A survey of the iterative methods for the solution of linear systems by extrapolation, relaxation and other techniques*, J. Comput. Appl. Math., 20 (1987), pp. 213–230.

[5] A. Hadjidimos, E. Houstis, J. Rice, and E. Vavalis, *Iterative line cubic spline collocation methods for elliptic partial differential equations in several dimensions*, SIAM J. Sci. Comput., 14 (1993), pp. 715–734.

[6] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[7] E. Houstis, J. Rice, and E. Vavalis, *Spline collocation methods for elliptic partial differential equations*, in Advances in Computer Methods for Partial Differential Equations, R. Vichnevetsky and R. Stepleman, eds., vol. V, IMACS, New Brunswick, NJ, 1984, pp. 191–194.

[8] E. Houstis, J. Rice, and E. Vavalis, *Parallelization of a new class of cubic spline collocation methods*, in Advances in Computer Methods for Partial Differential Equations, R. Vichnevetsky and R. Stepleman, eds., vol. VI, IMACS, New Brunswick, NJ, 1987, pp. 167–174.

[9] E. Houstis, E. Vavalis, and J. Rice, *Convergence of $O(h^4)$ cubic spline collocation methods for elliptic partial differential equations*, SIAM J. Numer. Anal., 25 (1988), pp. 54–74.

[10] B. Kredell, *On complex successive overrelaxation*, BIT, 2 (1962), pp. 143–152.

[11] H. Shintani, *Direct solution of partial differential equations for a rectangle*, J. Sci. Hiroshima Univ. Ser. A–I, 32 (1968), pp. 17–53.

[12] R. Varga, *Matrix Iterative Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1962.

[13] D. Young, *Iterative Solution of Large Linear Systems*, Academic Press, New York, 1971.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 522–536
© 1999 Society for Industrial and Applied Mathematics

MATRICES WITH LOW-RANK-PLUS-SHIFT STRUCTURE: PARTIAL SVD AND LATENT SEMANTIC INDEXING*

HONGYUAN ZHA† AND ZHENYUE ZHANG‡

Abstract. We present a detailed analysis of matrices satisfying the so-called *low-rank-plus-shift* property in connection with the computation of their partial singular value decomposition (SVD). The application we have in mind is latent semantic indexing for information retrieval, where the term-document matrices generated from a text corpus approximately satisfy this property. The analysis is motivated by developing more efficient methods for computing and updating partial SVD of large term-document matrices and gaining deeper understanding of the behavior of the methods in the presence of noise.

Key words. singular value decomposition, low-rank approximation, latent semantic indexing, perturbation analysis, large sparse matrix

AMS subject classifications. 15A18, 65F15, 65F50

PII. S0895479898344443

1. Introduction.

In many applications, such as compression of multiple-spectral image cubes, regularization methods for ill-posed problems, and latent semantic indexing (LSI) in information retrieval for large document collections, it is necessary to find a low-rank approximation of a given large and/or sparse matrix $A \in \mathcal{R}^{m\times n}$ [1, 11]. The theory of singular value decomposition (SVD) provides the following characterization of the best low-rank approximation of A in terms of Frobenius norm $\|\cdot\|_F$ [6].

THEOREM 1.1. *Let the singular value decomposition of $A \in \mathcal{R}^{m\times n}$ be $A = P\Sigma Q^T$ with $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_{\min(m,n)})$, $\sigma_1 \geq \cdots \geq \sigma_{\min(m,n)}$, and P and Q orthogonal. Then for $1 \leq j \leq \min(m, n)$,*

$$\sum_{i=j+1}^{\min(m,n)} \sigma_i^2 = \min\{\ \|A - B\|_F^2 \mid \operatorname{rank}(B) \leq j\}.$$

And the minimum is achieved with $A_j \equiv P_j \operatorname{diag}(\sigma_1, \ldots, \sigma_j) Q_j^T$, where P_j and Q_j are the matrices formed by the first j columns of P and Q, respectively.

It follows from Theorem 1.1 that once the SVD of A is available, the best rank-j approximation of A is readily computed. We call $A_j = P_j \operatorname{diag}(\sigma_1, \ldots, \sigma_j) Q_j^T$ a partial SVD of A. The state-of-the-art methods for computing the partial SVD of large

*Received by the editors September 3, 1998; accepted for publication (in revised form) by N. J. Higham May 20, 1999; published electronically December 3, 1999.

http://www.siam.org/journals/simax/21-2/34444.html

†Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802 (zha@cse.psu.edu). The work of this author was supported in part by the Director, Office of Energy Research, Office of Laboratory Policy and Infrastructure Management, Office of Computational and Technology Research, Division of Mathematical, Information, and Computational Sciences of the U.S. Department of Energy, under contract DE-AC03-76SF00098, and NSF grant CCR-9619452.

‡Center for Mathematical Sciences and Department of Applied Mathematics, Zhejiang University, Hangzhou 310027, People's Republic of China (zyzhang@math.zju.edu.cn). The work of this author was supported in part by NSFC project 19771073, the Zhejiang Provincial Natural Science Foundation of China, and the Scientific Research Foundation for Returned Overseas Chinese Scholars, State Education Commission.

and/or sparse matrices are based on variants of Lanczos algorithms and the core computation at each iterative step involves matrix-vector multiplications [9].[1] In order to deal effectively with large-scale problems, one is required to exploit various structures of the matrices. Despite its importance, the exploitation of structures so far has been restricted to (i) using the sparsity of a sparse matrix, and (ii) using displacement-rank structures such as Toeplitz or Hankel structures of the matrix to accelerate the matrix-vector multiplications used in the Lanczos process. In this paper, however, we propose to explore an alternative structure that is based on the singular value spectrum of a matrix. Specifically, we investigate matrices possessing the so-called *low-rank-plus-shift* structure, i.e., those matrices A (approximately) satisfying

$$A^T A = \text{a low-rank matrix} + \text{a multiple of the identity matrix}. \tag{1.1}$$

The particular application we have in mind is LSI for information retrieval, and we will show by way of examples that the term-document matrices generated from text corpora approximately satisfy (1.1). In large-scale LSI applications such as the World Wide Web, the term-document matrix generated is usually very large and cannot be kept in RAM or disk. In [15, 16] we have shown that the low-rank-plus-shift structure of the term-document matrix A allows us to compute its partial SVD in a block-wise fashion whereby partial SVD of submatrices of A are computed separately and then merged to obtain a partial SVD of A. The purpose of this paper is to further analyze the properties of matrices with low-rank-plus-shift structure, especially when (1.1) is only approximately satisfied. Our analysis will provide deeper insights into this special class of matrices which will enable us to develop more efficient methods for computing their partial SVD. We should emphasize that although our main results are motivated by problems arising from LSI in information retrieval, they may also be applied in other areas such as array signal processing and solving ill-posed problems.

The rest of the paper is organized as follows. In section 2, we provide a brief background on LSI and review some of the results in [14, 15, 16] related to computing the partial SVD of term-document matrices. In section 3, we discuss some matrix approximation problems associated with the low-rank-plus-shift structure and show by way of examples that the term-document matrices generated from text corpora approximately satisfy (1.1). In section 4 we prove a result on the partial SVD of a block-column partitioned matrix with low-rank-plus-shift structure. This result enables us to efficiently compute the partial SVD either with an incremental approach or a divide-and-conquer approach. We will also discuss the ramification of the result in dealing with LSI updating problems. In section 5 we provide a perturbation analysis of the result when the low-rank-plus-shift property is only satisfied approximately. In section 6 we conclude the paper with some remarks on future research.

2. Latent semantic indexing. The generic problem in information retrieval is the following: given a query, find all the relevant documents in a text collection. The text collection can be a homogeneous one such as the MEDLINE collection for medical information or it can be a heterogeneous one such as the World Wide Web. Variations of the standard keyword-matching methods are still employed by many of the commercial search engines. LSI is a concept-based automatic indexing method that aims at overcoming the two fundamental problems which plague traditional keyword-matching indexing schemes: synonymy and polysemy [2, 5]. Synonymy refers to the problem where several different words can be used to express a concept and the

[1] In the literature *truncated* SVD is also used for what we call partial SVD.

keywords in a user's query may not match those in the relevant documents, while polysemy means that words can have multiple meanings and the user's words may match those in irrelevant documents [8]. LSI is an extension of the vector space model for information retrieval [7, 10]. In the vector space model, the collection of text documents is represented by a *term-document* matrix $A = [a_{ij}] \in \mathcal{R}^{m\times n}$, where a_{ij} is the number of times term i appears in document j, m is the number of terms, and n is the number of documents in the collection. Consequently, a document becomes a column vector and a user's query can also be represented as a vector q of the same dimension. We then compute the score vector $s = q^T A$ and return to the user a list of ranked documents in decreasing order of the scores. Instead of using the raw term frequency for the entries of A, we can also apply many weighting schemes to change the relative importance of terms [10]. LSI extends this vector space model by modeling the term-document relationship using the SVD of the term-document matrix A. Specifically, using the notation in Theorem 1.1, we substitute A by its best rank-k approximation $A_k \equiv P_k \Sigma_k Q_k^T$, where Σ_k is the kth leading principal submatrix of Σ and the score vector is computed as $s = q^T A_k$. Corresponding to each of the k reduced dimensions is a latent concept which may not have any explicit semantic content yet helps to discriminate documents [2, 5].

Large text corpora such as those generated from World Wide Web give rise to very large term-document matrices, and the computation of their partial SVD poses a very challenging problem. Fortunately the term-document matrices possess certain useful properties besides sparsity that can be exploited for this matter. In [15, 16], we developed a theoretical foundation for LSI using the concept of subspaces, and we showed that the model we proposed imposes a so-called *low-rank-plus-shift* structure that is approximately satisfied by the cross-product of the term-document matrices.[2] Specifically, we showed that the term-document matrix $A \in \mathcal{R}^{m\times n}$ satisfies

$$A^T A/m \approx CWC^T + \sigma^2 I_n, \tag{2.1}$$

where $C \in \mathcal{R}^{n\times k}$ is the matrix whose columns represent latent concepts, $W \in \mathcal{R}^{k\times k}$ is a symmetric positive definite matrix, and σ is the variance of the noise. In LSI applications $k \ll \min\{m, n\}$, justifying the use of the terminology *low-rank-plus-shift* structure. As is shown before, we do not need to explicitly compute C, we only need to compute A_k in the factored form $A_k = U_k \Sigma_k V_k^T$.

In [14], we considered the updating problems for LSI: Let A be the term-document matrix for the original text collection and D represent a collection of new documents. The goal is to compute the partial SVD of $[A, D]$. However, in LSI applications, only A_k for some chosen k is available and the matrix A has been discarded. Since updating in this situation is based on a low-rank approximation of A, it has been argued in the literature that one will not be able to get an accurate partial SVD of $[A, D]$. In section 4, we show, however, that this is not the case since $[A, D]$ has the low-rank-plus-shift structure [14]. We will show that no retrieval accuracy degradation will occur if updating is done with a proper implementation. In [15, 16], we also discussed how to compute the partial SVD of a term-document matrix in a block-column partitioned form $A = [A_1, A_2]$ using a divide-and-conquer approach whereby the partial SVDs of A_1 and A_2 are first computed and the results are then merged into a partial SVD of A. This approach is rich in coarse-grain parallelism and can be used to handle very large term-document matrices. The justification for this divide-and-conquer approach will

[2] The low-rank-plus-shift structure was first discussed in the context of array signal processing [12, 13, 17].

be discussed in greater detail in section 4, and perturbation analysis will be provided to show that the approach is still valid even if the term-document matrix A only approximately satisfies the low-rank-plus-shift structure.

3. A matrix approximation problem. From the discussion in section 2 we know that the term-document matrix A approximately satisfies the low-rank-plus-shift property, and therefore A should have flat trailing singular values. In this section we use several example text collections to illustrate this issue. In order to assess whether a given matrix has the low-rank-plus-shift property, we investigate the following matrix approximation problem: Given a general rectangular matrix, what is the *closest* matrix that has the low-rank-plus-shift property? To proceed we first define a matrix set for a given $k > 0$,

$$\mathcal{J}_k = \{B \in \mathcal{R}^{m\times n} \mid \sigma_1(B) \geq \cdots \geq \sigma_{\min\{m,n\}}(B),\ \sigma_{k+1}(B) = \cdots = \sigma_{\min\{m,n\}}(B)\}.$$

With this notation, the matrix approximation problem reduces to finding the distance between a general matrix A and the set $\mathcal{J}_k$. In the following we consider the cases where distance is defined either by Frobenius norm $\|\cdot\|_F$ or by spectral norm $\|\cdot\|_2$.

THEOREM 3.1. *Let the SVD of A be $A = U\Sigma V^T$, $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_{\min\{m,n\}})$, and U and V orthogonal. Then for $k \leq \min\{m, n\}$,*

$$\arg\min_{J\in\mathcal{J}_k} \|A - J\|_p = U_k\Sigma_k V_k^T + \tau_p U_k^{\perp}(V_k^{\perp})^T, \tag{3.1}$$

where $\Sigma_k = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$, $U = [U_k, U_k^{\perp}]$, and $V = [V_k, V_k^{\perp}]$. Furthermore,

$$\tau_p = \begin{cases} \sum_{i=k+1}^{\min\{m,n\}} \sigma_i/(\min\{m, n\} - k), & p = F, \\ (\sigma_{k+1} + \sigma_{\min\{m,n\}})/2, & p = 2. \end{cases}$$

Proof. First define

$$\min_p = \begin{cases} \|a_k - \tau_p\|, & p = F, \\ \|a_k - \tau_p\|_\infty, & p = 2, \end{cases}$$

where $a_k = [\sigma_{k+1}, \ldots, \sigma_{\min\{m,n\}}]$. It is readily checked that $\min_p$ is achieved by the right-hand side of (3.1). Therefore, all we need to prove is $\|A - J\|_p \geq \min_p$ for $p = 2, F$ and for any $J \in \mathcal{J}_k$.

To this end we use standard perturbation analysis of singular values which states that [6]

$$|\sigma_i(A) - \sigma_i(J)| \leq \|A - J\|_2, \qquad i = 1, \ldots, \min\{m, n\},$$

and

$$\sum_{i=1}^{\min\{m,n\}} (\sigma_i(A) - \sigma_i(J))^2 \leq \|A - J\|_F^2.$$

It follows that

$$\max_{k+1\leq i\leq \min\{m,n\}} |\sigma_i(A) - \sigma_i(J)| \leq \max_{1\leq i\leq \min\{m,n\}} |\sigma_i(A) - \sigma_i(J)| \leq \|A - J\|_2$$

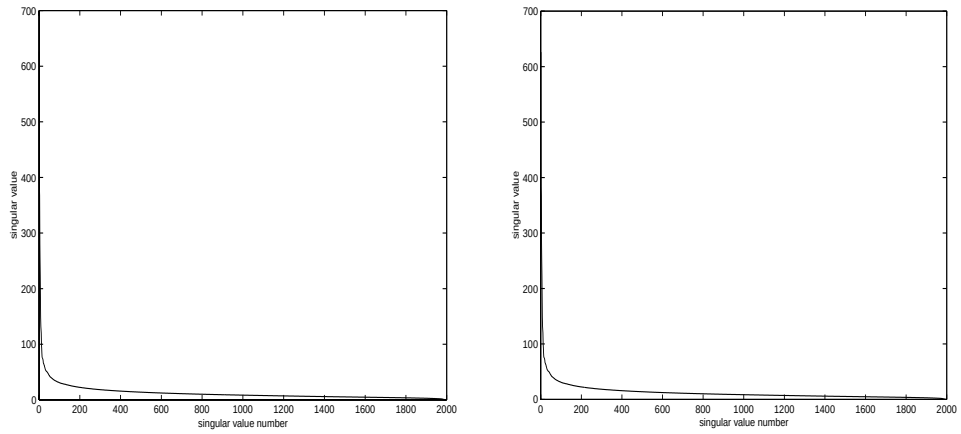


FIG. 3.1. *Singular value distributions:* 3681×1033 *term-document matrix of MEDLINE collection (left) and* 33583×1997 *NEWSGROUP collection (right).*

and

$$\sum_{i=k+1}^{\min\{m,n\}} (\sigma_i(A) - \sigma_i(J))^2 \leq \sum_{i=1}^{\min\{m,n\}} (\sigma_i(A) - \sigma_i(J))^2 \leq \|A - J\|_F^2.$$

Notice that since $\sigma_{k+1}(J) = \cdots = \sigma_{\min\{m,n\}}(J)$, it can be readily verified that the minima of the left-hand sides of the above two inequalities, i.e., $\min_p$, $p = 2, F$, are achieved by τ_2 and τ_F, respectively. □

Examples. In the following we will apply the above theorem to two example text collections and see how close the associated term-document matrices are to the set of matrices with low-rank-plus-shift structure. Our first example is the MEDLINE collection from the Cornell SMART system [3]. The term-document matrix is of size 3681×1033 and weighted by entropy weighting scheme [7]. The singular value distribution is plotted on the left of Figure 3.1. Our second example is from a collection consisting of news articles from 20 newsgroups [4]. The term-document matrix is of size 33583×1997. Its singular values are plotted on the right of Figure 3.1. From Theorem 3.1, the best approximation from $\mathcal{J}_k$ to A is given by

$$A^{(k)} \equiv U_k \Sigma_k V_k^T + \tau_p U_k^{\perp} (V_k^{\perp})^T.$$

For the MEDLINE collection we have $\|A - A^{(100)}\|_F / \|A\|_F = 0.2909$ and for the NEWSGROUP collection we have $\|A - A^{(100)}\|_2 / \|A\|_2 = 0.0491$. Several other text collections from the Cornell SMART system have also been tested and we observed similar singular value distributions: initially the singular values decrease rapidly and then the spectrum curve levels off, but the singular values are never close to zero (unless the sparse term-document matrix is *structurally* rank-deficient). The last point is very important: we usually should not treat those matrices simply as near rank-deficient and it is more appropriate that the more general low-rank-plus-shift structure with a nonzero σ be used (cf. (2.1)).

4. The low-rank-plus-shift structure. We start with an examination of the changes of the singular values of a matrix when its elements undergo certain types of modifications. If some of the elements of a general matrix are set to zero, generally it is

not possible to tell whether the singular values of the matrix will increase or decrease. However, a result we will show below states that the singular values of a matrix will always decrease if some submatrices of the matrix are replaced by its low-rank approximations. To proceed we introduce some notation: for any matrix $A \in \mathcal{R}^{m\times n}$, we will use $\text{best}_k(A)$ to denote its best rank-k approximation (cf. Theorem 1.1), and its singular values are assumed to be arranged in nonincreasing order,

$$\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_m(A).$$

As a convention, when we compare the singular values of two matrices with the same number of rows but different number of columns, we will count the singular values according to the number of rows. With the above preparation we present our first result. The proof is similar to that of a slightly special case presented in [14] and therefore is omitted.

THEOREM 4.1. *Let $A \in \mathcal{R}^{m\times n}$ and write $A = [A_1, A_2]$. Then for any k_1 and k_2, we have*

$$\sigma_i([\text{best}_{k_1}(A_1), \text{best}_{k_2}(A_2)]) \leq \sigma_i([A_1, A_2]), \qquad i = 1, \ldots, m.$$

Remark. It is not true that replacing *arbitrary* submatrices of a matrix by their low-rank approximations will result in the decrease of its singular values, as is illustrated in the following example: Let

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \qquad \hat{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Notice that $\hat{A}$ is obtained from A by replacing its $(1,1)$ and $(2,2)$ elements by zero (a best rank-zero approximation). Even though the largest singular value decreases, its smallest singular value increases.

It is rather easy to find examples for which strict *inequalities* hold in Theorem 4.1. In the following we show that this will not be the case if A has the low-rank-plus-shift structure.

THEOREM 4.2. *Let $A = [A_1, A_2] \in \mathcal{R}^{m\times n}$ with $m \geq n$. Furthermore assume that*

$$A^T A = X + \sigma^2 I, \qquad \sigma > 0, \tag{4.1}$$

where X is symmetric and positive semidefinite with $\text{rank}(X) = k$. *Then there are integers $k_1 \leq k$ and $k_2 \leq k$ with $k_1 + k_2 \geq k$ such that*

$$\text{best}_k([\text{best}_{k_1}(A_1), \text{best}_{k_2}(A_2)]) = \text{best}_k([A_1, A_2]). \tag{4.2}$$

Proof. The general idea of the proof is to show that what is discarded when A_1 is replaced by $\text{best}_k(A_1)$ and A_2 is replaced by $\text{best}_k(A_2)$ will also be discarded when $\text{best}_k([A_1, A_2])$ is computed from $[A_1, A_2]$. To this end write

$$A^T A - \sigma^2 I = \begin{bmatrix} A_1^T A_1 - \sigma^2 I & A_1^T A_2 \\ A_2^T A_1 & A_2^T A_2 - \sigma^2 I \end{bmatrix}.$$

Since $\text{rank}(X) = k$, it follows that $\text{rank}(A_1^T A_1 - \sigma^2 I) \leq k$ and $\text{rank}(A_2^T A_2 - \sigma^2 I) \leq k$. Let the eigendecompositions of

$$A_1^T A_1 - \sigma^2 I = V_{A_1} \,\text{diag}(\Sigma_{A_1}^2, 0) V_{A_1}^T, \qquad A_2^T A_2 - \sigma^2 I = V_{A_2} \,\text{diag}(\Sigma_{A_2}^2, 0) V_{A_2}^T,$$

where $\Sigma_{A_1} \in \mathcal{R}^{k_1 \times k_1}, \Sigma_{A_2} \in \mathcal{R}^{k_2 \times k_2}$ are nonsingular with $k_1 \leq k, k_2 \leq k$. We can write the SVD of A_1 and A_2 as follows:

$$A_1 = U_{A_1} \operatorname{diag}(\hat{\Sigma}_{A_1}, \sigma I_{t_1}) V_{A_1}^T = [U_{A_1}^{(1)}, U_{A_1}^{(2)}] \operatorname{diag}(\hat{\Sigma}_{A_1}, \sigma I_{t_1}) [V_{A_1}^{(1)}, V_{A_1}^{(2)}]^T, \tag{4.3}$$

$$A_2 = U_{A_2} \operatorname{diag}(\hat{\Sigma}_{A_2}, \sigma I_{t_2}) V_{A_2}^T = [U_{A_2}^{(1)}, U_{A_2}^{(2)}] \operatorname{diag}(\hat{\Sigma}_{A_2}, \sigma I_{t_2}) [V_{A_2}^{(1)}, V_{A_2}^{(2)}]^T, \tag{4.4}$$

where $\hat{\Sigma}_{A_1} = (\Sigma_{A_1}^2 + \sigma^2 I_{k_1})^{1/2}, \hat{\Sigma}_{A_2} = (\Sigma_{A_2}^2 + \sigma^2 I_{k_2})^{1/2}, U_{A_1}^{(1)} \in \mathcal{R}^{m \times k_1}, U_{A_2}^{(1)} \in \mathcal{R}^{m \times k_2}$, and $t_1 = n_1 - k_1, t_2 = n_2 - k_2$, respectively, where n_i is the column dimension of $A_i, i = 1, 2$. Now write $V_{A_1}^T A_1^T A_2 V_{A_2}$ in a partitioned form as

$$V_{A_1}^T A_1^T A_2 V_{A_2} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}, \qquad S_{11} \in \mathcal{R}^{k_1 \times k_2}. \tag{4.5}$$

Since $X = A^T A - \sigma^2 I$ is symmetric positive semidefinite and $\operatorname{rank}(X) = k$, it follows that $S_{12} = 0, S_{21} = 0, S_{22} = 0$, and $k_1 + k_2 \geq \operatorname{rank}(X) = k$. Using the SVD of A_1 and A_2 in (4.3) and (4.4), (4.5) becomes

$$[U_{A_1}^{(1)} \Sigma_{A_1}, \sigma U_{A_1}^{(2)}]^T [U_{A_2}^{(1)} \Sigma_{A_2}, \sigma U_{A_2}^{(2)}] = \begin{bmatrix} S_{11} & 0 \\ 0 & 0 \end{bmatrix},$$

which leads to[3]

$$U_{A_1}^{(1)} \perp U_{A_2}^{(2)}, \qquad U_{A_1}^{(2)} \perp U_{A_2}^{(1)}, \qquad U_{A_1}^{(2)} \perp U_{A_2}^{(2)}.$$

Let $\hat{U}$ be an orthonormal basis of $\mathcal{R}([U_{A_1}^{(1)}, U_{A_2}^{(1)}])$, where $\mathcal{R}(\cdot)$ denotes the column space of a matrix. Then we can write

$$[A_1, A_2] = [\hat{U}, U_{A_1}^{(2)}, U_{A_2}^{(2)}] \operatorname{diag}(\tilde{B}, \sigma I_{t_1}, \sigma I_{t_2}) \begin{bmatrix} (V_{A_1}^{(1)})^T & 0 \\ 0 & (V_{A_2}^{(1)})^T \\ (V_{A_1}^{(2)})^T & 0 \\ 0 & (V_{A_2}^{(2)})^T \end{bmatrix},$$

where $\tilde{B} \in \mathcal{R}^{\hat{k} \times (k_1 + k_2)}$ with all its singular values greater than σ, and $k \leq \hat{k} \leq k_1 + k_2$. Therefore,

$$[A_1, A_2] = \hat{U} \tilde{B} \begin{bmatrix} (V_A^{(1)})^T & 0 \\ 0 & (V_D^{(1)})^T \end{bmatrix} + \sigma [U_{A_1}^{(2)}, U_{A_2}^{(2)}] \begin{bmatrix} (V_{A_1}^{(2)})^T) & 0 \\ 0 & (V_{A_2}^{(2)})^T) \end{bmatrix}.$$

The first term in the right-hand side of the above is easily seen to be the matrix $[\text{best}_{k_1}(A_1), \text{best}_{k_2}(A_2)]$, and the relation in (4.2) therefore holds. □

Remark. Generically we will have $k_1 + k_2 = k$. In the following we give an example that shows the possibility of the case $k_1 + k_2 > k$. Given any two positive numbers a and b, choose θ_1 and θ_2 such that

$$\cos(\theta_1 - \theta_2) = \sqrt{ab/(a + \sigma^2)(b + \sigma^2)}.$$

[3]We use $S \perp T$ to denote $S^T T = 0$.

Construct two matrices U_1 and U_2 as follows,

$$U_1 = \begin{bmatrix} c_1 & 0 \\ s_1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad U_2 = \begin{bmatrix} c_2 & 0 \\ s_2 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix},$$

where $c_i = \cos(\theta_i)$ and $s_i = \sin(\theta_i)$ for $i = 1, 2$. Now construct the matrix $A = [A_1, A_2]$ with

$$A_1 = U_1 \operatorname{diag}(\sqrt{a + \sigma^2}, \sigma), \qquad A_2 = U_2 \operatorname{diag}(\sqrt{b + \sigma^2}, \sigma).$$

We obtain that

$$A^T A - \sigma^2 I = \begin{bmatrix} a & 0 & \sqrt{ab} & 0 \\ 0 & 0 & 0 & 0 \\ \sqrt{ab} & 0 & b & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

and therefore $k = 1$. However, we also have

$$A_1^T A_1 - \sigma^2 I = \begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix}, \qquad A_2^T A_2 - \sigma^2 I = \begin{bmatrix} b & 0 \\ 0 & 0 \end{bmatrix},$$

and thus $k_1 = k_2 = 1$. So we have the case $k_1 + k_2 > k$.

Remark. In essence the result in Theorem 4.2 states that if A has the low-rank-plus-shift structure, then an optimal low-rank approximation of A can be computed by merging the optimal low-rank approximations of its two submatrices A_1 and A_2. The result can be generalized to the case where A is partitioned into several blocks $A = [A_1, A_2, \ldots, A_s]$.

Remark. In general k_1 and k_2 are not available; they exist in the analysis in the proof of Theorem 4.2 but are never explicitly computed. However, since $k_i \leq k, i = 1, 2$, the relation in (4.2) still holds if we replace $k_i, i = 1, 2$, by k, i.e.,

$$\text{best}_k([\text{best}_k(A_1), \text{best}_k(A_2)]) = \text{best}_k([A_1, A_2]).$$

Referring back to our discussion on LSI updating problem in section 2, we see that Theorem 4.2 justifies the replacement of A by its best rank-k approximation because

$$\text{best}_k[\text{best}_k(A), D]) = \text{best}_k([A, D]),$$

assuming $[A, D]$ has the low-rank-plus-shift structure. That is to say, we will obtain the same best low-rank approximation even though A is replaced by $\text{best}_k(A)$. Numerical results conducted on several text collections show that no retrieval accuracy degradation occurs when updating is computed using a proper implementation [14].

On the other hand, Theorem 4.2 also leads to some novel approaches for computing a low-rank approximation of a large matrix. There are at least two general approaches to pursue ideas based on Theorem 4.2:

- *An incremental method.* The first is what we call an incremental approach, whereby we can use certain sampling methods to divide the whole collection of documents into several groups: Start with one group and compute its rank-k approximation, add the second group using the updating algorithm to produce a new rank-k approximation, then repeat the whole process. This incremental process can be very useful when the data collection is very large and the whole term-document matrix cannot reside completely in the main memory. Some computational results of this approach can be found in [14].

- *A divide-and-conquer method.* The second is what we call a divide-and-conquer approach. We can again divide the whole collection of documents into several groups and compute the rank-k approximation for *each group* and then combine the results together into a rank-k approximation for the whole data collection. Recursively, the rank-k approximation for each group can also be computed using this divide-and-conquer approach, and so on. The approach has the property that computation can be organized with a high degree of coarse-grain parallelism. A parallel implementation of this method is currently under investigation.

5. Perturbation analysis. In this section, we consider the case where A only approximately satisfies the low-rank-plus-shift property. Our main goal is to see to what extent the result in Theorem 4.2 still holds in the presence of perturbation. We first present some lemmas which are of their own interests as well. In what follows, $\|\cdot\|$ denotes 2-norm and $\|\cdot\|_F$ denotes Frobenius norm. We will use MATLAB notation for submatrices: `A(i:j,k:l)` denotes rows i to j and columns k to l of A.

LEMMA 5.1. *Assume that the matrix X defined below is symmetric positive semi-definite,*

$$X = \begin{bmatrix} A & B^T \\ B & C \end{bmatrix} - E.$$

Then we have

$$\|B\| \le \sqrt{(\|A\| + \|E\|)(\|C\| + \|E\|)}.$$

Proof. Without loss of generality we assume that the matrix B is diagonal. (The result still holds even if B is rectangular.) Write

$$B = \operatorname{diag}(\sigma_1, \ldots, \sigma_n), \qquad \sigma_1 \ge \cdots \ge \sigma_n.$$

Let a_{11} and c_{11} be the $(1,1)$ element of A and C, respectively. Then

$$\begin{bmatrix} a_{11} & \sigma_1 \\ \sigma_1 & c_{11} \end{bmatrix} = X_1 + E_1,$$

for some 2-by-2 symmetric positive semidefinite X_1 and E_1 with $\|E_1\| \le \|E\|$. Since the smallest eigenvalue of $X_1 + E_1$ is no smaller than $-\|E_1\|$, it follows that

$$\begin{aligned} \sigma_1^2 &\le ((a_{11} + c_{11})/2 + \|E_1\|)^2 - ((a_{11} - c_{11})/2)^2 \\ &\le (\|A\| + \|E\|)(\|C\| + \|E\|), \end{aligned}$$

thus completing the proof. ∎

LEMMA 5.2. *Let the matrix X be partitioned as*

$$X = \begin{bmatrix} A & B^T \\ B & C \end{bmatrix}.$$

Then $\|X\| \le \max\{\|A\|, \|C\|\} + \|B\|$.

Proof. The proof is straightforward and is therefore omitted. ∎

THEOREM 5.3. *Let $A = [A_1, A_2] \in \mathcal{R}^{m \times n}, m \ge n$ satisfy*

$$A^T A = X + \sigma^2 I + E,$$

where X is symmetric positive semidefinite with $\operatorname{rank}(X) = k$. *If*

$$\lambda_k(X) > 3\|E\| + 2\sqrt{\|E\|(\|X\| + 2\|E\|)},$$

then for some $k_1 \le k$, $k_2 \le k$, and $k_1 + k_2 \ge k$, we have

$$\operatorname{best}_k([A_1, A_2]) = \operatorname{best}_k([\operatorname{best}_{k_1}(A_1), \operatorname{best}_{k_2}(A_2)]) + \Delta$$

with

$$\|\Delta\| \le \alpha(\|E\| + 4\sqrt{\|E\|(\|X\| + 2\|E\|)})$$

and

$$\alpha = \frac{\sigma_1(A) + 2\sigma_{k+1}(A)}{\sigma_k^2(A) - \sigma^2}.$$

Proof. The proof is divided into several parts.

(1) We first write the eigendecomposition of the following matrices:

$$A^T A - \sigma^2 I = X + E = V \operatorname{diag}(\lambda_1, \ldots, \lambda_n) V^T,$$

$$A_i^T A_i - \sigma^2 I = X_i + E_i = V_i \operatorname{diag}(\lambda_1^{(i)}, \ldots, \lambda_{n_i}^{(i)}) V_i^T, \qquad i = 1, 2,$$

where X and X_i are symmetric positive semidefinite and V and V_i are orthogonal for $i = 1, 2$. The eigenvalues $\{\lambda_k\}$ and $\{\lambda_k^{(i)}\}$ are arranged in nonincreasing order. It follows that, for $i = 1, 2$, there are orthogonal matrices U_i such that

$$\begin{aligned} A_i &= U_i \operatorname{diag}(d_1^{(i)}, \ldots, d_{n_i}^{(i)}) V_i^T, \\ &= U_i \operatorname{diag}(D_{i1}, D_{i2}) V_i^T, \end{aligned}$$

where $d_j^{(i)} = \sqrt{\lambda_j^{(i)} + \sigma^2}$ and

$$\begin{aligned} D_{i1} &= \operatorname{diag}(d_1^{(i)}, \ldots, d_{k_i}^{(i)}), \\ D_{i2} &= \operatorname{diag}(d_{k_i+1}^{(i)}, \ldots, d_{n_i}^{(i)}), \end{aligned}$$

where $k_i = \operatorname{rank}(X_i), i = 1, 2$. The definition of best low-rank approximation leads to

$$\operatorname{best}_{k_i}(A_i) = U_i D_{i1} V_i, \qquad i = 1, 2.$$

(2) Using the above decompositions we now write the matrix A in several different forms: $A = [A_1, A_2] = B[W_1, W_2]^T = [B_1, B_2] W^T$, where

$$B_1 = [U_{11} D_{11}, U_{21} D_{21}], \qquad B_2 = [U_{12} D_{12}, U_{22} D_{22}],$$

and

$$W_1 = \begin{bmatrix} V_{11} & 0 \\ 0 & V_{21} \end{bmatrix}, \qquad W_2 = \begin{bmatrix} V_{12} & 0 \\ 0 & V_{22} \end{bmatrix}.$$

It can be readily verified that

$$B_1 W_1^T = [\operatorname{best}_{k_1}(A_1), \operatorname{best}_{k_2}(A_2)].$$

Now partition

$$B^TB - \sigma^2 I = \begin{bmatrix} B_1^TB_1 - \sigma^2 I & B_1^TB_2 \\ B_2^TB_1 & B_2^TB_2 - \sigma^2 I \end{bmatrix}. \tag{5.1}$$

(3) Let the eigendecomposition of $B_1^TB_1 - \sigma^2 I$ be

$$B_1^TB_1 - \sigma^2 I = G\,\mathrm{diag}(\alpha_1, \ldots, \alpha_s)G^T,$$

where $\alpha_1 \geq \cdots \geq \alpha_s, s = k_1 + k_2$. Now partition $\mathrm{diag}(G^T, I)B^TB\,\mathrm{diag}(G, I)$ as

$$\mathrm{diag}(G^T, I)B^TB\,\mathrm{diag}(G, I) = \begin{bmatrix} C & E_1^T \\ E_1 & E_2 \end{bmatrix},$$

where $C = \mathrm{diag}(\alpha_1, \ldots, \alpha_k)$ and the matrix E has the form

$$[E_1, E_2] = \begin{bmatrix} [0, \mathrm{diag}(\alpha_{k+1}, \ldots, \alpha_s)] & (G^TB_1^TB_2)\texttt{(k+1:m,:)} \\ B_2^TB_1G & B_2^TB_2 - \sigma^2 I \end{bmatrix}. \tag{5.2}$$

Furthermore, let the eigendecomposition of

$$\begin{bmatrix} C & E_1^T \\ E_1 & E_2 \end{bmatrix} = Q\Lambda Q^T, \qquad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n).$$

Now partition conformally,

$$Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}, \qquad \Lambda = \mathrm{diag}(\Lambda_1, \Lambda_2)$$

with $\Lambda_1 = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$. It can be verified that

$$Q_{21} = [E_1, E_2] \begin{bmatrix} Q_{11} \\ Q_{21} \end{bmatrix} \Lambda_1^{-1}. \tag{5.3}$$

It follows that we can find U orthogonal such that

$$B = [B_1, B_2] = U(\Lambda + \sigma^2 I)^{1/2} Q^T\,\mathrm{diag}(G^T, I),$$

which leads to

$$B_1W_1^T = U(\Lambda + \sigma^2 I)^{1/2} Q^T \begin{bmatrix} G^TW_1^T \\ 0 \end{bmatrix} = H + \Delta_1,$$

where

$$H = U\texttt{(:,1:k)}(\Lambda_1 + \sigma^2 I)^{1/2} Q_{11}^T G^T W_1^T$$

and

$$\Delta_1 = U\texttt{(:,K+1:n)}(\Lambda_2 + \sigma^2 I)^{1/2} Q_{12}^T G^T W_1^T.$$

Hence

$$\begin{aligned} \|\mathrm{best}_k(B_1W_1^T) - H\| &= \|\mathrm{best}_k(B_1W_1^T) - B_1W_1^T + \Delta_1\| \\ &\leq \|\mathrm{best}_k(B_1W_1^T) - B_1W_1^T\| + \|\Delta_1\| \leq 2\|\Delta_1\|. \end{aligned}$$

(4) On the other hand, let

$$\Delta = \text{best}_k(A) - \text{best}_k(B_1 W_1^T).$$

Then

$$\begin{aligned}
\|\Delta\| &\le \|\text{best}_k(A) - H\| + \|H - \text{best}_k(B_1 W_1^T)\| \\
&\le \|U(:,1:k)(\Lambda_1 + \sigma^2 I)^{1/2} Q_{21}^T W_2^T\| + 2\|\Delta_1\| \\
&\le \sqrt{\lambda_1 + \sigma^2}\|Q_{21}\| + 2\sqrt{\lambda_{k+1} + \sigma^2}\|Q_{12}\| \\
&= (\sqrt{\lambda_1 + \sigma^2} + 2\sqrt{\lambda_{k+1} + \sigma^2})\|Q_{21}\| \\
&\le (\sqrt{\lambda_1 + \sigma^2} + 2\sqrt{\lambda_{k+1} + \sigma^2})\|[E_1, E_2]\|/\lambda_k \\
&= \alpha\|[E_1, E_2]\|,
\end{aligned}$$

where

$$\alpha = \frac{\sqrt{\lambda_1 + \sigma^2} + 2\sqrt{\lambda_{k+1} + \sigma^2}}{\lambda_k} = \frac{\sigma_1(A) + 2\sigma_{k+1}(A)}{\sigma_k^2(A) - \sigma^2}.$$

Now we need to bound $\|[E_1, E_2]\|$, and this will be done in the following steps.

(5) Applying Lemma 5.1 to the matrix

$$B_2^T B_2 - \sigma^2 I = \begin{bmatrix} D_{12}^2 - \sigma^2 I & D_{11}^T U_{12}^T U_{22} D_{22} \\ D_{22}^T U_{22}^T U_{12} D_{11} & D_{22}^2 - \sigma^2 I \end{bmatrix},$$

we obtain the bound

$$\|D_{11}^T U_{12}^T U_{22} D_{22}\|^2 \le (\|D_{12}^2 - \sigma^2 I\| + \|E\|)(\|D_{22}^2 - \sigma^2 I\| + \|E\|) \le 4\|E\|^2,$$

where we have used $\|D_{i2}^2 - \sigma^2 I\| \le \|E\|, i = 1, 2$. By Lemma 5.2 we obtain

$$\|B_2^T B_2 - \sigma^2 I\| \le 3\|E\|.$$

(6) Now applying Lemma 5.1 to $B^T B - \sigma^2 I$ in (5.1) we obtain

$$\|B_2^T B_1\|^2 \le 4\|E\|(\|B_1^T B_1 - \sigma^2 I\| + \|E\|) \le 4\|E\|(\|X\| + 2\|E\|),$$

where we have bounded

$$\|B_1^T B_1 - \sigma^2 I\| \le \|B^T B - \sigma^2 I\| \le \|X\| + \|E\|.$$

(7) Using (5.2), the results in parts (5) and (6), and Lemma 5.2, we have

$$\|[E_1, E_2]\| \le \max\{|\alpha_{k+1}|, 3\|E\|\} + 2\sqrt{\|E\|(\|X\| + 2\|E\|)}.$$

(8) Now we are ready to complete the proof of the theorem by showing that

$$|\alpha_{k+1}| \le \|E\| + 2\sqrt{\|E\|(\|X\| + 2\|E\|)}.$$

In fact, by definition, $\lambda_j = \lambda_j(B^T B - \sigma^2 I)$, and therefore

$$|\lambda_{k+1} - \lambda_{k+1}(\text{diag}(B_1^T B_1 - \sigma^2 I, B_2^T B_2 - \sigma^2 I))| \le \|B_2^T B_1\|.$$

On the other hand, we have

$$|\lambda_{k+1}(B_2^T B_2 - \sigma^2 I)| \le \|B_2^T B_2 - \sigma^2 I\| \le 3\|E\|.$$

The assumption of the theorem implies that

$$\lambda_k > \|B_2^T B_1\| + 3\|E\|,$$

and therefore, for $j \le k$,

$$\lambda_j(\mathrm{diag}(B_1^T B_1 - \sigma^2 I, B_2^T B_2 - \sigma^2 I)) = \lambda_j(B_1^T B_1 - \sigma^2 I) = \alpha_j.$$

Now there is $i > k$ such that $|\alpha_{k+1} - \lambda_i| \le \|B_2^T B_1\|$ and thus

$$|\alpha_{k+1}| \le |\lambda_i| + \|B_2^T B_1\| \le \|E\| + 2\sqrt{\|E\|(\|X\| + 2\|E\|)},$$

completing the proof. □

Remark. In many of our numerical experiments, we observed that $\|\Delta\| = O(\|E\|)$, versus the bound $\|\Delta\| = O(\|E\|^{1/2})$ given in the above theorem. Here we give an example for which we do have $\|\Delta\| = O(\|E\|^{1/2})$. Let ϵ be small, and define $t = \sqrt{2\epsilon}$. Denote

$$\begin{aligned}
\mu_1 &= (1 + \epsilon + \sqrt{(1+\epsilon)^2 + 4\epsilon})/2 = 1 + 2\epsilon - 2\epsilon^2 + O(\epsilon^3),\\
\mu_2 &= (1 + \epsilon - \sqrt{(1+\epsilon)^2 + 4\epsilon})/2 = -\epsilon + 2\epsilon^2 + O(\epsilon^3),\\
a &= (\mu_1 - 1)/t = t(1 - \epsilon + O(\epsilon^2)).
\end{aligned}$$

Then it can be readily verified that

$$\begin{bmatrix} 1 & t \\ t & \epsilon \end{bmatrix} = (1/(1+a^2)) \begin{bmatrix} 1 & a \\ a & -1 \end{bmatrix} \begin{bmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{bmatrix} \begin{bmatrix} 1 & a \\ a & -1 \end{bmatrix}.$$

Now let

$$X = (\mu_1/(1+a^2)) \begin{bmatrix} 1 & 0 \\ 0 & a \\ 0 & 1 \\ a & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & a \\ 0 & a & 1 & 0 \end{bmatrix},$$

$$E = (\mu_2/(1+a^2)) \begin{bmatrix} 0 & a \\ 1 & 0 \\ -a & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 0 & 1 & -a & 0 \\ a & 0 & 0 & -1 \end{bmatrix}.$$

With $k = \mathrm{rank}(X) = 2$ and $\|E\| = |\mu_2| = \epsilon(1 - 2\epsilon + O(\epsilon^2))$, it is easy to see that

$$X + E = V\,\mathrm{diag}(\mu_1, \mu_1, \mu_2, \mu_2)V,$$

where

$$V = (1/\sqrt{1+a^2}) \begin{bmatrix} 1 & 0 & 0 & a \\ 0 & a & 1 & 0 \\ 0 & 1 & -a & 0 \\ a & 0 & 0 & -1 \end{bmatrix}$$

is orthogonal and symmetric. For a given $\sigma > 0$, define

$$c_1 = \sqrt{\mu_1 + \sigma^2}, \qquad c_2 = \sqrt{\mu_2 + \sigma^2}.$$

Now construct matrix A as $A = [A_1, A_2]$, where

$$A_1 = [e_1, e_3, e_2, e_4]DV\texttt{(:,1:2)}, \qquad A_2 = [e_1, e_3, e_2, e_4]DV\texttt{(:,3:4)},$$

with $D = \mathrm{diag}(c_1, c_1, c_2, c_2)$, and $I_4 = [e_1, e_2, e_3, e_4]$ is the identity matrix. It can be verified that $k_1 = k_2 = 2$. Then we have

$$\mathrm{best}_k(A) = (c_1/\sqrt{1+a^2})[e_1, e_3]V\texttt{(1:2,:)}.$$

Since $a < 1$, it can be verified that

$$\mathrm{best}_1(A_1) = (1/\sqrt{1+a^2}) \begin{bmatrix} c_1 & 0 \\ 0 & 0 \\ 0 & 0 \\ c_2 a & 0 \end{bmatrix}, \qquad \mathrm{best}_1(A_2) = (1/\sqrt{1+a^2}) \begin{bmatrix} 0 & 0 \\ -c_2 a & 0 \\ c_1 & 0 \\ 0 & 0 \end{bmatrix}.$$

It follows that

$$\mathrm{best}_k([\mathrm{best}_1(A_1), \mathrm{best}_1(A_2)]) = (1/\sqrt{1+a^2}) \begin{bmatrix} c_1 & 0 & 0 & 0 \\ 0 & 0 & -c_2 a & 0 \\ 0 & 0 & c_1 & 0 \\ c_2 a & 0 & 0 & 0 \end{bmatrix},$$

and therefore

$$\begin{aligned} \Delta &= \mathrm{best}_k(A) - \mathrm{best}_k([\mathrm{best}_1(A_1), \mathrm{best}_1(A_2)]) \\ &= a/\sqrt{1+a^2}\,\mathrm{diag}(c_1, c_2, c_1, c_2)[-e_4, e_3, e_2, e_1], \end{aligned}$$

which leads to $\|\Delta\| = ac_1/\sqrt{1+a^2}$. Then we see that

$$\lim_{\epsilon \to 0} \|\Delta\|/\sqrt{\|E\|} = \sqrt{2(1+\sigma^2+2\epsilon+O(\epsilon^2))}/[(1+O(\epsilon^2))(1-\epsilon+O(\epsilon^2)) = \sqrt{2(1+\sigma^2)}.$$

6. Concluding remarks. This paper presented a detailed analysis of matrices with a low-rank-plus-shift structure. We placed emphasis on justifying some novel methods for partial SVD computation and partial SVD updating problems arising from LSI in information retrieval. Our perturbation analysis demonstrated that the results we had derived were still valid even though the low-rank-plus-shift structure was approximately satisfied. The results we have proved provided theoretical justifications for the novel LSI updating algorithms and the incremental and divide-and-conquer approaches proposed in [14, 16].

REFERENCES

[1] M. Berry, *Large scale singular value computations*, Internat. J. Supercomputer Appl., 6 (1992), pp. 13–49.

[2] M. W. Berry, S. T. Dumais, and G. W. O'Brien, *Using linear algebra for intelligent information retrieval*, SIAM Rev., 37 (1995), pp. 573–595.

[3] Cornell SMART System, available from ftp://ftp.cs.cornell.edu/pub/smart.

[4] CMU Data, available from http://www.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes.html.
[5] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, *Indexing by latent semantic analysis*, J. Soc. Inform. Sci., 41 (1990), pp. 391–407.
[6] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., Johns Hopkins University Press, Baltimore, MD, 1989.
[7] G. Kowalski, *Information Retrieval System: Theory and Implementation*, Kluwer Academic Publishers, Boston, 1997.
[8] R. Krovetz and W. B. Croft, *Lexical ambiguity and information retrieval*, ACM Trans. Inform. Systems, 10 (1992), pp. 115–141.
[9] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.
[10] G. Salton, *Automatic Text Processing*, Addison-Wesley, New York, 1989.
[11] H. Simon and H. Zha, *Low Rank Matrix Approximation Using the Lanczos Bidiagonalization Process*, CSE Technical Report CSE-97-008, 1997; also LBNL Technical Report LBNL-40767-UC-405.
[12] G. Xu and T. Kailath, *Fast subspace decompsotion*, IEEE Trans. Signal Processing, 42 (1994), pp. 539–551.
[13] G. Xu, H. Zha, G. Golub, and T. Kailath, *Fast algorithms for updating signal subspaces*, IEEE Trans. Circuits and Systems, 41 (1994), pp. 537–549.
[14] H. Zha and H. D. Simon, *On updating problems in latent semantic indexing*, SIAM J. Sci. Comput., 21 (1999), pp. 782–791.
[15] H. Zha and H. Simon, *A subspace-based model for latent semantic indexing in information retrieval*, in Proceedings of the 30th Symposium on the Interface of Computer Science and Statistics, Berkeley, CA, Springer-Verlag, New York, 1998, pp. 315–320.
[16] H. Zha, O. Marques, and H. Simon, *Large-scale SVD and subspace-based methods for information retrieval*, in Proceedings of Irregular '98, Lecture Notes in Comput. Sci. 1457, 1998, pp. 29–42.
[17] H. Zha and Z. Zhang, *A sorted partial Jacobi method and its convergence analysis*, Linear Algebra Appl., 270 (1997), pp. 79–108.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 537–546
© 1999 Society for Industrial and Applied Mathematics

DISTRIBUTION OF ENTRIES IN A SUBSTOCHASTIC MATRIX HAVING EIGENVALUES NEAR 1*

D. J. HARTFIEL†

Abstract. This paper gives a quantitative result that if A is a substochastic matrix and has r eigenvalues which are sufficiently close to 1, then A has r disjoint principal submatrices which are nearly stochastic.

Key words. substochastic matrices, nearly stochastic blocks, canonical forms, condition

AMS subject classifications. 15A12, 15A18, 15A21, 15A51

PII. S0895479897323440

The condition of a Markov chain, having transition matrix A and stochastic eigenvector π, has been studied by numerous researchers. Condition numbers were given by Meyer (see [6]) and Seneta (see [7]). Meyer uses $K(A) = \max_{i,j} |a_{ij}^{\#}|$, where $A^{\#} = (I - A + e\pi)^{-1} - e\pi$, e denoting the column vector of 1's. (Here, $A^{\#}$ denotes the group inverse of $I - A$.) Seneta uses $\mathcal{T}(A^{\#})$, where $\mathcal{T}(B) = \frac{1}{2} \max_{i,j} \sum_{k=1}^{n} |b_{ik} - b_{jk}|$ for any $n \times n$ matrix B. Neither of these condition numbers is given in terms of the entries of A, although both papers provide bounds on the condition numbers in terms of the eigenvalues of A. Those bounds show that if a subdominant eigenvalue of A is close to 1, then A is ill-conditioned; the converse is also true.

In [4], Hartfiel and Meyer show qualitatively that if a subdominant eigenvalue of A is sufficiently close to 1, then A has two nearly uncoupled principal submatrices. Qualitatively, this links the condition of a Markov chain to the appearance of nearly uncoupled principal submatrices within A. In this paper, with a somewhat technically intricate argument, we show the first results quantifying these remarks.

Results. Our major result rests on a sequence of preliminary theorems. We provide these theorems below. Throughout the presentation, A is an $n \times n$ irreducible substochastic matrix (necessary for our inductive argument) with positive dominant eigenvalue λ_1 and remaining eigenvalues $\lambda_2, \ldots, \lambda_n$. We suppose $\min_{k \neq 1} |\lambda_1 - \lambda_k| = |\lambda_1 - \lambda_2|$. Let x denote the stochastic eigenvector belonging to λ_1. In addition, we use the terms defined below.

Let B be an $m \times m$ substochastic matrix and $\epsilon, \epsilon \geq 0$, a scalar. Define $r_i(B)$ as the ith row sum of B.

(i) If $1 - r_i(B) \leq \epsilon$ for all i, then B is called an ϵ *nearly stochastic matrix*.

(ii) If there is a permutation matrix P such that, in partitioned form,

$$PBP^t = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix},$$

and B_{11} is an ϵ nearly stochastic matrix, then we say that B_{11} is an ϵ nearly stochastic matrix of B.

*Received by the editors June 25, 1997; accepted for publication (in revised form) by G. Styan November 7, 1998; published electronically December 3, 1999.
http://www.siam.org/journals/simax/21-2/32344.html
†Department of Mathematics, Texas A&M University, College Station, TX 77843-3368 (hartfiel@math.tamu.edu).

(iii) If there is a permutation matrix Q such that, in partitioned form,

$$QBQ^t = \begin{bmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \\ B_{31} & B_{32} & B_{33} \end{bmatrix}$$

and B_{11} and B_{22} are ϵ nearly stochastic matrices, then we say that B has two *disjoint ϵ nearly stochastic matrices*. Of course, this definition can be extended to r disjoint ϵ nearly stochastic matrices.

(iv) Let $K = \{1, \ldots, m\}$. Let $R, C \subseteq K$ be such that $R \cup C = K$ and $R \cap C = \emptyset$. Define $\mu(B) = \min_{R,C}(\max_{i \in R, j \in C} b_{ij})$, called the *measure of irreducibility* of B [2], [3]. Note that $\mu(B) \leq 1$.

(v) Let P be a permutation matrix such that

$$PBP^t = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix},$$

where B_{11} is square and the largest entry in B_{12} is μ. Then B_{12} is called a μ-block of B.

We now begin with the preliminary work. The first theorem ensures the existence of a small μ-block in A. We need a lemma.

LEMMA 1. *Let λ be a complex number where $|\lambda| \leq 1$ and $|1 - \lambda| \leq \delta$ for some scalar δ. Then $(1 + \mu)^n - |\lambda + \mu|^n \leq n2^n\delta$ for all μ, $0 \leq \mu \leq 1$.*

Proof. First note that

$$\begin{aligned} 2 = |2 - \lambda + \lambda| &\leq |1 - \lambda| + |1 + \lambda| \text{ so} \\ 2 - |1 + \lambda| &\leq |1 - \lambda| < \delta. \\ \text{Set } \ 2 - |1 + \lambda| &= x \text{ so} \\ 2 - x &= |1 + \lambda|. \end{aligned}$$

Now, $0 \leq x \leq 2$ and

$$2^n - |1 + \lambda|^n = 2^n - (2 - x)^n.$$

Set $f(x) = 2^n - (2 - x)^n$ and $g(x) = n2^n x$. Then $f(0) = 0$ and $g(0) = 0$. Further,

$$f'(x) = n(2 - x)^{n-1} \quad \text{and} \quad g'(x) = n2^n.$$

Since $f'(x) \leq g'(x)$, it follows that $f(x) \leq g(x)$ for all x, $0 \leq x \leq 2$. In particular,

$$2^n - (2 - x)^n \leq n2^n x.$$

Thus, by substituting,

$$2^n - |1 + \lambda|^n \leq n2^n\delta.$$

Finally, since $(1 + \mu)^n - |\lambda + \mu|^n$ is nondecreasing in μ, the result follows. □

THEOREM 1. *Suppose $|1 - \lambda_2| < \delta$. Then, for some scalar μ, $0 \leq \mu \leq 1$, A has a μ-block where $\mu^n \leq n2^n\delta$. (Since we know that $\mu \leq 1$, this bound and those that follow are useful only if δ is sufficiently small.)*

Proof. A corollary to Theorem 2.2 in [5] is that if A is positive and m is the smallest entry in A, then

$$|\lambda_2| \leq \lambda_1 - m.$$

Applying the conversion technique described in [2] to this inequality shows that if μ is the measure of irreducibility of A,

$$|\lambda_2 + \mu|^n \leq (\lambda_1 + \mu)^n - \mu^n.$$

Thus,

$$\mu^n \leq (1 + \mu)^n - |\lambda_2 + \mu|^n.$$

By Lemma 1,

$$\mu^n \leq n2^n\delta. \qquad \square$$

Without loss of generality, we now assume that

$$A = \begin{bmatrix} A_1 & U \\ A_{21} & A_2 \end{bmatrix},$$

where A_1 is square and $\max_{i,j} u_{ij} = \mu$.

The next theorem shows the closeness of the eigenvalues of A_1, A_2 and those of A. We need a lemma.

LEMMA 2. *Let $B = (b_{ij})$ be an $n \times n$ matrix with $e = \max_{i<j} b_{ij}$ and $b = \max_{i>j} b_{ij}$. If λ is any eigenvalue of B, then*

$$|\lambda - b_{ii}| \leq \sqrt[n]{eb^{n-1}} + \cdots + \sqrt[n]{e^{n-1}b}$$

for some i.

Proof. Set $D = \operatorname{diag}(1, \sqrt[n]{(e/b)}, \ldots, \sqrt[n]{(e/b)^{n-1}})$ and $C = DBD^{-1}$. Since C is similar to B, C and B have the same list of eigenvalues. Applying Gershgorin's theorem [1] to B yields that for any eigenvalue λ of B,

$$\begin{aligned}
|\lambda - b_{ii}| &\leq \sum_{k \neq i} c_{ik} \\
&\leq \sum_{k<i} (\sqrt[n]{(e/b)^{i-1}}) b (\sqrt[n]{(e/b)^{k-1}})^{-1} \\
&\quad + \sum_{k>i} (\sqrt[n]{(e/b)^{i-1}}) \epsilon (\sqrt[n]{(e/b)^{k-1}})^{-1} \\
&= \sqrt[n]{e^{i-1}b^{n+1-i}} + \cdots + \sqrt[n]{e^1 b^{n-1}} \\
&\quad + \sqrt[n]{e^{n-1}b^1} + \cdots + \sqrt[n]{e^i b^{n-i}} \\
&= \sqrt[n]{eb^{n-1}} + \cdots + \sqrt[n]{e^{n-1}b}
\end{aligned}$$

for some i. $\square$

THEOREM 2. *Suppose $\mu \leq \frac{1}{n}$ and A has r eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_r$ within δ of 1. Then $\bar{A} = \begin{bmatrix} A_1 & 0 \\ A_{21} & A_2 \end{bmatrix}$ has r eigenvalues $\mu_1, \ldots, \mu_r$ within $n^3\mu^{\frac{1}{n}} + \delta$ of 1.*

Proof. Using Schur's theorem [1], let U_i be a unitary matrix so that $U_i^* A_i U_i = T_i$, where T_i is lower triangular for $i = 1, 2$. Set $\overline{U} = \operatorname{diag}(U_1, U_2)$ and

$$T = \overline{U}^* A \overline{U} = \begin{bmatrix} T_1 & T_{12} \\ T_{21} & T_2 \end{bmatrix}.$$

Since any unitary matrix preserves lengths, using $\beta(B) = (\sum_{i,j} b_{ij}^2)^{\frac{1}{2}}$ for any matrix B, we have for

$$\beta_{\bar{A}} = \beta\left(\begin{bmatrix} A_1 & 0 \\ A_{21} & A_2 \end{bmatrix}\right), \quad \beta\left(\begin{bmatrix} T_1 & 0 \\ T_{21} & T_2 \end{bmatrix}\right) = \beta_{\bar{A}}$$

and for

$$\beta_u = \beta\left(\begin{bmatrix} 0 & U \\ 0 & 0 \end{bmatrix}\right), \quad \beta\left(\begin{bmatrix} 0 & T_{12} \\ 0 & 0 \end{bmatrix}\right) = \beta_u.$$

Thus, if the main diagonal entries of T are $\alpha_1, \ldots, \alpha_n$ (these are also the eigenvalues of $\begin{bmatrix} A_1 & 0 \\ A_{21} & A_2 \end{bmatrix}$), it follows from Lemma 2 that if λ is an eigenvalue of A,

$$|\lambda - \alpha_i| \leq \sqrt[n]{\beta_u \beta_{\bar{A}}^{n-1}} + \cdots + \sqrt[n]{\beta_n^{n-1} \beta_{\bar{A}}}$$

for some i. Since $|\cdot|_2 < |\cdot|_1$,

$$|\lambda - \alpha_i| \leq \sqrt[n]{\sigma_u \sigma_{\bar{A}}^{n-1}} + \cdots + \sqrt{\sigma_u^{n-1} \sigma_{\bar{A}}},$$

where $\sigma_u = \sum_{i,j} u_{ij}$ and $\sigma_{\bar{A}} = \sum_{i,j} \bar{a}_{ij}$. Since A is substochastic, $\sigma_{\bar{A}} \leq n$ and $\sigma_u \leq n\mu$. Thus

$$\begin{aligned} |\lambda - \alpha_i| &\leq \sqrt[n]{(n\mu)n^{n-1}} + \cdots + \sqrt[n]{(n\mu)^{n-1} n} \\ &\leq n[(n\mu)n^{n-1}]^{\frac{1}{n}}, \quad \text{since} \quad \mu \leq \frac{1}{n}, \\ &= n^2 \mu^{\frac{1}{n}}. \end{aligned}$$

Each eigenvalue of A is in a disk

$$|\lambda - \alpha_i| \leq n^2 \mu^{\frac{1}{n}}.$$

And, by arguing as in Gershgorin's theorem, if there are k disks in a component of the union of these disks, then that component contains k eigenvalues of A. Thus, there are r eigenvalues, say, $\alpha_1, \ldots, \alpha_r$ of $\bar{A}$, such that

$$|\lambda_i - \alpha_i| \leq n(n^2 \mu^{\frac{1}{n}}) = n^3 \mu^{\frac{1}{n}} \quad \text{for} \quad i = 1, \ldots, r.$$

Finally, since $\lambda_1, \ldots, \lambda_r$ are within δ of 1, it follows $\alpha_1, \ldots, \alpha_r$ are within $n^3 \mu^{\frac{1}{n}} + \delta$ of 1. □

To obtain the major theorem of the paper requires an additional lemma. The lemma uses the notation

$$d = \min \frac{x_i}{x_j}.$$

LEMMA 3. *Let α be a scalar with $\alpha < \frac{1}{2}$ and suppose $2^n - (\lambda_1 + 1)^n = \alpha^2$. Then*

(i) *A has a μ-block with $\mu \leq \alpha^{1/n}$, or*

(ii) *A is an ϵ nearly stochastic matrix with $\epsilon \leq \alpha^2(1 + 2n)$.*

Proof. Using Corollary 7 in [3], we know that

$$(\lambda_1 + \mu)^n \leq (1+\mu)^n - \mu^n(1-d).$$

Rearranging yields

$$\begin{aligned}\mu^n(1-d) &\leq (1+\mu)^n - (\lambda_1+\mu)^n \\ &\leq 2^n - (\lambda_1+1)^n \\ &= \alpha^2.\end{aligned}$$

Thus, either $\mu^n \leq \alpha$ or $1-d \leq \alpha$. We look at the cases.

Case (i) $\mu^n \leq \alpha$.

In this case, $\mu \leq \alpha^{1/n}$ and so A has a μ-block with $\mu \leq \alpha^{\frac{1}{n}}$.

Case (ii) $1-d \leq \alpha$.

Here we have that

$$1-\alpha \leq \min \frac{x_i}{x_j}.$$

Now, using $e = (1,\ldots,1)^t$, since

$$\begin{aligned}xA &= \lambda_1 x, \\ xAe &= \lambda_1 xe.\end{aligned}$$

Using that $Ae = r = (r_1,\ldots,r_n)^t$, we get

$$xr = \lambda_1,$$

that is,

$$x_1r_1 + \cdots + x_nr_n = \lambda_1.$$

Now let

$$\begin{aligned}x_m &= \min x_i, \qquad & x_M &= \max x_i, \\ r_m &= \min r_i, \qquad & r_M &= \max r_i.\end{aligned}$$

Then

$$(1-x_m)r_M + x_m r_m \geq \lambda_1.$$

Thus,

$$r_M - \lambda_1 \geq x_m(r_M - r_m).$$

Since $\frac{x_m}{x_M} \geq 1-\alpha$ and $xe = 1$, it follows that $x_M \leq \frac{x_m}{1-\alpha}$ and so

$$1 = xe \leq n\frac{x_m}{1-\alpha},$$

that is,

$$(1-\alpha)\frac{1}{n} \leq x_m.$$

Substituting this expression yields

$$r_M - \lambda_1 \geq (1-\alpha)\frac{1}{n}(r_M - r_m).$$

Since $1 \geq r_M$, we have

$$1 - \lambda_1 \geq (1-\alpha)\frac{1}{n}(r_M - r_m).$$

Thus,

$$2 - (\lambda_1 + 1) \geq (1-\alpha)\frac{1}{n}(r_M - r_m)$$

and so

$$2^n - (\lambda_1 + 1)^n \geq (1-\alpha)\frac{1}{n}(r_M - r_m).$$

Hence, by hypothesis,

$$\alpha^2 \geq (1-\alpha)\frac{1}{n}(r_M - r_m),$$

that is,

$$\frac{\alpha^2}{1-\alpha} \geq \frac{1}{n}(r_M - r_m).$$

Continuing, since $\alpha < \frac{1}{2}$ we have

$$2\alpha^2 \geq \frac{1}{n}(r_M - r_m)$$

and so

$$2n\alpha^2 \geq r_M - r_m.$$

Thus,

$$r_m + 2n\alpha^2 \geq r_M.$$

Finally, since

$$\alpha^2 = 2^n - (\lambda_1 + 1)^n \geq 2 - (\lambda_1 + 1) = 1 - \lambda_1,$$

it follows that

$$\lambda_1 \geq 1 - \alpha^2.$$

Using the Perron–Frobenius theory, we have

$$r_M \geq \lambda_1;$$

thus

$$r_M \geq 1 - \alpha^2.$$

Substituting, we have

$$r_m + 2n\alpha^2 \geq 1 - \alpha^2,$$

that is,

$$r_m \geq 1 - \alpha^2(1 + 2n).$$

Thus, A is an ϵ nearly stochastic matrix where $\epsilon \leq \alpha^2(1 + 2n)$. □

THEOREM 3. *If* $1 - \lambda_1 < \delta$ *and* $4^n n^{3n} \delta^{\frac{1}{n^{2n}}} < \frac{1}{4}$, *then* A *has an* ϵ *nearly stochastic matrix where* $\epsilon \leq 4^n n^{3n} \delta^{\frac{1}{n^{2n}}}$.

Proof. The proof is by induction on n. If $n = 1$, then

$$A = [\lambda_1].$$

Thus, A is an ϵ nearly stochastic matrix where $\epsilon = 1 - \lambda_1$, and since $\epsilon \leq \delta$, the result follows.

Now suppose that the theorem is true for all $k < n$. Suppose that A is an $n \times n$ matrix satisfying the hypothesis. Then, since $1 - \lambda_1 < \delta$, by Lemma 1,

$$2^n - (\lambda_1 + 1)^n \leq n2^n\delta.$$

Thus, by Lemma 3,

(i) A has a μ-block with $\mu \leq \alpha^{1/n}$, or

(ii) A is an ϵ nearly stochastic matrix where $\epsilon \leq \alpha^2(1 + 2n)$, where $\alpha^2 \leq n2^n\delta$ and, from the hypothesis, $\sqrt{n2^n\delta} < \frac{1}{2}$.

Suppose (ii) holds. Then

$$\begin{aligned}
\alpha^2(1 + 2n) &\leq n2^n\delta(1 + 2n) \\
&\leq n2^n\delta(3n) \\
&\leq 3n^2 2^n\delta \text{ and since } n \geq 2 \\
&\leq n^4 2^n\delta \\
&\leq 4^n n^{3n} \delta^{\frac{1}{n^{2n}}}.
\end{aligned}$$

Thus, A has a ϵ nearly stochastic matrix where $\epsilon \leq 4^n n^{3n} \delta^{\frac{1}{n^{2n}}}$, and the theorem follows.

Now, suppose that (i) holds. Then A has an μ-block where $\mu \leq \alpha^{\frac{1}{n}}$ and where $\alpha^2 \leq n2^n\delta$. Let P be a permutation matrix where

$$PAP^t = \begin{bmatrix} A_1 & U \\ A_{21} & A_2 \end{bmatrix}$$

and $\mu(U) = \mu \leq [n2^n\delta]^{\frac{1}{2n}}$. Then, by Theorem 2, noting $\mu \leq \frac{1}{n}$ from the hypothesis, the dominant eigenvalue μ_1 of

$$\bar{A} = \begin{bmatrix} A_1 & 0 \\ A_{21} & A_2 \end{bmatrix}$$

is such that

$$1 - \mu_1 \leq n^3((n2^n\delta)^{\frac{1}{2n}})^{\frac{1}{n}} + \delta.$$

Since $n^{\frac{1}{n}} \leq 3^{\frac{1}{3}}$ for all n,

$$n^3((n2^n\delta)^{\frac{1}{2n}})^{\frac{1}{n}} + \delta \leq 4n^3\delta^{\frac{1}{2n^2}}.$$

Since μ_1 is an eigenvalue of A_1 or A_2, without loss of generality, we will assume that μ_1 is an eigenvalue of A_1 and that A_1 is $m \times m$.

Applying the induction hypothesis to A_1 yields that A_1 has an ϵ nearly stochastic matrix with

$$\begin{aligned}\epsilon &\leq 4^m m^{3m}[4n^3\delta^{\frac{1}{2n^2}}]^{\frac{1}{m^{2m}}} \\ &= 4^m m^{3m} 4^{\frac{1}{m^{2m}}} n^{\frac{3}{m^{2m}}} \delta^{\frac{1}{2n^2 m^{2m}}},\end{aligned}$$

and, using that $2n^2(n-1)^{2(n-1)} \leq n^{2n}$ (which follows by the binomial formula to show $2k^{2k} \leq (k+1)^{2k}$),

$$\begin{aligned}&\leq 4^m m^{3m} 4n^3 \delta^{\frac{1}{n^{2n}}} \\ &\leq 4^n n^{3n} \delta^{\frac{1}{n^{2n}}}.\end{aligned}$$

This concludes the induction. □

This result extends to the major result of the paper.

THEOREM 4. *If $|1-\lambda_k| < \delta$ for $k = 1, \ldots, r$ and $4^n n^{3n} \delta^{\frac{1}{n^{2n}}} < \frac{1}{4}$, then A has r disjoint ϵ nearly stochastic matrices with $\epsilon \leq 4^n n^{3n} \delta^{\frac{1}{n^{2n}}}$.*

Proof. The proof is by induction on n. For $n = 1$, the proof follows from Theorem 3.

Now, suppose the theorem is true for all $n \times n$ substochastic matrices where $m < n$. Let A denote an $n \times n$ substochastic matrix with $|1-\lambda_k| < \delta$ for $k = 1, \ldots, r$. If $r = 1$, the proof follows from Theorem 3. Thus, we can assume $r \geq 2$.

By Lemma 1, since $|1-\lambda_2| < \delta$, it follows that

$$2^n - |1+\lambda_2|^n \leq n2^n\delta.$$

By Theorem 1, A has a μ-block where $\mu \leq 2n^{\frac{1}{n}}\delta^{\frac{1}{n}}$. Let P be a permutation matrix such that

$$PAP^t = \begin{bmatrix} A_1 & U \\ A_{21} & A_2 \end{bmatrix},$$

where U is a μ-block, A_1 is $n_1 \times n_1$, and A_2 is $n_2 \times n_2$.

Now, by Theorem 2, since $\mu \leq \frac{1}{n}$, by using the hypothesis,

$$\bar{A} = \begin{bmatrix} A_1 & 0 \\ A_{21} & A_2 \end{bmatrix}$$

has r eigenvalues $\mu_1, \ldots, \mu_r$ such that

$$\begin{aligned}|1-\mu_i| &\leq n^3\mu^{\frac{1}{n}} + \delta \\ &\leq n^3[2n^{\frac{1}{n}}\delta^{\frac{1}{n}}]^{\frac{1}{n}} + \delta \\ &\leq 3n^3\delta^{\frac{1}{n^2}} + \delta \\ &\leq 4n^3\delta^{\frac{1}{n^2}}.\end{aligned}$$

Without loss of generality, suppose $\mu_1, \ldots, \mu_s$ are eigenvalues of A_1 while $\mu_{s+1}, \ldots, \mu_r$ are eigenvalues of A_2. Then, by the induction hypothesis, A_1 has s and A_2 has $r-s$ disjoint ϵ nearly stochastic matrices with

$$\epsilon \leq 4^{n_1} n_1^{3n_1} [4n^3 \delta^{\frac{1}{n^2}}]^{\frac{1}{n_1^{2n_1}}}, \qquad \epsilon \leq 4^{n_2} n_2^{3n_2} [4n^4 \delta^{\frac{1}{n^2}}]^{\frac{1}{n_2^{2n_2}}},$$

respectively. Thus A has r disjoint ϵ nearly stochastic matrices with $\epsilon \leq 4^n n^{3n} \delta^{\frac{1}{n^{2n}}}$, the result of the theorem. □

In [7], Seneta gave an upper bound for his condition number, namely,

$$\mathcal{T}(A^{\#}) \leq \frac{n}{|1-\lambda_2|}.$$

Theorem 4 assures us that

$$\epsilon \leq 4^n n^{3n} |1-\lambda_2|^{\frac{1}{n^{2n}}}$$

or

$$\Big[\frac{\epsilon}{4^n n^{3n}}\Big]^{n^{2n}} \leq |1-\lambda_2|.$$

Substitution yields

$$\mathcal{T}(A^{\#}) \leq \frac{n[4^n n^{3n}]^{n^{2n}}}{\epsilon^{n^{2n}}}.$$

Thus, for fixed n, if $\mathcal{T}(A^{\#})$ is sufficiently large, the appearance of two disjoint ϵ nearly stochastic occurs. Of course, we would like much better bounds. Using Theorem 1, in a like manner, we have

$$\mathcal{T}(A^{\#}) \leq \frac{n^2 2^n}{\mu^n},$$

where μ small implies one ϵ nearly stochastic matrix where $\epsilon \leq (n-1)\mu$. However, it is hoped that better bounds can still be found.

Finally, Seneta [7, p. 166] proves that

$$\mathcal{T}(A^{\#}) = \max_{i,j} \sum_{s=1}^{n} \max(a_{is}^{\#}, a_{js}^{\#}).$$

Using this expression, it is easy to see that

$$K(A) \leq \mathcal{T}(A^{\#}) \leq nK(A).$$

Thus, the results we have given in terms of $\mathcal{T}$ can be converted to ones involving K.

REFERENCES

[1] J. L. Franklin, *Matrix Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1968.
[2] D. J. Hartfiel, *Bounds for eigenvalues and eigenvectors of a nonnegative matrix which involve a measure of irreducibility*, SIAM J. Appl. Math., 24 (1973), pp. 6–8.
[3] D. J. Hartfiel, *Results on measures of irreducibility and full indecomposability*, Trans. Amer. Math. Soc., 202 (1975), pp. 357–368.

[4] D. J. Hartfiel and C. D. Meyer, *On the structure of stochastic matrices with a subdominant eigenvalue near* 1, Linear Algebra Appl., 272 (1998), pp. 193–203.
[5] M. S. Lynn and W. P. Timlake, *Bounds for Perron eigenvectors and subdominant eigenvalues of positive matrices*, Linear Algebra Appl., 2 (1969), pp. 143–152.
[6] C. D. Meyer, *Sensitivity of the stationary distribution of a Markov chain*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 715–728.
[7] E. Seneta, *Sensitivity of finite Markov chains under perturbation*, Statist. Probab. Lett., 17 (1993), pp. 163–168.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 547–561
© 1999 Society for Industrial and Applied Mathematics

NORMS OF LARGE TOEPLITZ BAND MATRICES*

A. BÖTTCHER†, S. GRUDSKY‡, A. KOZAK‡, AND B. SILBERMANN†

Abstract. Let $T(b_{jk})$ be a finite collection of infinite Toeplitz band matrices, let $T_n(b_{jk})$ denote their $(n+1)\times(n+1)$ truncations, and put

$$A_n = \sum_j \prod_k T_n(b_{jk}), \quad M_p = \lim_{n\to\infty} \|A_n\|_p,$$

where $\|\cdot\|_p$ stands for the operator norm associated with the l^p-norm ($1 \le p < \infty$) on $\mathbf{C}^{n+1}$. We establish tight two-sided estimates for the difference $M_p - \|A_n\|_p$. It is well known that if $T(b)$ is a Hermitian Toeplitz band matrix and $A_n = T_n(b)$, then $M_2 - \|A_n\|_2$ goes to zero with polynomial speed. We show that such a slow convergence rate is, in a sense, an exceptional case, and we prove that in the generic case $M_p - \|A_n\|_p$ approaches zero with exponential speed. Our results yield good error estimates when computing the norms of certain infinite matrices via their large truncations and, conversely, when determining the norms of certain large matrices by having recourse to known norms of infinite matrices.

Key words. Toeplitz matrix, band matrix, matrix norm

AMS subject classifications. 15A60, 47B35, 65F35

PII. S0895479898343012

1. Introduction and main results.

For $1 \le p < \infty$, we denote by $l^p := l^p(\mathbf{Z}_+)$ and $l^p[0,n]$ the usual l^p spaces over $\mathbf{Z}_+ := \{0,1,2,\ldots\}$ and $\{0,1,\ldots,n\}$, respectively. Given a trigonometric polynomial

$$b(t) = \sum_{j=-s}^{s} b_j\, t^j, \quad t = e^{i\theta} \in \mathbf{T} \tag{1.1}$$

on the complex unit circle $\mathbf{T}$, we let

$$T(b) := (b_{j-k})_{j,k=0}^{\infty} \quad \text{and} \quad T_n(b) := (b_{j-k})_{j,k=0}^{n}$$

stand for the infinite and $(n+1)\times(n+1)$ Toeplitz matrices generated by b. Clearly, $T(b)$ and $T_n(b)$ are band matrices. The matrices $T(b)$ and $T_n(b)$ induce bounded linear operators on l^p and $l^p[0,n]$ in the natural manner. The norms of these operators will be denoted by $\|T(b)\|_p$ and $\|T_n(b)\|_p$:

$$\|T(b)\|_p := \sup_{\substack{x\in l^p \\ \|x\|_p=1}} \|T(b)x\|_p, \quad \|T_n(b)\|_p := \sup_{\substack{x\in l^p[0,n] \\ \|x\|_p=1}} \|T_n(b)x\|_p.$$

*Received by the editors August 3, 1998; accepted for publication (in revised form) by L. Eldén March 22, 1999; published electronically December 9, 1999. This work was supported by DFG-Kooperationsprojekt 436 RUS 113/426 for German and Russian scientists, within the "Memorandum of Understanding" between DFG and RFFI.

http://www.siam.org/journals/simax/21-2/34301.html

†Department of Mathematics, TU Chemnitz, 09107 Chemnitz, Germany (aboettch@mathematik.tu-chemnitz.de).

‡Department of Mechanics and Mathematics, Rostov-on-Don State University, Bolshaya Sadovaya 105, 344 711 Rostov-on-Don, Russia (grudsk@ns.math.rsu.ru). The second and third authors were partially supported by RFFI grant 96-01-01195a. The second author was also partially supported by RFFI grant 98-01-01023.

We may think of $l^p[0,n]$ as a subspace of l^p by identifying $l^p[0,n]$ with the image of the canonical projection

$$P_n : l^p \to l^p, \ \ (x_0, x_1, x_2, \ldots) \mapsto (x_0, \ldots, x_n, 0, 0, \ldots).$$

Accordingly, we can write $T_n(b) = P_n T(b) P_n|\mathrm{Im}\, P_n$. Since $\|P_n\|_p = 1$ and $P_n \to I$ strongly on l^p as $n \to \infty$, it follows that $\|T_n(b)\|_p \le \|T(b)\|_p$ for all n and that $\|T(b)\|_p \le \liminf_{n\to\infty} \|T_n(b)\|_p$. Consequently,

$$\lim_{n\to\infty} \|T_n(b)\|_p = \|T(b)\|_p.$$

The purpose of this paper is to establish tight estimates for the difference $\|T(b)\|_p - \|T_n(b)\|_p$ and the difference of the norms of related operators.

The norm of the operator A induced by the matrix (a_{jk}) on l^1 is given by

$$\|A\|_1 = \sup_k \sum_j |a_{jk}|. \tag{1.2}$$

This implies that

$$\|T_n(b)\|_1 = \|T(b)\|_1 = \sum_{j=-s}^{s} |b_j| \ \text{ for } \ n \ge 2s+1.$$

Hence, in the case $p = 1$, our problem is trivial. For $p = 2$, we have

$$\|T(b)\|_2 = \|b\|_\infty := \max_{t\in\mathbf{T}} |b(t)|,$$

but in general $\|T_n(b)\|_2$ is not explicitly available. We prove the following.

THEOREM 1.1. *There exists a constant c depending only on b such that*

$$\|b\|_\infty (1 - c/n^2) \le \|T_n(b)\|_2 \le \|b\|_\infty \ \textit{for all } n.$$

Thus, $\|T_n(b)\|_2$ approaches $\|T(b)\|_2 = \|b\|_\infty$ at least as $O(1/n^2)$. The eigenvalues of the Hermitian matrix $T_n(b)$ generated by $b(t) = 3 - t - t^{-1}$ are

$$\lambda_k^{(n)} = 3 + 2\cos\frac{k\pi}{n+2} \quad (k = 1, 2, \ldots, n+1).$$

Since $\|T_n(b)\| = \lambda_1^{(n)}$ and $\|b\|_\infty = 5$, it follows that

$$\|b\|_\infty - \|T_n(b)\|_2 = 2\left(1 - \cos\frac{\pi}{n+2}\right) = \frac{\pi^2}{(n+2)^2} + O\left(\frac{1}{n^4}\right),$$

which shows that Theorem 1.1 is sharp.

In case we know more about b, we can say more about the speed with which $\|T_n(b)\|_2$ converges to $\|b\|_\infty$. The function $\|b\|_\infty - |b(t)|$ has finitely many zeros on $\mathbf{T}$, and each zero is of even order.

THEOREM 1.2. *Let 2γ be the maximal order of the zeros of $\|b\|_\infty - |b(t)|$; that is, let $\gamma \ge 1$ be the maximal integer for which there are $t_0 \in \mathbf{T}$ and $D \in (0,\infty)$ such that*

$$\|b\|_\infty - |b(t)| \le D|t - t_0|^{2\gamma} \ \textit{for all } t \in \mathbf{T}. \tag{1.3}$$

Then there exist constants $0 < d_1 < d_2 < \infty$ *independent of* n *such that*

$$\|b\|_\infty(1 - d_2/n^{2\gamma}) \le \|T_n(b)\|_2 \le \|b\|_\infty(1 - d_1/n^{2\gamma}) \text{ for all } n. \tag{1.4}$$

Since b is a trigonometric polynomial, estimate (1.3) is always true with $\gamma = 1$, if only $|b(t_0)| = \|b\|_\infty$. Thus, Theorem 1.1 is a direct consequence of Theorem 1.2. We remark that Theorem 1.2 is well known where case b is real-valued and, hence, $T_n(b)$ is Hermitian; see Kac, Murdock, and Szegő [6]; Grenander and Szegő [4]; Widom [10]; Parter [7]; and Serra [8], [9], for example. In this case, $\|T_n(b)\|_2$ is the maximal eigenvalue of $T_n(b)$, and in the aforementioned works it is shown that if 2γ is the maximal order of the zeros of $\|b\|_\infty - b(t)$, then estimate (1.4) holds.

The matrices one has to deal with in practice are frequently not pure Toeplitz band matrices but perturbed Toeplitz band matrices. Let R_n be the operator

$$R_n : \ l^p \to l^p, \ (x_0, x_1, x_2, \ldots) \mapsto (x_n, x_{n-1}, \ldots, x_0, 0, 0, \ldots).$$

We consider band matrices of the form

$$B_n = T_n(b) + P_nKP_n + R_nLR_n, \tag{1.5}$$

where b is a trigonometric polynomial of the form (1.1) and where the infinite matrices K and L have only a finite number of nonzero entries, which means that

$$P_{n_0}KP_{n_0} = K, \quad P_{n_0}LP_{n_0} = L \tag{1.6}$$

for some n_0. Of course, in (1.5) we identify the images of P_n and R_n with $l^p[0, n]$, while in (1.6) we think of P_n and R_n as acting in l^p. Thus, matrix (1.5) differs from $T_n(b)$ by the $(n_0 + 1) \times (n_0 + 1)$ block K in the upper left and the $(n_0 + 1) \times (n_0 + 1)$ block $R_{n_0}LR_{n_0}$ in the lower right corners. For example, if b_{jk} is a finite collection of trigonometric polynomials, then $B_n = \sum_j \prod_k T_n(b_{jk})$ can be shown to be of the form (1.1), (1.5), and (1.6).

Define $\tilde{b}$ by $\tilde{b}(t) := b(1/t)$ $(t \in \mathbf{T})$. If B_n is given by (1.5) and (1.6), then

$$B_n \to T(b) + K =: B, \quad R_nB_nR_n \to T(\tilde{b}) + L =: \tilde{B} \tag{1.7}$$

strongly as $n \to \infty$. Put

$$M_p^0 := \|T(b)\|_p, \quad M_p := \max\Big(\|B\|_p, \ \|\tilde{B}\|_p\Big).$$

Since $\|T(a)+K\|_p \ge \|T(a)\|_p$ for every trigonometric polynomial a and every compact operator K, and since $\|T(a)\|_p = \|T(\tilde{a})\|_p$ for every trigonometric polynomial a (see, e.g., [3, 6.2(a),(b),(c)]), we always have $M_p^0 \le M_p$.

With s and n_0 given by (1.1) and (1.6), put $m_0 := \max(s, n_0)$. Using (1.2), it is easy to see that

$$\|B_n\|_1 = M_1 \text{ for all } n \ge 2m_0 + 1,$$

so that the case $p = 1$ is again uninteresting. For general p, we will prove the following.

THEOREM 1.3. *Suppose* $M_p^0 < M_p$. *If* $s \ge 1$ *and* $n \ge 4(m_0 + s + 1)$, *then*

$$M_p(1 - c_p\varrho_p^n) \le \|B_n\|_p \le M_p(1 + c_p\varrho_p^n), \tag{1.8}$$

where

$$c_p := 2s\left(\frac{M_p}{M_p^0}\right)^{\frac{m_0+2s+2}{2s}}, \quad \varrho_p := \left(\frac{M_p^0}{M_p}\right)^{\frac{1}{4s}}.$$

As $0 < \varrho_p < 1$, it follows that $\|B_n\|_p$ converges to M_p with exponential speed.

To have a simple concrete example, let

$$B = T(b) + K := \begin{pmatrix} 3 & 84 & 0 & \dots \\ -1 & 3 & -1 & \dots \\ 0 & -1 & 3 & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}$$

and put $B_n := T_n(b) + P_nKP_n$. We have

$$M_2^0 = \|T(b)\|_2 = \|T(\tilde{b})\|_2 = \|b\|_\infty = 5, \quad \|K\|_2 = 85,$$

whence

$$80 = \|K\|_2 - \|T(b)\|_2 \le \|T(b) + K\|_2 \le \|K\|_2 + \|T(b)\|_2 = 90,$$

and thus $M_2^0 < M_2 = \max\Big(\|T(b) + K\|_2, \|T(\tilde{b})\|_2\Big)$. Furthermore,

$$c_2 = 2\left(\frac{M_2}{M_2^0}\right)^{\frac{m_0+2s+2}{2s}} \le 2\left(\frac{90}{5}\right)^{\frac{1+2+2}{2}} < 2750,$$
$$\varrho_2 = \left(\frac{M_2^0}{M_2}\right)^{\frac{1}{4s}} \le \left(\frac{5}{80}\right)^{\frac{1}{4}} = \frac{1}{2}$$

and therefore, by Theorem 1.3,

$$\Big|\, \|B_n\|_2 - \|B\|_2 \Big| \le M_2 c_2 \varrho_2^n \le 90 \cdot 2750 \cdot \left(\frac{1}{2}\right)^n = \frac{247\,500}{2^n}.$$

In particular,

$$\Big|\, \|B_{31}\|_2 - \|B\|_2 \Big| < 0.00011526.$$

Matlab gives $\|B_{31}\|_2 = 84.11171884\ldots$, and hence we arrive at the following fairly precise information on the norm of the infinite matrix B:

$$\|B\|_2 = 84.11171884 \pm 0.00011526.$$

The crucial assumption of Theorem 1.3 is the strict inequality $M_p^0 < M_p$. Let $\mathcal{A}_p$ be the set of all operators on l^p that are of the form $T(a) + K$ with a trigonometric polynomial a and an operator K whose matrix has only finitely many nonzero entries. We remark that if $T(a) + K = T(b) + L \in \mathcal{A}_p$, then necessarily $a = b$ and $K = L$. Furthermore, let $\mathcal{A}_p^0$ be the set of all operators $T(a)+K \in \mathcal{A}_p$ such that $\|T(a)+K\|_p > \|T(a)\|_p$.

PROPOSITION 1.4. *The set $\mathcal{A}_2^0$ is an open and dense subset of $\mathcal{A}_2$.*

If $p = 2$, the operators B and $\tilde{B}$ defined by (1.7) belong to $\mathcal{A}_2$. Proposition 1.4 therefore tells us that the strict inequality $M_2^0 < M_2$ represents the generic case,

whereas the equality $M_2^0 = M_2$ is the exceptional case. Thus, Theorem 1.3 says that generically $\|B_n\|_2 - M_2$ approaches zero with exponential speed.

In concrete situations, we nevertheless often encounter the exceptional case $M_p^0 = M_p$. For instance, we are in this case whenever $B_n = T_n(b)$. The following theorem does certainly not give sharp estimates, but it provides us with a universal estimate with good constants.

THEOREM 1.5. *Suppose $M_p^0 = M_p$. If $s \geq 1$ and $n \geq 4\max(n_0, 72s)$, then*

$$M_p\Big(1 - \frac{291}{p}\frac{s}{n}\Big) \leq \|B_n\|_p \leq M_p\Big(1 + \frac{8}{p}\frac{s}{n}\Big).$$

Full proofs of the results cited above will be given in sections 2, 3, and 4.

The present paper emerged from our papers [1], [2], [5], in which we studied the behavior of the norms of the inverses of (not necessarily band) Toeplitz matrices. In [5], we gave estimates for the speed with which $\|T_n^{-1}(a)\|_1$ approaches $\|T^{-1}(a)\|_1$, provided $T(a)$ is invertible (notice that in the $p = 1$ case we can have recourse to (1.2)). In [1], we considered the case where $T(a)$ is not invertible but still semidefinite, and we established results which show that the spectral condition numbers $\kappa(T_n(a)) := \|T_n(a)\|_2\,\|T_n^{-1}(a)\|_2$ can grow like $\log n$, n^α, or even $e^{\alpha n}$ as $n \to \infty$. Finally, in [2] we were able to prove that $\|T_n(a)\|_p\|T_n^{-1}(a)\|_p$ converges to $\|T^{-1}(a)\|_p\,\|T^{-1}(a)\|_p$. To estimate the *speed* of convergence, we need know the speed with which $\|T_n(a)\|_p$ converges to $\|T(a)\|_p$ and the speed with which $\|T_n^{-1}(a)\|_p$ goes to $\|T^{-1}(a)\|_p$. The former problem (for band matrices) is the subject of the present paper. The latter question can be answered by combining the results of this paper with appropriate approximation arguments, which will be dealt with in a forthcoming paper.

To prove estimates like (1.4) and (1.8), we need sufficiently precise techniques. The techniques employed in sections 2, 3, and 4 are essential refinements of those developed in [1], [2], [5]. These techniques allow us to give nearly self-contained proofs of the results quoted above.

2. Pure Toeplitz matrices. This section is devoted to the proof of Theorem 1.2.

Proof. Let b be given by (1.1). We have

$$T_n(b)T_n(\overline{b}) = T_n(|b|^2) + P_nKP_n + R_nLR_n$$

where $K = -H(b)H(\tilde{\overline{b}})$, $L = -H(\tilde{b})H(\overline{b})$, and $H(a)$ is the Hankel matrix $H(a) := (a_{j+k+1})_{j,k=0}^\infty$; also recall that $\tilde{a}(t) := a(1/t)$.

Since $(Kx, x) \leq 0$ and $(Lx, x) \leq 0$ for all $x \in l^2$, it follows that $(T_n(b)T_n(\overline{b})x, x) \leq (T_n(|b|^2)x, x)$ for all $x \in l^2$, whence

$$\|T_n(b)\|_2^2 = \|T_n(b)T_n(\overline{b})\|_2 \leq \|T_n(|b|^2)\|_2.$$

As the estimate (1.4) is known for Hermitian matrices (see Serra [8], [9]), we obtain

$$\|T_n(b)\|_2^2 \leq \|\,|b|^2\|_\infty\Big(1 - d/n^{2\gamma}\Big) = \|b\|_\infty^2\Big(1 - d/n^{2\gamma}\Big),$$

which implies that $\|T_n(b)\|_2 \leq \|b\|_\infty(1 - d_1/n^{2\gamma})$ with $d_1 := d/2 \in (0, \infty)$.

We are left with the estimate $\|T_n(b)\|_2 \geq \|b\|_\infty(1 - d_2/n^{2\gamma})$. Obviously, (1.6) is satisfied with $n_0 = s - 1$. Put $M = \|b\|_\infty^2$. Since

$$T_n\Big(b(t/t_0)\Big) = \operatorname{diag}(1, t_0^{-1}, \ldots, t_0^{-n})T_n(b)\,\operatorname{diag}(1, t_0, \ldots, t_0^n),$$

we can, without loss of generality, assume that $M = |b(1)|^2$. By virtue of (1.3), there exists a constant $C \in (0,\infty)$ such that

$$M - |b(e^{i\theta})|^2 \le C|\theta|^{2\gamma} \quad \text{for all} \quad \theta \in [-\pi,\pi]. \tag{2.1}$$

For $f \in l^2[0,n]$,

$$\begin{aligned} \|T_n(b)\|_2^2\|f\|_2^2 &= \|T_n(b)T_n(\overline{b})\|_2\|f\|_2^2 \ge |(T_n(b)T_n(\overline{b})f,f)| \\ &\ge M(f,f) - \Big|\Big((M - T_n(|b|^2))f,f\Big)\Big| - |(P_nKP_nf,f)| - |(R_nLR_nf,f)|. \end{aligned} \tag{2.2}$$

We identify $l^2[0,n]$ with the polynomials of degree at most n and, accordingly,

$$(f,g) = \frac{1}{2\pi}\int\limits_{-\pi}^{\pi} f(e^{i\theta})\,\overline{g(e^{i\theta})}\,d\theta \quad \text{for} \quad f,g \in l^2[0,n].$$

Put $j = \gamma+1$, let m be the integer satisfying $mj \le n < (m+1)j$, and define $f \in l^2[0,n]$ by

$$f(e^{i\theta}) = (1 + e^{i\theta} + \cdots + e^{im\theta})^j. \tag{2.3}$$

It is easily seen that

$$f(e^{i\theta}) = e^{imj\theta/2}\left(\frac{\sin((m+1)\theta/2)}{\sin(\theta/2)}\right)^j, \tag{2.4}$$

and in [1] it was shown that

$$\|f\|_2 > \frac{4}{3\sqrt{\pi}}j^{-1/4}(m+1)^{j-1/2}. \tag{2.5}$$

We now estimate the terms on the right of (2.2). By (1.6),

$$|(P_nKP_nf,f)| = |(P_{n_0}KP_{n_0}f,f)| = |(KP_{n_0}f,P_{n_0}f)| \le \|K\|_2\|P_{n_0}f\|_2^2,$$

and from (2.3) we infer that $\|P_{n_0}f\|_2$ is independent on n if only n (and thus m) is large enough. Consequently, by (2.5),

$$\frac{|(P_nKP_nf,f)|}{\|f\|_2^2} \le \frac{C_1}{(m+1)^{2j-1}} < \frac{C_1j^{2j-1}}{n^{2j-1}} = \frac{C_1(\gamma+1)^{2\gamma+1}}{n^{2\gamma+1}} \tag{2.6}$$

with some $C_1 \in (0,\infty)$ independent of n. Analogously,

$$\frac{|(R_nLR_nf,f)|}{\|f\|_2^2} \le \frac{C_2(\gamma+1)^{2\gamma+1}}{n^{2\gamma+1}} \tag{2.7}$$

where $C_2 \in (0,\infty)$ does not depend on n. From (2.1) and (2.4) we get

$$\begin{aligned} \Big|\Big((M - T_n(|b|^2))f,f\Big)\Big| &= \Big|\Big((M - |b|^2)f,f\Big)\Big| \\ &= \frac{1}{2\pi}\int\limits_{-\pi}^{\pi}\Big(M - |b(e^{i\theta})|^2\Big)\left|\frac{\sin((m+1)\theta/2)}{\sin(\theta/2)}\right|^{2j}d\theta \end{aligned}$$

$$\leq \frac{1}{2\pi}\int\limits_{|\theta|<1/(m+1)} C|\theta|^{2\gamma}\left(\frac{(m+1)|\theta|}{2}\right)^{2j}\left(\frac{\pi}{2}\frac{2}{|\theta|}\right)^{2j}d\theta$$
$$+\frac{1}{2\pi}\int\limits_{|\theta|>1/(m+1)} C|\theta|^{2\gamma}\left(\frac{\pi}{2}\frac{2}{|\theta|}\right)^{2j}d\theta$$
$$\leq C_3\frac{(m+1)^{2j}}{(m+1)^{2\gamma+1}}+C_4\frac{1}{(m+1)^{2j-2\gamma}}$$
$$= C_3\frac{(m+1)^{2\gamma+2}}{(m+1)^{2\gamma+1}}+C_4\frac{1}{(m+1)^2}\leq C_5 n$$

with constants $C_3, C_4, C_5 \in (0,\infty)$ independent on n. Taking into account (2.5) we therefore obtain

$$\frac{|((M-T_n(|b|^2))f,f)|}{\|f\|^2}\leq C_6\frac{n}{n^{2j-1}}=C_6\frac{1}{n^{2\gamma}}, \tag{2.8}$$

where $C_6 \in (0,\infty)$ is independent on n. Combining (2.2), (2.6), (2.7), and (2.8), we arrive at the estimate

$$\|T_n(b)\|_2^2\geq M\left(1-C_7\frac{1}{n^{2\gamma}}\right),$$

where $C_7 \in (0,\infty)$ is independent of n, whence

$$\|T_n(b)\|_2\geq\|b\|_\infty\left(1-C_7\frac{1}{n^{2\gamma}}\right) \tag{2.9}$$

whenever $n^{2\gamma} > C_7$. This completes the proof of Theorem 1.2. □

3. The exceptional case. The purpose of this section is to prove Theorem 1.5 and to prepare the proof of Theorem 1.3.

The following result extends a weakened version of Theorem 1.1 to general p.

PROPOSITION 3.1. *Let b given by* (1.1). *If $n \geq 144s$, then*

$$\|T(b)\|_p\left(1-\frac{144}{p}\frac{s}{n}\right)\leq\|T_n(b)\|_p\leq\|T(b)\|_p.$$

Proof. The case $s = 0$ is trivial. So let $s \geq 1$. Put $M_0 := \|T(b)\|_p$ and $\varepsilon_n := M_0 - \|T_n(b)\|_p$. It is clear that ε_n decreases monotonously to zero. Given n, pick $x \in l^p[0,n]$ so that

$$\|x\|_p=\left(\sum_{j=0}^{n}|x_j|^p\right)^{1/p}=1 \text{ and } \|T_n(b)x\|_p=M_0-\varepsilon_n. \tag{3.1}$$

Let $l := [n/4]$, where here and in what follows $[\alpha]$ is the integral part of the number α. We claim that there is a natural number m such that

$$l+s\leq m\leq 3l-s \text{ and } \sum_{j=m-s+1}^{m+s}|x_j|^p<\left[\frac{l}{s}\right]^{-1}. \tag{3.2}$$

Indeed, we have

$$1 \geq \sum_{j=l+1}^{3l} |x_j|^p \geq \sum_{k=0}^{d-1} \Big(\sum_{j=l+2ks+1}^{l+2(k+1)s} |x_j|^p \Big),$$

where $d := [l/s]$, which shows that there exists a k_0 such that $0 \leq k_0 \leq d-1$ and the inner sum does not exceed d^{-1}. Letting $m := l + (2k_0+1)s$, we obtain our claim (3.2).

Define $P_{[j,k]} := P_k - P_{j-1}$. Thus, $P_{[j,k]}$ acts by the rule

$$P_{[j,k]} : l^p \to l^p,\ (x_0, x_1, x_2, \ldots) \mapsto (0, \ldots, 0, x_j, \ldots, x_k, 0, 0, \ldots).$$

Since $T_n(b)$ has only $2s+1$ nonzero diagonals, we get

$$\begin{aligned} T_n(b)x &= P_m T_n(b)x + P_{[m+1,n]} T_n(b)x \\ &= P_m T_n(b)(x_j)_{j=0}^{m+s} + P_{[m+1,n]} T_n(b)(x_j)_{j=m-s+1}^{n}, \end{aligned}$$

where $m = l + (2k_0+1)s$ is as above. Hence

$$\|T_n(b)x\|_p^p = \|P_m T_n(b)(x_j)_{j=0}^{m+s}\|_p^p + \|P_{[m+1,n]} T_n(b)(x_j)_{j=m-s+1}^{n}\|_p^p.$$

It can be directly verified that

$$R_n P_{[m+1,n]} T_n(b) P_{[m-s+1,n]} R_n = P_{n-m-1} T_{n-m+s-1}(\tilde{b}).$$

Because R_n is isometric on $l^p[0,n]$ and $P_m = P_m P_{m+s}$, it follows that

$$\begin{aligned} \|T_n(b)x\|_p^p &\leq \|T_{m+s}(b)(x_j)_{j=0}^{m+s}\|_p^p + \|T_{n-m+s-1}(\tilde{b})(x_{n-j})_{j=0}^{n-m+s-1}\|_p^p \\ &\leq (M_0 - \varepsilon_{m+s})^p \sum_{j=0}^{m+s} |x_j|^p + (M_0 - \varepsilon_{n-m+s-1})^p \sum_{j=m-s+1}^{n} |x_j|^p. \end{aligned}$$

As $\max(m+s, n-m+s-1) \leq n-l$ and as ε_n goes monotonously to zero, we see that

$$\begin{aligned} \|T_n(b)x\|_p^p &\leq (M_0 - \varepsilon_{n-l})^p \Big(\sum_{j=0}^{m+s} |x_j|^p + \sum_{j=m-s+1}^{n} |x_j|^p \Big) \\ &\leq (M_0 - \varepsilon_{n-l})^p + M_0^p [l/s]^{-1}, \end{aligned}$$

the last inequality resulting from (3.1) and (3.2). The second equality of (3.1) now gives

$$(M_0 - \varepsilon_n)^p \leq (M_0 - \varepsilon_{n-l})^p + M_0^p [l/s]^{-1}$$

or, equivalently,

$$\Big(1 - \frac{\varepsilon_n}{M_0}\Big)^p - \Big(1 - \frac{\varepsilon_{n-l}}{M_0}\Big)^p \leq \Big[\frac{l}{s}\Big]^{-1}. \tag{3.3}$$

Recall that $l = [n/4]$ and put $n_1 = [4n/3]$. Taking into account the identity

$$n = [4n/3] - \Big[\frac{[4n/3]}{4}\Big]$$

and the inequality

$$\left[\left[\frac{[4n/3]}{4}\right]\Big/ s\right]^{-1} \le \frac{12s}{n} \quad \text{for} \quad n > 5 + 4s,$$

and replacing in (3.3) the number n by n_1, we arrive at the inequality

$$\left(1 - \frac{\varepsilon_{n_1}}{M_0}\right)^p - \left(1 - \frac{\varepsilon_n}{M_0}\right)^p \le \frac{12s}{n} \quad \text{for} \quad n > 5 + 4s. \tag{3.4}$$

Now substitute the n in (3.4) consecutively by

$$n_1 = \left[\frac{4}{3}n\right],\ n_2 = \left[\frac{4}{3}n_1\right], \ldots, n_j = \left[\frac{4}{3}n_{j-1}\right],$$

add the corresponding inequalities, and then pass to the limit $j \to \infty$. What results is

$$1 - \left(1 - \frac{\varepsilon_n}{M_0}\right)^p \le \frac{12s}{n} + \sum_{j=1}^{\infty} \frac{12s}{n_j}.$$

As $[4n/3] > 6n/5$ for $n \ge 8$, we get

$$1 - \left(1 - \frac{\varepsilon_n}{M_0}\right)^p \le \frac{12s}{n} \sum_{j=0}^{\infty} \left(\frac{5}{6}\right)^j = \frac{72s}{n}.$$

Thus, if $n \ge 72s$, then

$$\varepsilon_n \le \left(1 - \left(1 - \frac{72s}{n}\right)^{1/p}\right) M_0.$$

Because $(1-x)^{1/p} \ge 1 - 2x/p$ for $0 \le x \le 1/2$, we finally obtain $\varepsilon_n \le 144 M_0 s/(pn)$ for $n \ge 144s$. Since $n > 5 + 4s$ and $n \ge 8$ whenever $n \ge 144s$ and $s \ge 1$, the proof is complete. □

Now let B_n be given by (1.1), (1.5), and (1.6). The following result gives a universal estimate of $\|B_n\|_p$ from above and shows that the right estimate of Theorem 1.5 is true without the additional hypothesis $M_p^0 = M_p$.

PROPOSITION 3.2. *If $s \ge 1$ and $n \ge 4\max(n_0, 2s+2)$, then*

$$\|B_n\|_p \le M_p\left(1 + \frac{8}{p}\frac{s}{n}\right).$$

Proof. Choose $x = (x_j)_{j=0}^n$ so that $\|x\|_p = 1$. Put $l := [n/4]$. By assumption, $l \ge \max(n_0, s)$. As in the proof of Proposition 3.1, there is a natural number m satisfying (3.2) and we obtain

$$\|B_n x\|_p^p = \|P_m B_n (x_j)_{j=0}^{m+s}\|_p^p + \|P_{[m+1,n]} B_n (x_j)_{j=m-s+1}^n\|_p^p.$$

We have

$$R_n P_{[m+1,n]} B_n P_{[m-s+1,n]} R_n = P_{n-m-1} R_n B_n R_n P_{n-m+s-1},$$

which shows that

$$\begin{aligned}\|B_n x\|_p^p &= \Big\|P_m\Big(T(b)+K\Big)(x_j)_{j=0}^{m+s}\Big\|_p^p \\ &\quad +\Big\|P_{n-m-1}\Big(T(\tilde b)+L\Big)(x_{n-j})_{j=0}^{n-m+s-1}\Big\|_p^p \\ &\le \|B\|_p^p \sum_{j=0}^{m+s}|x_j|^p + \|\tilde B\|_p^p \sum_{j=m-s+1}^{n}|x_j|^p \\ &\le \max\Big(\|B\|_p^p,\ \|\tilde B\|_p^p\Big)\Big(1+\sum_{j=m-s+1}^{m+s}|x_j|^p\Big) \\ &\le M_p^p\Big(1+[l/s]^{-1}\Big);\end{aligned}$$

here we made use of the notation (1.7), of the equality $\|x\|_p = 1$, and of (3.2). If $n \ge 8+8s$ (which is satisfied by assumption), then $[l/s]^{-1} \le 8s/n$, whence

$$\|B_n\|_p \le M_p\Big(1+\frac{8s}{n}\Big)^{1/p} \le M_p\Big(1+\frac{8}{p}\frac{s}{n}\Big).$$

This completes the proof. □

Here now is the *proof of Theorem* 1.5.

The upper estimate is immediate from Proposition 3.2. We are thus left with the estimate from below. By Proposition 3.1, there is an $x \in l^p[0,[n/2]]$ such that $\|x\|_p = 1$ and

$$\|T_{[n/2]}(b)x\|_p \ge M_p^0\Big(1-\frac{144}{p}\frac{s}{[n/2]}\Big).$$

Let $l = [n/4]$ and define $y \in l^p[0,n]$ by

$$y = (\underbrace{0,\dots,0}_{l}, x_0, x_1, \dots, x_{[n/2]}, \underbrace{0,\dots,0}_{l'}).$$

By assumption, $l \ge n_0$ and $l' \ge n_0$. From (1.6) we therefore deduce that $B_n y = T_n(b)y$, whence

$$\begin{aligned}\|B_n y\|_p &= \|T_n(b)y\|_p \ge \|P_{[l+1,l+[n/2]+1]}T_n(b)y\|_p \\ &= \|T_{[n/2]}(b)x\|_p \ge M_p^0\Big(1-\frac{144}{p}\frac{s}{[n/2]}\Big) \\ &\ge M_p^0\Big(1-\frac{288}{p}\frac{s}{n-2}\Big) \ge M_p^0\Big(1-\frac{291}{p}\frac{s}{n}\Big),\end{aligned}$$

the last estimate resulting from the assumption that $n \ge 288s \ge 288$. As $M_p^0 = M_p$, the proof of Theorem 1.5 is complete.

4. The generic case. We begin with the proof of Proposition 1.4.

Proof. It is clear that $\mathcal{A}_2^0$ is an open subset of $\mathcal{A}_2$. To show that $\mathcal{A}_2^0$ is dense in $\mathcal{A}_2$, pick $T(a)+K \in \mathcal{A}_2$ and suppose $\|T(a)+K\|_2 = \|T(a)\|_2 = \|a\|_\infty$. Given any number $\varepsilon > 0$, we can find a complex number λ such that $|\lambda| < \varepsilon$ and $\|a+\lambda\|_\infty > \|a\|_\infty$. Then

$$\|T(a)+K+\lambda I\|_2 = \|T(a+\lambda)+K\|_2 \ge \|T(a+\lambda)\|_2 = \|a+\lambda\|_\infty > \|a\|_\infty,$$

and since $T(a)+K+\lambda P_n$ converges strongly to $T(a)+K+\lambda I$, it follows that

$$\liminf_{n\to\infty}\|T(a)+K+\lambda P_n\|_2 \geq \|T(a)+K+\lambda I\|_2 > \|a\|_\infty.$$

Consequently, letting $K_0 := K+\lambda P_n$, we see that $\|T(a)+K_0\|_2 > \|T(a)\|_2$ for all sufficiently large n, while

$$\Big\|\Big(T(a)+K_0\Big)-\Big(T(a)+K\Big)\Big\|_2 \leq |\lambda|\,\|P_n\|_2 = |\lambda| < \varepsilon.$$

This completes the proof of Proposition 1.4. □

To prove Theorem 1.3, we need the following lemma, which is of interest on its own. Recall that $B = T(b)+K$, where b is given by (1.1) and K satisfies (1.6). Also, let $s \geq 1$ without loss of generality. Put $m_0 := \max(s, n_0)$.

LEMMA 4.1. *If $\|T(b)\|_p < \|B\|_p$, then there exists an $x_0 = \big(x_j^{(0)}\big)_{j=0}^{\infty} \in l^p$ such that*

$$\|x_0\|_p = 1, \quad \|Bx_0\|_p = \|B\|_p \tag{4.1}$$

and

$$\Big(\sum_{j=m+1}^{\infty} |x_j^{(0)}|^p\Big)^{1/p} < \Big(\|T(b)\|_p \Big/ \|B\|_p\Big)^{\frac{m-(m_0+s)}{2s}} \tag{4.2}$$

for every $m > m_0 + s$.

Proof. Let $M_0 := \|T(b)\|_p$ and $M := \|B\|_p$. There are $x_k = \big(x_j^{(k)}\big)_{j=0}^{\infty} \in l^p$ such that

$$\|x_k\|_p = 1 \ \text{ and } \ \|Bx_k\|_p^p = M^p - \delta_k, \tag{4.3}$$

where δ_k goes monotonously to zero as $k \to \infty$. Pick any natural number $m \geq m_0$. By (1.6),

$$Bx_k = P_m\Big(T(b)+K\Big)\Big(x_j^{(k)}\Big)_{j=0}^{m+s} + P_{[m+1,\infty)}T(b)\Big(x_j^{(k)}\Big)_{j=m-s+1}^{\infty},$$

whence

$$M^p - \delta_k = \|Bx_k\|_p^p \leq M^p \sum_{j=0}^{m+s} |x_j^{(k)}|^p + M_0^p \sum_{j=m-s+1}^{\infty} |x_j^{(k)}|^p. \tag{4.4}$$

Defining

$$O_m^{(k)} := \sum_{j=m+1}^{\infty} |x_j^{(k)}|^p,$$

we can rewrite (4.4) in the form

$$1 - \gamma_k \leq \Big(1 - O_{m+s}^{(k)}\Big) + qO_{m-s}^{(k)},$$

where $\gamma_k := \delta_k/M^p$ and $q := (M_0/M)^p$. Thus,

$$O_{m+s}^{(k)} \leq qO_{m-s}^{(k)} + \gamma_k \ \text{ for all } \ m \geq m_0. \tag{4.5}$$

Application of (4.5) to $m = m_0, m_0 + 2s, \dots, m_0 + 2js$ gives

$$\begin{aligned} O^{(k)}_{m_0+(2j+1)s} &\le q^{j+1} O^{(k)}_{m_0-s} + (q^j + q^{j-1} + \cdots + 1)\gamma_k \\ &= q^{j+1} O^{(k)}_{m_0-s} + \frac{1-q^{j+1}}{1-q}\gamma_k. \end{aligned}$$

As $O^{(k)}_{m_0-s} \le 1$ and $q \in (0,1)$, we get

$$O^{(k)}_{m_0+(2j+1)s} \le q^{j+1} + \frac{\gamma_k}{1-q}. \tag{4.6}$$

Analogously, putting $m = m_0 + 1 + 2js, m_0 + 2 + 2js, \dots, m_0 + (2s-1) + 2js$ in (4.5), we arrive at the inequalities

$$O^{(k)}_{m_0+v+(2j+1)s} \le q^{j+1} + \frac{\gamma_k}{1-q} \tag{4.7}$$

for $v \in \{1, 2, \dots, 2s-1\}$. Letting $r := q^{1/(2s)}$, the inequalities (4.6) and (4.7) can be rewritten in the form

$$O^{(k)}_{m_0+s+j} \le r^j + \frac{\gamma_k}{1-q} \quad (j = 0, 1, 2, \dots). \tag{4.8}$$

We now show that $\{x_k\}_{k=1}^\infty$ is a Cauchy sequence in l^p. Without loss of generality, we can assume that $\{P_m x_k\}_{k=1}^\infty$ is a convergent sequence in $l^p[0,m]$ for each m (if this is not the case, we can pass to a subsequence of $\{x_k\}$ by a diagonalization process). Given $\varepsilon > 0$, choose natural numbers N, k_1, and k_2 so that

$$r^{(N-(m_0+s))/p} < \frac{\varepsilon}{4}, \quad \sum_{j=0}^{N} |x_j^{(k_1)} - x_j^{(k_2)}|^p < \left(\frac{\varepsilon}{4}\right)^p, \quad \frac{\gamma_{k_1} + \gamma_{k_2}}{1-q} < \frac{\varepsilon}{4}.$$

Then, by (4.8), $\|x_{k_1} - x_{k_2}\|_p$ is at most

$$\begin{aligned} &\left(\sum_{j=0}^{N} |x_j^{(k_1)} - x_j^{(k_2)}|^p\right)^{1/p} + \left(\sum_{j=N+1}^{\infty} |x_j^{(k_1)}|^p\right)^{1/p} + \left(\sum_{j=N+1}^{\infty} |x_j^{(k_2)}|^p\right)^{1/p} \\ &\quad \le \frac{\varepsilon}{4} + \left(O_N^{(k_1)}\right)^{1/p} + \left(O_N^{(k_2)}\right)^{1/p} \\ &\quad \le \frac{\varepsilon}{4} + r^{(N-(m_0+s))/p} + r^{(N-(m_0+s))/p} + \frac{\gamma_{k_1} + \gamma_{k_2}}{1-q} \\ &\quad < \frac{\varepsilon}{4} + \frac{\varepsilon}{4} + \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \varepsilon. \end{aligned}$$

This shows that $\{x_k\}$ is a Cauchy sequence.

Let $x_0 \in l^p$ be the limit of x_k as $k \to \infty$. Passing to the limit $k \to \infty$ in (4.3) we arrive at (4.1). Repeating the above arguments with x_k replaced by x_0, we get the analogue of (4.8) for x_0; that is,

$$\sum_{i=m_0+s+j+1}^{\infty} |x_i^{(0)}|^p \le r^j \quad (j = 0, 1, 2, \dots),$$

which is equivalent to (4.2). ∎

Here now is the *proof of Theorem* 1.3.

Put $M_0 := M_p^0$ and $M := M_p$. Proposition 3.2 implies that $\limsup_{n\to\infty} \|B_n\|_p \le M$, and from (1.7) we infer that $M \le \liminf_{n\to\infty} \|B_n\|_p$. Consequently,

$$\lim_{n\to\infty} \|B_n\|_p = M.$$

It follows that there are $x_n = (x_j^{(n)})_{j=0}^n \in l^p[0,n]$ such that

$$\|x_n\|_p = 1,\ \|B_n x_n\|_p = \|B_n\|_p,\ \|B_n x_n\|_p^p = M^p + \varepsilon_n, \tag{4.9}$$

where $\varepsilon_n \in \mathbf{R}$ and $\varepsilon_n \to 0$ as $n \to \infty$. Again set $l = [n/4]$ and note that $l > m_0 := \max(s, n_0)$ by assumption. For $l < m < 2l$, we have

$$\begin{aligned}\|B_n x_n\|_p^p = &\Big\|P_m\Big(T_n(b) + K\Big)\Big(x_j^{(n)}\Big)_{j=0}^{m+s}\Big\|_p^p \\ &+ \Big\|P_{[m+1,n-(m+1)]}T_n(b)\Big(x_j^{(n)}\Big)_{j=m-s+1}^{n-(m+1)+s}\Big\|_p^p \\ &+ \Big\|P_{[n-m,n]}\Big(T_n(b) + R_nLR_n\Big)\Big(x_j^{(n)}\Big)_{j=n-m-s}^{n}\Big\|_p^p.\end{aligned}$$

Since R_n is isometric on $l^p[0,n]$, we therefore obtain from (4.9) that

$$M^p + \varepsilon_n \le M^p\Big(\sum_{j=0}^{m+s} |x_j^{(n)}|^p\Big) + M_0^p\Big(\sum_{j=m-s+1}^{n-(m+1)+s} |x_j^{(n)}|^p\Big) + M^p\Big(\sum_{j=n-m-s}^{n} |x_j^{(n)}|^p\Big).$$

With

$$O_m^{(n)} := \sum_{j=m+1}^{n-(m+1)} |x_j^{(n)}|^p$$

we get $M^p + \varepsilon_n \le M^p(1 - O_{m+s}^{(n)}) + M_0^p O_{m-s}^{(n)}$, whence

$$O_{m+s}^{(n)} \le qO_{m-s}^{(n)} - \gamma_n,$$

where $q := (M_0/M)^p$ and $\gamma_n := \varepsilon_n/M^p$. The last inequality is of the form (4.5); consequently, analogous to (4.8), we have

$$O_{m_0+s+j}^{(n)} \le r^j - \frac{1-r^j}{1-q}\gamma_n, \quad r := (M_0/M)^{p/(2s)} \tag{4.10}$$

as long as $O_{m_0+s+j}^{(n)}$ is well defined, i.e., for all j satisfying $2(m_0 + s + j + 1) \le n$ or, equivalently, for

$$j \le n/2 - 1 - m_0 - s. \tag{4.11}$$

As the left-hand side of (4.10) is nonnegative provided (4.11) holds, we get

$$\gamma_n \le \frac{1-q}{1-r} r^{n/2-2-m_0-s} \le 2s\left(\frac{M_0}{M}\right)^{\frac{p(n-2m_0-2s-4)}{4s}}.$$

Thus,

$$\varepsilon_n = M^p\gamma_n \le M^p 2s\Big(\frac{M_0}{M}\Big)^{\frac{p(n-2m_0-2s-4)}{4s}}.$$

Put $\eta_n := \|B_n\|_p - M$. By (4.9), $\eta_n = (M^p+\varepsilon_n)^{1/p} - M$. Hence, if $\varepsilon_n \ge 0$, then $\eta_n \le \varepsilon_n/(M^{p-1}p)$, while $\eta_n < 0$ if $\varepsilon_n < 0$. In summary, we arrive at the estimate

$$\eta_n \le \frac{M}{p}2s\Big(\frac{M_0}{M}\Big)^{\frac{p(n-2m_0-2s-4)}{4s}} \le 2sM\Big(\frac{M_0}{M}\Big)^{\frac{n-2m_0-2s-4}{4s}} = Mc_p\varrho_p^n,$$

which is the estimate from above in (1.8).

To prove the left inequality of (1.8), we make use of Lemma 4.1. For the sake of definiteness, suppose $M = \|T(b)+K\|_p$. (The proof is analogous to the case $M = \|T(\tilde{b})+L\|_p$.) Let x_0 satisfy (4.1) and (4.2). Put $v := [n/2]$ ($> m_0$) and $y_n := P_v x_0$. By (1.6),

$$\begin{aligned}
\|B_n y_n\|_p &= \|P_n\Big(T(b)+K\Big)y_n\|_p \\
&= \Big\|\Big(T(b)+K\Big)\Big(x_j^{(0)}\Big)_{j=0}^{\infty} - P_{[n+1,\infty)}\Big(T(b)+K\Big)\Big(x_j^{(0)}\Big)_{j=v-s+1}^{v} \\
&\quad - \Big(T(b)+K\Big)\Big(x_j^{(0)}\Big)_{j=v+1}^{\infty}\Big\|_p \\
&\ge \Big\|\Big(T(b)+K\Big)x_0\Big\|_p - \Big\|P_{[n+1,\infty)}\Big(T(b)+K\Big)\Big(x_j^{(0)}\Big)_{j=v-s+1}^{v} \\
&\quad + \Big(T(b)+K\Big)\Big(x_j^{(0)}\Big)_{j=v+1}^{\infty}\Big\|_p \\
&\ge M - 2M\Big(\sum_{j=v-s+1}^{\infty}|x_j^{(0)}|^p\Big)^{1/p}.
\end{aligned}$$

From (4.2) and the equality $v = [n/2]$ we now infer that

$$\|B_n y_n\|_p \ge M - 2M\Big(\frac{M_0}{M}\Big)^{\frac{v-s-m_0-s}{2s}} \ge M - 2M\Big(\frac{M_0}{M}\Big)^{\frac{n-2m_0-4s-4}{2s}} \ge M - Mc_p\varrho_p^n.$$

Because $\|y_n\|_p \le 1$, we finally obtain

$$\|B_n\|_p \ge \|B_n y_n\|_p/\|y_n\|_p \ge \|B_n y_n\|_p \ge M - Mc_p\varrho_p^n,$$

and this is the lower estimate in (1.8).

REFERENCES

[1] A. Böttcher and S. Grudsky, *On the condition numbers of large semi-definite Toeplitz matrices*, Linear Algebra Appl., 279 (1998), pp. 285–301.

[2] A. Böttcher, S. Grudsky, and B. Silbermann, *Norms of inverses, spectra, and pseudospectra of large truncated Wiener-Hopf operators and Toeplitz matrices*, New York J. Math., 3 (1997), pp. 1–31.

[3] A. Böttcher and B. Silbermann, *Analysis of Toeplitz Operators*, Akademie-Verlag, Berlin, 1989.

[4] U. Grenander and G. Szegő, *Toeplitz Forms and Their Applications*, University of California Press, Berkeley, Los Angeles, 1958.

[5] S. Grudsky and A. Kozak, *On the convergence speed of the norms of inverses of truncated Toeplitz operators*, in Integro-Differential Equations Appl., Rostov-on-Don University Press, Rostov-on-Don, Russia, 1995, pp. 45–55 (in Russian).

[6] M. Kac, W.L. Murdock, and S. Szegő, *On the eigenvalues of certain Hermitian forms*, J. Rational Mech. Anal., 2 (1953), pp. 767–800.

[7] S. V. Parter, *Extreme eigenvalues of Toeplitz forms and applications to elliptic difference equations*, Trans. Amer. Math. Soc., 99 (1966), pp. 153–192.

[8] S. Serra, *On the extreme spectral properties of Toeplitz matrices generated by L^1 functions with several minima/maxima*, BIT, 36 (1996), pp. 135–142.

[9] S. Serra, *On the extreme eigenvalues of Hermitian (block) Toeplitz matrices*, Linear Algebra Appl., 270 (1997), pp. 109–129.

[10] H. Widom, *On the eigenvalues of certain Hermitian operators*, Trans. Amer. Math. Soc., 88 (1958), pp. 491–522.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 562–580
© 1999 Society for Industrial and Applied Mathematics

ACCURATE SINGULAR VALUE DECOMPOSITIONS OF STRUCTURED MATRICES*

JAMES DEMMEL†

Abstract. We present new $O(n^3)$ algorithms to compute very accurate singular value decompositions of Cauchy matrices, Vandermonde matrices, and related "unit-displacement-rank" matrices. These algorithms compute all the singular values with guaranteed relative accuracy, independent of their dynamic range. In contrast, previous $O(n^3)$ algorithms can potentially lose all relative accuracy in the tiniest singular values.

Key words. singular value decomposition, high accuracy, Cauchy matrix, Vandermonde matrix, displacement rank

AMS subject classifications. 65F15, 65F20, 65G05

PII. S0895479897328716

1. Introduction.

A *singular value decomposition (SVD)* of a real matrix G is the factorization $G = U\Sigma V^T$, where U and V are orthogonal matrices and Σ is nonnegative and diagonal. If G is m-by-n, with $m \geq n$ (otherwise transpose G), then U is m-by-n, $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ with $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$, and V is n-by-n. We call the columns u_i of $U = [u_1, \ldots, u_n]$ the *left singular vectors*, the columns v_i of $V = [v_1, \ldots, v_n]$ the *right singular vectors*, and the σ_i the *singular values*. When G is complex we write instead $G = U\Sigma V^H$ and take U and V unitary.

Our goal is to compute the SVD (i.e., the u_i, v_i, and σ_i) to *high relative accuracy* and with cost $O(n^3)$, i.e., roughly the same cost as prior, less accurate dense matrix algorithms. By high relative accuracy we mean the following:

- The error $|\sigma_i - \hat{\sigma}_i|$ in the computed singular value $\hat{\sigma}_i$ is bounded by $O(\varepsilon)\sigma_i$, where ε is machine precision, i.e., the relative error is small.
- The angle $\theta(u_i, \hat{u}_i)$ between the true left singular vector u_i and the computed vector $\hat{u}_i$ is bounded by $O(\varepsilon)/\mathrm{relgap}_i$, where $\mathrm{relgap}_i = \min_{j \neq i} |\sigma_j - \sigma_i|/\sigma_i$ is the relative gap between σ_i and the nearest other singular value. An analogous statement holds for the computed right singular vectors $\hat{v}_i$.

In contrast, conventional numerical algorithms (QR iteration, divide and conquer, bisection and inverse iteration, traditional Jacobi, etc.) only compute the SVD with *high absolute accuracy*. This means that the error in the singular value is bounded only by $O(\varepsilon)\sigma_1$ instead of $O(\varepsilon)\sigma_i$. Thus the largest singular values are computed with high relative accuracy, but the tiniest singular values (those near or less than $\varepsilon\sigma_1$) may have no relative accuracy at all. Similarly, the error in the computed singular vectors can also be σ_1/σ_i times as large.

In an earlier paper [8] we showed that if we were given a *rank-revealing decomposition* (RRD) $G = XDY^T$, i.e., a decomposition where D is diagonal and X and

*Received by the editors October 10, 1997; accepted for publication (in revised form) by I. Ipsen August 7, 1998; published electronically December 15, 1999. This material is based in part upon work supported by Advanced Research Projects Agency contract DAAH04-95-1-0077 (via subcontract ORA4466.02 with the University of Tennessee), the Department of Energy grant DE-FG03-94ER25219 and contract W-31-109-Eng-38 (via subcontracts 20552402 and 941322401 with Argonne National Laboratory), National Science Foundation grant ASC-9313958, and NSF Infrastructure grants CDA-8722788 and CDA-9401156.

http://www.siam.org/journals/simax/21-2/32871.html

†Computer Science Division and Mathematics Department, University of California, Berkeley, CA 94720 (demmel@cs.berkeley.edu).

Y are full column rank and well conditioned (but otherwise arbitrary), then we could compute the SVD of G to high relative accuracy from X, D, and Y in $O(n^3)$ time. For this to work, the computed $\hat{X}$, $\hat{D}$, and $\hat{Y}$ must be sufficiently accurate:

- Each entry of $\hat{D}$ is known to high relative accuracy: $|D_{ii} - \hat{D}_{ii}| = O(\varepsilon)|D_{ii}|$.
- $\hat{X}$ and $\hat{Y}$ have small norm errors, i.e., $\|X - \hat{X}\| = O(\varepsilon)\|X\|$ and $\|Y - \hat{Y}\| = O(\varepsilon)\|Y\|$.

To summarize, our high accuracy SVD algorithm consists of two steps:

1. Compute the RRD $G = XDY^T$ sufficiently accurate.
2. Compute the SVD of XDY^T using the algorithm from [8] (see also section 2 below).

In [8] a variety of matrix classes were described that permitted the RRD $G = XDY^T$ to be computed sufficiently accurately. In particular we considered the RRD provided by Gaussian elimination with complete pivoting (GECP): X and Y are both unit triangular matrices with off-diagonals bounded by one in magnitude. However, the usual implementation of GECP will often not compute X, D, and Y as accurately as required; a new way to compute the triangular factors is needed.

In particular, suppose C is a square Cauchy matrix: $C_{ij} = 1/(x_i + y_j)$, where the x_i and y_j are given floating point numbers. There is a well-known formula for the determinant of C:

$$\det(C) = \frac{\prod_{1 \le i < j \le n}(x_j - x_i)(y_j - y_i)}{\prod_{1 \le i,j \le n}(x_i + y_j)} \quad . \tag{1}$$

Every factor $x_j - x_i$, $y_j - y_i$, or $x_i + y_j$ is computable to high relative accuracy, as well as their products and quotients, in floating point arithmetic. Thus $\det(C)$ is computable to high relative accuracy. Since every submatrix of a Cauchy matrix is Cauchy, every minor of C can be computed to high relative accuracy. The same is true of the *quasi-Cauchy matrix* $G = D_1 C D_2$, where D_1 and D_2 are diagonal. Finally, since every entry of every Schur complement of G is the quotient of minors of G (or just a minor), all the intermediate and final results of GECP on G can be computed to high relative accuracy using (1). In other words, we can compute the triangular factors of G, and thus its SVD, to high relative accuracy, where G is defined by the x_i, y_j, and diagonal entries of D_1 and D_2. This extends straightforwardly to rectangular Cauchy matrices.

This does *not* mean that small relative changes in these parameters cause small relative changes in the singular values. Indeed, the 1-by-1 example with $x_1 = 1$ and $y_1 = -1 + \varepsilon$ shows that this is not true. It does mean that our algorithm is forward stable, and simply gets an accurate answer for the floating point values of the parameters stored in the machine, whether or not the answer is sensitive to small changes in the parameters. Section 7 addresses the separate perturbation theory question of how small changes in the input data can affect the SVD. We will see, for example, that the Hilbert matrix is nearly perfectly conditioned, in the sense that small relative changes in its Cauchy parameters $x_i = i$ and $y_j = j - 1$ cause only small relative changes in its SVD. Ours is the only algorithm not using expensive high precision arithmetic that can accurately compute the SVD of matrices like the Hilbert matrix, which are very ill conditioned from the conventional point of view. Also, there are some rational interpolation problems where the data is known exactly, in which case our algorithm is most appropriate [17].

Unfortunately, using (1) in the most straightforward way to implement GECP would cost $O(n^5)$ operations by using it to compute each of the $O(n^3)$ intermediate

results that would be produced by Gaussian elimination. The first main result in this paper is an algorithm that reduces the cost from $O(n^5)$ to $\frac{4}{3}n^3$. Similar observations were made for several closely related (noneigenvalue and non-SVD) factorizations in work on predictive pivoting for low displacement rank matrices [17, 4, 15, 3, 30, 21], where the goal was generally to reduce the complexity to $O(n^2)$. In particular, see Theorem 4.1 in [15]. But with complete pivoting it appears impossible to reduce the complexity below $\Theta(n^3)$. Our algorithm can be derived straightforwardly from (1), which is the approach we take here.

Among other examples, we compute the SVD of a 100-by-100 Hilbert matrix. The computed singular values range over 150 orders of magnitude and agree to 15 decimal digits with the singular values computed using a conventional algorithm with 200-decimal digit arithmetic; the singular vectors agree to 14 or more digits. This result is presented in section 4.

The second result is a similar high accuracy RRD for Vandermonde matrices. It is based on the observation that if V is Vandermonde, so that $V_{ij} = x_i^{j-1}$, and $F_{ij} = \exp(2\sqrt{-1}\pi(i-1)(j-1)/n)/\sqrt{n}$ is the discrete Fourier transform matrix, then VF is (essentially) quasi-Cauchy, and the parameters determining the quasi-Cauchy structure can be computed to sufficient accuracy to use the previous algorithm. The observation that VF is quasi-Cauchy was also exploited in the displacement rank literature [16, (3.1)], [15, Section 7], but we again derive it from first principles. This result is presented in section 5.

We observe more generally that if G is defined by the Sylvester equation (also called *displacement equation*)

$$XG + GY = d_1 d_2^T, \tag{2}$$

where

- X and Y are diagonalizable matrices with accurately known eigendecompositions, and
- d_1 and d_2 are column vectors,

then G can be accurately transformed into a quasi-Cauchy matrix, and the above techniques applied. (If X and Y are normal matrices, this process is slightly simpler than if they are not.) This result is presented in section 6.

A matrix satisfying (2) is said to have *displacement rank* equal to 1, because the right-hand side has rank 1. It is natural to ask about the possibility of computing accurate SVDs of matrices with higher displacement rank. For example, general Toeplitz matrices have displacement rank 2. We do not currently see any way to extend our results to higher displacement rank than 1. In section 6 we argue that no high accuracy SVD for general Toeplitz matrices can exist, in the sense that they exist for Cauchy and Vandermonde matrices, and describe conditions under which such algorithms can exist.

In section 7 we describe a simple estimator for the relative error in the SVD we compute, in the case where the input data is uncertain.

2. Computing an accurate SVD of $G = XDY^T$. For completeness, we restate Algorithm 3.1 from [8] (see also [10]) for computing a high accuracy SVD of the RRD $G = XDY^T$.

ALGORITHM 1. Compute a high accuracy SVD of $G = XDY^T$.

1. Perform QR factorization with pivoting on XD to get $XD = QRP$, where P is a permutation. Thus $G = QRPY^T$.

2. Multiply to get $W = RPY^T$. This must be *conventional* matrix multiplication, e.g., Strassen's method [22] may not be used. Thus $G = QW$.

3. Compute the SVD of $W = \bar{U}\Sigma V^T$ using one-sided Jacobi [9]. Thus $G = Q\bar{U}\Sigma V^T$.

4. Multiply $U = Q\bar{U}$. Thus $G = U\Sigma V^T$ is the desired SVD.

The statements of Theorems 3.1 and 3.2 from [8] which describes the accuracy of this algorithm are as follows.

THEOREM 2.1 (see [8, Theorem 3.1]). *Let D' be a diagonal matrix, chosen so that $R' = D'^{-1}R$ is as well conditioned as possible. Then in floating point arithmetic with machine precision ε, Algorithm* 1 *computes the SVD of G with relative accuracy $\eta = O(\varepsilon\kappa(R') \cdot \max(\kappa(X), \kappa(Y)))$.*

In practice $\kappa(R')$ is usually close to 1 so that the error bound is close to the minimum $O(\varepsilon \max(\kappa(X), \kappa(Y)))$ given by the perturbation bound in [8, Theorem 2.1]. But the best we can prove is as shown below.

THEOREM 2.2 (see [8, Theorem 3.2]). *D' can be chosen so that $\kappa(R') = O(\min(n2^n, n^{3/2}\kappa(X)))$.*

In theory [10, 29, 9, 28, 11] the Jacobi rotations during the one-sided Jacobi in step (3) must be applied to the right side of W to guarantee the bounds. But in practice, applying them from the left is significantly faster, and has never been found to be significantly less accurate (see the comments on speed in section 4.1 below).

The factor $\kappa(R')$ in the error bound depends on how well the pivoting during the QR decomposition of XD "reveals the rank" of XD. The bound $O(n2^n)$ comes from the standard column-pivoting algorithm [18] and choosing $D'_{ii} = R_{ii}$, but better alternatives are available [34, 1, 5, 6, 20, 24, 32]. For example, M. Gu has a pivoting scheme that reduces $O(2^n)$ to $O(n^{1+(1/4)\log_2 n})$, analogous to the pivot growth bound for GECP [19] (see also [31]). For our numerical experiments in section 4.1, $\kappa(R')$ never exceeded 5.

We can eliminate the factor $\kappa(R')$, yielding the relative error bound $\eta = O(\varepsilon \cdot \max(\kappa(X), \kappa(Y)))$, by using the more expensive Algorithm 3.2 from [8].

ALGORITHM 2. Compute a high accuracy SVD of $G = XDY^T$.

1. Compute the SVD of $XD = \bar{U}\bar{\Sigma}\bar{V}^T$ using one-sided Jacobi. Thus $G = \bar{U}\bar{\Sigma}\bar{V}^T Y^T$.

2. Multiply $W = \bar{\Sigma}(\bar{V}^T Y^T)$, respecting parentheses. Strassen may be used inside the parentheses. Thus $G = \bar{U}W$.

3. Compute the SVD $W = \bar{\bar{U}}\Sigma V^T$ using one-sided Jacobi. Thus $G = \bar{U}\bar{\bar{U}}\Sigma V^T$.

4. Multiply $U = \bar{U}\bar{\bar{U}}$. Thus $G = U\Sigma V^T$ is the desired SVD.

However, Algorithm 2 is likely to be up to twice as expensive as Algorithm 1 because it does two Jacobi SVDs instead of one (although we note recent work that has made Jacobi-based SVDs about as fast as the QR-based algorithm [13]). Since R' is unlikely to be large, we cannot recommend Algorithm 2 for general use.

3. Assumptions about floating point arithmetic. We use the conventional error model for floating point arithmetic:

$$\mathrm{fl}(a \odot b) = (a \odot b)(1 + \delta), \tag{3}$$

where $\odot \in \{+, -, \times, /\}$, and $|\delta| \le \varepsilon$, where ε is machine precision. In particular, we assume that neither overflow nor underflow occur, since both can destroy relative accuracy in the result.

This assumption implies that expressions like the one in (1) can be evaluated to high relative accuracy. Products and quotients of many floating point numbers

such as in (1) can in principle be quite susceptible to over/underflow; this was not a problem in any test that we ran (once we replaced expressions like $(\prod_i a_i)/(\prod_i b_i)$ with $\prod_i (a_i/b_i)$) but could be guarded against by testing and scaling, or use of IEEE over/underflow flags.

Furthermore, for our algorithm for Vandermonde matrices we further assume that floating point arithmetic is implemented with a guard digit, so that $\mathrm{fl}(a \pm b)$ is exactly $a \pm b$ when cancellation occurs, i.e., when $.5 < a/b < 2$ in binary arithmetic. Almost all existing machines (except the Cray C90 and its predecessors and emulators) satisfy these conditions.

When a and b are complex numbers, we assume that (3) continues to hold, with δ a tiny complex number bounded by $|\delta| = O(\varepsilon)$ [22]. This is true for any reasonable implementation of complex arithmetic. Note that this is weaker than requiring that both real and imaginary parts be computed to high relative accuracy.

4. Quasi-Cauchy matrices. We begin with the simple observation that GECP on G can be implemented as follows. We write this decomposition as $G = XDY^T$, where X and Y are (possibly row permuted) unit lower triangular matrices. Suppose without loss of generality that the first k pivot rows and columns have been chosen as the leading k rows and columns. Then the kth Schur complement $S^{(k)} = G_{22} - G_{21}G_{11}^{-1}G_{12}$ satisfies

$$G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} = \begin{bmatrix} I & 0 \\ G_{21}G_{11}^{-1} & I \end{bmatrix} \cdot \begin{bmatrix} G_{11} & G_{12} \\ 0 & S^{(k)} \end{bmatrix}.$$

The next pivot row and column is chosen by finding the largest entry in $S^{(k)}$; assume without loss of generality that it is in the (1,1) position of $S^{(k)}$. Finally, $D_{k+1,k+1} = S_{11}^{(k)}$, the $(k+1)$st column of X below the diagonal is first column of $S^{(k)}$ divided by $S_{11}^{(k)}$, and the $(k+1)$st row of Y^T to the right of the diagonal is the first row of $S^{(k)}$ divided by $S_{11}^{(k)}$. Therefore, if we can derive a formula for computing all the entries of $S^{(k)}$ in c operations per entry, we will have an algorithm requiring $\sum_{i=1}^{n} ci^2 = \frac{c}{3}n^3 + O(n^2)$ operations to compute the same triangular factorization as GECP.

We use the fact that

$$S_{rs}^{(k)} = \det(G([1:k, r], [1:k, s]))/\det(G(1:k, 1:k)),$$

where we use Matlab notation for submatrices of $G = D_1 C D_2$. Substituting (1) into this expression and canceling terms, we get that

$$\begin{aligned} S_{rs}^{(k)} &= D_{1,r} \cdot D_{2,s} \cdot \frac{\prod_{1 \le i \le k}(x_r - x_i)(y_s - y_i)}{(x_r + y_s)\prod_{1 \le i \le k}(x_i + y_s)(x_r + y_i)} \\ &= S_{rs}^{(k-1)} \frac{(x_r - x_k)(y_s - y_k)}{(x_k + y_s)(x_r + y_k)}, \end{aligned}$$

where $S_{rs}^{(0)} = D_{1,r}D_{2,s}/(x_r + y_s)$ is the original matrix entry [15, Eq. (2.10)]. Thus, updating each entry of the Schur complement from the previous Schur complement costs $c = 8$ operations, or few as $c = 4$ operations if all $2n^2$ possible values of $x_r - x_s$, $y_r - y_s$, and $1/(x_r + y_s)$ are precomputed. This makes the overall cost of the algorithm range from $\frac{8}{3}n^3$ (when done in place, with no extra storage) to $\frac{4}{3}n^3$ (with an extra $2n^2$ storage). The $\frac{8}{3}n^3$ operations include $\frac{2}{3}n^3$ divisions, which are very expensive,

whereas the $\frac{4}{3}n^3$ operations are all multiplications. To complete the algorithm, each time the Schur complement is updated, it must be pivoted so that $S_{11}^{(k)}$ is the largest entry; this requires $\frac{1}{3}n^3$ comparisons. Also, after completing the factorization the kth row and column must be divided by the kth diagonal entry. The entire algorithm is shown below.

ALGORITHM 3. HIGH ACCURACY GECP OF A QUASI-CAUCHY MATRIX.

for $r = 1 : n$ and $s = 1 : n$
  $G_{rs} = D_{1,r}D_{2,s}/(x_r + y_s)$
endfor
for $k = 1 : n - 1$
  Find the largest absolute entry in $G(k : n, k : n)$
  Swap rows and columns of G, and entries of x and y, so that G_{kk} is largest
  for $r = k + 1 : n$ and $s = k + 1 : n \ldots$ overwrite $G(k + 1 : n, k + 1 : n)$ by $S^{(k)}$
    $G_{rs} = G_{rs}(x_r - x_k)(y_s - y_k)/[(x_k + y_s)(x_r + y_k)]$
  endfor
endfor
$D = \mathrm{diag}(G)$
$X = \mathrm{tril}(G) * D^{-1} + I$ ... tril means strict lower triangle
$Y = (D^{-1} * \mathrm{triu}(G) + I)^T$ ... triu means strict upper triangle

All operations involve multiplication, division, and sums and differences of input data and so are accurate to high relative accuracy.

4.1. Numerical experiments. Our challenge is to test the accuracy of a routine that we claim is more accurate than any other. We do this in several ways:

1. We can compute the SVD using a conventional algorithm implemented in very high precision arithmetic, and compare the answers to our new algorithm (implemented in Matlab on a SPARC 10, running IEEE double precision arithmetic with $\varepsilon = 2^{-53} \approx 10^{-16}$). We have only performed this rather expensive test for the 100-by-100 Hilbert matrix, using Mathematica [37] with 200-decimal digit software floating-point, and rounding the final answers back to 16 decimal digits, using the command

```
N[SingularValues[N[Table[1/(i+j-1),{i,100},{j,100}],200],
   Tolerance->10^(-200)],16].
```

The singular values range from about 2.2 down to $5.8 \cdot 10^{-151}$, as shown in the top of Figure 1. If we were to plot the pivots from Algorithm 3 (the $D_{i,i}$) on the same graph, they would be visually nearly indistinguishable; the ratios $D_{i,i}/\sigma_i$ of pivots to singular values all lie in the range [1/16, 16], and 86% are in [1/4,4]. The relative gaps all exceed .6, so we expect all singular vectors from both Mathematica and from our new algorithm to be very accurate. The largest relative difference between a singular value from Mathematica and from our new algorithm was less than $4 \cdot 10^{-15}$ or about 34ε. The largest angle between a right singular vector from Mathematica and from our new algorithm was less than $6.5 \cdot 10^{-15}$, or 58ε. The largest angle between a left singular vector from Mathematica and from our algorithm was less than $2.8 \cdot 10^{-14}$, or 255ε. A surface plot of the singular vector (or equivalently eigenvector) matrix is shown in the bottom of Figure 1, which reveals interesting patterns of "parallel valleys." This is shown in a different way in Figure 2, which plots dark dots for positive eigenvector components and leaves white space for the negative eigenvectors components. This amounts to a low resolution contour plot of the eigenvector matrix. The top of Figure 2 shows

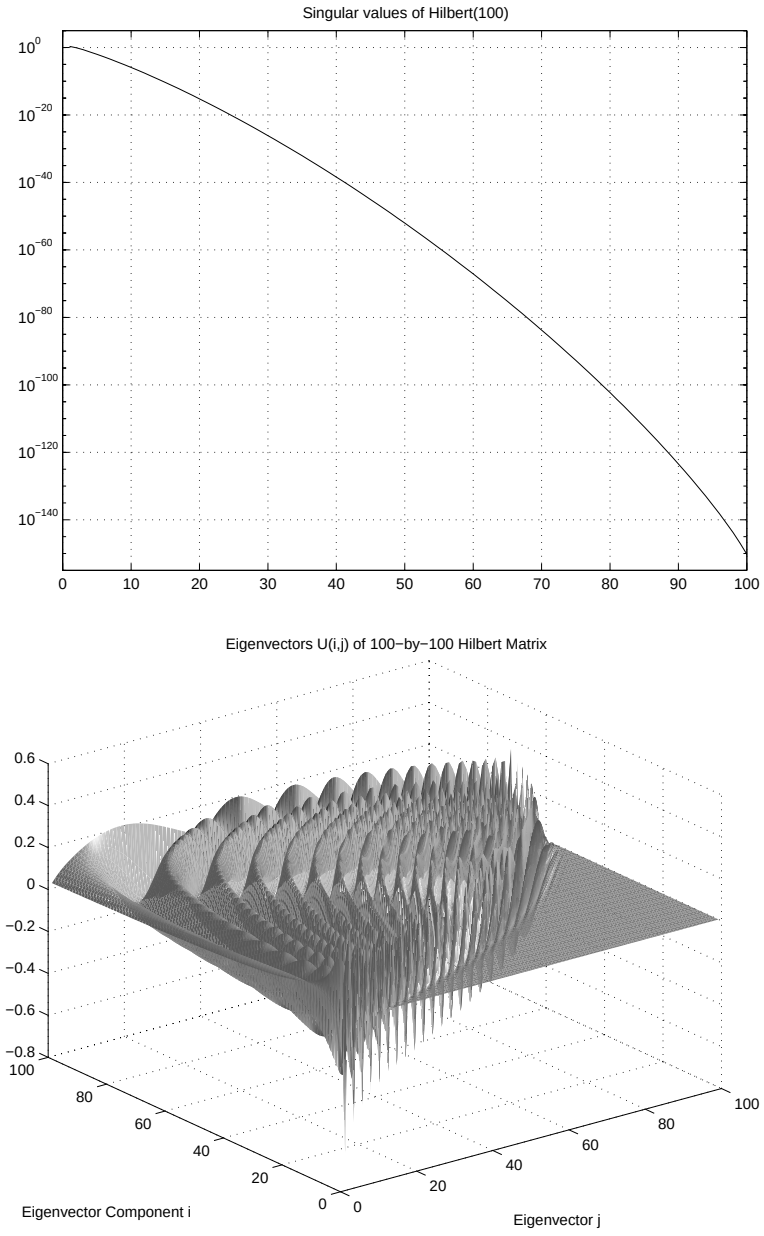


FIG. 1. *Singular values and vectors of the* 100*-by-*100 *Hilbert matrix.*

the results from Mathematica. Although it is somewhat difficult to see, the number of sign changes in eigenvector j (the number of transitions from a positive (black) component to a negative (white) component, or white to black, in column j) is equal to $j-1$. For example, the first column is entirely black (no sign changes) and the last column alternates signs (99 sign changes). Furthermore, the sign changes in eigenvector j interlace those in eigenvector $j+1$ (in the sense that the zero crossings of the piecewise linear curve with corners $(i, U_{i,j})$ interlace the zero crossings of the curve with corners

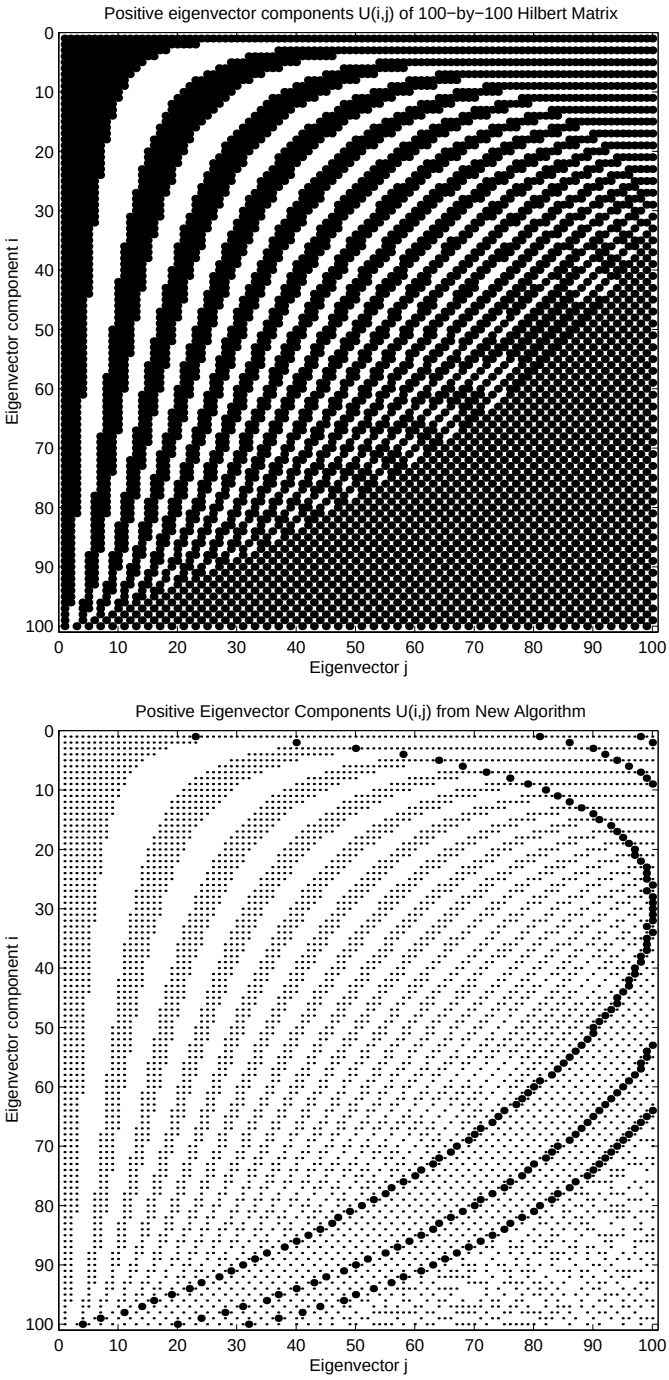


FIG. 2. *Positive components of the singular vectors of the* 100-*by*-100 *Hilbert matrix. The top picture shows the results from Mathematica, and the lower picture shows the results from the new algorithm. In the lower picture, the gray entries are the positive components. The leftmost contour of black dots corresponds to entries of size* .2*, the next contour of black dots corresponds to entries of size* 10^{-7}*, and the rightmost contour of black dots corresponds to entries of size* 10^{-14}*. The* $O(10^{-14})$ *errors can make the smallest entries, those below the rightmost contour, have incorrect signs, as can be seen by the irregular pattern of gray dots below the rightmost contour.*

$(i, U_{i,j+1})$). These properties are as predicted by the theory of totally positive (or oscillation) matrices [14, 26], of which the Hilbert matrix is an example. (Totally nonnegative (positive) matrices are those matrices whose every minor is nonnegative (positive). Oscillation matrices are totally nonnegative matrices A such that A^k is totally positive for some positive k.) The bottom of Figure 2 shows the results from the new algorithm. They agree everywhere except for the components whose magnitudes are less than 10^{-14}, which are on the bottom right, below the contour of black dots. The magnitudes of the components quickly drop from 10^{-14} to near 10^{-74} in this part of the eigenvector matrix. The 200-decimal digit computation with Mathematica computed all these correctly, but the new algorithm has an $O(\varepsilon)$ absolute error in each component, so the signs of such small components are uncertain. In contrast, a conventional SVD algorithm in double precision arithmetic (as implemented in Matlab) loses all significant figures for the 79 singular values less than ε, and most of these are wrong by many orders of magnitude.
The next seven tests use only conventional floating point.

2. We confirm that we have computed an accurate SVD in the conventional sense by measuring

$$Q_1 = \frac{\|G - U\Sigma V^T\|}{\varepsilon\|G\|} + \frac{\|UU^T - I\|}{\varepsilon} + \frac{\|VV^T - I\|}{\varepsilon}$$

and confirming that it is $O(1)$ or perhaps $O(n)$.

3. We compute the SVD of G using a conventional algorithm and confirm that the singular values agree with the singular values computed by our new algorithm to within $\pm O(\varepsilon)\sigma_1$. We measure this as follows:

$$Q_2 = \max_i \frac{|\sigma_{conventional,i}(G) - \sigma_{new,i}(G)|}{\varepsilon\sigma_{conventional,1}(G)} \quad .$$

Note that Q_2 should be $O(1)$ to confirm that our new algorithm computes the largest singular values (those between σ_1 and $\varepsilon\sigma_1$) at least as accurately as a conventional algorithm. We also confirm that the singular vectors are at least as accurate as those from a conventional algorithm by measuring whether Q_3 is $O(1)$:

$$Q_3 = \max_i \frac{\max(\theta(u_{conventional,i}(G), u_{new,i}(G)), \theta(v_{conventional,i}(G), v_{new,i}(G)))}{\varepsilon / \left(\frac{\min j \neq i |\sigma_{new,i}(G) - \sigma_{new,j}(G)|}{\sigma_{new,1}(G)}\right)} \quad .$$

4. We compute each entry of G^{-1} to high relative accuracy using a well-known formula similar to (1) and compute the SVD of G^{-1} using a conventional algorithm. We confirm that these singular values agree with the reciprocals of the singular values computed by our new algorithm to within $\pm O(\varepsilon)/\sigma_n$, by measuring

$$Q_4 = \max_i \frac{|1/\sigma_{new,i}(G) - \sigma_{conventional,i}(G^{-1})|}{\varepsilon/\sigma_{new,n}(G)}$$

to see if it is $O(1)$. This confirms that the smallest singular values of G (those between σ_n and σ_n/ε) are computed accurately. We also confirm that

the singular vectors are at least as accurate as those from the conventional algorithm, by confirming that the following quantity is $O(1)$:

$$Q_5 = \max_i \frac{\max(\theta(u_{conventional,i}(G^{-1}), v_{new,i}(G)), \theta(v_{conventional,i}(G^{-1}), u_{new,i}(G)))}{\varepsilon / \left(\frac{\min j \neq i |\sigma^{-1}_{new,i}(G) - \sigma^{-1}_{new,j}(G)|}{\sigma^{-1}_{new,n}(G)} \right)}.$$

5. We use the fact that G^{-1} is quasi-Cauchy to compute its SVD using our new algorithm, and confirm that it agrees (after reciprocating the singular values) with the SVD computed from G. This is measured by

$$Q_6 = \max_i \frac{|\sigma_{new,i}(G) - \sigma^{-1}_{new,i}(G^{-1})|}{\varepsilon \sigma_{new,i}(G)} \quad .$$

We also confirm that the singular vectors agree to within the error bound determined by the relative gap, via

$$Q_7 = \max_i \frac{\max(\theta(u_{new,i}(G^{-1}), v_{new,i}(G)), \theta(v_{new,i}(G^{-1}), u_{new,i}(G)))}{\varepsilon / \left(\frac{\min j \neq i |\sigma^{-1}_{new,i}(G) - \sigma^{-1}_{new,j}(G)|}{\sigma^{-1}_{new,i}(G)} \right)} \quad .$$

Here are our test cases:

1. 10 Hilbert matrices of dimensions $10, 20, 30, \ldots, 100$ ($x_i = i$, $y_i = i - 1$, $D_{1,i} = D_{2,i} = 1$);
2. 10 Hilbert matrices of dimensions $10, 20, \ldots, 100$, with strong diagonal scaling ($x_i = i$, $y_i = i - 1$, $D_{1,i} = 10^{20u}$, $D_{2,i} = 10^{20u}$, where u is a random number uniformly distributed from 0 to 1);
3. 10 random Cauchy matrices of dimensions $10, 20, \ldots, 100$ ($x_i = u$, $y_i = -u$, $D_{1,i} = D_{2,i} = 1$, where u is a random number uniformly distributed from 0 to 1);
4. 10 random Cauchy matrices of dimensions $10, 20, \ldots, 100$ ($x_i = 10^{10u}$, $y_i = -10^{10u}$, $D_{1,i} = D_{2,i} = 1$, where u is a random number uniformly distributed from 0 to 1);
5. 10 random Cauchy matrices of dimensions $10, 20, \ldots, 100$ with strong diagonal scaling ($x_i = 10^{10u}$, $y_i = -10^{10u}$, $D_{1,i} = 10^{10u}$, $D_{2,i} = 10^{10u}$, where u is a random number uniformly distributed from 0 to 1).

The results were as follows (given as the maximum values over all 50 test cases). As shown, the results were about as accurate as expected.

$\max Q_1$	$\max Q_2$	$\max Q_3$	$\max Q_4$	$\max Q_5$	$\max Q_6$	$\max Q_7$
2752	21	30	21	101	98	137

In addition, we measured some other quantities of numerical interest. The condition numbers of the matrices tested ranged from about 10^8 to 10^{182}, a very wide range. The relative gaps all exceeded 10^{-3}, meaning that all singular vectors were computed with small norm errors, about $10^3\varepsilon$. The condition numbers of the triangular factors X and Y, which limit the accuracy of Algorithm 1, never exceeded 379. Indeed, in tens of thousands of other examples run using a direct search technique [23] in an attempt to maximize $\mu = \max(\max_{ij} |X^{-1}|_{ij}, \max_{ij} |Y^{-1}|_{ij})$, μ never exceeded 7 for $n \leq 20$. Another factor that limits the accuracy of Algorithm 1, the condition number of R' in the statement of Theorem 3.1 from [8], never exceeded 5. The number of Jacobi sweeps required for convergence in Algorithm 1 depends very much on how

we do one-sided Jacobi. The matrix W to which we apply one-sided Jacobi consists of a well conditioned matrix premultiplied by a diagonal matrix with widely varying entries, so that the rows of W vary widely in norm. If we apply Jacobi rotations to W from the left (the algorithm analyzed in [9]), then convergence is rapid, with never more than 8 Jacobi sweeps required, and an average of 4.6 sweeps (including one just to confirm convergence, which does not apply any rotations). If we apply Jacobi rotations to W from the right, which has a more satisfying theoretical analysis [10, 29, 28], then convergence is much slower, with the number of sweeps being 50 for modest sized examples and growing rapidly with n. The above results were run with Jacobi rotations applied on the left.

4.2. The symmetric positive definite case. Significant simplifications are possible in this case, although the overall algorithm is still $O(n^3)$. If G is symmetric and positive definite, the largest pivot is always on the diagonal. Thus, only the diagonal entries of the $S^{(k)}$ need to be computed to determine the pivot order. This leads to the following algorithm:

Algorithm 4. High Accuracy SVD of a Symmetric Positive Definite Quasi-Cauchy Matrix G.

1. Compute only the diagonals of each $S^{(k)}$, and use this to determine the pivot order. This costs just $O(n^2)$.

2. Reorder the data determining G in the correct order, and compute the reordered G's LDU factorization using Algorithm 2.5 of [3]. This costs $O(n^2)$, and is forward stable for the same reasons as Algorithm 3.

3. Rewrite $LDU = \hat{L}\hat{L}^T$, the Cholesky factorization. This also costs $O(n^2)$.

4. Apply one-sided Jacobi to $\hat{L}$ to compute its SVD. The left singular vectors of L are the eigenvectors of G. This costs $O(n^3)$.

Algorithm 2.5 in [3] and Algorithm 3 above are quite similar, but Algorithm 2.5 takes advantage of a fixed pivot order to compute the final entries of L column by column, and final entries of U row by row, rather than computing all the $S^{(k)}$. To see how, note that all the following factors of entries of $S^{(k)}$ can be computed and then combined in $O(n^2)$ time:

$$\begin{aligned}
&\prod_{1\le i\le k}(x_r - x_i) && \text{for all } 1 \le k < r \le n,\\
&\prod_{1\le i\le k}(y_s - y_i) && \text{for all } 1 \le k < s \le n,\\
&x_r + y_s && \text{for all } 1 \le r, s \le n,\\
&\prod_{1\le i\le k}(x_i + y_s) && \text{for all } 1 \le k < s \le n,\\
&\prod_{1\le i\le k}(x_r + y_i) && \text{for all } 1 \le k < r \le n.
\end{aligned}$$

Algorithm 2.5 uses only $O(n)$ extra storage, rather than $O(n^2)$.

It is possible to recognize symmetric positive definite quasi-Cauchy matrices quite easily. Suppose $D_1 = D_2$ and $x_i = y_i$ for all i, so that G is symmetric. (This last condition can be weakened slightly, since replacing each x_i by $x_i - c$ and each y_i by $y_i + c$ leaves G unchanged. But this must be done without rounding error to guarantee accuracy. This is possible for the Hilbert matrix, with $c = .5$.) Then G is positive definite if and only if all $x_i > 0$, as can be seen from (1) applied to all leading principal submatrices. Furthermore, G's rows and columns can be symmetrically permuted (so

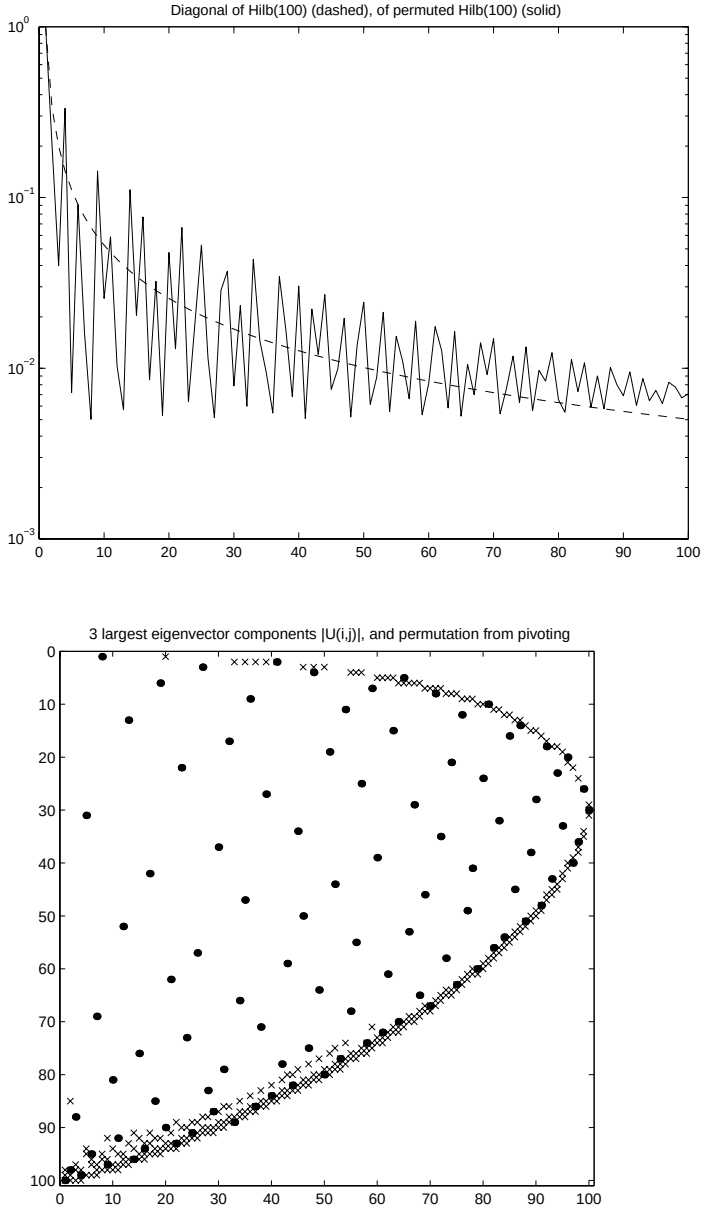


FIG. 3. *Complete pivoting of the* 100-*by*-100 *Hilbert matrix. The top plot shows the diagonal of the Hilbert matrix before (dashed line) and after (solid line) pivoting. The bottom plot superimposes a "spy-plot" of the locations of the three largest entries of each eigenvector (as x's) and the permutation matrix corresponding to complete pivoting (as black dots).*

$x_i < x_{i+1}$) and multiplied by -1 (if any $D_{1,i} < 0$) to make G totally positive. We note that this order is in general not the same as the order determined by pivoting during the algorithm.

A similar approach was also applied to the Hilbert matrix in [29], but using a

formula for the Cholesky factor of the Hilbert matrix without pivoting. When $n = 100$, the corresponding L has a condition number in excess of 10^{20}, so relative accuracy is not guaranteed. Complete pivoting is essential. The top of Figure 3 shows both the diagonal of the unpivoted Hilbert matrix H (dashed line) and the diagonal of the Hilbert matrix PHP^T after applying the permutation P corresponding to complete pivoting; they differ greatly. To better understand the permutation P, in the bottom of Figure 3 we superimpose a "spy-plot" of the nonzero entries of P and the three largest entries in absolute value of each eigenvector; the permutation P frequently corresponds to the location of one of the three largest matrix entries, and otherwise usually corresponds to one of the larger eigenvector entries (which fall to the left of the curve of x's). We do not entirely understand this phenomenon.

5. Vandermonde matrices. Let V be an n-by-n Vandermonde matrix, so that $V_{ij} = x_i^{j-1}$. Let F be the n-by-n discrete Fourier transform matrix, so that $F_{ij} = \omega^{(i-1)(j-1)}/\sqrt{n}$, where $\omega = \exp(2\pi\sqrt{-1}/n)$ is a primitive nth root of unity. Assume for now that no x_i is equal to one of the roots of unity in F; we eliminate this assumption later. Then it is easy to see that VF is quasi-Cauchy by computing

$$\begin{aligned}(VF)_{ij} &= \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} x_i^k \omega^{(j-1)k} \\ &= \frac{1}{\sqrt{n}} \frac{1-(x_i\omega^{j-1})^n}{1-x_i\omega^{j-1}} \\ &= \frac{1-x_i^n}{\sqrt{n}\omega^{j-1}} \frac{1}{\omega^{1-j}-x_i} \quad .\end{aligned}$$

Thus, provided we can compute all of $D_{1,i} = 1 - x_i^n$, $D_{2,j} = \omega^{1-j}/\sqrt{n}$, $\omega^{1-j} - x_i$, and $\omega^{1-j} - \omega^{1-k}$ to high relative accuracy, we can use Algorithm 3 to compute the triangular factorization of VF to high relative accuracy, and then the SVD $VF = U\Sigma\widetilde{V}^T$ to high relative accuracy. We compute the final SVD via $V = U\Sigma(F\widetilde{V})^T$; multiplying $F\widetilde{V}$ increases the angular error in the right singular vectors by only $O(\varepsilon)$.

Now consider that case when some x_i is a root of unity appearing in F. Since x_i is a (real or complex) floating point number, i.e., rational, the only possible roots of unity that x_i can equal are ± 1 and $\pm\sqrt{-1}$. In this case, the ith row of VF will contain all zeros except for one entry equal to $\sqrt{n}$; thus VF is not quasi-Cauchy. Here is how to incorporate such rows into the GECP process. Since all entries but one are zero, the row will not participate in any operations until either 1. $\sqrt{n}$ is itself selected as the pivot or 2. $\sqrt{n}$ is eliminated by another larger pivot in the same column. In case 1, this simply results in eliminating all other entries in its column, without updating any other nonzero entries. The remaining matrix is quasi-Cauchy (with possibly other rows with a single nonzero), so the elimination can proceed as before. In case 2, the zero row entries are filled in, but it is straightforward to show that the resulting overall matrix with the newly filled row is still quasi-Cauchy, so elimination can again proceed as before.

Now we consider how to compute $\omega^{1-j} - \omega^{1-k}$ to high relative accuracy. Given a table of powers of ω, we can simply compute $\omega^{1-j} - \omega^{1-k}$ in the obvious way. Since $|\omega^{1-j} - \omega^{1-k}| \geq 2\sin(\pi/n)$ the difference cannot be too small (unless n is truly enormous), so it is computed with relative accuracy $O(n\epsilon)$. One can do better by using formulas like $\cos(\alpha) - \cos(\beta) = 2\sin(\frac{\beta+\alpha}{2})\sin(\frac{\beta-\alpha}{2})$.

Next we discuss how to compute $x_i^n - 1$ and $\omega^{1-j} - x_i$ to high relative accuracy. We distinguish between the cases of x_i real and x_i complex, and deal with the easier

real case first. First consider $x_i^n - 1$ when x_i is real. When $x_i^n \leq 0$, we just compute $x_i^n - 1$ straightforwardly since there is no cancellation. When $x_i^n > 0$, we can either use the formula $(|x_i| - 1)\sum_{j=0}^{n-1} |x_i|^j$ for an $O(n)$ cost or use a repeated-squaring-like algorithm for $O(\log n)$ cost [25], or use $\mathrm{expm1}(n \cdot \log_e(x_i))$ for $O(1)$ cost, provided that good implementations of $\mathrm{expm1}(z) \equiv e^z - 1$ and $\log_e(z)$ are available [35, 36]. Next consider $\omega^{1-j} - x_i$. Assume that we have a table of values of ω^{1-j} which are accurate to high relative accuracy in the sense of (3) (with a tiny complex δ). Assume further that the real values of ω^{1-j} ($+1$ or -1) are *exact* in the table; the complex values can be computed using any reasonable implementations of sine and cosine. Then $\omega^{1-j} - x_i$ can be computed straightforwardly to high relative accuracy for the following reason: if ω^{1-j} is real, then it is exact, and the difference is computed to high relative accuracy. If ω^{1-j} is complex, its imaginary part is at least $\sin(2\pi/n)$, so $\omega^{1-j} - x_i$ cannot be any smaller in magnitude; therefore the difference is again small in the sense of a tiny complex δ in (3).

When x_i is complex, the situation is more complicated. We assume that we have a table of powers of ω, represented as follows: for each ω^k we store $w_{1,k} = \mathrm{round}(\omega^k)$ (the real and imaginary components being rounded to the *nearest* floating point numbers; a slightly larger error can be tolerated too) and $w_{2,k} = \mathrm{round}(\omega^k - w_{1,k})$ (rounding here may be less careful; a few units of error in the last place is tolerable). Later we will return to the computation of this table.

Using this table, we compute $\omega^k - x_i$ as $d_{k,i} = \mathrm{round}(\mathrm{round}(w_{1,k} - x_i) + w_{2,k})$, i.e., with two floating point operations. We claim that $d_{k,i}$ is accurate to high relative accuracy in both real and imaginary components independently. To see this, there are three cases. (The argument should be applied to the real and imaginary parts separately.)

1. If $x_i = w_{1,k}$ exactly, then $\mathrm{round}(w_{1,k} - x_i) = 0$. Then $d_{k,i} = w_{2,k}$ is correct to as many digits as $w_{2,k}$.
2. If $.5 < w_{1,k}/x_i < 2$ but $x_i \neq w_{1,k}$, then cancellation occurs in $\mathrm{round}(w_{1,k} - x_i)$. Because we assume arithmetic has a guard digit, this nonzero difference is exact, and at least about twice as large in magnitude as $w_{2,k}$ (we use the fact that $w_{1,k} = \mathrm{round}(\omega^k)$ is (nearly) the *nearest* floating point number to ω^k). This means that there is no significant cancellation in $\mathrm{round}(\mathrm{round}(w_{1,k} - x_i) + w_{2,k}) = \mathrm{round}((w_{1,k} - x_i) + w_{2,k})$, so it is computed to high relative accuracy.
3. If $w_{1,k}/x_i \leq .5$ or $2 \leq w_{1,k}/x_i$, then $\mathrm{round}(w_{1,k} - x_i)$ is accurate to high relative accuracy, and since it is at least half as large in magnitude as $w_{1,k}$, adding $w_{2,k}$ does not change it significantly.

To compute $x_i^n - 1$ accurately, we can evaluate it as $\prod_{i=0}^{n-1}(x_i - \omega^i)$, evaluating each difference as described above.

Now we return to the computation of the table of values of $w_{1,k}$ and $w_{2,k}$. The obvious algorithm is to compute $w = \mathrm{round_to_double}(\omega^k)$ to double precision, let $w_{1,k} = \mathrm{round_to_single}(w)$, and $w_{2,k} = \mathrm{round_to_single}(w - w_{1,k})$. The trouble is that $w_{2,k}$ may be much smaller than εw, so that it may have few or no significant digits when computed this way, whereas we rely on it being accurate to single precision. For example, $\cos(2 \cdot 99 \cdot \pi/3565) \approx .9848162531852694$. Rounding this to single yields the real part of $w_{1,99} \approx .9848162531852722$, and subtracting to get the real part of $w_{2,99}$ yields $\approx -2.776 \cdot 10^{-15}$, which only has 5 significant bits. In other words, high precision arithmetic may be needed to compute the table. But since we can empirically confirm that the above example is the worst case up to $n = 10000$, we can

assert that *triple* precision is enough up to $n = 10000$ for input data in IEEE single precision.

Since this table needs only to be computed once and for all for each n, and *not* for each input matrix, its cost should not be considered part of the algorithm. Since its cost is only $O(n)$ as opposed to $O(n^3)$ for the rest of the algorithm anyway, one could in practice recompute the table for each input. A similar issue arises for the FFT itself, where one can precompute and store the ω^k, or compute them on the fly (the latter of which may actually be faster because it uses less memory bandwidth).

The following observation [2] indicates why we expect triple precision to be enough except perhaps in rare cases. By the theory of continued fractions [27, Thm. 19], a rational approximation $p/q \approx \Re\omega^k$ (the real part of ω^k) can satisfy only $|p/q - \omega^k| < 1/(2q^2)$ if p/q is a *convergent* of $\Re\omega^k$, i.e., a truncation of the infinite continued fraction expansion of $\Re\omega^k$ (the same is true for the imaginary part). Therefore, unless the floating point approximation $\Re w_{1,k}$ happens to equal this convergent, which seems unlikely, we will have

$$|\Re w_{2,k}| \approx |\Re w_{1,k} - \Re\omega^k| \geq 1/(2q^2).$$

Since $\Re w_{1,k} = p/q$ is a floating point number, $q = 2^e$, where $e \approx |\log_2 \Re\omega^k| + b$, where b is the number of bits in the floating point fraction, and $p \approx 2^b$, since the floating point fraction is normalized. The nontrivial values of $\Re\omega^k$ satisfy $|\Re\omega^k| \gtrapprox 2\pi/n$, so

$$\frac{|\Re w_{2,k}|}{|\Re\omega^k|} \gtrapprox \frac{1}{(2q^2)} \cdot \frac{q}{p} = \frac{1}{2pq} \gtrapprox 2^{-2b}(\pi/n).$$

This says that computing ω^k to triple precision means that $\Re w_{2,k}$ will be determined with relative error at most $(n/\pi)2^{-b}$, i.e., nearly single precision.

It is an open problem to avoid the need for this high precision table entirely.

6. Unit displacement rank matrices. The most general situation for which the above high accuracy SVD technique works is described as follows. Assume that G satisfies the Sylvester equation (or *displacement equation* [4, 15, 3, 30])

$$XG + GY = d_1 d_2^T,$$

where X and Y are diagonal matrices with diagonal entries x_i and y_j, respectively, and d_1 and d_2 are column vectors. Then it is immediate to see that G is quasi-Cauchy with $G_{ij} = d_{1i}d_{2j}/(x_i + y_j)$.

Now suppose that X and Y are normal matrices with eigendecompositions $X = U_X D_X U_X^T$ and $Y = U_Y D_Y U_Y^T$. Then we can premultiply the above equation by U_X^T and postmultiply it by U_Y to transform it to

$$D_X(U_X^T G U_Y) + (U_X^T G U_Y)D_Y = (U_X^T d_1)(U_Y^T d_2)^T$$

or

$$D_X G' + G' D_Y = d_1' d_2'^T$$

so that G' is quasi-Cauchy. Thus, if this transform can be performed accurately enough, then we can compute the SVD of $G' = U\Sigma V^T$, and backtransform using U_X and U_Y to get the SVD of $G = (U_X U)\Sigma(U_Y V)^T$.

For example, Vandermonde matrices satisfy the displacement equation with X diagonal and Y a circular shift. The eigenvector matrix of Y is the discrete Fourier

transform matrix F, and its eigenvalues are roots of unity. The entries of d_1 are $x_i^n - 1$. This is the transformation used in the previous section.

One can imagine other possibilities besides X and Y diagonal or circular shifts, but it is important that $d_1' = U_X^T d_1$ and $d_2' = U_Y^T d_2$ be computable to high relative accuracy (alternatively, one could say that d_1' and d_2' are the parameters of G that determine its SVD to high relative accuracy).

Now consider the case of X and Y diagonalizable, but not necessarily normal, so that $X = U_X D_X U_X^{-1}$ and $Y = U_Y D_Y U_Y^{-1}$. Then given the SVD $G' = U\Sigma V^T$, $G = (U_X U)\Sigma(U_Y^{-T} V)^T$ is *not* the SVD of G, because U_X and U_Y are not unitary. Instead, we simply convert the RRD $G' = XDY^T$ obtained by Algorithm 3 into the RRD $G = (U_X X)D(U_Y^{-T} Y)^T$ and then apply Algorithm 1 [12].

What about matrices satisfying displacement equations with higher than rank-1 matrices on the right-hand side? For example, Toeplitz matrices satisfy such an equation with rank 2. We argue that no high relative accuracy algorithm like ours can exist for general Toeplitz matrices, and we believe it is unlikely that they exist for higher displacement rank in general.

It is tricky to argue impossibility results in floating point arithmetic, because one can simulate arbitrary precision floating point using just standard floating point, without "bit-fiddling" [33]. Still, we may argue informally as follows. Consider just a 2-by-2 Toeplitz matrix:

$$T = \begin{bmatrix} a & b \\ c & a \end{bmatrix}.$$

If we could compute the singular values of T to high relative accuracy, then by multiplying them we could compute the determinant of T to high relative accuracy (actually just its absolute value, but that is enough for this argument). In other words, we would have a floating point algorithm for evaluating $a^2 - b \cdot c$ to high relative accuracy, without any tricks like splitting a, b, or c into half-precision parts and combining them separately. But since $a^2 - b \cdot c$ does not factor into products and quotients of factors we can compute to high relative accuracy (like $x_i - x_j$ for Cauchy and Vandermonde matrices) – indeed, it does not factor at all – we believe that there is no forward stable SVD algorithm for Toeplitz matrices. Furthermore, it appears that n-by-n Toeplitz determinants have irreducible factors whose degrees grow with n, which means that higher precision would be needed as n grew.

To illustrate the difficulties in formalizing this argument, consider a matrix of the form $G_{ij} = d_{1i} d_{2j}/(x_i w_j + y_j z_i)$. This is quasi-Cauchy, since we can also write $G_{ij} = (d_{1i}/z_i)(d_{2j}/w_j)/((x_i/z_i) + (y_j/w_j))$. The sums and differences that we would need to compute to high relative accuracy in the algorithm of section 4 look like $(x_i/z_i) + (y_j/w_j) = (x_i w_j + y_j z_i)/(w_j z_i)$. These can be evaluated to high relative accuracy using double precision, or in single precision using the techniques in [33].

More generally, if *every* minor of G factors into products and quotients of 1 or 2 term polynomials of degree at most d in the original parameters ($d = 1$ for the quasi-Cauchy matrices in section 4, $d = 1$ for acyclic matrices [8, 7], and $d = 2$ for the matrix in the last paragraph), then a high accuracy SVD algorithm requiring at most d-tuple precision (for the LU factorization) exists. As stated above, d appears to grow with dimension of Toeplitz matrices, so no fixed precision would suffice in the Toeplitz case. Whether an acceptably fast algorithm ($O(n^3)$ or better) exists to compute the LU factorization this way is a different question that depends on another structure of the matrix, like its sparsity pattern or whether it has low displacement rank.

One might object that our required table of high accuracy roots of unity in section 5 violates the spirit of our impossibility argument. The difference is that the table need only be computed once and for all for all complex Vandermonde matrices, that the precision (two significant words) is independent of the Vandermonde data, and that the cost is $O(n)$, in contrast to the $O(n^3)$ overall cost. In contrast, for Toeplitz matrices $O(n^3)$ extra precise operations are required for each matrix, real or complex, and the precision seems to grow with dimension.

Matrices which are simultaneously Cauchy and Toeplitz deserve special mention. The accuracy question depends on which parameters are used to represent these matrices. Either the Cauchy data alone (x_i and y_j), or the Toeplitz data alone (the entries $a_{i,j} \equiv t_{j-i}$ of the first row and column) is redundant. It is not hard to see that two parameters, α and β, form a minimal parameterization of such matrices: $a_{i,j} = 1/(\alpha + \beta(j-i))$. A quasi-Cauchy-Toeplitz matrix is determined by three parameters: $a_{i,j} = \gamma^{j-i}/(\alpha + \beta(j-i))$. Either parameterization determines $x_i - x_j = (j-i)\beta = y_j - y_i$ to as much relative accuracy as β is known. But it may or may not determine $x_i + y_j = \gamma^{j-i}/t_{j-i} = \alpha + (j-i)\beta$ accurately, depending on cancellation. A better parameterization would be the *smallest* $\alpha + k\beta$ in absolute value for any $-n < k < n$, the value k (which is exact) at which the minimum is attained, β, and γ. This way all values of $1/t_{j-i}$ can be determined to high relative accuracy, and small relative changes in $\alpha + k\beta$, β, and γ can cause only small relative changes in the SVD.

7. Estimating the error when the input is uncertain. So far we have considered only exact inputs, and shown that we compute the SVD to high relative accuracy, as defined in the introduction. When the parameters x_i, y_i, D_{1i}, and D_{2i} defining a quasi-Cauchy matrix are themselves uncertain, then we may easily compute a relative error bound on the SVD using Theorem 2.1 from [8].

THEOREM 7.1 (see [8, Theorem 2.1]). *Let $G = XDY^T$ be an RRD with SVD $G = U\Sigma V^T$, and let $\hat{G} = \hat{X}\hat{D}\hat{Y}^T$ with SVD $\hat{G} = \hat{U}\hat{\Sigma}\hat{V}^T$, where $\hat{X}$, $\hat{D}$, and $\hat{Y}$ are defined as follows:*

$$\hat{X} = X + \delta X, \quad \textit{where} \quad \frac{\|\delta X\|}{\|X\|} \le \epsilon,$$

$$\hat{D} = D + \delta D, \quad \textit{where} \quad \delta D \ \textit{is diagonal and} \ \frac{|\delta D_{ii}|}{|D_{ii}|} \le \epsilon,$$

$$\hat{Y} = Y + \delta Y, \quad \textit{where} \quad \frac{\|\delta Y\|}{\|Y\|} \le \epsilon,$$

where $0 \le \epsilon < 1$. Let $\eta = \epsilon(2+\epsilon)\max(\kappa(X), \kappa(Y))$ and $\eta' = 2\eta + \eta^2$, where $\kappa(Z) = \sigma_{\max}(Z)/\sigma_{\min}(Z)$ is the condition number of Z. Then the difference between the singular values of G and $\hat{G}$ is bounded as follows:

$$\frac{|\sigma_i - \hat{\sigma}_i|}{\sigma_i} \le \eta'. \tag{4}$$

Furthermore, the angle θ between u_i and $\hat{u}_i$ (or between v_i and $\hat{v}_i$) is bounded by

$$\sin\theta \le \sqrt{2}\left(\frac{1+\eta'}{1-\eta'} \cdot \frac{\eta'}{\min(\text{relgap}_i, 2) - \eta'} + \eta\right) \tag{5}$$

provided that $\min(\text{relgap}_i, 2) \ge \eta'$.

It is easy to see that relative errors of size ϵ in the D_{1i} and D_{2i} can increase the relative errors in the singular values by at most ϵ, and the errors in the singular vectors by at most about ϵ divided by the relative gap. Given uncertainty bounds $x_i \pm \delta x_i$ and $y_i \pm \delta y_i$, we can also easily bound the maximum relative error in any factor like $x_i + y_j$ (namely $\varepsilon + (|\delta x_i| + |\delta y_j|)/|x_i + y_j|$) appearing in the algorithm in section 4. The maximum such factor-wise relative error can be used as an estimate of ϵ in Theorem 2.1 above. (This ignores factors depending on the dimension n, which are often pessimistic anyway.)

From this point of view the Hilbert matrix is very well conditioned, since small relative changes in the values of $x_i = i$ and $y_i = i-1$ cause only small relative changes in the SVD.

More generally, consider a totally positive Cauchy matrix, where $0 < x_1 < \cdots < x_m$ and $0 < y_1 < \cdots < y_n$. In this case small relative changes in the x_i and y_j cause only small relative changes in the $x_i + y_j$, so the relative accuracy depends only on the *relative gaps* $(x_{i+1} - x_i)/(x_{i+1} + x_i) < 1$ and $(y_{j+1} - y_j)/(y_{j+1} + y_j) < 1$; if these are not small, small relative changes in the x_i and y_j cause only small relative changes in the SVD. (The most general characterization of totally positive Cauchy matrices reduces the requirements $0 < x_1$ and $0 < y_1$ to just $0 < x_1 + y_1$ [14], but this allows cancellation, which is why we used the stronger conditions $0 < x_1$ and $0 < y_1$.)

Acknowledgments. The author thanks Ming Gu and Vadim Olshevsky for discussion of and pointers to relevant displacement rank literature, Doug Priest, W. Kahan, and Peter Tang for discussions of floating point issues, Zlatko Drmač for simplifying and improving some of the algorithms, David Blackston for pointing out the connection with continued fractions, and the three referees for many useful comments.

REFERENCES

[1] C. Bischof and G. Quintana-Orti, *Computing rank-revealing QR factorizations of dense matrices*, ACM Trans. Math. Software, 24 (1998), pp. 226–253.
[2] D. Blackston, *private communication*, 1997.
[3] T. Boros, T. Kailath, and V. Olshevsky, *Pivoting and backward stability of fast algorithms for solving Cauchy linear equations*, SIAM J. Matrix Anal. Appl., submitted.
[4] T. Boros, T. Kailath, and V. Olshevsky, *The Fast Björck-Pereyra-type Algorithm for Parallel Solution of Cauchy Linear Equations*, Linear Algebra Appl., to appear.
[5] T. Chan, *Rank revealing QR factorizations*, Linear Algebra Appl., 88/89 (1987), pp. 67–82.
[6] S. Chandrasekaran and I. Ipsen, *On rank-revealing QR factorizations*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 592–622.
[7] J. Demmel and W. Gragg, *On computing accurate singular values and eigenvalues of acyclic matrices*, Linear Algebra Appl., 185 (1993), pp. 203–218.
[8] J. Demmel, M. Gu, S. Eisenstat, I. Slapničar, K. Veselić, and Z. Drmač, *Computing the singular value decomposition with high relative accuracy*, Linear Algebra Appl., (1997), to appear.
[9] J. Demmel and K. Veselić, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1204–1246.
[10] Z. Drmač, *Accurate computation of the product-induced singular value decomposition with applications*, SIAM J. Numer. Anal., 35 (1998), pp. 1969–1994.
[11] Z. Drmač, *On the condition behaviour in the Jacobi method*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 509–514.
[12] Z. Drmač, *Fast and accurate algorithms for canonical correlations, weighted least squares and related generalized eigenvalue and singular value decompositions*, SIAM J. Matrix Anal. Appl., to appear.
[13] Z. Drmač, *A posteriori computation of the singular vectors in a preconditioned Jacobi SVD algorithm*, IMA J. Numer. Anal., 19 (1998), pp. 191–213.
[14] F. Gantmacher and M. Krein, *Oszillationsmatrizen, Oszillationskerne, und kleine Schwingungen mechanischer Systeme*, Akademie-Verlag, Berlin, 1960.

[15] I. Gohberg, T. Kailath, and V. Olshevsky, *Fast Gaussian elimination with partial pivoting for matrices with displacement structure*, Math. Comp., 64 (1995), pp. 1557–1576.
[16] I. Gohberg and V. Olshevsky, *Fast algorithms with preprocessing for matrix-vector multiplication problems*, J. Complexity, 10 (1994), pp. 411–427.
[17] I. Gohberg and V. Olshevsky, *Fast state space methods for matrix Nehari and Nehari–Takagi interpolation problems*, Integral Equations Operator Theory, 20 (1994), pp. 44–83.
[18] G. Golub and C. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[19] M. Gu, *Some New Rank-Revealing Factorizations*, Tech. report CAM Report 98-33, Department of Mathematics, University of California at Los Angeles, Los Angeles, 1998.
[20] M. Gu and S. Eisenstat, *Efficient algorithms for computing a strong rank–revealing QR factorization*, SIAM J. Sci. Comput., 17 (1996), pp. 848–869.
[21] N. J. Higham, *Stability analysis of algorithms for solving confluent Vandermonde-like systems*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 23–41.
[22] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.
[23] N. J. Higham, *Optimization by direct search in matrix computations*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 317–333.
[24] P. Hong and C. T. Pan, *The rank revealing QR and SVD*, Math. Comp., 58 (1992), pp. 575–232.
[25] W. Kahan, *private communication*, 1997.
[26] S. Karlin, *Total Positivity*, Stanford University Press, Stanford, CA, 1968.
[27] A. Khinchin, *Continued Fractions*, The University of Chicago Press, Chicago, IL, 1964.
[28] W. Mascarenhas, *A note on Jacobi being more accurate than QR*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 215–218.
[29] R. Mathias, *Accurate eigensystem computations by Jacobi methods*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 977–1003.
[30] V. Olshevsky, *Pivoting on structured matrices and applications*, Linear Algebra Appl., to appear.
[31] C. T. Pan, *On the existence and computation of rank-revealing LU factorizations*, Linear Algebra Appl., (1996), submitted.
[32] C. T. Pan and P. Tang, *Bounds on singular values revealed by QR factorization*, BIT, 39 (1999), to appear.
[33] D. Priest, *Algorithms for arbitrary precision floating point arithmetic*, in Proceedings of the 10th Symposium on Computer Arithmetic, P. Kornerup and D. Matula, eds., Grenoble, France, 1991, IEEE Computer Society Press, Los Alamitos, CA, 1991, pp. 132–145.
[34] G. W. Stewart, *Updating a rank-revealing ULV decomposition*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 494–499.
[35] P. T. P. Tang, *Table-driven implementation of the logarithm function for IEEE floating-point arithmetic*, ACM Trans. Math. Software, 16 (1990), pp. 378–400.
[36] P. T. P. Tang, *Table-driven implementation of the Expm1 function for IEEE floating-point arithmetic*, ACM Trans. Math. Software, 18 (1992), pp. 211–222.
[37] S. Wolfram, *Mathematica: A System for Doing Mathematics by Computer*, Addison–Wesley, Reading, MA, 1988.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 581–592
© 1999 Society for Industrial and Applied Mathematics

INTERLACING INEQUALITIES AND CARTAN SUBSPACES OF CLASSICAL REAL SIMPLE LIE ALGEBRAS*

TIN-YAU TAM†

Abstract. The well-known interlacing inequalities are examined in the framework of classical real simple Lie algebras.

Key words. interlacing inequalities, Cartan subspaces, real simple Lie algebras

AMS subject classifications. 15A42, 17B20

PII. S0895479898343528

1. Introduction.

Let A be an $n \times n$ Hermitian matrix with eigenvalues

$$\alpha_1 \geq \cdots \geq \alpha_n \tag{1.1}$$

and let B be a $k \times k$ principal submatrix of A, having eigenvalues

$$\beta_1 \geq \cdots \geq \beta_k. \tag{1.2}$$

The well-known Cauchy interlacing inequalities [7] are

$$\alpha_i \geq \beta_i \geq \alpha_{i+n-k}, \qquad i = 1, \ldots, k. \tag{1.3}$$

When $k = n - 1$, (1.3) becomes

$$\alpha_1 \geq \beta_1 \geq \alpha_2 \geq \beta_2 \geq \cdots \geq \alpha_{n-1} \geq \beta_{n-1} \geq \alpha_n, \tag{1.4}$$

which are genuinely interlacing inequalites. The interlacing inequalites (1.3) are closely related to the Poincaré inequality and the Courant–Fischer–Weyl minimax principle. Indeed they can be derived from each other [1], [2]. The following converse is true. Given two sets of real numbers (1.1) and (1.2) satisfying the interlacing inequalites (1.3), there exists an $n \times n$ real symmetric (thus Hermitian) matrix A with a $k \times k$ principal submatrix B, having eigenvalues (1.1) and (1.2), respectively. Fan and Pall [4] were apparently the first to publish a proof of the converse, and according to them, in 1953 H. Wielandt had an unpublished proof of the same result based on a different idea; see also [8] and [9]. The essential case is (1.4), as observed by Fan and Pall [4, p. 299], in the sense that (1.3) and its converse will then follow. For if (1.3) holds, sequences of eigenvalues can be inserted such that two consecutive sequences are genuinely interlaced as in (1.4). Then, by the result for $k = n-1$, there exists a chain of Hermitian matrices having prescribed eigenvalues such that each is a principal submatrix of the next.

Thompson [12] obtained similar interlacing inequalites for an arbitrary $p \times q$ complex matrix A, but now the role of the eigenvalues is played by the singular values. Let B be a $k_1 \times k_2$ submatrix of A. Let

$$\alpha_1 \geq \cdots \geq \alpha_{\min\{p,q\}} \tag{1.5}$$

*Received by the editors August 16, 1998; accepted for publication (in revised form) by R. Bhatia March 16, 1999; published electronically December 15, 1999.
http://www.siam.org/journals/simax/21-2/34352.html
†Department of Mathematics, Auburn University, Auburn, AL 36849-5310 (tamtiny@auburn.edu).

be the singular values of A and let

$$\beta_1 \geq \cdots \geq \beta_{\min\{k_1,k_2\}} \tag{1.6}$$

be the singular values of B. Then [12, p. 4]

$$\alpha_i \geq \beta_i, \qquad i = 1, \ldots, \min\{k_1, k_2\} \tag{1.7}$$

$$\beta_i \geq \alpha_{i+(p-k_1)+(q-k_2)}, \qquad i \leq \min\{k_1 + k_2 - p, k_1 + k_2 - q\}. \tag{1.8}$$

When $p = q$ and $k_1 = k_2 = p - 1$, i.e., we have the singular values of a $p \times p$ matrix and its $(p-1) \times (p-1)$ submatrices, (1.7) and (1.8) become

$$\begin{gathered} \alpha_1 \geq \beta_1 \geq \alpha_3, \\ \alpha_2 \geq \beta_2 \geq \alpha_4, \\ \vdots \\ \alpha_{p-2} \geq \beta_{p-2} \geq \alpha_p, \\ \alpha_{p-1} \geq \beta_{p-1}. \end{gathered} \tag{1.9}$$

The index drops by two in (1.9) instead of one. However, if we just delete one row or one column at a time, then we do have the genuine interlacing inequalities

$$\alpha_1 \geq \beta_1 \geq \alpha_2 \geq \beta_2 \geq \cdots. \tag{1.10}$$

Similar to the Hermitian case, given (1.5) and (1.6) satisfying (1.7) and (1.8), there exists a $p \times q$ real (indeed, the proof of Thompson [12, Theorem 2] also works for the real case) matrix A with a $k_1 \times k_2$ submatrix B having singular values (1.5) and (1.6), respectively.

There are orbital descriptions of the above two results.

THEOREM 1.1 (see Cauchy–Fan–Pall [3], [4]). *Let A be an $n \times n$ Hermitian (respectively, real symmetric) matrix with eigenvalues* (1.1). *The eigenvalues of $(UAU^*)[1, \ldots, k|1, \ldots, k]$ fill out the values permitted by* (1.3) *as U runs through the unitary group $U(n)$ (respectively, the orthogonal group $O(n)$).*

THEOREM 1.2 (Thompson [12]). *If A is a $p \times q$ complex (respectively, real) matrix, then the singular values of $(UAV)[1, \ldots, k_1|1, \ldots, k_2]$ fill out the values permitted by* (1.7) *and* (1.8) *as U and V run through the unitary groups $U(p)$ and $U(q)$ (respectively, $O(p)$ and $O(q)$) independently.*

This orbital point of view turns out to be very useful when we study a similar problem in a more general setting.

Thompson [13] suggested that the index on the lower side of each inequality (1.9) drops by two instead of one, as in (1.4), due to the fact that the number of degrees of freedom in the singular value case is doubled, since one considers UAV when A is a complex matrix, in contrast to UAU^* when A is Hermitian. But the explanation is not satisfactory since Thompson [13] also discovered the following result for a complex symmetric matrix somewhat unexpectedly. Thus, counting the degrees of freedom is not always a reliable guide [14, p. 119] in the study of interlacing inequalities.

THEOREM 1.3 (see Thompson [13]). *Let A be an $n \times n$ complex symmetric matrix having singular values* (1.1). *The singular values of $(UAU^T)[1, \ldots, n-1|1, \ldots, n-1]$, $U \in U(n)$, fill out all values permitted by* (1.9) *where $p = n$.*

We will see that the number of deleted rows and columns is a contributing factor to the above phenomenon when we encounter $\mathfrak{sp}_{2n}(\mathbb{R})$.

Let $\mathfrak{g} = \mathfrak{k} + \mathfrak{p}$ be a Cartan decomposition of a real semisimple Lie algebra $\mathfrak{g}$ whose analytic group is G. Let K be the analytic subgroup of G (it is unique once we fix G), whose Lie algebra is $\mathfrak{k}$. Here $\mathfrak{p}$, which will be called the Cartan subspace of $\mathfrak{g}$, is the orthogonal complement of $\mathfrak{k}$ in $\mathfrak{g}$ with respect to the Killing form $B(\cdot, \cdot)$. The group K acts on $\mathfrak{p}$ via adjoint representation Ad, i.e., $x \mapsto \mathrm{Ad}(k)x$, if $x \in \mathfrak{p}$. Denote by $O(x)$ the orbit of x under the action of K, i.e., $O(x) = \{\mathrm{Ad}(k)x : k \in K\}$. It is known that

$$O(x) \cap \mathfrak{a}_+ \neq \phi, \qquad x \in \mathfrak{p}, \tag{1.11}$$

where $\mathfrak{a}_+$ is a fundamental Weyl chamber of the maximal abelian subalgebra $\mathfrak{a}$ in $\mathfrak{p}$. See [5]. Indeed the set $O(x) \cap \mathfrak{a}_+$ is merely a singleton set whose element will be denoted by $\mathfrak{a}_+(x)$. Since we will study only the classical real simple Lie algebras, the adjoint action is just conjugation, i.e., $\mathrm{Ad}(k)x = kxk^{-1}$. The idea of recovering interlacing inequalities is that we pick any element $x \in \mathfrak{p}$ and consider an appropriate compression P from $\mathfrak{p}$ to a lower dimensional space $\mathfrak{p}'$, which is a Cartan subspace of a simple $\mathfrak{g}'$, where $\mathfrak{g}'$ and $\mathfrak{g}$ are of the same classical series. Then we consider the compression of the orbit $O(x)$, $O'(x) := PO(x) \subset \mathfrak{p}'$. Now we want to describe $\mathfrak{a}'_+(O'(x))$. It turns out that the appropriate interlacing inequalities are the complete description.

The classical real simple Lie algebras [10, p. 233] are isomorphic to one of the following real forms $\mathfrak{h} \subset \mathfrak{g}$ and $\mathfrak{g}^{\mathbb{R}}$ (the realification of $\mathfrak{g}$).

1. $\mathfrak{g} = \mathfrak{sl}_n(\mathbb{C})$, $n \geq 2$
   (a) $\mathfrak{h} = \mathfrak{sl}_n(\mathbb{R})$,
   (b) $\mathfrak{h} = \mathfrak{sl}_m(\mathbb{H})$, $n = 2m$,
   (c) $\mathfrak{h} = \mathfrak{su}_{p,q}$ $(p = 0, 1, \ldots, [\frac{n}{2}]$, $p + q = n)$.
2. $\mathfrak{g} = \mathfrak{so}_{2n+1}(\mathbb{C})$, $n \geq 2$
   (a) $\mathfrak{h} = \mathfrak{so}_{p,q}$ $(p = 0, 1, \ldots, n$, $p + q = 2n + 1)$.
3. $\mathfrak{g} = \mathfrak{sp}_{2n}(\mathbb{C})$, $n = 2m$, $m \geq 3$
   (a) $\mathfrak{h} = \mathfrak{sp}_{2n}(\mathbb{R})$, $n = 2m$,
   (b) $\mathfrak{h} = \mathfrak{sp}_{p,q}$ $(p = 0, 1, \ldots, [m/2]$, $p + q = m)$.
4. $\mathfrak{g} = \mathfrak{so}_{2n}(\mathbb{C})$, $n \geq 4$
   (a) $\mathfrak{h} = \mathfrak{so}_{p,q}$ $(p = 0, 1, \ldots, n$, $p + q = 2n)$,
   (b) $\mathfrak{h} = \mathfrak{so}^*(2n)$.

2. The real forms of $\mathfrak{sl}_n(\mathbb{C})$. We will see that the real forms of $\mathfrak{sl}_n(\mathbb{C})$ are related to the interlacing results for real symmetric, Hermitian, quaternionic Hermitian, and complex matrices. The group $SL(n, \mathbb{F}) \subset GL(n, \mathbb{F})$ is the special linear group consisting of $n \times n$ matrices with determinant 1. The Lie algebra is $\mathfrak{sl}_n(\mathbb{F})$, i.e., the traceless $n \times n$ matrices with elements in $\mathbb{F}$. The Cartan decomposition of $\mathfrak{sl}_n(\mathbb{F})$ is $\mathfrak{sl}_n(\mathbb{F}) = \mathfrak{k} + \mathfrak{p}$, where $\mathfrak{p}$ is the space of traceless real symmetric, Hermitian, and quaternionic Hermitian matrices (i.e., $A = A^*$, where $A^* = \overline{A}^T$ and $\overline{a_1 + ia_2 + ja_3 + ka_4} = a_1 - ia_2 - ja_3 - ka_4$) when $\mathbb{F} = \mathbb{R}$, $\mathbb{C}$, and $\mathbb{H}$, respectively. The group K is $SU_n(\mathbb{F})$ and $\mathfrak{k}$ is the algebra of real skew symmetric, traceless skew Hermitian, and traceless quaternionic skew Hermitian matrices accordingly; $\mathfrak{a} \subset \mathfrak{p}$ is the subset of real diagonal matrices; $\mathfrak{a}_+ \subset \mathfrak{a}$ can be chosen as the subset of real diagonal matrices with decreasing diagonal entries. Then (1.11) is simply the spectral decomposition, though the quaternionic spectral decomposition is less well known. Now (1.11) means that a quaternionic Hermitian matrix is quaternionic unitarily similar to a real diagonal matrix whose diagonal elements are called the (right) eigenvalues of A.

When $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$, we have Cauchy–Fan–Pall's result. Of course one can consider general Hermitian or real symmetric matrices which correspond to the reductive algebras $\mathfrak{gl}_n(\mathbb{F}) = \mathfrak{sl}_n(\mathbb{F}) + \mathbb{R}I$, $\mathbb{F} = \mathbb{C}$, $\mathbb{R}$, respectively. The groups $SU_n(\mathbb{C})$ and $SU_n(\mathbb{R})$ act trivially on the center $\mathbb{R}I$. So the study of traceless cases is sufficient. Indeed one can replace $SU_n(\mathbb{F})$ by $U_n(\mathbb{F})$, as they will give the same orbit $O(A)$. The following result is for the quaternionic case, due to Thompson [14], and we will give another proof.

THEOREM 2.1 (see [14, pp. 132–133]). *Let A be an $n \times n$ quaternionic Hermitian matrix with eigenvalues* (1.1); *then the eigenvalues of $(UAU^*)[1, \dots, k|1, \dots, k]$ fill out the values permitted by* (1.3) *as U runs through the quaternionic unitary group $U_n(\mathbb{H})$.*

Proof. It suffices to consider $A \in \mathfrak{sl}_n(\mathbb{H})$. We identify $\mathfrak{sl}_n(\mathbb{H})$ with $\mathfrak{su}^*(2n)$ via the standard isomorphism $\mathbb{H} = \mathbb{C} + \mathbb{C}j = \mathbb{C}^2$, such that $a + bj \mapsto (a, b)$, $a, b \in \mathbb{C}$ ([6], pp. 26–27), i.e., $A + Bj \in \mathfrak{sl}_n(\mathbb{H})$ is identified with

$$\varphi(A + Bj) := \begin{pmatrix} A & -\overline{B} \\ B & \overline{A} \end{pmatrix} \in \mathfrak{su}^*(2n).$$

Notice that $\varphi(XY) = \varphi(X)\varphi(Y)$ and $\varphi(X^*) = \varphi(X)$ for any $X, Y \in \mathbb{H}_{n\times n}$. With this isomorphism, $SU_n(\mathbb{H})$ is identified with $Sp(n) \cap SU(2n)$. Since $U(n)$ acts on Hermitian matrices in the same way as $SU(n)$, we can simply use $Sp(n)$ instead of $Sp(n) \cap SU(2n)$. Indeed,

$$\begin{aligned}
\mathfrak{su}^*(2n) &= \left\{ \begin{pmatrix} X & Y \\ -\overline{Y} & \overline{X} \end{pmatrix} : \operatorname{Re} \operatorname{tr} X = 0,\ X, Y \in \mathbb{C}_{n\times n} \right\}, \\
K = Sp(n) &= \left\{ \begin{pmatrix} U & V \\ -\overline{V} & \overline{U} \end{pmatrix}, U, V \in \mathbb{C}_{n\times n} \right\} \cap U(2n), \\
\mathfrak{k} = \mathfrak{sp}(n) &= \left\{ \begin{pmatrix} X & Y \\ -\overline{Y} & \overline{X} \end{pmatrix} : X^* = -X,\ Y^T = Y \right\}, \\
\mathfrak{p} &= \left\{ \begin{pmatrix} X & Y \\ -\overline{Y} & \overline{X} \end{pmatrix} : X^* = X,\ \operatorname{tr} X = 0,\ Y^T = -Y \right\}, \\
\mathfrak{a} &= \left\{ \operatorname{diag}(a_1, \dots, a_n, a_1, \dots, a_n), a_1, \dots, a_n \in \mathbb{R},\ \sum_{i=1}^{n} a_i = 0 \right\}.
\end{aligned}$$

The study of $(U(A + Bj)U^*)[1, \dots, k|1, \dots, k]$ is reduced to that of

$$(\varphi(U)\varphi(A + Bj)\varphi(U)^*)[1, \dots, k, n+1, \dots, n+k|1, \dots, k, n+1, \dots, n+k],$$

a Hermitian case. We notice that the eigenvalues of $\varphi(A + Bj)$ appear in pairs, i.e., $\alpha_1, \alpha_1, \dots, \alpha_n, \alpha_n$, but the number of deleted rows and columns is $2(n - k)$. So we have (1.3) by Cauchy's interlacing inequality. For the converse, we just consider the orbit $\{(U \oplus U)\Lambda(U^T \oplus U^T) : U \in O(n)\}$, where $\Lambda = \operatorname{diag}(\alpha_1, \dots, \alpha_n, \alpha_1, \dots, \alpha_n)$ and apply Fan–Pall's result. ∎

Summarizing the above results we arrive at the following theorem.

THEOREM 2.2 (Cauchy–Fan–Pall–Thompson). *Let A be an $n \times n$ Hermitian matrix over the field $\mathbb{F}$, where $\mathbb{F} = \mathbb{R}$, $\mathbb{C}$, or $\mathbb{F}$, with $\mathfrak{a}_+(A) = (\alpha_1, \dots, \alpha_n)$. Then*

$$\{\mathfrak{a}_+((UAU^*)[1, \dots k|1, \dots, k]) : U \in SU_n(\mathbb{F})\} = \prod_{i=1}^{k} [\alpha_{i+n-k}, \alpha_i].$$

We remark that the set in the above theorem is a "rectangular solid" in $\mathbb{R}^k$.

Let us consider the last real form of the $\mathfrak{sl}_n(\mathbb{C})$, namely, $\mathfrak{su}_{p,q}$, which corresponds to the case of $p \times q$ complex matrices. It is known that [10]

$$\begin{aligned}
\mathfrak{su}_{p,q} &= \left\{ \begin{pmatrix} X_1 & Y \\ Y^* & X_2 \end{pmatrix} : X_1^* = -X_1,\ X_2^* = -X_2,\ \operatorname{tr} X_1 = \operatorname{tr} X_2 = 0,\ Y \in \mathbb{C}_{p\times q} \right\}, \\
K = SU(p,q) &= \left\{ \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix} : U \in U(p), V \in U(q),\ \det U \det V = 1 \right\}, \\
\mathfrak{k} &= \mathfrak{su}_{p,q},\ \text{i.e.,}\ Y = 0, \\
\mathfrak{p} &= \left\{ \begin{pmatrix} 0 & Y \\ Y^* & 0 \end{pmatrix}, Y \in \mathbb{C}_{p\times q} \right\},\ \text{i.e.,}\ X_1 = X_2 = 0, \\
\mathfrak{a} &= \oplus_{1\le j\le p} \mathbb{R}(E_{j,p+j} + E_{p+j,j}), \\
\mathfrak{a}_+ &= \{\oplus_{1\le j\le p} a_j(E_{j,p+j} + E_{p+j,j}) : a_1 \ge \cdots \ge a_p \ge 0\}.
\end{aligned}$$

Now the orbit of an element in $\mathfrak{p}$ under the adjoint action of K is

$$\begin{aligned}
&\left\{ \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix} \begin{pmatrix} 0 & A \\ A^* & 0 \end{pmatrix} \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix}^* : U \in U(p), V \in U(q),\ \det U \det V = 1 \right\} \\
&\quad = \left\{ \begin{pmatrix} 0 & UAV^* \\ VA^*U^* & 0 \end{pmatrix} : U \in U(p), V \in U(q),\ \det U \det V = 1 \right\} \\
&\quad = \left\{ \begin{pmatrix} 0 & UAV \\ (UAV)^* & 0 \end{pmatrix} : U \in U(p), V \in U(q) \right\}.
\end{aligned}$$

The last equality comes from

$$\begin{aligned}
&\{UAV^* : U \in U(p), V \in U(q), \det U \det V = 1\} \\
&= \{UAV^* : U \in U(p), V \in U(q), \det U = \det V^*\} \\
&= \{UAV : U \in U(p), V \in U(q)\}.
\end{aligned}$$

So (1.11) is the singular value decomposition of a $p \times q$ complex matrix. Notice that

$$\begin{aligned}
&\begin{pmatrix} 0 & A[1,\dots,k_1|1,\dots,k_2] \\ A[1,\dots,k_1|1,\dots,k_2]^* & 0 \end{pmatrix} \\
&\quad = \begin{pmatrix} 0 & A \\ A^* & 0 \end{pmatrix} [1,\dots,k_1,p+1,\dots,p+k_2|1,\dots,k_1,p+1,\dots p+k_2].
\end{aligned}$$

The eigenvalues of the matrix

$$\begin{pmatrix} 0 & A \\ A^* & 0 \end{pmatrix} \tag{2.1}$$

are $\alpha_1 \ge \alpha_2 \ge \cdots \ge \alpha_{\min\{p,q\}} \ge 0 \ge \cdots \ge 0 \ge -\alpha_{\min\{p,q\}} \ge \cdots \ge -\alpha_2 \ge -\alpha_1$, where α's are the singular values of A and there are $p+q-2\min\{p,q\}$ zeros. So the missing lower bound in the last inequality of (1.9) is indeed $-\alpha_p$, which is too superficial to write down. The matrix (2.1) is also used in [12] as a means to deduce the interlacing inequalities. Theorem 1.2 is the following corresponding interlacing result:

$$\begin{aligned}
&\{\mathfrak{a}_+((UAV)[1,\dots k_1|1,\dots,k_2]) : U \in U(p), V \in U(q)\} \\
&\quad = \prod_{i \le \min\{k_1+k_2-p,\, k_1+k_2-q\}} [\alpha_{i+(p-k_1)+(q-k_2)}, \alpha_i] \\
&\qquad \times \prod_{\min\{k_1+k_2-p,\, k_1+k_2-q\} < i \le \min\{k_1,k_2\}} [0, \alpha_i].
\end{aligned}$$

It is interesting to notice that the proof of Theorem 1.2 in [12] also relies on Theorem 1.1. However, [11] provides another proof without using Theorem 1.

3. The real forms of $\mathfrak{sp}_{2n}(\mathbb{C})$. The Cartan decomposition of

$$\mathfrak{sp}_{2n}(\mathbb{C}) = \left\{ \begin{pmatrix} X & Y_1 \\ Y_2 & -X^T \end{pmatrix} : \ X, Y_1, Y_2 \in \mathbb{C}_{n\times n},\ Y_1^T = Y_1, Y_2^T = Y_2 \right\}$$

is $\mathfrak{sp}_{2n}(\mathbb{C})^{\mathbb{R}} = \mathfrak{sp}(n) + i\mathfrak{sp}(n)$, i.e., $\mathfrak{k} = \mathfrak{sp}(n)$ and $\mathfrak{p} = i\mathfrak{sp}(n)$, where

$$\mathfrak{sp}(n) = \left\{ \begin{pmatrix} A_1 & -\overline{A}_2 \\ A_2 & \overline{A}_1 \end{pmatrix} : A_1, A_2 \in \mathbb{C}_{n\times n},\ A_1^* = -A_1,\ A_2^T = A_2 \right\},$$

and K is the symplectic group $Sp(n)$. Now

$$\begin{aligned} \mathfrak{a} &= \{\mathrm{diag}\,(a_1, \ldots, a_n, -a_1, \ldots, -a_n) : a_1, \ldots, a_n \in \mathbb{R}\} \\ \mathfrak{a}_+ &= \{\mathrm{diag}\,(a_1, \ldots, a_n, -a_1, \ldots, -a_n) : a_1 \geq \cdots \geq a_n \geq 0\}, \end{aligned}$$

which will be identified with $\mathbb{R}^n$ and $\{(a_1, \ldots, a_n) : a_1 \geq \cdots \geq a_n \geq 0\}$, respectively. Thus (1.11) means that if $C \in i\mathfrak{sp}(n)$, then there exists $U \in Sp(n)$ such that

$$U^*CU = \mathrm{diag}\,(c_1, \ldots, c_n, -c_1, \ldots, -c_n),$$

where $c_i \geq 0$ are the singular values of C, $i = 1, \ldots, n$. In the following we identify $i\mathfrak{sp}(n)$ and $\mathfrak{sp}(n)$ in the obvious way.

THEOREM 3.1. *Let $A \in \mathfrak{sp}(n)$ with $\mathfrak{a}_+(A) = (\alpha_1, \cdots, \alpha_n)$. Then*

$$\begin{aligned} &\{\mathfrak{a}_+((UAU^*)[1, \ldots, k, n+1, \ldots, n+k | 1, \ldots, k, n+1, \ldots, n+k]) : U \in Sp(n)\} \\ &\quad = \prod_{i \leq 2k-n} [\alpha_{i+(n-k)+(n-k)}, \alpha_i] \times \prod_{2k-n < i \leq k} [0, \alpha_i]. \end{aligned}$$

Proof. By Theorem 1.1 we have the inequalities since the matrix A is Hermitian. By considering the set

$$\{(U \oplus U)\mathrm{diag}\,(\alpha_1, \ldots, \alpha_n, -\alpha_1, \ldots, -\alpha_n)(U^T \oplus U^T) : U \in O(n)\}$$

and using Theorem 1.1 again, we have the desired result. □

We now consider $\mathfrak{sp}_{2n}(\mathbb{R})$, which will yield the interlacing inequalities for complex symmetric matrices. It is known that

$$\begin{aligned} \mathfrak{sp}_{2n}(\mathbb{R}) &= \left\{ \begin{pmatrix} X & Y_1 \\ Y_2 & -X^T \end{pmatrix} : Y_1^T = Y_1,\ Y_2^T = Y_2,\ X, Y_1, Y_2 \in \mathbb{R}_{n\times n} \right\}, \\ K &= \left\{ \begin{pmatrix} A & B \\ -B & A \end{pmatrix} : A^TA + B^TB = I,\ A^TB = B^TA,\ A, B \in \mathbb{R}_{n\times n} \right\}, \\ \mathfrak{k} &= \left\{ \begin{pmatrix} X & Y \\ -Y & X \end{pmatrix} : Y^T = Y,\ X^T = -X,\ X, Y \in \mathbb{R}_{n\times n} \right\}, \\ \mathfrak{p} &= \left\{ \{ \begin{pmatrix} X & Y \\ Y & -X \end{pmatrix} : Y^T = Y,\ X^T = X,\ X, Y \in \mathbb{R}_{n\times n} \right\}, \\ \mathfrak{a} &= \{\mathrm{diag}\,(a_1, \ldots, a_n, -a_1, \ldots, -a_n) : a_1, \ldots, a_n \geq 0\}. \end{aligned}$$

Notice that

$$\begin{pmatrix} U & V \\ -V & U \end{pmatrix} \in K$$

if and only if $U+iV \in U(n)$. Hence we identify K with $U(n)$ via the map

$$\begin{pmatrix} U & V \\ -V & U \end{pmatrix} \mapsto U+iV.$$

Now we identify $\mathfrak{p}$ with the space of $n \times n$ complex symmetric matrices via the map

$$\begin{pmatrix} X & Y \\ Y & -X \end{pmatrix} \mapsto Y+iX, \qquad X,Y \in \mathbb{R}_{n\times n},\ X^T=X,\ Y^T=Y.$$

Notice that with this identification, the group K acting on $\mathfrak{p}$ contains just those transformations of the form $x \mapsto cxc^T$. Hence the maximal abelian subalgebra $\mathfrak{a}$ can then be identified with the space of real diagonal matrices. We then pick $\mathfrak{a}_+$ to be the set of diagonal matrices $\operatorname{diag}(a_1,\dots,a_n)$, where $a_1 \geq \cdots \geq a_n$. Thus (1.11) means Takagi decomposition: a complex symmetric matrix is unitarily congruent to a diagonal matrix whose diagonal entries are the singular values of the matrix. The following is apparently an extension of Theorem 1.3. But it can also be proved by using the same technique in the proof [13] for Theorem 1.3, or it follows from Theorem 1.3 by the remark of Fan and Pall, analogously. We now give a proof according to our setting.

THEOREM 3.2 (see Thompson [13]). *Let A be an $n \times n$ complex symmetric matrix having singular values* (1.1). *The singular values of $(UAU^T)[1,\dots,k|1,\dots,k]$, $U \in U(n)$, denoted by* (1.2), *fill out all values permitted by* (1.7) *and* (1.8).

Proof. The eigenvalues of the real symmetric matrix

$$\begin{pmatrix} X & Y \\ Y & -X \end{pmatrix}$$

appear in pairs but in opposite signs, i.e, $\alpha_1 \geq \cdots \geq \alpha_n \geq -\alpha_n \geq \cdots \geq -\alpha_1$. But now taking a $k\times k$ principal submatrix of an $n\times n$ complex symmetric matrix $Y+iX$, say, $(Y+iX)[1,\dots,k|1,\dots,k]$, amounts to taking a $(2k)\times(2k)$ principal submatrix of the $(2n)\times(2n)$ matrix

$$\begin{pmatrix} X & Y \\ Y & -X \end{pmatrix}$$

corresponding to the rows and columns $1,\dots,k,n+1,\dots,n+k$. So we have (1.7) and (1.8) by Theorem 1.1, with $p=q=k$ and $m=n$. The converse can be established by using Theorem 1.1 again while considering the following set:

$$\{(U\oplus U)\operatorname{diag}(\alpha_1,\dots,\alpha_n,-\alpha_1\dots,-\alpha_n)(U^T\oplus U^T) : U \in O(n)\}. \qquad \square$$

The deficiency of 2 in Theorem 1.3 can now be viewed as a result of deleting 2 rows and 2 columns of the real symmetric matrix

$$\begin{pmatrix} X & Y \\ Y & -X \end{pmatrix}.$$

The last real form of $\mathfrak{sp}(n,\mathbb{C})$ is $\mathfrak{sp}_{p,q}$. Without loss of generality, we may assume that $p \leq q$. It is known that

$$\mathfrak{sp}_{p,q} = \left\{ \begin{pmatrix} X_{11} & X_{12} & X_{13} & X_{14} \\ X_{12}^* & X_{22} & X_{14}^T & X_{24} \\ -\overline{X}_{13} & \overline{X}_{14} & \overline{X}_{11} & -\overline{X}_{12} \\ X_{14}^* & -\overline{X}_{24} & -X_{12}^T & \overline{X}_{22} \end{pmatrix} : X_{11}^* = -X_{11} \in \mathbb{C}_{p\times p},\right.$$

$$X_{22}^* = -X_{22} \in \mathbb{C}_{q\times q},\ X_{13}^T = X_{13},\ X_{24}^T = X_{24} \Bigg\},$$

$$\mathfrak{k} = \left\{ \begin{pmatrix} X_{11} & 0 & X_{13} & 0 \\ 0 & X_{22} & 0 & X_{24} \\ -\overline{X}_{13} & 0 & \overline{X}_{11} & 0 \\ 0 & -\overline{X}_{24} & 0 & \overline{X}_{22} \end{pmatrix} : X_{11}^* = -X_{11} \in \mathbb{C}_{p\times p},\right.$$

$$\left. X_{22}^* = -X_{22} \in \mathbb{C}_{q\times q},\ X_{13}^T = X_{13},\ X_{24}^T = X_{24} \right\},$$

$$\mathfrak{p} = \left\{ \begin{pmatrix} 0 & X_{12} & 0 & X_{14} \\ X_{12}^* & 0 & X_{14}^T & 0 \\ 0 & \overline{X}_{14} & 0 & -\overline{X}_{12} \\ X_{14}^* & 0 & -X_{12}^T & 0 \end{pmatrix} \right\},$$

$$\mathfrak{a} = \oplus_{1\le j\le p}\mathbb{R}(E_{j,p+j} + E_{p+j,j} - E_{p+q+j,2p+q+j} - E_{2p+q+j,p+q+j}),$$

$$\mathfrak{a}_+ = \{\oplus_{1\le j\le p} a_j(E_{j,p+j} + E_{p+j,j} - E_{p+q+j,2p+q+j} - E_{2p+q+j,p+q+j}) : a_1 \ge \cdots \ge a_p \ge 0\}.$$

We then identify $\mathfrak{a}$ with $\mathbb{R}$ and $\mathfrak{a}_+$ with $\{(a_1,\dots,a_n) : a_1 \ge \cdots \ge a_n \ge 0\}$, respectively. The group K is

$$\left\{ \begin{pmatrix} U_1 & 0 & -\overline{V}_1 & 0 \\ 0 & U_2 & 0 & -\overline{V}_2 \\ V_1 & 0 & \overline{U}_1 & 0 \\ 0 & V_2 & 0 & \overline{U}_2 \end{pmatrix} : \begin{pmatrix} U_1 & -\overline{V}_1 \\ V_1 & \overline{U}_1 \end{pmatrix} \in Sp(p),\ \begin{pmatrix} U_2 & -\overline{V}_2 \\ V_2 & \overline{U}_2 \end{pmatrix} \in Sp(q) \right\},$$

which can be viewed as the image of $Sp(p)\times Sp(q)$ under the map γ:

$$\left(\begin{pmatrix} A_1 & B_1 \\ C_1 & D_1 \end{pmatrix}, \begin{pmatrix} A_2 & B_2 \\ C_2 & D_2 \end{pmatrix} \right) \mapsto \begin{pmatrix} A_1 & 0 & B_1 & 0 \\ 0 & A_2 & 0 & B_2 \\ C_1 & 0 & D_1 & 0 \\ 0 & C_2 & 0 & D_2 \end{pmatrix}.$$

Given $C \in \mathfrak{sp}_{p,q}$, there exists $W \in Sp(p,q)$ such that W^*CW is of the following form:

$$\begin{pmatrix} 0 & C_1 & 0 & 0 \\ C_1^T & 0 & 0 & 0 \\ 0 & 0 & 0 & -C_1 \\ 0 & 0 & -C_1^T & 0 \end{pmatrix},$$

where $C_1 = c_1E_{11} \oplus \cdots \oplus c_pE_{pp}$ with $c_i \ge 0$ for all $i = 1,\dots,p$. So the eigenvalues of $C \in \mathfrak{p}$ appear as double pairs but opposite in signs.

THEOREM 3.3. *With the above notations, let $A \in \mathfrak{p}$ and let*

$$\mathfrak{a}_+(A) = (\alpha_1,\dots,\alpha_{\min\{p,q\}}).$$

Then

$$\{\mathfrak{a}_+((UAU^*)[1,\dots,k_1,p+q+1,\dots,p+q+k_2|1,\dots,k_1,p+q+1,\dots,p+q+k_2]) : U \in K\}$$

consists of the elements $(\beta_1, \ldots, \beta_{\min\{k_1,k_2\}})$ *permitted by* (1.7) *and* (1.8).

Proof. Without loss of generality, we may assume that $p \le q$. Apply Thompson's interlacing inequalities (1.7) and (1.8) to show one inclusion. Then consider the set

$$\{W\hat{A}W^{-1} : W = \gamma((U \oplus U), (V \oplus V)), U \in O(p), V \in O(q)\},$$

where

$$\hat{A} = \begin{pmatrix} 0 & A_1 & 0 & 0 \\ A_1^T & 0 & 0 & 0 \\ 0 & 0 & 0 & -A_1 \\ 0 & 0 & -A_1^T & 0 \end{pmatrix},$$

and $A_1 = \alpha_1 E_{11} \oplus \cdots \oplus \alpha_p E_{pp}$ with $\alpha_i \ge 0$ for all $i = 1, \ldots, p$. Apply Theorem 1.2 to obtain the desired conclusion. □

4. The real forms of $\mathfrak{so}_{2n}(\mathbb{C})$ and $\mathfrak{so}_{2n+1}(\mathbb{C})$. We will consider the even and odd cases together in this section. The Cartan decomposition of the algebra of complex skew symmetric matrices $\mathfrak{so}_n(\mathbb{C})^{\mathbb{R}}$ is $\mathfrak{so}_n(\mathbb{C})^{\mathbb{R}} = \mathfrak{so}(n) + i\mathfrak{so}(n)$, where $\mathfrak{so}(n)$ is the algebra of real skew symmetric matrices. The group K is $SO(n)$, the special orthogonal group. The orbit of $C \in \mathfrak{so}(n)$ is $O(C) = \{OCO^T : O \in SO(n)\}$. Thus,

$$\mathfrak{p} = i\mathfrak{so}(n)$$

and

$$\mathfrak{a} = i\mathbb{R}(E_{12} - E_{21}) \oplus i\mathbb{R}(E_{34} - E_{43}) \oplus \cdots \oplus i\mathbb{R}(E_{[\frac{n}{2}]-1,[\frac{n}{2}]} - E_{[\frac{n}{2}],[\frac{n}{2}]-1}) \oplus (0),$$

where the trivial block exists for the odd case. For the odd case $\mathfrak{so}_{2n+1}(\mathbb{C})$,

$$\mathfrak{a}_+ = \{i(a_1(E_{12} - E_{21}) \oplus a_2(E_{34} - E_{43}) \oplus \cdots \oplus a_n(E_{2n-1,2n} - E_{2n,2n-1}) \oplus 0) : a_1 \ge \cdots \ge a_n \ge 0\},$$

which can then be identified with $\{(a_1, \ldots, a_n) : a_1 \ge \cdots \ge a_n \ge 0\}$. However, for the even case $\mathfrak{so}_{2n}(\mathbb{C})$,

$$\mathfrak{a}_+ = \{i(a_1(E_{12} - E_{21}) \oplus a_2(E_{34} - E_{43}) \oplus \cdots \oplus a_n(E_{2n-1,2n} - E_{2n,2n-1})) : a_1 \ge \cdots \ge a_{n-1} \ge |a_n|\},$$

which can then be identified with $\{(a_1, \ldots, a_n) : a_1 \ge \cdots \ge a_{n-1} \ge |a_n|\}$. In either case, the absolute values of α's (each appears twice) are the singular values of the real skew symmetric matrix. We shall identify $\mathfrak{so}(n)$ and $i\mathfrak{so}(n)$ in the obvious way. The following theorem has been established by Thompson [13, p. 251] when $k = n - 1$.

THEOREM 4.1. *Let A be an $n \times n$ real skew symmetric matrix with singular values $\alpha_1 \ge \alpha_1 \ge \alpha_2 \ge \alpha_2 \ge \cdots \ge \alpha_{[\frac{n}{2}]} \ge \alpha_{[\frac{n}{2}]} \ge (0)$ where the zero is a singular value if n is odd, i.e.,*

$$\mathfrak{a}_+(A) = \begin{cases} (\alpha_1, \ldots, \alpha_{[\frac{n}{2}]}) & \text{if } n \text{ is odd,} \\ (\alpha_1, \ldots, \alpha_{\frac{n}{2}-1}, \pm\alpha_{\frac{n}{2}}) & \text{if } n \text{ is even.} \end{cases}$$

Then the set $\{\mathfrak{a}_+((UAU^T)[1, \ldots, k|1, \ldots, k]) : U \in SO(n)\}$ consists of the elements $(\beta_1, \ldots, \beta_{[\frac{k}{2}]})$, where

1. *If k is odd,*

$$\alpha_i \geq \beta_i, \quad i = 1, \ldots, \left[\frac{k}{2}\right], \qquad \beta_i \geq \begin{cases} \alpha_{i+n-k} & \text{if } i+n-k \leq [\frac{n}{2}], \\ 0 & \text{otherwise.} \end{cases}$$

2. *If k is even,*

$$\alpha_i \geq \beta_i, \quad i = 1, \ldots, \frac{k}{2} - 1, \quad \alpha_{\frac{k}{2}} \geq |\beta_{\frac{k}{2}}|,$$

and, when $i \neq \frac{k}{2}$,

$$\beta_i \geq \begin{cases} \alpha_{i+n-k} & \text{if } i+n-k \leq [\frac{n}{2}], \\ 0 & \text{otherwise.} \end{cases}$$

Proof. Notice that the special orthogonal group $SO(n)$ in the following set

$$\{\mathfrak{a}_+((UAU^T)[1, \ldots, k|1, \ldots, k] : U \in SO(n)\}$$

can be replaced by $O(n)$, providing that $k < n$. Notice also that the orbits of the real skew symmetric matrix under the actions of $SO(n)$ and $O(n)$ are not the same, but the compressions are identical. The eigenvalues of the Hermitian iA and the eigenvalues of its $k \times k$ principal submatrix are

$$\alpha_1 \geq \cdots \geq \alpha_{[\frac{n}{2}]-1} \geq \alpha_{[\frac{n}{2}]} \geq (0) \geq -\alpha_{[\frac{n}{2}]} \geq -\alpha_{[\frac{n}{2}]-1} \cdots \geq -\alpha_1,$$
$$\beta_1 \geq \cdots \geq \beta_{[\frac{k}{2}]-1} \geq \beta_{[\frac{k}{2}]} \geq (0) \geq -\beta_{[\frac{k}{2}]} \geq -\beta_{[\frac{k}{2}]-1} \cdots \geq -\beta_1,$$

respectively, where the zero exists if n (k) is odd. We will denote them by $\alpha'_1 \geq \cdots \geq \alpha'_n$ and $\beta'_1 \geq \cdots \geq \beta'_k$. By Theorem 1.1, $\alpha'_i \geq \beta'_i \geq \alpha'_{i+n-k}$, $i = 1, \ldots, [\frac{k}{2}]$. But $\alpha'_i = \alpha_i$ and $\beta'_i = \beta_i$ when $i = 1, \ldots, [\frac{k}{2}]$. Moreover, if $1 \leq i+n-k \leq [\frac{n}{2}]$, then $\alpha'_{i+n-k} = \alpha_{i+n-k}$; if $i+n-k > [\frac{n}{2}]$, then $\alpha'_{i+n-k} = -\alpha_{k-i+1}$ when $i+n-k > [\frac{n}{2}]+1$ and $\alpha'_{[\frac{n}{2}]+1} = 0$. The rest follows from a result of Thompson [13] (which is the case $k = n-1$; see the remark on p. 251 and the proof of Theorem 3 in [13]) and the remark of Fan and Pall analogously. The converse of the interlacing inequalities corresponding to the even k needs special care and can be done easily by suitably adjusting $U \in O(n)$. □

We notice that if $n-k$ is even, then the submatrix is in the Cartan subspace associated with the same series. But if $n-k$ is odd, the submatrix belongs to the Cartan subspace of another series, i.e., switches from B-series to D-series or from D-series to B-series.

The real forms $\mathfrak{so}_{p,q}$ ($q+p$ can be $2n$ or $2n+1$) give the interlacing result for real $p \times q$ matrices (with loss of generality, we assume that $p \leq q$) and the treatment is exactly the same as the $\mathfrak{su}_{p,q}$ case once we see that

$$\mathfrak{so}_{p,q} = \left\{ \begin{pmatrix} X_1 & Y \\ Y^T & X_2 \end{pmatrix} : X_1^T = -X_1,\ X_2^T = -X_2,\ Y \in \mathbb{R}_{p\times q} \right\},$$
$$K = SO(p) \times SO(q),$$
$$\mathfrak{k} = \mathfrak{so}(p) \oplus \mathfrak{so}(q),\ \text{i.e.,}\ Y = 0,$$
$$\mathfrak{p} = \{ \begin{pmatrix} 0 & Y \\ Y^T & 0 \end{pmatrix}, Y \in \mathbb{R}_{p\times q} \},\ \text{i.e.,}\ X_1 = X_2 = 0,$$
$$\mathfrak{a} = \oplus_{1\leq j\leq p} \mathbb{R}(E_{j,p+j} + E_{p+j,j}).$$

If $p < q$, then

$$\mathfrak{a}_+ = \{\oplus_{1\le j\le p} a_j(E_{j,p+j} + E_{p+j,j}) : a_1 \ge \cdots \ge a_p \ge 0\},$$

which can then be identified with $\{(a_1, \ldots, a_p) : a_1 \ge \cdots \ge a_p \ge 0\}$. If $p = q$, then

$$\mathfrak{a}_+ = \{\oplus_{1\le j\le p} a_j(E_{j,p+j} + E_{p+j,j}) : a_1 \ge \cdots \ge a_p \ge 0\},$$

which can then be identified with $\{(a_1, \ldots, a_p) : a_1 \ge \cdots \ge \alpha_{p-1} \ge |a_p| \ge 0\}$.

THEOREM 4.2. *Let A be a $p \times q$ real matrix with*

$$\mathfrak{a}_+(A) = \begin{cases} (\alpha_1, \ldots, \alpha_{\min\{p,q\}}) & \text{if } p \ne q, \\ (\alpha_1, \ldots, \alpha_{p-1}, \mathit{sign}(\det\ A)\alpha_p) & \text{if } p = q, \end{cases}$$

where $\alpha_1 \ge \cdots \alpha_{\min\{p,q\}} \ge 0$ are the singular values of A.

1. *If $k_1 \ne k_2$, then $\{\mathfrak{a}_+((UAV)[1, \ldots, k_1|1, \ldots, k_2]) : U \in SO(p), V \in SO(q)\}$ consists of the elements $(\beta_1, \ldots, \beta_{\min\{k_1,k_2\}})$ permitted by* (1.7) *and* (1.8).
2. *If $k_1 = k_2 = k$, then $\{\mathfrak{a}_+((UAV)[1, \ldots, k|1, \ldots, k]) : U \in SO(p), V \in SO(q)\}$ consists of the elements $(\beta_1, \ldots, \beta_k)$ permitted by* (1.7) *and* (1.8)*, where the last inequality of* (1.8) *is replaced by $\alpha_k \ge |\beta_k|$.*

Proof. Notice that the special orthogonal groups $SO(p)$ and $SO(q)$ in the following set

$$\{\mathfrak{a}_+((UAV)[1, \ldots, k_1|1, \ldots, k_2]) : U \in SO(p), V \in SO(q)\}$$

can be replaced by $O(p)$ and $O(q)$, respectively, provided that $(k_1, k_2) \ne (p, q)$. The rest follows from the real case of Theorem 1.2. ∎

Theorem 4.2, when $k_1 = k_2 = k$, says that a real $p \times q$ matrix can be constructed so that the $k \times k$ submatrix has the prescribed singular values, β's, and nonnegative or nonpositive determinants if the interlacing inequalities are satisfied.

The real form $\mathfrak{so}^*(2n)$ of $\mathfrak{so}_{2n}(\mathbb{C})$ yields the interlacing result for the complex skew symmetric matrices. It is known that

$$\begin{aligned}
\mathfrak{so}^*(2n) &= \left\{ \begin{pmatrix} X & Y \\ -\overline{Y} & \overline{X} \end{pmatrix} : X^T = -X,\ Y^* = Y,\ X, Y \in \mathbb{C}_{n\times n} \right\}, \\
K &= \left\{ \begin{pmatrix} A & B \\ -B & A \end{pmatrix} : A^TA + B^TB = I,\ A^TB = B^TA,\ A, B \in \mathbb{R}_{n\times n} \right\}, \\
\mathfrak{k} &= \left\{ \begin{pmatrix} X & Y \\ -Y & X \end{pmatrix} : X^T = -X,\ Y^T = Y,\ X, Y \in \mathbb{R}_{n\times n} \right\}, \\
\mathfrak{p} &= \left\{ \begin{pmatrix} X & Y \\ Y & -X \end{pmatrix} : X^T = -X,\ Y^T = -Y,\ X, Y \in i\mathbb{R}_{n\times n} \right\}, \\
\mathfrak{a} &= i\mathbb{R}((E_{12} - E_{21}) - (E_{n+1,n+2} - E_{n+2,n+1})) \\
&\quad \oplus\ i\mathbb{R}((E_{34} - E_{43}) - (E_{n+3,n+4} - E_{n+4,n+3})) \oplus \cdots.
\end{aligned}$$

Analogous to $\mathfrak{sp}_{2n}(\mathbb{R})$, we identify K with $U(n)$ and identify $\mathfrak{p}$ with the space of $n \times n$ complex skew symmetric matrices via the map

$$\begin{pmatrix} X & Y \\ Y & -X \end{pmatrix} \mapsto Y + iX, \qquad X, Y \in i\mathbb{R}_{n\times n},\ X^T = -X,\ Y^T = -Y.$$

Then $\mathfrak{a}$ is identified with $i \oplus_{1\le j\le [\frac{n}{2}]} \mathbb{R}(E_{2j-1,2j} - E_{2j,2j-1})$ and the action of K on $\mathfrak{p}$ is $A \to UAU^T$. Now (1.11) is the Autonne decomposition for a complex skew symmetric matrix.

THEOREM 4.3 (see Thompson [13]). *Let A be an $n \times n$ complex skew symmetric matrix having singular values $\alpha_1 \geq \alpha_1 \geq \alpha_2 \geq \alpha_2 \geq \cdots \alpha_{[\frac{n}{2}]} \geq \alpha_{[\frac{n}{2}]} \geq (0)$. (The zero is a singular value if n is odd.) The singular values of $(UAU^T)[1,\ldots,k|1,\ldots,k]$, $U \in U(n)$, denoted by $\beta_1 \geq \beta_1 \geq \beta_2 \geq \beta_2 \geq \cdots \geq \beta_{[\frac{k}{2}]} \geq \beta_{[\frac{k}{2}]} \geq (0)$ (the zero is a singular value if k is odd), fill out all values permitted by*

$$\alpha_i \geq \beta_i, \quad i = 1,\ldots,\left[\frac{k}{2}\right], \qquad \beta_i \geq \begin{cases} \alpha_{i+n-k} & \text{if } i+n-k \leq [\frac{n}{2}], \\ 0 & \text{otherwise.} \end{cases}$$

Proof. The case $k = n-1$ has been established by Thompson [13, Theorem 2]. The other cases follow by the remark of Fan and Pall analogously. The following is a reason for interlacing inequalities. Each element in $\mathfrak{p}$ is a $(2n) \times (2n)$ Hermitian matrix. The eigenvalues appear in double pairs but opposite in signs, i.e., $\alpha_1 \geq \alpha_1 \geq \cdots \geq \alpha_{[\frac{n}{2}]} \geq \alpha_{[\frac{n}{2}]} \geq (0) \geq (0) \geq -\alpha_{[\frac{n}{2}]} \geq -\alpha_{[\frac{n}{2}]} \geq \cdots \geq -\alpha_1 \geq -\alpha_1$. (If n is odd, the two zeros are included as eigenvalues.) But now taking a $k \times k$ principal submatrix of an $n \times n$ complex skew symmetric matrix $Y + iX$, say, $(Y+iX)[1,\ldots,k|1,\ldots,k]$, amounts to taking a $(2k) \times (2k)$ principal submatrix of the $(2n) \times (2n)$ Hermitian matrix

$$\begin{pmatrix} X & Y \\ -Y & X \end{pmatrix}$$

corresponding to the rows and columns $1,\ldots,k,n+1,\ldots,n+k$. Hence we have the desired interlacing inequalities. □

Acknowledgment. The author thanks Professor Queiró for his careful reading and stimulating suggestions and for bringing reference [11] to the author's attention.

REFERENCES

[1] R. BHATIA, *Matrix Analysis*, Springer-Verlag, New York, 1997.
[2] D. CARLSON, *Minimax and interlacing theorems for matrices*, Linear Algebra Appl., 54 (1983), pp. 153–172.
[3] A. CAUCHY, *Sur l' Équation à l' Aide de Laquelle on Détermine les Inegalités Séculaires des Mouvements des Planètes,* 1829*; Oeuvres Complètes, Seconde Série*, Vol. IX, Gauthier–Villars, Paris, pp. 174–195.
[4] K. FAN AND G. PALL, *Imbedding conditions for Hermitian and normal matrices*, Canad. J. Math., 9 (1957), pp. 298–304.
[5] S. HELGASON, *Differential Geometry, Lie Groups and Symmetric Spaces*, Academic Press, New York, 1978.
[6] A. W. KNAPP, *Representation Theory of Semisimple Groups*, Princeton University Press, Princeton, NJ, 1986.
[7] A. W. MARSHALL AND I. OLKIN, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York, 1979.
[8] L. MIRSKY, *Matrices with prescribed characteristic roots and diagonal elements*, J. London Math. Soc., 33 (1958), pp. 14–21.
[9] G. N. DE OLIVEIRA, *Matrices with prescribed characteristic polynomial and a prescribed submatrix*, Pacific J. Math., 29 (1969), pp. 653–661.
[10] A. L. ONISHCHIK AND E. B. VINBERG, *Lie Groups and Algebraic Groups*, Springer-Verlag, Berlin, 1990.
[11] J. F. QUEIRÓ, *On the interlacing property for singular values and eigenvalues*, Linear Algebra Appl., 97 (1987), pp. 23–28.
[12] R. C. THOMPSON, *Principal submatrices* IX: *Interlacing inequalities for singular values of submatrices*, Linear Algebra Appl., 5 (1972), pp. 1–12.
[13] R. C. THOMPSON, *Principal minors of complex symmetric and skew matrices*, Linear Algebra Appl., 28 (1979), pp. 249–255.
[14] R. C. THOMPSON, *Matrix Spectral Inequalities*, Johns Hopkins Lecture Series, unpublished manuscript.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 593–612
© 1999 Society for Industrial and Applied Mathematics

MORE RESULTS ON EIGENVECTOR SADDLEPOINTS AND EIGENPOLYNOMIALS*

MARK A. MENDLOVITZ†

Abstract. This paper extends the author's earlier work which proved that, for a certain class of Hermitian eigenproblems $AP = \lambda BP$ (including, most notably, the Toeplitz eigenproblem), the eigenvectors can be represented as saddlepoints of a special form of the Rayleigh quotient. Each saddlepoint solves a min-max/max-min optimization problem whose optimal value is the eigenvalue. The zeros of the eigenpolynomial factors are located with respect to the unit circle. These results were proved, in part, by using an inertia theorem for Stein equations. For another class of Hermitian eigenproblems $AP = \lambda BP$, an analogous set of results are derived here using an inertia theorem for Lyapunov equations. In this case, the zeros of the eigenpolynomial factors are located with respect to the imaginary axis, and sufficient conditions for the eigenpolynomial factors to have only extended imaginary zeros are established. The highlight of the article is an important theorem that yields a general factorization of saddlepoint eigenpolynomials and parameterizes all eigenvector saddlepoint representations for an m-dimensional eigenspace. Several key extensions to the theory are also presented.

Key words. eigenvectors, eigenpolynomials, saddlepoints, zeros, Lyapunov equation, Rayleigh quotient, Toeplitz matrix, inertia, perturbation

AMS subject classification. 15A18

PII. S0895479897324482

NOTATION.

$\mathbb{R}^{n\times m}$	the space of real $n \times m$ matrices ($\mathbb{R}^n \triangleq \mathbb{R}^{n\times 1}$)
$\mathbb{C}^{n\times m}$	the space of complex $n \times m$ matrices ($\mathbb{C}^n \triangleq \mathbb{C}^{n\times 1}$)
$\mathbb{C}_e$	the extended complex plane, $\mathbb{C} \cup \{\infty\}$
$\overline{[G]}_m$	the upper left $m \times m$ submatrix of the $n \times n$ matrix G, i.e., $[g_{ij}]$ $i = 1, \ldots, m$, $j = 1, \ldots, m$
$\underline{[G]}_m$	the lower right $m \times m$ submatrix of the $n \times n$ matrix G, i.e., $[g_{ij}]$ $i = n-m+1, \ldots, n$, $j = n-m+1, \ldots, n$
$G\rceil$	the upper right principal submatrix of the $n \times n$ matrix G, i.e., $[g_{ij}]$ $i = 1, \ldots, n-1$, $j = 2, \ldots, n$
$\lfloor G$	the lower left principal submatrix of the $n \times n$ matrix G, i.e., $[g_{ij}]$ $i = 2, \ldots, n$, $j = 1, \ldots, n-1$
H^*	the complex conjugate transpose of H
$\det(\cdot)$	the determinant of $(\cdot)$
I	the identity matrix
e_i	the ith column of the identity matrix
$\mathcal{I}(G(\lambda))$	an open interval in $\mathbb{R}$ containing $\{\lambda \mid \det G(\lambda) = 0\}$
$H > 0$ $(H < 0)$	H is a positive (negative) definite matrix
$H \geq 0$ $(H \leq 0)$	H is a positive (negative) semidefinite matrix
$\tilde{P} * \hat{P}$	the convolution of $\tilde{P}$ and $\hat{P}$
$\mathcal{P}_m(R)$	the $(n+m) \times m$ convolution matrix formed from $R = [\ r(0) \quad r(1) \quad \cdots \quad r(n)\ ]^T$; the ijth element is $r(i-j)$ for $0 \leq i-j \leq n$, and 0 otherwise
$\arg\min_x f(x)$	an argument x which minimizes $f(x)$
$\lvert z\rvert < 1$ (> 1)	the region of the complex plane interior (exterior) to the unit circle

*Received by the editors July 14, 1997; accepted for publication (in revised form) by V. Mehrmann May 5, 1999; published electronically December 28, 1999.
http://www.siam.org/journals/simax/21-2/32448.html

†Former address: Mission Research Corporation, P.O. Drawer 719, Santa Barbara, CA 93102. Current address: LinCom Corporation, 5120 W. Goldleaf Circle, Suite 400, Los Angeles, CA 90056 (mendlovitz@prodigy.net).

$\mathrm{Re}[s] > 0 \ (< 0)$	the region of the complex plane where the real part of the complex variable s is positive (negative)
$\rho(P)$	$\stackrel{\triangle}{=} P^*AP/P^*BP$ for $P \in \mathbb{C}^{N+1} - \{0\}$, $A = A^*, B = B^* \in \mathbb{C}^{(N+1)\times(N+1)}$
$R(z)$	$\stackrel{\triangle}{=} r(0) + r(1)z^{-1} + \cdots + r(n)z^{-n}$ when $R = [\ r(0) \quad r(1) \quad \cdots \quad r(n)\ ]^T$
$R(s)$	$\stackrel{\triangle}{=} r(0)s^n + r(1)s^{n-1} + \cdots + r(n)$ when $R = [\ r(0) \quad r(1) \quad \cdots \quad r(n)\ ]^T$
C_R	$\stackrel{\triangle}{=} \begin{bmatrix} -r(1) & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -r(n-1) & 0 & \dots & 1 \\ -r(n) & 0 & \dots & 0 \end{bmatrix}$ for $R = [\ 1 \quad r(1) \quad \cdots \quad r(n)\ ]^T$
$X_{(i:j)}$	$\stackrel{\triangle}{=} [\ x(i) \quad \cdots \quad x(j)\ ]^T$

1. Introduction. The Hermitian eigenproblem is one of the most widely studied topics in linear algebra [5]. This important problem can be stated as follows: For Hermitian matrices $A, B \in \mathbb{C}^{(N+1)\times(N+1)}$, where $B \neq 0$, determine a basis set of finite eigenvalue/eigenvector pairs or *eigenpairs* $\{(\lambda_k, P_k) \in \mathbb{R} \times \mathbb{C}^{N+1} \mid k = 1, \dots, j \leq N+1\}$ which solve

$$AP = \lambda BP. \tag{1}$$

The eigenvalues are the zeros of the jth order λ-polynomial $\det(A - \lambda B)$.[1] It can be shown that if an eigenvector P_k satisfies (1) and $P_k^* B P_k \neq 0$, then P_k is a stationary point of the Rayleigh quotient $\rho(P) = P^*AP/P^*BP$, i.e., the gradient $\nabla\rho(P_k) = 0$, and the eigenvalue $\lambda_k = \rho(P_k)$ is the corresponding stationary value. Conversely, all stationary points and stationary values of $\rho(P)$ are eigenvectors and eigenvalues, respectively. In general, a stationary point may be a maximizing point, a minimizing point, a saddlepoint, or none of these. An example of the last case is a *point of inflection*, such as the one at $(x, y) = (0, 0)$ for the function $f(x, y) = x^3 + y^3$ defined on $\{(x, y) \in \mathbb{R} \times \mathbb{R}\}$.

For eigenvector P_k to be a *saddlepoint eigenvector*, there must exist a point $(\tilde{P}_k, \hat{P}_k)$ contained in a real or complex Cartesian product space $\tilde{\mathcal{S}}_k \times \hat{\mathcal{S}}_k$ and a map $\mathcal{F} : \tilde{\mathcal{S}}_k \times \hat{\mathcal{S}}_k \longrightarrow \mathbb{C}^{N+1}$ such that $P_k = \mathcal{F}(\tilde{P}_k, \hat{P}_k)$, and

$$\rho(\mathcal{F}(\tilde{P}_k, \hat{P})) \leq \rho(\mathcal{F}(\tilde{P}_k, \hat{P}_k)) \leq \rho(\mathcal{F}(\tilde{P}, \hat{P}_k))$$

for all $\tilde{P} \in \tilde{\mathcal{S}}_k$ and $\hat{P} \in \hat{\mathcal{S}}_k$. Any point $(\tilde{P}_k, \hat{P}_k)$ satisfying this inequality is said to be a *saddlepoint representation* of the eigenvector or, simply, an *eigenvector saddlepoint*. It is easy to see from the inequality that an eigenvector saddlepoint is a stationary point of a differentiable function $\rho(\mathcal{F}(\cdot, \cdot))$. An equivalent definition is given by the so-called *minimax theorem* [14], which states that $(\tilde{P}_k, \hat{P}_k)$ is a saddlepoint if and only if

$$\max_{\hat{P} \in \hat{\mathcal{S}}_k} \min_{\tilde{P} \in \tilde{\mathcal{S}}_k} \rho(\mathcal{F}(\tilde{P}, \hat{P})) = \rho(\mathcal{F}(\tilde{P}_k, \hat{P}_k)) = \min_{\tilde{P} \in \tilde{\mathcal{S}}_k} \max_{\hat{P} \in \hat{\mathcal{S}}_k} \rho(\mathcal{F}(\tilde{P}, \hat{P})),$$

$$\tilde{P}_k = \arg \min_{\tilde{P} \in \tilde{\mathcal{S}}_k} \left\{ \max_{\hat{P} \in \hat{\mathcal{S}}_k} \rho(\mathcal{F}(\tilde{P}, \hat{P})) \right\},$$

[1] $\det(A - \lambda B)$ of degree j implies that there are j finite eigenvalues and $N + 1 - j$ infinite eigenvalues. If $\det(A - \lambda B)$ is identically zero, every number is an eigenvalue.

and

$$\hat{P}_k = \arg\max_{\hat{P}\in\hat{\mathcal{S}}_k} \left\{ \min_{\tilde{P}\in\tilde{\mathcal{S}}_k} \rho(\mathcal{F}(\tilde{P},\hat{P})) \right\}.$$

The optimal value of the objective function at the eigenvector saddlepoint solution is the eigenvalue, i.e., $\lambda_k = \rho(\mathcal{F}(\tilde{P}_k,\hat{P}_k))$.

Now let $\rho(\tilde{P},\hat{P})$ denote the special Rayleigh quotient obtained by choosing $\mathcal{F}$ to be the convolution map, i.e., $\rho(\mathcal{F}(\tilde{P},\hat{P})) = \rho(\tilde{P}*\hat{P}) \stackrel{\triangle}{=} \rho(\tilde{P},\hat{P})$ or, equivalently,

$$\begin{aligned}\rho(\tilde{P},\hat{P}) &= \frac{\tilde{P}^*\left[\mathcal{P}^*(\hat{P})A\mathcal{P}(\hat{P})\right]\tilde{P}}{\tilde{P}^*\left[\mathcal{P}^*(\hat{P})B\mathcal{P}(\hat{P})\right]\tilde{P}} \\ &= \frac{\hat{P}^*\left[\mathcal{P}^*(\tilde{P})A\mathcal{P}(\tilde{P})\right]\hat{P}}{\hat{P}^*\left[\mathcal{P}^*(\tilde{P})B\mathcal{P}(\tilde{P})\right]\hat{P}} \\ &= \rho(\hat{P},\tilde{P}),\end{aligned}$$

where $\tilde{P}\in\mathbb{C}^k$, $\hat{P}\in\mathbb{C}^{N-k+2}$,[2] and $\mathcal{P}(\cdot)$ is the matrix representation of the convolution operator. Since convolution is commutative, $\mathcal{P}(\hat{P})\tilde{P} = \mathcal{P}(\tilde{P})\hat{P}$. The subscript of $\mathcal{P}$, which indicates column dimension, is dropped to reduce notational complexity whenever the dimension is clear from the context, as it is here.

Let us express the Hermitian eigenproblem as

$$G(\lambda)P \stackrel{\triangle}{=} (A-\lambda B)P = 0, \tag{2}$$

with $A, B\in\mathbb{C}^{(N+1)\times(N+1)}$ and $B>0$, and let $\underline{[G(\lambda)]}_m$ and $\overline{[G(\lambda)]}_m$ denote the lower right and upper left $m\times m$ submatrices of $G(\lambda)$, respectively. In the earlier work [9], it was proved, using an inertia theorem for Stein equations and some basic matrix perturbation theory, that when $\underline{[G(\lambda)]}_N - \overline{[G(\lambda)]}_N \le 0$ or $\underline{[G(\lambda)]}_N - \overline{[G(\lambda)]}_N \ge 0$ for all $\lambda\in\mathcal{I}(G(\lambda))$ (an open interval in $\mathbb{R}$ containing all of the eigenvalues), an eigenvector P_k corresponding to the kth largest eigenvalue of (2) has a *factorization* $P_k = \tilde{P}_k * \hat{P}_k$ such that $(\tilde{P}_k,\hat{P}_k)\in\mathbb{C}^k\times\mathbb{C}^{N-k+2}$ is a saddlepoint of the Rayleigh quotient $\rho(\tilde{P},\hat{P})$. (Note that since convolution and polynomial multiplication are equivalent operations, we can also say that the *eigenpolynomial* $P_k(z)\stackrel{\triangle}{=} p_k(0)+\cdots+p_k(N)z^{-N}$ has a factorization $P_k(z) = \tilde{P}_k(z)\hat{P}_k(z)$.) If $\underline{[G(\lambda)]}_N - \overline{[G(\lambda)]}_N \le 0$ (≥ 0),[3] the eigenpolynomial factor $\tilde{P}_k(z)$ has all of its zeros in $|z|\ge 1$ ($|z|\le 1$), and the factor $\hat{P}_k(z)$ has all of its zeros in $|z|\le 1$ ($|z|\ge 1$). Additionally, it was shown that real $\tilde{P}_k$ and $\hat{P}_k$ exist whenever A and B are real. Eigenvector saddlepoints for this class of problems are called *saddlepoints of the first kind* or *type*-1 *saddlepoints*. Included in this class of eigenproblems are the important special cases in which A is a Hermitian Toeplitz or a Hankel-conjugate–transpose-Hankel matrix [3, 4]. Sufficient conditions for the eigenpolynomial factors to have all of their zeros on $|z|=1$ when A and B are Toeplitz were also developed.

[2] For the sake of notational brevity, we shall implicitly assume that the spaces containing the arguments of ρ exclude the possibility of $\rho(0)$, since ρ is not defined at the origin. For example, $\min_{\tilde{P}\in\mathbb{C}^k}\rho(\tilde{P},\hat{P})$ actually means $\min_{\tilde{P}\in\mathbb{C}^k-\{0\}}\rho(\tilde{P},\hat{P})$.

[3] A convention frequently used here is the placement of a parallel condition and its corresponding result in parentheses.

Here we study another set of eigenproblems for which $\rho(\tilde{P}, \hat{P})$ has eigenvector saddlepoints, and results analogous to the ones summarized above are obtained by applying an inertia theorem for Lyapunov equations instead. These eigenvector saddlepoints, termed *saddlepoints of the second kind* or *type-2 saddlepoints*, are shown to exist when the sum of the lower left principal submatrix of $G(\lambda)$, $\lfloor G(\lambda) \in \mathbb{C}^{N\times N}$, and the upper right principal submatrix of $G(\lambda)$, $G(\lambda)\rceil \in \mathbb{C}^{N\times N}$, satisfies either $\lfloor G(\lambda) + G(\lambda)\rceil \leq 0$ or $\lfloor G(\lambda)+G(\lambda)\rceil \geq 0$ for all $\lambda \in \mathcal{I}(G(\lambda))$. The zeros of the eigenpolynomial factors in this case are located with respect to the imaginary axis rather than the unit circle, and sufficient conditions for the eigenpolynomial factors to have all of their zeros on the extended imaginary axis $\{\mathrm{Re}[s] = 0\} \cup \{\infty\}$ are proved. Again, perturbation methods are needed to prove existence of the saddlepoints. This paper also serves to resolve some subtle issues not dealt with in [9], such as the correct way to define the eigenpolynomial zeros. The highlight of the paper is an important theorem that gives a general factorization of any saddlepoint eigenpolynomial and parameterizes all eigenvector saddlepoint representations for an m-dimensional eigenspace. This result is significant because it provides a very elegant and useful characterization of the eigenstructure for many problems of practical interest, including the case where A and B are Hermitian Toeplitz matrices. It also leads to a new set of bounds on the number of zeros that an eigenpolynomial factor may have on the extended imaginary axis or unit circle.

2. Preliminaries. Some foundational theory and definitions, primarily derived from [2, 8, 9], are presented in this section.

LEMMA 2.1. *The eigenvalues of the $n \times n$ companion matrix*

$$C_R \triangleq \begin{bmatrix} -r(1) & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -r(n-1) & 0 & \dots & 1 \\ -r(n) & 0 & \dots & 0 \end{bmatrix}$$

are the zeros of the monic polynomial $R(s) = s^n + r(1)s^{n-1} + \cdots + r(n)$.

A more detailed discussion of the "zeros" of polynomials $R(s)$ and $R(z)$ is dealt with later. For the moment, however, as long as $r(0) \neq 0$, it is safe to assume the zeros of $R(s)$ are the set of points $\{s_i \in \mathbb{C} \mid R(s_i) = 0, \ \ i = 1, \dots, n\}$.

DEFINITION 2.2. *$(U, V) \in \mathbb{C}^{n\times n} \times \mathbb{C}^{n\times m}$ is said to be a controllable pair if and only if*

$$[\ U - \lambda I \quad V\ ]$$

has full row rank for all $\lambda \in \mathbb{C}$.

The pair (U, V) is always controllable if V is definite or invertible or has full row rank.

DEFINITION 2.3. *If $H \in \mathbb{C}^{n\times n}$, then $\pi(H)$, $\nu(H)$, and $\delta(H)$ denote the number of eigenvalues of H with positive, negative, and zero real parts, respectively, and $\pi(H) + \nu(H) + \delta(H) = n$. The ordered triple $\{\pi(H), \nu(H), \delta(H)\}$ is called the inertia of H with respect to the imaginary axis.*

If $r(0) \neq 0$, the inertia of a polynomial $R(s)$ with respect to the imaginary axis is similarly defined to be the triple containing the number of zeros of $R(s)$ in $\mathrm{Re}[s] > 0$, $\mathrm{Re}[s] < 0$, and on $\mathrm{Re}[s] = 0$, respectively, while the inertia of polynomial $R(z)$ with respect to the unit circle is given by the triple containing the number of zeros of $R(z)$ in $|z| > 1$, $|z| < 1$, and on $|z| = 1$, respectively.

THEOREM 2.4. *Let* $(U^*, V) \in \mathbb{C}^{n\times n} \times \mathbb{C}^{n\times n}$ *be a controllable pair, and assume the Lyapunov equation* $U^*H + HU = V$ *holds for some Hermitian* $H \in \mathbb{C}^{n\times n}$.

(a) *If* $V \geq 0$, *then* $\det H \neq 0$, $\delta(U) = \delta(H) = 0$, $\pi(U) = \pi(H)$, *and* $\nu(U) = \nu(H)$.

(b) *If* $V \leq 0$, *then* $\det H \neq 0$, $\delta(U) = \delta(H) = 0$, $\pi(U) = \nu(H)$, *and* $\nu(U) = \pi(H)$.

Theorem 2.4 is an inertia theorem for Lyapunov equations which originates from the work of Chen [1] and Wimmer [13].

DEFINITION 2.5. $G(\lambda, \tilde{P}_k) \triangleq \mathcal{P}^*_{N-k+2}(\tilde{P}_k)(A - \lambda B)\mathcal{P}_{N-k+2}(\tilde{P}_k)$ *for* $\tilde{P}_k \in \mathbb{C}^k$, $k = 1, \ldots, N+1$.

DEFINITION 2.6. $G(\lambda, \hat{P}_k) \triangleq \mathcal{P}^*_k(\hat{P}_k)(A - \lambda B)\mathcal{P}_k(\hat{P}_k)$ *for* $\hat{P}_k \in \mathbb{C}^{N-k+2}$, $k = 1, \ldots, N+1$.

It is clear from the definitions that the stationary points $\tilde{P}$ of $\rho(\tilde{P}, \hat{P}_k)$ are the eigenvector solutions of $G(\lambda, \hat{P}_k)\tilde{P} = 0$ and vice versa, and the stationary points $\hat{P}$ of $\rho(\hat{P}, \tilde{P}_k)$ are the eigenvector solutions of $G(\lambda, \tilde{P}_k)\hat{P} = 0$ and vice versa. Also, $G(\lambda_k)P_k = 0$ and $P_k = \tilde{P}_k * \hat{P}_k$ imply $G(\lambda_k, \hat{P}_k)\tilde{P}_k = 0$ and $G(\lambda_k, \tilde{P}_k)\hat{P}_k = 0$.

LEMMA 2.7. *Let* (λ_*, R_*) *be an eigenpair satisfying the eigenproblem* $G'(\lambda)R = 0$, *where* $G'(\lambda) \in \mathbb{C}^{n\times n}$ *and* R_* *is scaled so that* $r_*(0) = 1$. *Then*

$$\underline{[G'(\lambda_*)]}_{n-1} C_{R_*} = \lfloor G'(\lambda_*).$$

Proof. The result follows from the definition of C_{R_*} and the equation $G'(\lambda_*)R_* = 0$. □

In a similar way, $C^*_{R_*}\underline{[G'(\lambda_*)]}_{n-1} = G'(\lambda_*)\rceil$.

As in [9], basic matrix perturbation theory is used in some of the proofs. A good introduction to this subject can be found in [8]; however, a more extensive treatment is given by Rellich [12]. Classical matrix perturbation theory generally deals with perturbations which are *analytic functions* of a real parameter ϵ. For our purposes, an analytic function is a function which has a power series expansion convergent on a neighborhood of $\epsilon = 0$. Thus, analytic functions are a proper subset of the set of infinitely differentiable functions. An example of a function which is infinitely differentiable at $\epsilon = 0$ but nonanalytic is $f(\epsilon) = e^{-1/\epsilon^2}\sin(2/\epsilon)$ [12].

It can be shown that if $A(\epsilon), B(\epsilon) \in \mathbb{C}^{n\times n}$ are Hermitian analytic matrix functions of a real parameter ϵ, and $B(\epsilon) > 0$ on a neighborhood of $\epsilon = 0$, then there exists a complete set of analytic eigenpairs $\{(\lambda_{k\epsilon}, P_{k\epsilon}) \in \mathbb{R} \times \mathbb{C}^n \mid k = 1, \ldots, n\}$ satisfying the eigenproblem $(A(\epsilon) - \lambda B(\epsilon))P = 0$. As a result, $P_{k\epsilon}$ and any of its factors are continuous at $\epsilon = 0$. This simple but important fact is invoked in subsequent arguments.

Finally, we shall continue to assume that an eigenvalue's (first) subscript indicates its relative rank, i.e., $\lambda_{k-1} \geq \lambda_k \geq \lambda_{k+1}$ for $k = 2, \ldots, N$. If an eigenvalue is not equal to any other eigenvalues, it is said to be *distinct* (its multiplicity equals one); otherwise, it is *repeated* (its multiplicity is greater than one). If a stationary point of a Rayleigh quotient is an element of a one-dimensional eigenspace, i.e., it corresponds to a distinct stationary value, the stationary point is said to be *unique*.

3. Main results. The first theorem in this section is fundamental.

THEOREM 3.1. *If* (λ_k, P_k) *is an eigenpair which satisfies* $G(\lambda)P = 0$, *where* $B > 0$, $p_k(0) = 1$,

$$\lfloor G(\lambda) + G(\lambda)\rceil \geq 0 \ \ (\leq 0) \quad \forall\, \lambda \in \mathcal{I}(G(\lambda)),$$

and $(C^*_{P_k}, \lfloor G(\lambda_k) + G(\lambda_k)\rceil)$ *is a controllable pair, then*

(a) λ_k *is distinct;*
(b) $P_k(s)$ *has* $k-1$ *zeros in* $\mathrm{Re}[s] > 0$ ($\mathrm{Re}[s] < 0$) *and* $N-k+1$ *zeros in* $\mathrm{Re}[s] < 0$ ($\mathrm{Re}[s] > 0$).

Proof. The proof closely follows the proof of Theorem 5 in [9], so it is shortened here. From Lemma 2.7 and our assumption,

$$C_{P_k}^* \underline{[G(\lambda_k)]}_N + \underline{[G(\lambda_k)]}_N C_{P_k} = \lfloor G(\lambda_k) + G(\lambda_k) \rceil \geq 0 \ \ (\leq 0). \tag{3}$$

The Lyapunov equation (3) satisfies the hypothesis of Theorem 2.4(a) (Theorem 2.4(b)), so $\mathrm{rank}\{\underline{[G(\lambda_k)]}_N\} = N$. This implies that $N \leq \mathrm{rank}\{G(\lambda_k)\} \leq N+1$, but since $G(\lambda_k)$ is singular by definition, $\mathrm{rank}\{G(\lambda_k)\} = N$. Thus, λ_k is distinct.

It can be shown that $\pi(\underline{[G(\lambda_k)]}_N) = k-1$ and $\nu(\underline{[G(\lambda_k)]}_N) = N-k+1$ [9, Theorem 5]. Again, by Theorem 2.4(a) (Theorem 2.4(b)), C_{P_k} satisfying (3) must have $k-1$ eigenvalues in $\mathrm{Re}[s] > 0$ ($\mathrm{Re}[s] < 0$) and $N-k+1$ eigenvalues in $\mathrm{Re}[s] < 0$ ($\mathrm{Re}[s] > 0$). Lemma 2.1 implies (b). ∎

If the hypotheses of Theorem 3.1 are satisfied for every eigenpair, then each of the eigenpolynomials has a different inertia. An important consequence of this is that the inertia of an eigenpolynomial can be used to determine the relative rank of its corresponding eigenvalue with respect to the other eigenvalues. For example, knowing only that an eigenpolynomial has zeros entirely in $\mathrm{Re}[s] < 0$ when $\lfloor G(\lambda) + G(\lambda) \rceil > 0$ for all $\lambda \in \mathbb{R}$ indicates that the eigenvector must correspond to the largest eigenvalue; thus it maximizes the Rayleigh quotient. Note that both the controllability and semidefiniteness conditions in Theorem 3.1 are satisfied for all of the eigenpairs if $\lfloor G(\lambda) + G(\lambda) \rceil$ is definite on an interval of λ containing the eigenvalues, e.g., $\mathcal{I}(G(\lambda))$. (See the remark after Definition 2.2.) The corollary is now apparent.

COROLLARY 3.2. *Let* $G'(\lambda)R \triangleq (A' - \lambda B')R = 0$, *where* $A', B' \in \mathbb{C}^{n\times n}$ *are Hermitian,* $B' > 0$, *and assume*

$$\lfloor G'(\lambda) + G'(\lambda) \rceil > 0 \ \ (< 0) \ \ \forall\, \lambda \in \mathcal{I}(G'(\lambda)).$$

Also, let $\rho'(R) \triangleq R^* A' R / R^* B' R$ *and assume* $(\rho'(R_*), R_*)$ *is an eigenpair, with* $r_*(0) \neq 0$. *If* $R_*(s)$ *has all of its zeros in* $\mathrm{Re}[s] < 0$ ($\mathrm{Re}[s] > 0$), *then* R_* *is a unique maximizer of* $\rho'(R)$. *If* $R_*(s)$ *has all of its zeros in* $\mathrm{Re}[s] > 0$ ($\mathrm{Re}[s] < 0$), *then* R_* *is a unique minimizer of* $\rho'(R)$.

Even if an eigenproblem fails to satisfy the controllability or definiteness conditions in Theorem 3.1 and Corollary 3.2, *we can still find eigenproblems that are arbitrarily close to the original which do satisfy these conditions*, i.e., we can analytically perturb the problem with an arbitrarily small perturbation matrix in a way which satisfies the definiteness (and in turn the controllability) condition. The ultimate purpose of introducing a perturbation here is to enable us to draw conclusions about the original eigenproblem, which may be difficult to analyze directly, by examining the characteristics of a more easily analyzed perturbed problem as the perturbation approaches zero.

Consider the perturbed matrix

$$G_\epsilon(\lambda) \triangleq G(\lambda) + \Delta(\epsilon),$$

where the perturbation matrix $\Delta(\epsilon)$ is a Hermitian analytic matrix function of the real scalar parameter ϵ and $\lim_{\epsilon\to 0} \Delta(\epsilon) = 0$. If we choose the elements of $\Delta(\epsilon)$ to be

$$\Delta_{ij}(\epsilon) \triangleq \begin{cases} \epsilon & |i-j| = 1, \\ 0 & \text{otherwise}, \end{cases}$$

then

$$\lfloor \Delta(\epsilon) + \Delta(\epsilon) \rceil = \begin{cases} 2\epsilon & i = j, \\ \epsilon & |i-j| = 2, \\ 0 & \text{otherwise}, \end{cases}$$
$$= \mathcal{P}_N^T(R)\mathcal{P}_N(R),$$

where $R = [\ \sqrt{\epsilon}\ \ 0\ \ \sqrt{\epsilon}\ \ 0\ \ \cdots\ \ 0\ ]^T$. Note that $\lfloor \Delta(\epsilon) + \Delta(\epsilon) \rceil = \mathcal{P}_N^T(R)\mathcal{P}_N(R)$ is positive definite when $\epsilon > 0$ and negative definite when $\epsilon < 0$. Thus, if $\lfloor G(\lambda)+G(\lambda) \rceil \geq 0$ (≤ 0) on $\lambda \in \mathcal{I}(G(\lambda))$,

$$\lfloor G_\epsilon(\lambda) + G_\epsilon(\lambda) \rceil = (\lfloor G(\lambda) + G(\lambda) \rceil) + (\lfloor \Delta(\epsilon) + \Delta(\epsilon) \rceil) > 0\ (< 0)$$

for $\lambda \in \mathcal{I}(G(\lambda))$ and $\epsilon > 0$ $(\epsilon < 0)$. Since the eigenvalues of $G_\epsilon(\lambda)P = 0$ are continuously variable with respect to ϵ, and $\mathcal{I}(G(\lambda))$ is open (by assumption), there exists an $\mathcal{I}(G_\epsilon(\lambda)) \subset \mathcal{I}(G(\lambda))$ for $|\epsilon| > 0$ sufficiently small, so

$$\lfloor G_\epsilon(\lambda) + G_\epsilon(\lambda) \rceil > 0\ (< 0) \tag{4}$$

for $\lambda \in \mathcal{I}(G_\epsilon(\lambda))$ and sufficiently small $|\epsilon| > 0$.

It is worth noting that the choice of $\Delta(\epsilon)$ above is by no means the only one possible. The perturbation matrix

$$\Delta(\epsilon) = \epsilon\Delta_1 + \sum_{i=2}^{\infty} \epsilon^i \Delta_i \quad \in \mathbb{C}^{(N+1)\times(N+1)}$$

meets the criteria, for sufficiently small $|\epsilon| > 0$, if $\epsilon\Delta_1$ is any Hermitian matrix satisfying the condition $\lfloor \epsilon\Delta_1 + \epsilon\Delta_1 \rceil > 0$ (< 0) and Δ_i are *arbitrary*, bounded Hermitian matrices for $i \geq 2$. The Δ_i for $i \geq 2$ can be arbitrary because $\epsilon\Delta_1$ dominates the higher-order terms for sufficiently small $|\epsilon| > 0$.

The following theorem proves the existence of eigenvector saddlepoints of the second kind and characterizes some properties of the eigenpolynomial factors.

THEOREM 3.3. *Consider the Hermitian eigenproblem* (2) *in which* $\lfloor G(\lambda) + G(\lambda) \rceil \geq 0$ (≤ 0) *for all* $\lambda \in \mathcal{I}(G(\lambda))$ *and* $B > 0$. *Let* $G_\epsilon(\lambda)P = [G(\lambda) + \Delta(\epsilon)]\,P = 0$ *be an analytically perturbed version of this problem such that for some* ϵ' *and each* $\epsilon \in (0, \epsilon')$,

1. $\lfloor \Delta(\epsilon) + \Delta(\epsilon) \rceil > 0\ (< 0)$;
2. $G_\epsilon(\lambda_{k\epsilon})P_{k\epsilon} = 0$, $\lambda_{k\epsilon}$ *is the* k*th largest eigenvalue, and* $\lim_{\epsilon\to 0} p_{k\epsilon}(0) \neq 0$.

For the eigenpair of $G(\lambda)P = 0$,

$$(\lambda_k, P_k) \stackrel{\triangle}{=} \lim_{\epsilon\to 0} (\lambda_{k\epsilon}, P_{k\epsilon}),$$

there exists a factorization

$$P_k(s) = \tilde{P}_k(s)\hat{P}_k(s),$$

where $\tilde{P}_k \in \mathbb{C}^k$ *and* $\hat{P}_k \in \mathbb{C}^{N-k+2}$, *such that*

(a)

$$\rho(\tilde{P}_k, \hat{P}) \leq \rho(\tilde{P}_k, \hat{P}_k) \leq \rho(\tilde{P}, \hat{P}_k) \quad \forall\ \tilde{P} \in \mathbb{C}^k, \hat{P} \in \mathbb{C}^{N-k+2}, \tag{5}$$

or, equivalently,

$$\max_{\hat{P}\in\mathbb{C}^{N-k+2}} \min_{\tilde{P}\in\mathbb{C}^{k}} \rho(\tilde{P},\hat{P}) = \lambda_k = \min_{\tilde{P}\in\mathbb{C}^{k}} \max_{\hat{P}\in\mathbb{C}^{N-k+2}} \rho(\tilde{P},\hat{P}),$$

where

$$\tilde{P}_k = \arg\min_{\tilde{P}\in\mathbb{C}^{k}} \left\{ \max_{\hat{P}\in\mathbb{C}^{N-k+2}} \rho(\tilde{P},\hat{P}) \right\}$$

and

$$\hat{P}_k = \arg\max_{\hat{P}\in\mathbb{C}^{N-k+2}} \left\{ \min_{\tilde{P}\in\mathbb{C}^{k}} \rho(\tilde{P},\hat{P}) \right\};$$

(b) $\tilde{P}_k(s)$ *has all of its zeros in* $\mathrm{Re}[s] \geq 0$ $(\mathrm{Re}[s] \leq 0)$ *and* $\hat{P}_k(s)$ *has all of its zeros in* $\mathrm{Re}[s] \leq 0$ $(\mathrm{Re}[s] \geq 0)$;

(c) $A,\ B \in \mathbb{R}^{(N+1)\times(N+1)} \Longrightarrow \exists \tilde{P}_k \in \mathbb{R}^k, \hat{P}_k \in \mathbb{R}^{N-k+2}$.

Proof. We can begin the proof of part (a) by using Theorem 3.1 to factor the perturbed eigenpolynomial $P_{k\epsilon}(s)$ as $P_{k\epsilon}(s) = \tilde{P}_{k\epsilon}(s)\hat{P}_{k\epsilon}(s)$, such that $\tilde{P}_{k\epsilon}(s)$ has all $k-1$ of its zeros in $\mathrm{Re}[s] > 0$ $(\mathrm{Re}[s] < 0)$ and $\hat{P}_{k\epsilon}(s)$ has all $N-k+1$ of its zeros in $\mathrm{Re}[s] < 0$ $(\mathrm{Re}[s] > 0)$. (Note that for $k=1$, $\tilde{P}_{k\epsilon}$ is a scalar, and for $k=N+1$, $\hat{P}_{k\epsilon}$ is a scalar.) The assumption $\lim_{\epsilon\to 0} p_{k\epsilon}(0) \neq 0$ implies that $p_{k\epsilon}(0)$, $\tilde{p}_{k\epsilon}(0)$, and $\hat{p}_{k\epsilon}(0)$ are all nonzero on a neighborhood of $\epsilon = 0$.

Now $G_\epsilon(\lambda,\tilde{P}_{k\epsilon})$ and $G_\epsilon(\lambda,\hat{P}_{k\epsilon})$ are defined by Definitions 2.5 and 2.6, but with $A+\Delta(\epsilon)$ replacing A. The corresponding Rayleigh quotients are $\rho_\epsilon(\cdot,\tilde{P}_{k\epsilon})$ and $\rho_\epsilon(\cdot,\hat{P}_{k\epsilon})$, respectively. Assumption 1 implies (4), so that

$$\begin{aligned} \lfloor G_\epsilon(\lambda,\hat{P}_{k\epsilon}) + G_\epsilon(\lambda,\hat{P}_{k\epsilon})\rceil &= \mathcal{P}^*(\hat{P}_{k\epsilon})\lfloor G_\epsilon(\lambda)\mathcal{P}(\hat{P}_{k\epsilon}) + \mathcal{P}^*(\hat{P}_{k\epsilon})G_\epsilon(\lambda)\rceil\mathcal{P}(\hat{P}_{k\epsilon}) \\ &= \mathcal{P}^*(\hat{P}_{k\epsilon})\left(\lfloor G_\epsilon(\lambda) + G_\epsilon(\lambda)\rceil\right)\mathcal{P}(\hat{P}_{k\epsilon}) \\ &> 0 \ \ (<0) \end{aligned} \tag{6}$$

for $k = 2,\ldots,N+1$, $\lambda \in \mathcal{I}(G_\epsilon(\lambda))$, and $\epsilon \in (0,\epsilon'')$, where $|\epsilon''| \leq |\epsilon'|$. Since

$$\min_{P\in\mathbb{C}^{N+1}} \rho_\epsilon(P) \leq \min_{\tilde{P}\in\mathbb{C}^{k}} \rho_\epsilon(\tilde{P},\hat{P}_{k\epsilon}) \leq \max_{\tilde{P}\in\mathbb{C}^{k}} \rho_\epsilon(\tilde{P},\hat{P}_{k\epsilon}) \leq \max_{P\in\mathbb{C}^{N+1}} \rho_\epsilon(P),$$

it follows that the range of eigenvalues for $G_\epsilon(\lambda,\hat{P}_{k\epsilon})\tilde{P} = 0$ are contained within the range of eigenvalues for $G_\epsilon(\lambda)P = 0$. In other words, there exists an $\mathcal{I}(G_\epsilon(\lambda,\hat{P}_{k\epsilon})) \subset \mathcal{I}(G_\epsilon(\lambda))$, so (6) can be rewritten as

$$\lfloor G_\epsilon(\lambda,\hat{P}_{k\epsilon}) + G_\epsilon(\lambda,\hat{P}_{k\epsilon})\rceil > 0 \ \ (<0) \tag{7}$$

for $k = 2,\ldots,N+1$, $\lambda \in \mathcal{I}(G_\epsilon(\lambda,\hat{P}_{k\epsilon}))$, and $\epsilon \in (0,\epsilon'')$. In a similar way, we can show that

$$\lfloor G_\epsilon(\lambda,\tilde{P}_{k\epsilon}) + G_\epsilon(\lambda,\tilde{P}_{k\epsilon})\rceil > 0 \ \ (<0) \tag{8}$$

for $k = 1,\ldots,N$, $\lambda \in \mathcal{I}(G_\epsilon(\lambda,\tilde{P}_{k\epsilon}))$, and $\epsilon \in (0,\epsilon'')$.

Since $(\lambda_{k\epsilon},\tilde{P}_{k\epsilon})$ is an eigenpair satisfying $G_\epsilon(\lambda,\hat{P}_{k\epsilon})\tilde{P} = 0$, $\tilde{P}_{k\epsilon}(s)$ has all of its zeros in $\mathrm{Re}[s] > 0$ $(\mathrm{Re}[s] < 0)$ for $k = 2,\ldots,N+1$, $\tilde{p}_{k\epsilon}(0) \neq 0$, and (7) is true, then by Corollary 3.2, $\tilde{P}_{k\epsilon}$ is a unique minimizer of $\rho_\epsilon(\tilde{P},\hat{P}_{k\epsilon})$, i.e.,

$$\rho_\epsilon(\tilde{P}_{k\epsilon},\hat{P}_{k\epsilon}) \leq \rho_\epsilon(\tilde{P},\hat{P}_{k\epsilon}) \tag{9}$$

for all $\tilde{P} \in \mathbb{C}^k$, $k = 1, \ldots, N+1$ and $\epsilon \in (0, \epsilon'')$. (For $k = 1$, (9) is trivially true since $\tilde{P}$ is a scalar.) Since $(\lambda_{k\epsilon}, \hat{P}_{k\epsilon})$ is an eigenpair satisfying $G_\epsilon(\lambda, \tilde{P}_{k\epsilon})\hat{P} = 0$, $\hat{P}_{k\epsilon}(s)$ has all of its zeros in $\mathrm{Re}[s] < 0$ ($\mathrm{Re}[s] > 0$) for $k = 1, \ldots, N$, $\hat{p}_{k\epsilon}(0) \neq 0$, and (8) is true, then by Corollary 3.2, $\hat{P}_{k\epsilon}$ is a unique maximizer of $\rho_\epsilon(\hat{P}, \tilde{P}_{k\epsilon})$, i.e.,

$$\rho_\epsilon(\hat{P}_{k\epsilon}, \tilde{P}_{k\epsilon}) \geq \rho_\epsilon(\hat{P}, \tilde{P}_{k\epsilon}) \tag{10}$$

for all $\hat{P} \in \mathbb{C}^{N-k+2}$, $k = 1, \ldots, N+1$ and $\epsilon \in (0, \epsilon'')$. (For $k = N+1$, (10) is trivially true since $\hat{P}$ is a scalar.) Reordering the arguments of (10) and combining with (9) yields

$$\rho_\epsilon(\tilde{P}_{k\epsilon}, \hat{P}) \leq \rho_\epsilon(\tilde{P}_{k\epsilon}, \hat{P}_{k\epsilon}) \leq \rho_\epsilon(\tilde{P}, \hat{P}_{k\epsilon}) \tag{11}$$

for all $\tilde{P} \in \mathbb{C}^k$, $\hat{P} \in \mathbb{C}^{N-k+2}$ and $\epsilon \in (0, \epsilon'')$. Since (11) holds for all $\epsilon \in (0, \epsilon'')$, and the limit of each term exists as $\epsilon \to 0$, then (5) is true and (a) follows.

Part (b) is fairly obvious from the definitions of $\tilde{P}_{k\epsilon}$ and $\hat{P}_{k\epsilon}$ and their continuity at $\epsilon = 0$, though it should be noted that at $\epsilon = 0$, the zeros of the eigenpolynomial and its factors may lie on the imaginary axis. Since $\tilde{p}_k(0)$, $\hat{p}_k(0)$, and $p_k(0)$ are nonzero, all of the eigenpolynomial zeros must be finite. Allowing these leading coefficients to be zero complicates matters a bit, so this case is considered separately.

To prove (c), we can assume that $G_\epsilon(\lambda)$ is real, without loss of generality. Thus there is a real eigenpolynomial $P_{k\epsilon}(s)$ whose complex zeros (if there are any) must exist in conjugate pairs. Since the zeros of $\tilde{P}_{k\epsilon}(s)$ and $\hat{P}_{k\epsilon}(s)$ are separated by the imaginary axis, their complex zeros must be in conjugate pairs; thus they are real polynomials. By continuity of the factors at $\epsilon = 0$, (c) follows. □

Without perturbation in Theorem 3.3, many important problems for which the controllability condition of Theorem 3.1 naturally fails to hold would be unjustly excluded since factorization of the unperturbed eigenpolynomial into recognizable maximizing and minimizing factors could be difficult or impossible (e.g., if any eigenpolynomial zeros are on the imaginary axis). An obvious example is when $\lfloor G(\lambda) + G(\lambda) \rceil = 0$. Clearly, perturbation is critical to the scope of the theorem.

Because the eigenproblem is perturbed, however, special care is required in the statement of Theorem 3.3. Note that while every eigenvector (in the original problem) corresponding to a distinct eigenvalue is the limit of an analytically perturbed eigenvector as $\epsilon \to 0$, an arbitrary element of the eigenspace of a repeated eigenvalue may not necessarily be the limit of an analytically perturbed eigenvector [12], so the eigenvectors cannot, in general, be specified a priori. Rather, they are determined by the perturbation matrix, and this is the reason for defining $P_k = \lim_{\epsilon \to 0} P_{k\epsilon}$.

Theorem 3.3 and its analogue, [9, Theorem 6], should not be confused with the well-known Courant–Fischer theorem [8], which also contains a min-max equals max-min result, nor do they follow from it. The Courant–Fischer theorem, in contrast, is applicable to a larger set of eigenproblems, but it is not a saddlepoint theorem in general. Also, Theorem 3.3 and [9, Theorem 6] clearly imply that each saddlepoint eigenvector is, in principle, *independently computable*. In other words, computation of the saddlepoint eigenvector P_k need not rely on the computation of any other eigenvectors or any larger eigenspaces containing P_k.

In both Theorem 3.3 and [9, Theorem 6], it is assumed that $p_k(0) = \lim_{\epsilon \to 0} p_{k\epsilon}(0) \neq 0$. Proving the existence of an eigenvector saddlepoint $(\tilde{P}_k, \hat{P}_k)$ when $p_k(0) = 0$ simply involves choosing an admissible perturbation (for either a type-1 or type-2 saddlepoint problem) such that $p_{k\epsilon}(0) \neq 0$ for $\epsilon \in (0, \epsilon')$. Since a perturbed eigenvector $P_{k\epsilon}$ with $p_{k\epsilon}(0) = 0$ would be an element of the set of vectors

$[0 \ \ (\mathbb{C}^N)^T]^T$ which has *zero measure* in $\mathbb{C}^{N+1}$, then $p_{k\epsilon}(0) \neq 0$ for almost every $P_{k\epsilon}$. Thus, almost every perturbation will ensure that this condition is satisfied. The steps in proving Theorem 3.3(a) (or [9, Theorem 6(a)]) for this case are the same otherwise. An example of an eigenproblem having saddlepoint eigenvectors of the first kind with coefficients $p_k(0) = 0$ for $k \leq N$ is

$$\left(\begin{bmatrix} e_1 & 2e_2 & \cdots & (N+1)e_{N+1} \end{bmatrix} - \lambda I \right) P = 0,$$

where e_i denotes the ith column of the identity matrix. It is easily verified that $\lambda_k = N - k + 2$, $P_k = e_{N-k+2} \in \mathbb{R}^{N+1}$, $\hat{P}_k = e_{N-k+2} \in \mathbb{R}^{N-k+2}$, and $\tilde{P}_k = e_1 \in \mathbb{R}^k$ for $k = 1, \ldots, N+1$.

The issue that remains to be addressed when $p_k(0) = 0$ concerns the definition of the "zeros" of the eigenpolynomial and its factors. In [9], it was implicitly assumed that for $R \in \mathbb{C}^{n+1}$ with $r(0) \neq 0$, $R(z) \stackrel{\triangle}{=} r(0) + \cdots + r(n)z^{-n}$ has exactly n poles at $z = 0$ and n zeros in the complex plane $\mathbb{C}$, i.e., $R(z)$ is an nth order *rational function* [6]. In this paper, the zeros are again understood to mean the zeros of a rational function. To include the possibility of $r(0) = 0$, however, the zero domain of $R(z)$ must be expanded to the *extended complex plane* $\mathbb{C}_e \stackrel{\triangle}{=} \mathbb{C} \cup \{\infty\}$ [6]. Thus, if j is the largest integer such that $r(j) \neq 0$, then the zeros of the rational function $R(z)$ (with n poles at $z = 0$) are the set of (possibly nondistinct) points given by either

$$\{z_i \in \mathbb{C}_e \mid R(z_i) = 0, \ i = 1, \ldots, n\} \quad \text{if} \quad j = n$$

or

$$\{z_i \in \mathbb{C}_e \mid R(z_i) = 0, \ i = 1, \ldots, j\} \cup \{z_i = 0 \mid i = j+1, \ldots, n\} \quad \text{if} \quad 0 < j < n$$

or

$$\{z_i = 0 \mid i = 1, \ldots, n\} \quad \text{if} \quad j = 0.$$

Given the same R, an identical set of zeros can be obtained by the following equivalent definition: Let j be the smallest number such that $r(j) \neq 0$. Then the zeros of the rational function $R(s)$ (with n poles at $s = \infty$) are the set of (possibly nondistinct) points given by either

$$\{s_i \in \mathbb{C} \mid R(s_i) = 0, \ i = 1, \ldots, n\} \quad \text{if} \quad j = 0$$

or

$$\{s_i = \infty \mid i = 1, \ldots, j\} \cup \{s_i \in \mathbb{C} \mid R(s_i) = 0, \ i = j+1, \ldots, n\} \quad \text{if} \quad 0 < j < n$$

or

$$\{s_i = \infty \mid i = 1, \ldots, n\} \quad \text{if} \quad j = n.$$

The assumption of rationality is critical for several reasons. First, it is entirely valid to allow a zero to exist at a pole location (in this case, either $z = 0$ or $s = \infty$), particularly when convenience and consistency necessitate it. Pole-zero cancellation is a well-established concept in linear system theory [2]. Second, the assumption that the number of zeros is the same for all $R \in \mathbb{C}^{n+1}$ is especially important as R varies within $\mathbb{C}^{n+1}$, e.g., from perturbation of the coefficients. For instance, a zero of $R(z)$

which would otherwise "disappear" as $r(n) \to 0$ is instead viewed as moving to the origin. Third, by our definition of the zeros, we ensure that the eigenvalues of the companion matrix C_R are always considered zeros of $R(z)$ or $R(s)$. If, instead, the zeros were to be defined as the points where $R(\cdot) = 0$, then an awkward exception would have to be made for $R(z)$ when C_R has a zero eigenvalue, since $R(z) \neq 0$ at $z = 0$. Finally, and perhaps most importantly, it ensures that for a given eigenvector R, the locations of the zeros of the eigenpolynomial are *independent of whether the eigenpolynomial is realized as* $R(z)$ *or* $R(s)$.[4] Using either definition, we can see, for example, that a polynomial $R(\cdot)$ generated from $R = e_k \in \mathbb{R}^{n+1}$ has $k-1$ zeros at infinity and $n-k+1$ zeros at the origin, regardless of whether we write $R(z) = z^{-(k-1)}$ or $R(s) = s^{n-k+1}$.[5]

Irrespective of whether the zeros are defined in the manner just discussed or by a more restrictive definition, the results in Theorem 3.3(b) and [9, Theorem 6(b)] remain correct for eigenvectors where $p_k(0) \neq 0$. With the extension made for eigenvectors with $p_k(0) = 0$, however, part (b) in each of these theorems must be modified to allow for zeros at infinity. Specifically, $\mathrm{Re}[s] \geq 0$, $\mathrm{Re}[s] \leq 0$, and $|z| \geq 1$ are replaced by $\{\mathrm{Re}[s] \geq 0\} \cup \{\infty\}$, $\{\mathrm{Re}[s] \leq 0\} \cup \{\infty\}$, and $\{|z| \geq 1\} \cup \{\infty\}$, respectively. Additionally, with the general definition of a zero, Theorem 3.3 and [9, Theorem 6] can be further strengthened to say that $\tilde{P}_k(\cdot)$ has exactly $k-1$ zeros and $\hat{P}_k(\cdot)$ has exactly $N-k+1$ zeros.

The next theorem provides one set of sufficient conditions for the zeros of the eigenpolynomial factors to lie on the extended imaginary axis.

THEOREM 3.4. *Let* $(\tilde{P}_k, \hat{P}_k) \in \mathbb{C}^k \times \mathbb{C}^{N-k+2}$ *be an eigenvector saddlepoint for* $G(\lambda)P = 0$ *corresponding to eigenvalue* λ_k *and assume that* $\lfloor G(\lambda) + G(\lambda) \rceil = 0$.

(a) *If* λ_k *is a distinct eigenvalue of* $G(\lambda, \hat{P}_k)\tilde{P} = 0$, *then all* $k-1$ *zeros of* $\tilde{P}_k(s)$ *are on* $\{\mathrm{Re}[s] = 0\} \cup \{\infty\}$.

(b) *If* λ_k *is a distinct eigenvalue of* $G(\lambda, \tilde{P}_k)\hat{P} = 0$, *then all* $N-k+1$ *zeros of* $\hat{P}_k(s)$ *are on* $\{\mathrm{Re}[s] = 0\} \cup \{\infty\}$.

Proof. The proofs to (a) and (b) are similar, so only (a) is proved. Assume that j is the smallest integer such that $\tilde{p}_k(j) \neq 0$. By our definition of the eigenpolynomial zeros, there are j zeros at ∞. If $j = k-1$, all zeros are at ∞. If $j < k-1$, then let $\tilde{P}_{k(j:k-1)} \stackrel{\triangle}{=} [\ \tilde{p}_k(j) \ \cdots \ \tilde{p}_k(k-1)\ ]^T$, and assume $\tilde{P}_{k(j:k-1)}$ is scaled so that $\tilde{p}_{k(j:k-1)}(0) = \tilde{p}_k(j) = 1$. Since the zeros of $\tilde{P}_{k(j:k-1)}(s)$ are the finite zeros of $\tilde{P}_k(s)$, our goal is to show that $\tilde{P}_{k(j:k-1)}(s)$ has zeros only on $\mathrm{Re}[s] = 0$.

From (5), λ_k is the minimum $\rho(\tilde{P}, \hat{P}_k)$, i.e., $\rho(\tilde{P}, \hat{P}_k) \geq \lambda_k$ for all $\tilde{P} \in \mathbb{C}^k$. This implies that

$$\tilde{P}^* G(\lambda_k, \hat{P}_k)\tilde{P} = \tilde{P}^* \left[\mathcal{P}^*(\hat{P}_k) A \mathcal{P}(\hat{P}_k) - \lambda_k \mathcal{P}^*(\hat{P}_k) B \mathcal{P}(\hat{P}_k)\right] \tilde{P} \geq 0 \ \ \forall \tilde{P} \in \mathbb{C}^k.$$

Thus, $G(\lambda_k, \hat{P}_k) \geq 0$. Since $\tilde{P}_k^* G(\lambda_k, \hat{P}_k)\tilde{P}_k = 0$ and λ_k is distinct by assumption, then $X^* G(\lambda_k, \hat{P}_k) X > 0$ for all nontrivial X not in the eigenspace containing $\tilde{P}_k$. An arbitrary $\bar{X} \in \mathbb{C}^k$ of the form

$$\bar{X} = [\ 0 \ \cdots \ 0 \ \ \bar{x}(j+1) \ \cdots \ \bar{x}(k-1)\ ]^T = [\ 0 \ \cdots \ 0 \ \ \bar{X}^T_{(j+1:k-1)}\ ]^T$$

[4] Whether R is realized as $R(z)$ or $R(s)$ is essentially arbitrary. It is traditionally assumed, however, that complex functions of z (or z^{-1}) and complex functions of s have different inertial definitions.

[5] Henrici redefined the meaning of the term "polynomial" to be entirely consistent with our definition of the term "zeros" [6, p. 464].

cannot be in the eigenspace of $\tilde{P}_k$ since $\bar{x}(j) = 0$ and $\tilde{p}_k(j) \neq 0$, so

$$\bar{X}^* G(\lambda_k, \hat{P}_k)\bar{X} = \bar{X}^*_{(j+1:k-1)} \underline{[G(\lambda_k, \hat{P}_k)]}_{k-j-1} \bar{X}_{(j+1:k-1)} > 0$$

for all nontrivial $\bar{X}_{(j+1:k-1)} \in \mathbb{C}^{k-j-1}$. Thus $\underline{[G(\lambda_k, \hat{P}_k)]}_{k-j-1} > 0$.

The equation

$$\begin{aligned} \lfloor G(\lambda_k, \hat{P}_k) + G(\lambda_k, \hat{P}_k) \rceil &= \mathcal{P}^*(\hat{P}_k) \lfloor G(\lambda_k)\mathcal{P}(\hat{P}_k) + \mathcal{P}^*(\hat{P}_k) G(\lambda_k) \rceil \mathcal{P}(\hat{P}_k) \\ &= \mathcal{P}^*(\hat{P}_k) \left(\lfloor G(\lambda_k) + G(\lambda_k) \rceil \right) \mathcal{P}(\hat{P}_k) \\ &= 0 \end{aligned}$$

is equivalent to the condition that the sum of the (i, j)th and $(i+1, j-1)$th elements of $G(\lambda_k, \hat{P}_k) \in \mathbb{C}^{k \times k}$ is zero. The same relationship must hold for the elements in a submatrix of $G(\lambda_k, \hat{P}_k)$, so

$$\left\lfloor \underline{[G(\lambda_k, \hat{P}_k)]}_{k-j} + \underline{[G(\lambda_k, \hat{P}_k)]}_{k-j} \right\rceil = 0. \tag{12}$$

$G(\lambda_k, \hat{P}_k)\tilde{P}_k = 0$ implies $\underline{[G(\lambda_k, \hat{P}_k)]}_{k-j} \tilde{P}_{k(j:k-1)} = 0$; hence Lemma 2.7 and (12) allow us to obtain the Lyapunov equation

$$\underline{[G(\lambda_k, \hat{P}_k)]}_{k-j-1} C_{\tilde{P}_{k(j:k-1)}} + C^*_{\tilde{P}_{k(j:k-1)}} \underline{[G(\lambda_k, \hat{P}_k)]}_{k-j-1} = 0. \tag{13}$$

Now let $(\sigma, Y) \in \mathbb{C} \times \mathbb{C}^{k-j-1}$ be an *arbitrary* eigenpair of $C_{\tilde{P}_{k(j:k-1)}}$. Premultiplication and postmultiplication of (13) by Y^* and Y, respectively, yields

$$Y^* \underline{[G(\lambda_k, \hat{P}_k)]}_{k-j-1} Y\sigma + \sigma^* Y^* \underline{[G(\lambda_k, \hat{P}_k)]}_{k-j-1} Y = 0.$$

Since $\underline{[G(\lambda_k, \hat{P}_k)]}_{k-j-1} > 0$, then $Y^* \underline{[G(\lambda_k, \hat{P}_k)]}_{k-j-1} Y > 0$; hence $\frac{1}{2}(\sigma + \sigma^*) = \mathrm{Re}[\sigma] = 0$. Finally, the proof is completed by noting that σ is a zero of $\tilde{P}_{k(j:k-1)}(s)$ and thus $\tilde{P}_k(s)$. □

A few comments about Theorem 3.4 are in order:

- Part (a) (part (b)) is satisfied when λ_{N+1} (λ_1) is a distinct eigenvalue of $G(\lambda)P = 0$ since $\hat{P}_{N+1}$ ($\tilde{P}_1$) is a scalar.
- Theorem 3.4 is analogous to [9, Theorem 7], which provides sufficient conditions for the zeros of eigenpolynomial factors to exist on the unit circle.[6] The difference is that here we replace $\underline{[G(\lambda)]}_N - \overline{[G(\lambda)]}_N = 0$ with the parallel condition $\lfloor G(\lambda) + G(\lambda) \rceil = 0$.
- The Hermitian matrices $W \in \mathbb{C}^{n \times n}$ which satisfy $\lfloor W + W \rceil = 0$ are of the general pattern

$$W = \begin{bmatrix} w_1 & jw_2 & w_3 & jw_4 & & & \cdots & \\ -jw_2 & -w_3 & -jw_4 & & \cdots & & & \vdots \\ w_3 & jw_4 & & \cdots & & & & \\ -jw_4 & & \cdots & & & & \cdots & \\ & & & & \cdots & & & w_{2n-3} \\ \vdots & & & \cdots & & -w_{2n-3} & jw_{2n-2} & \\ & & \cdots & & w_{2n-3} & -jw_{2n-2} & w_{2n-1} & \end{bmatrix},$$

[6] The unit circle and extended imaginary axis are clearly related since there is a one-to-one mapping between the two sets of points via the bilinear transformation (also known as the linear fractional or Moebius transformation [6]) pair $z = (1+s)/(1-s)$ and $s = (1-z^{-1})/(1+z^{-1})$ [10]. The point $z = -1$ corresponds to the point $s = \infty$.

where $j \triangleq \sqrt{-1}$ and all w_i are real. Matrices having the form of W (*W-matrices*) are to eigenvector saddlepoint problems of the second kind what Hermitian Toeplitz matrices are to eigenvector saddlepoint problems of the first kind. Of course, a Toeplitz matrix $T \in \mathbb{C}^{n\times n}$ satisfies the analogous condition $\underline{[T]}_{n-1} - \overline{[T]}_{n-1} = 0$. Interestingly, the connection between W-matrices and Hermitian Toeplitz matrices extends further. In fact, the covariance matrix of a stationary, continuous random process is a W-matrix, while the covariance matrix of a stationary, discrete random process is a Hermitian Toeplitz matrix. Covariance matrices arise, for example, in the subjects of spectral estimation and linear prediction of continuous or discrete-time autoregressive processes [7, 11].

The remaining theorems relate to both types of saddlepoint eigenproblems. The first of these relaxes an earlier condition that B be positive definite.

THEOREM 3.5. *Consider the eigenproblem* (1)*, in which* $B \geq 0$*. If there exists a Hermitian* $\bar{B}$ *such that* $B+\epsilon\bar{B} > 0$ *and* $AP = \lambda(B+\epsilon\bar{B})P$ *has saddlepoint eigenvectors of the first or second kind for all* $\epsilon \in (0, \epsilon')$*, then the eigenvectors of* (1) *corresponding to finite eigenvalues are also saddlepoint eigenvectors.*

Proof. Let $\rho_\epsilon(P) \triangleq P^*AP/P^*(B+\epsilon\bar{B})P$. By assumption, for each $\epsilon \in (0,\epsilon')$, there is a set of saddlepoints $\{(\tilde{P}_{k\epsilon}, \hat{P}_{k\epsilon}) \mid k = 1, \ldots, N+1\}$ which satisfy (11) for all $\tilde{P} \in \mathbb{C}^k$ and $\hat{P} \in \mathbb{C}^{N-k+2}$. Since the perturbation is analytic,

$$\lim_{\epsilon\to 0}(\tilde{P}_{k\epsilon}, \hat{P}_{k\epsilon}) = (\tilde{P}_k, \hat{P}_k) \in \mathbb{C}^k \times \mathbb{C}^{N-k+2}$$

for $k = 1, \ldots, N+1$, but

$$\lim_{\epsilon\to 0} \rho_\epsilon(\tilde{P}_{k\epsilon}, \hat{P}_{k\epsilon}) = \rho(\tilde{P}_k, \hat{P}_k) \in \mathbb{R}$$

if and only if $(\tilde{P}_k * \hat{P}_k)^* B(\tilde{P}_k * \hat{P}_k) \neq 0$ or, equivalently, $\tilde{P}_k * \hat{P}_k$ is not an eigenvector corresponding to an infinite eigenvalue.

Now for each k such that $\rho(\tilde{P}_k, \hat{P}_k) \in \mathbb{R}$, the continuity of ρ implies there is a neighborhood of $\tilde{P}_k$, denoted by $\tilde{\mathcal{N}}_k$, such that $\rho(\tilde{P}, \hat{P}_k)$ is bounded for all $\tilde{P} \in \tilde{\mathcal{N}}_k$, and a neighborhood of $\hat{P}_k$, denoted by $\hat{\mathcal{N}}_k$, such that $\rho(\tilde{P}_k, \hat{P})$ is bounded for all $\hat{P} \in \hat{\mathcal{N}}_k$. Thus, the limit of (11)

$$\rho(\tilde{P}_k, \hat{P}) \leq \rho(\tilde{P}_k, \hat{P}_k) \leq \rho(\tilde{P}, \hat{P}_k)$$

holds for all $\tilde{P} \in \tilde{\mathcal{N}}_k$ and $\hat{P} \in \hat{\mathcal{N}}_k$, so $(\tilde{P}_k, \hat{P}_k)$ is an eigenvector saddlepoint corresponding to the finite eigenvalue of (1), $\rho(\tilde{P}_k, \hat{P}_k)$. ∎

As an example, when $B \geq 0$ is Hermitian Toeplitz and $\underline{[A]}_N - \overline{[A]}_N \leq 0$ (≥ 0), the eigenvectors in (1) corresponding to finite eigenvalues must be saddlepoint eigenvectors since the eigenvectors of $AP = \lambda(B+\epsilon I)P$ for $\epsilon > 0$ are clearly type-1 saddlepoint eigenvectors under the original criteria. Note that the eigenvector saddlepoints in Theorem 3.5 were demonstrated only to be *local* rather than *global*, i.e., the saddlepoint inequalities were shown to hold over a neighborhood of the saddlepoint, $\tilde{\mathcal{N}}_k \times \hat{\mathcal{N}}_k$, rather than the entire domain $\mathbb{C}^k \times \mathbb{C}^{N-k+2}$. A local saddlepoint is analogous to a local maximum or a local minimium. An example is the saddlepoint $(0,0)$ of the function $f(x,y) = (x^2-2)^2 - (y^2-2)^2$, which satisfies the inequalities $f(x,0) \leq f(0,0) \leq f(0,y)$ for all $x \in [-2,2]$ and $y \in [-2,2]$. Of course, every global saddlepoint is a local one as well.

The next theorem reveals the existence of a general factorization for eigenpolynomials corresponding to eigenvalues of multiplicity m and parameterizes all eigenvector saddlepoint representations.

THEOREM 3.6. *If $G(\lambda)P = 0$ is a type-1 or type-2 saddlepoint eigenproblem with an eigenvalue of multiplicity m, e.g., $\lambda_{k+1} = \cdots = \lambda_{k+m}$, then the m-dimensional eigenspace can be represented as the set*

$$\Xi \stackrel{\triangle}{=} \{P = \tilde{P}_{k+1} * \hat{P}_{k+m} * \bar{P} \mid \bar{P} \in \mathbb{C}^m - \{0\}\},$$

where

$$\tilde{P}_{k+1} \stackrel{\triangle}{=} \arg \min_{\tilde{P} \in \mathbb{C}^{k+1}} \left\{ \max_{\hat{P} \in \mathbb{C}^{N-k+1}} \rho(\tilde{P}, \hat{P}) \right\}$$

and

$$\hat{P}_{k+m} \stackrel{\triangle}{=} \arg \max_{\hat{P} \in \mathbb{C}^{N-k-m+2}} \left\{ \min_{\tilde{P} \in \mathbb{C}^{k+m}} \rho(\tilde{P}, \hat{P}) \right\}.$$

In other words, all of the eigenpolynomials generated from the m-dimensional eigenspace share a common polynomial factor, of degree no greater than $N - m + 1$, given by

$$P_{k+1,k+m}(\cdot) \stackrel{\triangle}{=} \tilde{P}_{k+1}(\cdot)\hat{P}_{k+m}(\cdot).$$

Furthermore, every eigenvector in Ξ has saddlepoint representations of the form

$$(\tilde{P}_{k+1} * \bar{P}_1, \hat{P}_{k+m} * \bar{P}_2) \in \mathbb{C}^{k+j+1} \times \mathbb{C}^{N-k-j+1},$$

where $\bar{P}_1 \in \mathbb{C}^{j+1}$ and $\bar{P}_2 \in \mathbb{C}^{m-j}$, for $j = 0, \ldots, m-1$.

Proof. To prove that the m-dimensional set Ξ is actually the eigenspace, our goal is to show that every vector $P \in \Xi$ has saddlepoint representations (in the form shown) for each of the Cartesian product spaces and that $\rho(P) = \lambda_{k+1} = \cdots = \lambda_{k+m}$. Since λ_{k+1} corresponds to an eigenvector saddlepoint with a minimizing factor $\tilde{P}_{k+1}$,

$$\begin{aligned} \min_{\tilde{P} \in \mathbb{C}^{k+1}} \max_{\hat{P} \in \mathbb{C}^{N-k+1}} \rho(\tilde{P}, \hat{P}) &= \max_{\hat{P} \in \mathbb{C}^{N-k+1}} \rho(\tilde{P}_{k+1}, \hat{P}) \\ &= \lambda_{k+1} \\ &\geq \rho(\tilde{P}_{k+1}, \hat{P}) \ \ \forall \hat{P} \in \mathbb{C}^{N-k+1}. \end{aligned} \tag{14}$$

Using the assumption that λ_{k+m} corresponds to an eigenvector saddlepoint with a maximizing factor $\hat{P}_{k+m}$,

$$\begin{aligned} \max_{\hat{P} \in \mathbb{C}^{N-k-m+2}} \min_{\tilde{P} \in \mathbb{C}^{k+m}} \rho(\tilde{P}, \hat{P}) &= \min_{\tilde{P} \in \mathbb{C}^{k+m}} \rho(\tilde{P}, \hat{P}_{k+m}) \\ &= \lambda_{k+m} \\ &\leq \rho(\tilde{P}, \hat{P}_{k+m}) \ \ \forall \tilde{P} \in \mathbb{C}^{k+m}. \end{aligned} \tag{15}$$

With an *arbitrary* nontrivial $\bar{P} \in \mathbb{C}^m$, let $\hat{P} \stackrel{\triangle}{=} \hat{P}_{k+m} * \bar{P}$ and $\tilde{P} \stackrel{\triangle}{=} \tilde{P}_{k+1} * \bar{P}$. Substitution into (14) and (15) yields

$$\rho(\tilde{P}_{k+1}, \hat{P}_{k+m} * \bar{P}) \leq \lambda_{k+1} = \cdots = \lambda_{k+m} \leq \rho(\tilde{P}_{k+1} * \bar{P}, \hat{P}_{k+m}). \tag{16}$$

Since $\rho(\tilde{P}_{k+1}, \hat{P}_{k+m} * \bar{P}) = \rho(\tilde{P}_{k+1} * \bar{P}, \hat{P}_{k+m})$, (16) is actually an equality, so clearly

$$\rho(P) = \lambda_{k+1} = \cdots = \lambda_{k+m} \quad \forall P \in \Xi. \tag{17}$$

Now let $\bar{P} = \bar{P}_1 * \bar{P}_2$ be *any* factorization of $\bar{P}$, where $\bar{P}_1 \in \mathbb{C}^{j+1}$, $\bar{P}_2 \in \mathbb{C}^{m-j}$, and $j \in \{0, \ldots, m-1\}$. From (14) and (17) we have

$$\rho(\tilde{P}_{k+1}, \hat{P}' * \bar{P}_1) \leq \rho(\tilde{P}_{k+1} * \bar{P}_1, \hat{P}_{k+m} * \bar{P}_2) \quad \forall \hat{P}' \in \mathbb{C}^{N-k-j+1}, \tag{18}$$

and from (15) and (17) we have

$$\rho(\tilde{P}_{k+1} * \bar{P}_1, \hat{P}_{k+m} * \bar{P}_2) \leq \rho(\tilde{P}' * \bar{P}_2, \hat{P}_{k+m}) \quad \forall \tilde{P}' \in \mathbb{C}^{k+j+1}. \tag{19}$$

Combining (18) and (19) and rearranging terms yields

$$\rho(\tilde{P}_{k+1} * \bar{P}_1, \hat{P}') \leq \rho(\tilde{P}_{k+1} * \bar{P}_1, \hat{P}_{k+m} * \bar{P}_2) \leq \rho(\tilde{P}', \hat{P}_{k+m} * \bar{P}_2) \tag{20}$$

for all $\tilde{P}' \in \mathbb{C}^{k+j+1}$ and $\hat{P}' \in \mathbb{C}^{N-k-j+1}$. Since every $P = \tilde{P}_{k+1} * \bar{P}_1 * \hat{P}_{k+m} * \bar{P}_2 \in \Xi$ satisfies the saddlepoint inequality (20), then every $P \in \Xi$ is a stationary point of ρ, and since (17) is also true, then Ξ must be the m-dimensional eigenspace corresponding to eigenvalues $\lambda_{k+1} = \cdots = \lambda_{k+m}$. ∎

There are some other interesting items to note from the last theorem and its proof:

- $\tilde{P}_{k+1}$ and $\hat{P}_{k+m}$ are the lowest dimensional minimizing and maximizing saddlepoint factors, respectively.
- In the case where $m = 1$, $\bar{P}$ is a scalar, and the eigenpolynomial factorization reduces to $P_{k+1}(\cdot) = \tilde{P}_{k+1}(\cdot)\hat{P}_{k+1}(\cdot)$.
- In the case where $m > 1$, a particular eigenvector may have multiple saddlepoint representations in each one of the Cartesian product spaces. Also, since $\bar{P}$ is arbitrary, an eigenpolynomial can have $m-1$ arbitrary zeros.
- Assuming the hypotheses are satisfied, Theorem 3.3 and [9, Theorem 6] only guarantee the existence of a set of eigenvector saddlepoints

$$\mathcal{S} \triangleq \{(\tilde{P}_j, \hat{P}_j) \in \mathbb{C}^j \times \mathbb{C}^{N-j+2} \mid j = k+1, \ldots, k+m\}$$

  for the m-dimensional eigenspace corresponding to eigenvalues $\lambda_{k+1}, \ldots, \lambda_{k+m}$. On the other hand, Theorem 3.6 above proves that the existence of the set $\mathcal{S}$ implies that *all* of the eigenvectors in the eigenspace are saddlepoint eigenvectors, i.e., they all have saddlepoint representations.
- Finally, Theorem 3.6 extends the known results on eigenpolynomial factorization for Hermitian Toeplitz and Hankel-conjugate–transpose-Hankel eigenproblems [3, 4].

As we will demonstrate, Theorem 3.6 can be used to derive estimates of the number of zeros on the extended imaginary axis or the unit circle for each eigenpolynomial factor when $\lfloor G(\lambda) + G(\lambda) \rceil = 0$ or when $\underline{[G(\lambda)]}_N - \overline{[G(\lambda)]}_N = 0$, respectively. The following preliminary theorem is also needed.

THEOREM 3.7. *Let $\{\lambda_i \mid \lambda_1 \geq \cdots \geq \lambda_{N+1}\}$ be the set of eigenvalues for $G(\lambda)P = 0$, and let $\{\mu_i \mid \mu_1 \geq \cdots \geq \mu_j\}$ be the set of eigenvalues for a related eigenproblem $Q^*G(\mu)QR = 0$, where $Q \in \mathbb{C}^{(N+1)\times j}$ is a constant matrix with full column rank. Then*

$$\lambda_i \geq \mu_i \geq \lambda_{i+(N+1-j)} \quad \forall\, i = 1, \ldots, j.$$

We will not give a formal proof to this theorem, which appears in a slightly different form in [8, p. 292], but the essence of the proof is contained in the relationships

$$\mu_i = \max_{\mathcal{Q}_i} \min_{P \in \mathcal{Q}_i} \rho(P) \leq \max_{\mathcal{L}_i} \min_{P \in \mathcal{L}_i} \rho(P) = \lambda_i$$

and

$$\mu_i = \min_{\mathcal{Q}_{j-i+1}} \max_{P \in \mathcal{Q}_{j-i+1}} \rho(P) \geq \min_{\mathcal{L}_{j-i+1}} \max_{P \in \mathcal{L}_{j-i+1}} \rho(P) = \lambda_{i+(N+1-j)},$$

where $\mathcal{L}_i$ denotes an arbitrary i-dimensional subspace and $\mathcal{Q}_i$ denotes an i-dimensional subspace in the range of Q.

THEOREM 3.8. *Assume $\lambda_{k+1} = \cdots = \lambda_{k+m}$ is an eigenvalue of multiplicity m for the saddlepoint eigenproblem $G(\lambda)P = 0$ and that $\tilde{P}_{k+1}$ and $\hat{P}_{k+m}$ are defined as in Theorem* 3.6.

1. *If $\lfloor G(\lambda) + G(\lambda) \rceil = 0$,*
   (a) *$\tilde{P}_{k+1}(s)$ has at least $\max\{0, -N+2k+m-1\}$ zeros and at most k zeros on $\{\mathrm{Re}[s] = 0\} \cup \{\infty\}$. If λ_{k+m} is a minimum eigenvalue, i.e., $k+m = N+1$, then all $N-m+1$ zeros of $\tilde{P}_{k+1}(s)$ are on $\{\mathrm{Re}[s] = 0\} \cup \{\infty\}$.*
   (b) *$\hat{P}_{k+m}(s)$ has at least $\max\{0, N-2k-m+1\}$ zeros and at most $N-k-m+1$ zeros on $\{\mathrm{Re}[s] = 0\} \cup \{\infty\}$. If λ_{k+1} is a maximum eigenvalue, i.e., $k = 0$, then all $N-m+1$ zeros of $\hat{P}_{k+m}(s)$ are on $\{\mathrm{Re}[s] = 0\} \cup \{\infty\}$.*
2. *If $\underline{[G(\lambda)]}_N - \overline{[G(\lambda)]}_N = 0$,*
   (a) *$\tilde{P}_{k+1}(z)$ has at least $\max\{0, -N+2k+m-1\}$ zeros and at most k zeros on $|z| = 1$. If λ_{k+m} is a minimum eigenvalue, i.e., $k+m = N+1$, then all $N-m+1$ zeros of $\tilde{P}_{k+1}(z)$ are on $|z| = 1$.*
   (b) *$\hat{P}_{k+m}(z)$ has at least $\max\{0, N-2k-m+1\}$ zeros and at most $N-k-m+1$ zeros on $|z| = 1$. If λ_{k+1} is a maximum eigenvalue, i.e., $k = 0$, then all $N-m+1$ zeros of $\hat{P}_{k+m}(z)$ are on $|z| = 1$.*

Proof. We begin with the proof of 1(a). From the inequality

$$\lambda_{k+m} = \rho(\tilde{P}_{k+m}, \hat{P}_{k+m}) \leq \rho(\tilde{P}, \hat{P}_{k+m})$$

obtained from (5), we can see that the eigenproblem

$$G(\sigma, \hat{P}_{k+m})\tilde{P} = \mathcal{P}^*_{k+m}(\hat{P}_{k+m})G(\sigma)\mathcal{P}_{k+m}(\hat{P}_{k+m})\tilde{P} = 0 \tag{21}$$

has a minimum eigenvalue $\sigma = \lambda_{k+m}$ of multiplicity, say, $\tilde{m}$. Since (21) must obey Theorem 3.6, the $\tilde{m}$-dimensional eigenspace of the minimum eigenvalue can be given by

$$\tilde{\Xi}' \triangleq \{\tilde{P} = \tilde{P}' * \bar{P}' \in \mathbb{C}^{k+m} \mid \bar{P}' \in \mathbb{C}^{\tilde{m}} - \{0\}\}$$

for some $\tilde{P}' \in \mathbb{C}^{k+m-\tilde{m}+1}$. It should also be apparent from Theorem 3.6 that the m-dimensional set

$$\tilde{\Xi} \triangleq \{\tilde{P} = \tilde{P}_{k+1} * \bar{P} \in \mathbb{C}^{k+m} \mid \bar{P} \in \mathbb{C}^{m} - \{0\}\}$$

satisfies (21) for $\sigma = \lambda_{k+m}$, so it follows that $\tilde{\Xi}' \supseteq \tilde{\Xi}$, $m \leq \tilde{m}$, and $\tilde{P}'$ is a factor of $\tilde{P}_{k+1}$.

The next step is to prove that all of the zeros of $\tilde{P}'(s)$ are on the extended imaginary axis. First, define j to be the lowest integer such that $\tilde{p}'(j) \neq 0$, and let

$$\tilde{P}'_{(j:k+m-\tilde{m})} \triangleq [\ \tilde{p}'(j) \quad \cdots \quad \tilde{p}'(k+m-\tilde{m})\ ]^T \in \mathbb{C}^{k+m-\tilde{m}-j+1}.$$

Then $\tilde{P}'(s)$ has j zeros at $s=\infty$ and $k+m-\tilde{m}-j$ finite zeros which are the zeros of $\tilde{P}'_{(j:k+m-\tilde{m})}(s)$. Next, define $\tilde{\Xi}'_0$ to be the subset of $\tilde{\Xi}'$ containing vectors of the form

$$Y = [\ Y^T_{(0:\tilde{m}+j-2)} \quad Y^T_{(\tilde{m}+j-1:k+m-1)}\ ]^T \in \mathbb{C}^{k+m}$$

such that $Y_{(0:\tilde{m}+j-2)}$ is the zero vector or the empty vector, i.e.,

$$Y_{(0:\tilde{m}+j-2)} = \begin{cases} 0 & \text{if } \tilde{m}+j-2 \geq 0, \\ [\ ] & \text{if } \tilde{m}+j-2 < 0. \end{cases}$$

Since $X_{(0:\tilde{m}+j-2)}$ is zero or empty when $X = \tilde{P}' * e_{\tilde{m}} \in \tilde{\Xi}'$, then

$$X = \alpha_{\tilde{m}} \tilde{P}' * e_{\tilde{m}} \Longrightarrow X \in \tilde{\Xi}'_0 \tag{22}$$

for any $\alpha_{\tilde{m}} \neq 0$. The set $\{\tilde{P}' * e_1, \ldots, \tilde{P}' * e_{\tilde{m}} \mid e_i \in \mathbb{R}^{\tilde{m}}\}$ is a basis for $\tilde{\Xi}'$, so

$$X \in \tilde{\Xi}'_0 \subseteq \tilde{\Xi}' \implies X = \alpha_1 \tilde{P}' * e_1 + \cdots + \alpha_{\tilde{m}} \tilde{P}' * e_{\tilde{m}}$$

for some set of coefficients $\{\alpha_1, \ldots, \alpha_{\tilde{m}}\}$, not all of which are zero. If $\alpha_i \neq 0$ for any $i \neq \tilde{m}$, then $X_{(0:\tilde{m}+j-2)} \neq 0$; therefore

$$X \in \tilde{\Xi}'_0 \Longrightarrow X = \alpha_{\tilde{m}} \tilde{P}' * e_{\tilde{m}} \tag{23}$$

for some nontrivial $\alpha_{\tilde{m}}$. Thus, (22) and (23) imply

$$X \in \tilde{\Xi}'_0 \Longleftrightarrow X = \alpha_{\tilde{m}} \tilde{P}' * e_{\tilde{m}},$$

so $\tilde{\Xi}'_0$ has dimension one. Since $G(\lambda_{k+m}, \hat{P}_{k+m}) \geq 0$ and $\underline{[G(\lambda_{k+m}, \hat{P}_{k+m})]}_{k+m-\tilde{m}-j+1} \geq 0$, $(\lambda_{k+m}, X_{(\tilde{m}+j-1:k+m-1)})$ is an eigenpair of

$$\underline{[G(\sigma, \hat{P}_{k+m})]}_{k+m-\tilde{m}-j+1} R = 0 \tag{24}$$

if and only if $X \in \tilde{\Xi}'_0$. This means the eigenspace of (24) corresponding to eigenvalue $\sigma = \lambda_{k+m}$ is also one-dimensional, and $\sigma = \lambda_{k+m}$ is distinct. Theorem 3.7 implies that the minimum eigenvalue of (24) cannot be less than the minimum eigenvalue of (21); therefore, $\sigma = \lambda_{k+m}$ is the distinct, minimum eigenvalue of (24). By the same argument used to get (12), we obtain

$$\left\lfloor \underline{[G(\sigma, \hat{P}_{k+m})]}_{k+m-\tilde{m}-j+1} + \underline{[G(\sigma, \hat{P}_{k+m})]}_{k+m-\tilde{m}-j+1} \right\rfloor = 0.$$

It now follows from Theorem 3.4(a) that, for the eigenproblem (24), the eigenpolynomial $\tilde{P}'_{(j:k+m-\tilde{m})}(s)$ must have all $k+m-\tilde{m}-j$ of its zeros on $\mathrm{Re}[s]=0$. Thus, $\tilde{P}'(s)$ has all $k+m-\tilde{m}$ zeros on $\{\mathrm{Re}[s]=0\} \cup \{\infty\}$.

We now want to determine usable bounds on the number of zeros of $\tilde{P}'(s)$,

$$k+m-\tilde{m} \stackrel{\triangle}{=} \tilde{n}'. \tag{25}$$

To do that, we need to derive an upper bound on $\tilde{m}$. (The lower bound is known.) Thus, if $\{\sigma_i \mid \sigma_1 \geq \cdots \geq \sigma_{k+m}\}$ are the eigenvalues of (21), we must find the *lowest* index i such that $\sigma_i = \cdots = \sigma_{k+m} = \lambda_{k+1} = \cdots = \lambda_{k+m}$ is possible. If that index is denoted by i_{min}, then

$$m \leq \tilde{m} \leq (k+m) - (i_{min} - 1). \tag{26}$$

From Theorem 3.7, we know that $\sigma_i \geq \lambda_{i+N+1-(k+m)}$, so we can write

$$\sigma_{i_{min}} \geq \lambda_{i_{min}+N-k-m+1} = \lambda_{k+1},$$

which yields $i_{min} + N - k - m + 1 = k + 1$, or

$$i_{min} = 2k - N + m. \tag{27}$$

Therefore, by (26) and (27),

$$m \leq \tilde{m} \leq N - k + 1. \tag{28}$$

By combining (25), (28), and the fact that $\tilde{n}' \geq 0$, we obtain

$$\max\{0, -N + 2k + m - 1\} \leq \tilde{n}' \leq k.$$

Since $\tilde{P}'(s)$ is a factor of $\tilde{P}_{k+1}(s)$, and $\tilde{P}_{k+1}(s)$ has exactly k zeros, the bounds on $\tilde{n}'$ are also the bounds on the number of extended imaginary zeros of $\tilde{P}_{k+1}(s)$. Finally, note that when $k + m = N + 1$, the upper and lower bounds on $\tilde{n}'$ are both equal to $k = N - m + 1$; hence the estimate is exact for that case.

To prove 1(b), we proceed in the same fashion as before. From the inequality

$$\rho(\tilde{P}_{k+1}, \hat{P}) \leq \rho(\tilde{P}_{k+1}, \hat{P}_{k+1}) = \lambda_{k+1}$$

obtained from (5), we can see that the eigenproblem

$$G(\theta, \tilde{P}_{k+1})\hat{P} = \mathcal{P}^*_{N-k+1}(\tilde{P}_{k+1})G(\theta)\mathcal{P}_{N-k+1}(\tilde{P}_{k+1})\hat{P} = 0 \tag{29}$$

has a maximum eigenvalue $\theta = \lambda_{k+1}$ of multiplicity, say, $\hat{m}$. Since (29) must obey Theorem 3.6, the $\hat{m}$-dimensional eigenspace of the maximum eigenvalue can be given by

$$\hat{\Xi}' \triangleq \{\hat{P} = \hat{P}' * \bar{P}' \in \mathbb{C}^{N-k+1} \mid \bar{P}' \in \mathbb{C}^{\hat{m}} - \{0\}\}$$

for some $\hat{P}' \in \mathbb{C}^{N-k-\hat{m}+2}$. It should also be apparent from Theorem 3.6 that the m-dimensional set

$$\hat{\Xi} \triangleq \{\hat{P} = \hat{P}_{k+m} * \bar{P} \in \mathbb{C}^{N-k+1} \mid \bar{P} \in \mathbb{C}^{m} - \{0\}\}$$

satisfies (29) for $\theta = \lambda_{k+1}$, so it follows that $\hat{\Xi}' \supseteq \hat{\Xi}$, $m \leq \hat{m}$, and $\hat{P}'$ is a factor of $\hat{P}_{k+m}$.

The procedure needed to prove that $\hat{P}'(s)$ has all of its zeros on the extended imaginary axis is much like the one used for $\tilde{P}'(s)$, so we will continue with the step which bounds the number of zeros of $\hat{P}'(s)$,

$$N - k - \hat{m} + 1 \triangleq \hat{n}'. \tag{30}$$

To do that, we need to derive an upper bound on $\hat{m}$. (The lower bound is known.) Thus, if $\{\theta_i \mid \theta_1 \geq \cdots \geq \theta_{N-k+1}\}$ are the eigenvalues of (29), we must find the *highest* index i such that $\theta_1 = \cdots = \theta_i = \lambda_{k+1} = \cdots = \lambda_{k+m}$ is possible. If that index is denoted by i_{max}, then

$$m \leq \hat{m} \leq i_{max}. \tag{31}$$

From Theorem 3.7, we know that $\lambda_i \geq \theta_i$, so we can write

$$\lambda_{k+m} = \lambda_{i_{max}} \geq \theta_{i_{max}},$$

which yields

$$i_{max} = k + m. \tag{32}$$

Therefore, by (31) and (32),

$$m \leq \hat{m} \leq k + m. \tag{33}$$

By combining (30), (33), and the fact that $\hat{n}' \geq 0$, we obtain

$$\max\{0, N - 2k - m + 1\} \leq \hat{n}' \leq N - k - m + 1.$$

Since $\hat{P}'(s)$ is a factor of $\hat{P}_{k+m}(s)$, and $\hat{P}_{k+m}(s)$ has exactly $N - k - m + 1$ zeros, the bounds on $\hat{n}'$ are also the bounds on the number of extended imaginary zeros of $\hat{P}_{k+m}(s)$. Finally, note that when $k = 0$, the upper and lower bounds on $\hat{n}'$ are both equal to $N - m + 1$; hence the estimate is exact for that case.

The proof of 2(a) is similar to the proof of 1(a) except for one step. Instead, we need to show that all of the zeros of $\tilde{P}'(z)$ are on the unit circle. Because $G(\lambda_{k+m}, \hat{P}_{k+m})$ is a positive semidefinite Toeplitz matrix, it contains $\tilde{m} + 1$ overlapping, positive semidefinite Toeplitz submatrices identical to $\underline{[G(\lambda_{k+m}, \hat{P}_{k+m})]}_{k+m-\tilde{m}}$ embedded along the diagonal. If we assume $\underline{[G(\lambda_{k+m}, \hat{P}_{k+m})]}_{k+m-\tilde{m}}$ is singular and that

$$\bar{V}^* \underline{[G(\lambda_{k+m}, \hat{P}_{k+m})]}_{k+m-\tilde{m}} \bar{V} = 0,$$

then the set of vectors $\{\bar{V} * e_1, \ldots, \bar{V} * e_{\tilde{m}+1} \mid e_i \in \mathbb{R}^{\tilde{m}+1}\}$ generates an $(\tilde{m} + 1)$-dimensional null space of $G(\lambda_{k+m}, \hat{P}_{k+m})$. Observe, however, that $G(\lambda_{k+m}, \hat{P}_{k+m})$ cannot have a null space of dimension $\tilde{m}+1$, since it would contradict the fact that $\sigma = \lambda_{k+m}$ is an eigenvalue of (21) with multiplicity $\tilde{m}$. Therefore, $\underline{[G(\lambda_{k+m}, \hat{P}_{k+m})]}_{k+m-\tilde{m}}$ is nonsingular, which implies that

$$k + m - \tilde{m} \leq \operatorname{rank}\{\underline{[G(\lambda_{k+m}, \hat{P}_{k+m})]}_{k+m-\tilde{m}+1}\} \leq k + m - \tilde{m} + 1. \tag{34}$$

Also, since $(\lambda_{k+m}, \tilde{P}' * e_{\tilde{m}})$ is an eigenpair of (21), then

$$\underline{[G(\lambda_{k+m}, \hat{P}_{k+m})]}_{k+m-\tilde{m}+1} \tilde{P}' = 0. \tag{35}$$

It follows from (34) and (35) that

$$\operatorname{rank}\{\underline{[G(\lambda_{k+m}, \hat{P}_{k+m})]}_{k+m-\tilde{m}+1}\} = k + m - \tilde{m}.$$

Thus, $\sigma = \lambda_{k+m}$ is the distinct, minimum eigenvalue of

$$\underline{[G(\sigma, \hat{P}_{k+m})]}_{k+m-\tilde{m}+1} R = 0 \tag{36}$$

and $\tilde{P}'$ is the corresponding eigenvector. Since $G(\sigma, \hat{P}_{k+m})$ and all of its submatrices along the diagonal are Toeplitz,

$$\underline{\left[\underline{[G(\sigma, \hat{P}_{k+m})]}_{k+m-\tilde{m}+1}\right]}_{k+m-\tilde{m}} - \overline{\left[\underline{[G(\sigma, \hat{P}_{k+m})]}_{k+m-\tilde{m}+1}\right]}_{k+m-\tilde{m}} = 0.$$

It now follows from [9, Theorem 7(a)] that, for the eigenproblem (36), the eigenpolynomial $\tilde{P}'(z)$ must have all $k+m-\tilde{m}$ of it zeros on $|z|=1$. The remainder of the proof of 2(a) follows that of 1(a). The proof of 2(b) follows those of 1(b) and 2(a). □

For the product $\tilde{P}_{k+1}(\cdot)\hat{P}_{k+m}(\cdot)$, the lower bound on the number of zeros located on the unit circle or extended imaginary axis is $|N-2k-m+1|$, and the upper bound is $N-m+1$. This lower bound is consistent with the one given in [4] for eigenpolynomials of Toeplitz eigenproblems.

4. Summary. Using an inertia theorem for Lyapunov equations, among other techniques, a new class of eigenproblems was shown to have saddlepoint eigenvectors. Sufficient conditions were given for the eigenpolynomial factors to have zeros in $\mathrm{Re}[s]>0\,(\geq 0)$, $\mathrm{Re}[s]<0\,(\leq 0)$, and on $\{\mathrm{Re}[s]=0\}\cup\{\infty\}$. An important theorem that gives a general factorization for an m-dimensional saddlepoint eigenspace was also proved. In the final result, estimates of the number of zeros located on the extended imaginary axis or unit circle were derived for the eigenpolynomial factors in W-matrix and Hermitian Toeplitz eigenproblems.

This paper complements the earlier work in [9], and it provides new perspective on the material there. The usefulness of classical inertia theory in the study of certain types of eigenproblems has been exhibited in both of the discussions as well.

REFERENCES

[1] C. T. Chen, *A generalization of the inertia theorem*, SIAM J. Appl. Math., 25 (1973), pp. 158–161.
[2] C. T. Chen, *Linear System Theory and Design*, Holt, Rinehart and Winston, New York, 1984.
[3] D. Clark, *On the spectra of bounded, Hermitian, Hankel matrices*, Amer. J. Math., 90 (1968), pp. 627–656.
[4] Y. Genin, *A survey of the eigenstructure properties of finite Hermitian Toeplitz matrices*, Integral Equations Operator Theory, 10 (1987), pp. 621–639.
[5] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.
[6] P. Henrici, *Applied and Computational Complex Analysis*, Vol. 1, John Wiley, New York, 1974.
[7] S. M. Kay, *Modern Spectral Estimation: Theory and Application*, Prentice–Hall, Englewood Cliffs, NJ, 1988.
[8] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, Academic Press, New York, 1985.
[9] M. A. Mendlovitz, *Eigenvector saddlepoints and zero properties of the eigenpolynomial factors*, Linear Algebra Appl., 259 (1997), pp. 281–298.
[10] A. V. Oppenheim and R. W. Schaefer, *Discrete-Time Signal Processing*, Prentice–Hall, Englewood Cliffs, NJ, 1989.
[11] D. T. Pham and A. Le Breton, *Levinson-Durbin-type algorithms for continuous-time autoregressive models and applications*, Math. Control Signals Systems, 4 (1991), pp. 69–79.
[12] F. Rellich, *Perturbation Theory of Eigenvalue Problems*, Gordon and Breach, New York, 1969.
[13] H. Wimmer, *Inertia theorems for matrices, controllability, and linear vibrations*, Linear Algebra Appl., 8 (1974), pp. 337–343.
[14] W. I. Zangwill, *Nonlinear Programming: A Unified Approach*, Prentice–Hall, Englewood Cliffs, NJ, 1969.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 613–628
© 1999 Society for Industrial and Applied Mathematics

ITERATIVE REGULARIZATION AND MINRES*

MISHA KILMER† AND G. W. STEWART‡

Abstract. In this paper we present three theorems which give insight into the regularizing properties of MINRES. While our theory does not completely characterize the regularizing behavior of the algorithm, it provides a partial explanation of the observed behavior of the method. Unlike traditional attempts to explain the regularizing properties of Krylov subspace methods, our approach focuses on convergence properties of the residual rather than on convergence analysis of the harmonic Ritz values. The import of our analysis is illustrated by two examples. In particular, our theoretical and numerical results support the following important observation: in some circumstances the dimension of the optimal Krylov subspace can be much smaller than the number of the components of the truncated spectral solution that must be computed to attain comparable accuracy.

Key words. regularization, ill-posed problem, MINRES, iterative regularization

AMS subject classification. 65F10

PII. S0895479898348623

1. Introduction.

The discretization of a compact self-adjoint operator equation with errors in the right-hand side results in a linear system whose exact solution bears no relation to the solution of the original error-free system. Specifically, in the spectral domain — that is, in the coordinate system of the eigenvectors of the matrix — the exact linear system can be written in the form

$$\Lambda x = b,$$

where

$$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n), \qquad 1 = \lambda_1 > \cdots > \lambda_n > 0.$$

Note that we assume that the system is positive definite and has been scaled so that $\lambda_1 = 1$. To the extent that the quantities in this equation track the corresponding quantities in the original operator, they have the following properties:

(1) If the operator is smooth, the eigenvalues λ_i of the discrete operator will eventually decay rapidly to zero.
(2) If the solution of the original problem is square integrable, then the components ξ_i of x decay to zero.
(3) Since the components β_i of the right-hand side satisfy $\beta_i = \xi_i \lambda_i$, they decay faster than either the components of the solution or the eigenvalues of the operator.

The system with error is

$$\Lambda \tilde{x} = b + e \equiv \tilde{b}.$$

*Received by the editors December 2, 1998; accepted for publication (in revised form) by M. Hanke April 26, 1999; published electronically December 28, 1999.
http://www.siam.org/journals/simax/21-2/34862.html

†Mathematics Department, Tufts University, Medford, MA 02155 (mkilme01@tufts.edu). This research was supported in part by the Army Research Office under grant DAAG55-97-1-0013.

‡Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 (stewart@cs.umd.edu). This research was supported by the National Science Foundation under grant CCR-95-03126.

If we model the error as white noise, then we can append the following condition to the above list.

(4) The components ϵ_i of e are random variables with mean zero and standard deviation ϵ.

The components of the solution for the perturbed right-hand side are

$$\tilde{\xi}_i = \frac{\beta_i + \epsilon_i}{\lambda_i}. \tag{1.1}$$

Since the β_i decay rapidly, they soon fall below the error level. The subsequent components of the solution are effectively random variables with standard deviation ϵ/λ_i. The division by λ_i magnifies the originally small error so that it dominates the solution.

As long as the β_i are larger than the errors ϵ_i, the $\tilde{\xi}_i$ contain useful information. The process of extracting this information is called *regularization.* There are many regularization schemes. A natural one is to stop computing components of x when the β_i get near the error level. We will call this procedure *truncated spectral regularization.*[1]

In this paper we will be concerned with regularization based on the iterative method MINRES [9]. The method can be described briefly as follows. Let

$$K_k = (\tilde{b}\ \Lambda\tilde{b}\ \cdots\ \Lambda^{k-1}\tilde{b})$$

be the kth *Krylov matrix*. Then we seek an approximation to x in the form

$$x_k = K_k a_k,$$

where a_k is determined so that

$$\rho_k^2 \equiv \|\tilde{b} - \Lambda x_k\|^2 = \min.$$

Here $\|\cdot\|$ denotes the usual Euclidean norm.

The residual $r_k = \tilde{b} - \Lambda x_k$ has an alternate expression that is used in the analysis of the algorithm. Given a vector $a_k = (\alpha_1, \ldots, \alpha_k)^{\mathrm{T}}$, let a polynomial p be defined by

$$p(t) = 1 - \alpha_1 t - \alpha_2 t^2 - \cdots - \alpha_k t^k.$$

Then the residual $r_k = \tilde{b} - \Lambda K_k a_k$ can be written in the form $r_k = p_k(\Lambda)\tilde{b}$, where p_k is the polynomial associated with a_k. Since MINRES minimizes $\|r_k\|$, we have

$$\rho_k \equiv \|r_k\| = \min_{\substack{\deg(p_k)=k \\ p_k(0)=1}} \|p(\Lambda)\tilde{b}\|. \tag{1.2}$$

Because the ith component of the residual is the value of a polynomial at λ_i, we will often call it the value of the residual at λ_i.

It has been widely observed that up to a certain index k_0 the MINRES solutions x_k are increasingly accurate approximates to the exact solution x, after which their accuracy rapidly deteriorates. Thus if we can determine k_0 (a difficult problem), MINRES can be used as a regularization method. What seems less well known is that

[1] In the nonsymmetric case this process is also known as truncated SVD regularization (see [7] and the references therein).

in some circumstances the critical index k_0 can be quite small — much smaller than the number of components of the truncated spectral solution that must be computed to attain the same accuracy.[2]

In this paper we give three theorems that help explain the regularizing properties of MINRES. The phrase "help explain" is a deliberate warning to the reader not to expect a complete analysis. The theorems do not say that MINRES has to behave in a particular way. Nonetheless, they give considerable insight into the method.

There is a vast body of literature on regularization, much of which is devoted to determining the asymptotic behavior of the error in the optimally regularized solution as the error in the right-hand side approaches zero. (This literature has been admirably surveyed by Engl, Hanke, and Neubauer [4].) There is also a body of literature for well-posed problems in which the asymptotic convergence of Krylov methods like MINRES are analyzed (see, for example, the recent book by Greenbaum [5]). Our approach, which consists of leaving the error in b fixed and determining how the solution behaves as the regularization parameter varies, is essentially nonasymptotic and fits in neither of these categories. On the other hand, our work is closely related to analyses that attempt to find the optimal value of the regularization parameter. (For a survey of this literature see Hansen [8].)

In the next section we give a toy example that will be used to illustrate the subsequent results. We will also describe in a general way what is happening. The next three sections are devoted to establishing our three results. In section 6 we apply our results to another example. The paper concludes with a discussion of the results and suggestions for future work.

A little informal terminology will help in our discussions. We will call the part of the spectrum for which the $\tilde{\beta}_i$ are little affected by the error the *initial part* of the spectrum. We will call the part where they are fully contaminated the *terminal part* of the spectrum. The part of the spectrum lying between will be called the *intermediate part* or the *transitional part.*

2. An example. In this section we will introduce a simple-minded example from image processing. Let $\underline{x}(t)$ represent an "image" on $[0, 1]$, and consider the Gaussian blurring operator

$$\underline{b}(t) = \int_0^1 \underline{x}(s) e^{-\left(\frac{s-t}{\sigma}\right)^2} \, ds.$$

We will discretize this operator by choosing an integer n and setting

$$t_i = \frac{i-1}{n}, \qquad i = 1, \ldots, n+1,$$

and generating a matrix K whose elements are

$$\kappa_{ij} = e^{-\left(\frac{t_j - t_i}{\sigma}\right)^2}.$$

The resulting matrix is then scaled so that its dominant eigenvalue is one. Although this problem is highly simplified, its regularization properties are typical of more realistic problems.

In the following experiments, we took $n = 40$ and $\epsilon = 0.1$. Our function was $\underline{x}(t) = 1 + t + t^2$, and it was discretized by evaluating it at the points t_i to give the components x_i of x. The vector b was generated in the form Ax. Figure 2.1 is a plot

[2] Martin Hanke has also observed this phenomenon [6].

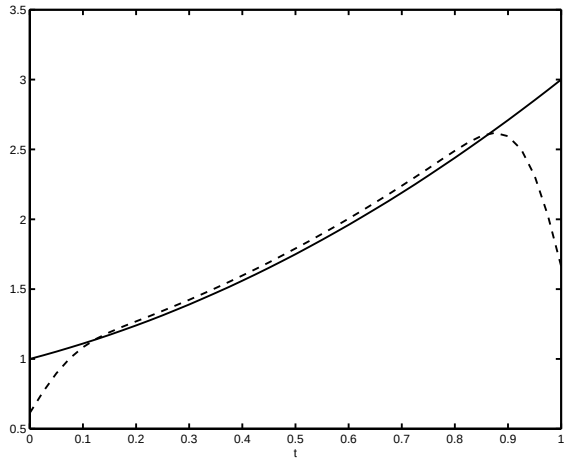


FIG. 2.1. *$x(t)$ (solid) and $b(t)$ (dashed).*

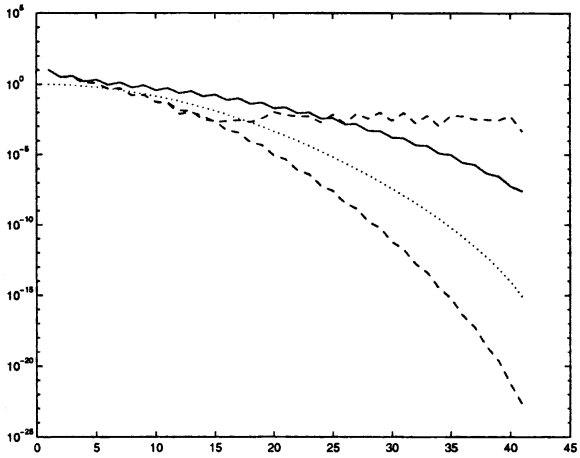


FIG. 2.2. *ξ_i (solid), λ_i (dotted), and β_i, $\tilde{\beta}_i$ (dashed).*

of $\underline{x}(t)$ and $\underline{b}(t)$. The divergence at the ends of the intervals is due to the fact that at those points the part of the blurring distribution that lies outside $[0, 1]$ becomes significant.

Figure 2.2 shows the vectors b, $\tilde{b}$, and x after they have been transformed to the spectral coordinates. Note that for $i > 15$ values $\tilde{\beta}_i$ are significantly contaminated with error. The plot also shows the eigenvalues of A, which decrease slowly at first and then rapidly plunge to about 10^{-15}. This rapid decrease is typical of smooth operators.

Truncated spectral regularization and MINRES regularization behave quite differently for this problem. Figure 2.3 plots the norms of the errors at the kth step of each method. The error for truncated spectral regularization decreases slowly until $k = 15$, after which it increases rapidly. The error in the MINRES iterates decreases rapidly until $k = 5$ and then rises. Truncated spectral regularization produces a bet-

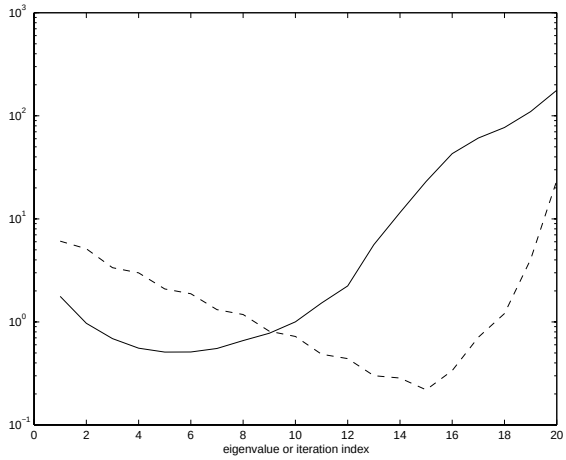


FIG. 2.3. *Errors from spectral regularization (dashed) and MINRES regularization (solid).*

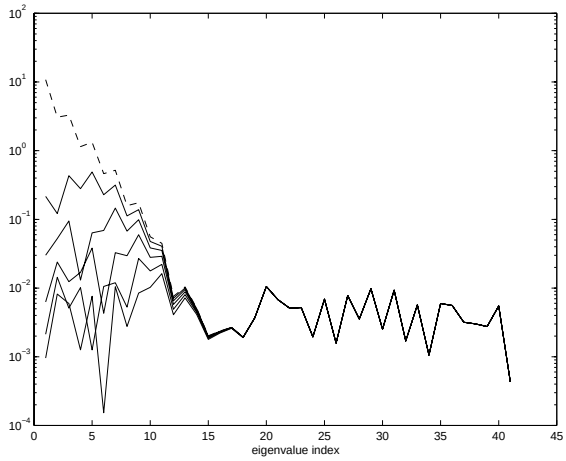


FIG. 2.4. *First five MINRES residuals and $\tilde{b}$ (dashed line).*

ter solution with about half the error but arrives at it more slowly.[3] In fact, not until $k = 11$ does it produce a solution as accurate as the MINRES solution for $k = 5$.

It is informative to examine the behavior of the MINRES residuals. Figure 2.4 displays the magnitudes of the residuals at the eigenvalues λ_i as a function of i for the first five MINRES approximations along with the magnitudes of the components of $\tilde{b}$. It is seen that they decrease in the initial part of the spectrum. At other points in the spectrum, they change very little. The plot in Figure 2.5 shows the same thing in terms of the residual polynomials that satisfy (1.2).[4] The polynomials are small in

[3]The relative quality of the solutions depends on the throw of the dice in constructing the random vector e. For ten other simulations the ratio of the minimum error for truncated spectral regularization to that for MINRES was the following: 0.80, 0.57, 0.76, 0.41, 1.14, 0.65, 0.67, 0.67, 0.82, 0.68. These numbers show the solutions to be usually comparable, with the MINRES solution actually better in a single case.

[4]The plots are not strictly the polynomials. The values between i and $i + 1$ were obtained by evaluating the polynomial at ten equally spaced points between λ_i and λ_{i+1}.

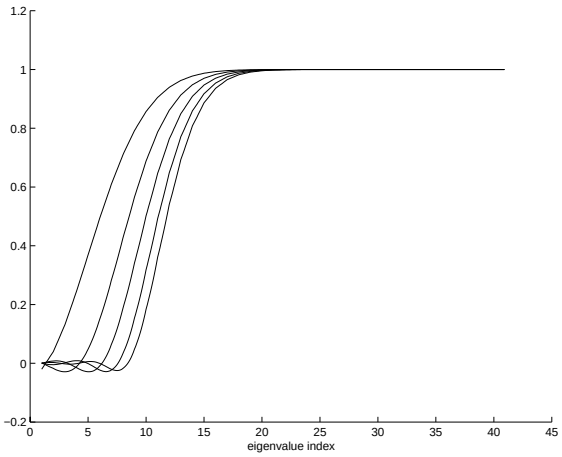


FIG. 2.5. *First five MINRES residual polynomials.*

the initial part of the spectrum and then rise swiftly to a value of one.

Stepping back from this example, we see that three factors are operating to give the fast regularization.

(1) From (1.1) it follows that the residual cannot be reduced indefinitely without harming the solution.
(2) The plot in Figure 2.4 shows that the residual decreases rapidly to a point where further reduction is harmful.
(3) Notice in Figure 2.4 that the values of the first five residuals corresponding to terminal eigenvalues λ_i coincide with $\tilde{\beta}_i$. Equation (1.1) implies that if we reduce the residual at λ_i from $\tilde{\beta}_i$ to $\tilde{\beta}_i(1-\alpha)$, where $0 < \alpha < 1$, then the corresponding component of the approximate solution is $\alpha\tilde{\beta}_i/\lambda_i$. Now in the terminal part of the spectrum $\tilde{\beta}_i \cong \epsilon_i$, and hence the component of the approximate solution is essentially $\alpha\epsilon_i/\lambda_i$. Because the terminal eigenvalues are extremely small compared to the error, even a very small reduction of the residual at a terminal eigenvalue must result in a large component in the approximate solution. Thus if MINRES is to produce an acceptably regularized solution, it must not reduce the residuals at the terminal part of the spectrum while it is still reducing the residuals at the initial part. As it turns out (see Figure 2.5), the first 5 residual polynomials for our running example are very near one on the terminal spectrum, so there is virtually no reduction of the residual there.

In the next three sections we will establish results that give quantitative substance to these observations.

3. The residual and the error. The first of our results concerns the relation between the size of the error in a purported solution and the size of the corresponding residual. We have already noted that if the residual is required to be too small, the error must be large. However, we can be more precise.

Specifically, let y be a purported solution with residual norm δ. There are many such solutions, and among them one must give minimum error; i.e., there must be a

vector y that solves the following problem. Given $\delta > 0$,

$$\begin{aligned} &\text{minimize } \|x - y\|^2 \\ &\text{subject to } \|\tilde{\beta} - \Lambda y\|^2 = \delta^2. \end{aligned} \tag{3.1}$$

The following theorem shows when the solution to this problem has a large error. Below we have assumed $\lambda_1 = 1$.

THEOREM 3.1. *If $\tau > -1$ is the solution of the equation*

$$\sum_i \frac{\epsilon_i^2}{(1+\tau\lambda_i^2)^2} = \delta^2, \tag{3.2}$$

then

$$\eta_i = \xi_i + \frac{\tau\lambda_i\epsilon_i}{1+\tau\lambda_i^2} \tag{3.3}$$

is a solution of (3.1). *In this case*

$$\|x - y\|^2 = \sum_i \left(\frac{\tau\lambda_i}{1+\tau\lambda_i^2}\right)^2 \epsilon_i^2. \tag{3.4}$$

Proof. Consider the Lagrangian

$$\sum_i (\xi_i - \eta_i)^2 + \tau\left[\sum_i (\tilde{\beta}_i - \lambda_i\eta_i)^2 - \delta^2\right],$$

where τ is the Lagrange multiplier. Differentiating with respect to η_i and setting the results to zero, we get

$$\xi_i - \eta_i + \tau\lambda_i(\tilde{\beta}_i - \lambda_i\eta_i) = 0.$$

Equation (3.3) follows on solving this equation for η_i and using the fact that $\tilde{\beta}_i = \lambda_i\xi_i + \epsilon_i$. Equation (3.2) defines the value of τ for which the solution satisfies the constraint. Finally, (3.4) follows by direct computation. ∎

The best way to understand this theorem is to consider the solution y_τ generated as τ varies from -1 to ∞. From (3.2) we see that the residual norm δ decreases monotonically from ∞ to 0. The error norm $\|x - y_\tau\|$ decreases from ∞ to 0 at $\tau = 0$, where $\delta = \|e\|$. Thereafter, small reductions in the residual cause large increases in the error. To see this, consider the term

$$\frac{\epsilon_i^2}{(1+\tau\lambda_i^2)^2}$$

in (3.2). For this term to be reduced by a factor of four from its value ϵ_i^2 at $\tau = 0$, a very modest reduction in the total, we must have $\tau\lambda_i^2 = 1$. The corresponding term in the square of the error norm (3.4) is

$$\left(\frac{\tau\lambda_i\epsilon_i}{1+\tau\lambda_i^2}\right)^2 = \frac{\epsilon_i^2}{4\lambda_i^2},$$

which for small λ_i is large. It is worth noting that these observations depend only on Λ and e and not on x or b.

The above comments suggest that an attempt to reduce the residual norm below $\|e\|$ will increase the error in the solution. This is certainly true of our example, for which $\|e\| = 0.033$. The error turns around at $k = 5$. The residual norms are $\rho_4 = 0.055$ $\rho_5 = 0.039$, and $\rho_6 = 0.031$. Thereafter the error increases sharply, as it must (see Figure 2.3).

It is worth noting that Theorem 3.1 supports the discrepancy principle of Morozov, which says that a regularization parameter should be chosen to make the norm of the residual approximately the size of the norm of the error (see [4, section 4.3]). However, the rapid increase of the error when the residual becomes too small suggests that methods based on the discrepancy principle will be extremely sensitive to variations in the estimate of the error norm (see the experiments in section 7.7.1 of [8]).

4. Reduction of the residual. We now turn to the reduction of the residual. The usual analysis of MINRES is based on (1.2). Specifically, given any polynomial p_k of degree k satisfying $p_k(0) = 1$, we have

$$\|r_k\|^2 \leq \sum_{i=1}^{n} p_k(\lambda_i)^2 \tilde{\beta}_i^2. \tag{4.1}$$

One convergence result amounts to choosing a sequence of polynomials p_k such that

$$\lambda \in [\lambda_n, \lambda_1] \implies p_k(\lambda) \to 0.$$

The rate at which the p_k approach zero on $[\lambda_n, \lambda_1]$ is an upper bound on the convergence rate of MINRES.

The key to analyzing the regularizing properties of MINRES is to note that we are not interested in the convergence of the method on the terminal part of the spectrum. In fact (1.1) shows that MINRES *must not* converge there until after a reasonable regularized solution has been formed. We can therefore restrict our attention to an interval, say, $[\lambda_m, \lambda_1]$ in which the $\tilde{\beta}_i$ are relatively error free. Since the interval is farther removed from the origin, the convergence will be faster.[5]

The polynomial $p_{k,m}$ for this analysis is constructed in the usual way from the Chebychev polynomials $c_k(\tau)$. Specifically, let

$$s_m(t) = -\frac{2t - \lambda_1 - \lambda_m}{\lambda_1 - \lambda_m} \tag{4.2}$$

and set

$$p_{k,m}(t) = \frac{c_k[s_m(t)]}{c_k[s_m(0)]}. \tag{4.3}$$

By bounding the values of these polynomials we obtain the following theorem.

THEOREM 4.1. *Let $m > 1$ and set*

$$\kappa = \frac{\lambda_1}{\lambda_m}$$

[5] For the standard analysis see [5, Chapter 3]. In fact, one reviewer has pointed out that Björck and Eldén [2] used polynomials of this kind to construct regularizing iterative methods. Here, on the other hand, we use them to analyze the regularizing properties of a particular iterative method.

and

$$\tilde{b}_m = (\tilde{\beta}_1, \ \ldots, \ \tilde{\beta}_m)^{\mathrm{T}}.$$

Then

$$\begin{aligned} \|r_k\|^2 &\le \|p_{k,m}(\Lambda)\tilde{b}\|^2 \\ &\le 4\left(\tfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k} \|\tilde{b}_m\|^2 \\ &\quad + \textstyle\sum_{i=m+1}^{n} \left(1 - 2\frac{\lambda_i}{(\sqrt{\lambda_1}+\sqrt{\lambda_i})^2}\right)^{2k} \tilde{\beta}_i^2. \end{aligned} \tag{4.4}$$

The proof of this theorem is highly technical and will be found in the appendix. The first term in (4.4) is of the same form as the bound from the usual analysis of MINRES mentioned above. However, for that bound κ is λ_1/λ_n, which is larger than λ_1/λ_m and implies slower convergence. The first factor in the remaining terms is essentially one for small eigenvalues. However, for eigenvalues near λ_m it has a reducing effect on the corresponding part of the residual.

The bound is too crude to use in applications, but it provides insight into the way MINRES converges. In our running example, when m is small, κ is near one, and the polynomials $p_{k,m}$ rapidly reduce the residual on $\lambda_1, \ldots, \lambda_m$. But this is precisely the place where there is a large residual to reduce. Since the reduction by $p_{k,m}$ bounds the MINRES reduction, we conclude that initially MINRES will rapidly reduce the residual.

But the bound implies more. As k increases, for fixed m, $p_{k,m}$ will reach a point of diminishing returns when there is not enough residual to reduce. This can be seen by looking at the upper bound in (4.4). Consider m fixed at 4. Since κ is effectively 1, the first term becomes insignificant for $k \ge 2$, and we have a tight bound on the residual for $k = 2$. Yet as k increases and the residual continues to be reduced, the bound stagnates because the second term decreases slowly. However, if we increase m a little, the first term remains small while the second term becomes smaller (since there are fewer indices over which to sum). Thus with increasing k one should increase m to get the smallest possible upper bound. The following table shows that this is precisely what happens in our example.

k	1	2	3	4	5
m	2	4	5	6	7
$\|p_{k,m}(\Lambda)\tilde{b}\|$	8.8e−01	2.8e−01	1.3e−01	7.1e−02	4.7e−02
$\max_{[\lambda_m,\lambda_1]} p_{k,m}$	8.0e−02	2.9e−02	1.0e−02	6.1e−03	5.2e−03
ρ_k	8.7e−01	2.5e−01	1.0e−01	5.5e−02	3.9e−02

For each value of k, the value of m for which $p_{k,m}$ minimizes the residual $\|p_{k,m}(\Lambda)\tilde{b}\|$ was determined. As the table shows, the value of m increases with k. This increase is accompanied by a modest decrease in the size of $p_{k,m}$ on $[\lambda_m, \lambda_1]$. Note that by holding m fixed one can obtain a dramatic decrease in the size of $p_{k,m}$ over $[\lambda_m, \lambda_1]$, but this strategy is not as effective at reducing the bound on the residual norm as the strategy of increasing m and putting up with a modest decrease in $p_{k,m}$. For example, if m is fixed at two, then $\max_{[\lambda_2,\lambda_1]} p_{5,2} = 1.5\text{e}{-09}$, but the residual norm is only 8.9e−02.

The last line in the table shows the MINRES residual norms. They are well approximated by the values in the third line, which suggests that MINRES is following a similar strategy. Figure 2.5 supports our contention that MINRES reduces the residual by increasing the interval in which the polynomial is small.

5. Nonreduction of the residual. Let $\alpha \in [0,1]$, and suppose we wish to reduce the residual at λ_i from $\tilde{\beta}_i$ to $\tilde{\beta}_i(1-\alpha)$. We have already noted that the corresponding component of $\tilde{x}$ is then

$$\tilde{\xi}_i = \frac{\alpha\tilde{\beta}_i}{\lambda_i}.$$

If λ_i is small, then $\tilde{\xi}_i$ is large and the solution will be unsatisfactory. Thus while MINRES is producing a regularized solution, it must not reduce the residuals on the terminal part of the spectrum. From Figure 2.5 we see that this is precisely what is happening in our example. The residual polynomials are effectively one at small λ_i.

The reason MINRES behaves in this manner is that if any polynomial is small at a small eigenvalue, it must be very large at the initial eigenvalues, say $\lambda_1, \ldots, \lambda_m$. Specifically, we have the following theorem.

THEOREM 5.1. *Let $p_{k,m}$ be defined by* (4.3). *If p is any polynomial of degree k satisfying $p(0) = 1$, $p(\lambda_i) = 1-\alpha$ for some $i > m$, and*

$$\frac{p_{k-1,m}(\lambda_i)\lambda_i}{p_{k-1,m}(\lambda_i) - 1 + \alpha} < \lambda_m, \tag{5.1}$$

then there is a point $t_{\text{big}} \in [\lambda_m, \lambda_1]$ such that

$$|p(t_{\text{big}})| \geq \left| 1 - \frac{p_{k-1,m}(\lambda_i) - 1 + \alpha}{p_{k-1,m}(\lambda_i)} \frac{\lambda_m}{\lambda_i} \right| |p_{k-1,m}(\lambda_m)|. \tag{5.2}$$

Proof. The proof is a variant of the standard proof of the theorem of de laVallée Poussin [1, p. 191]. Consider the polynomial

$$q(t) = \left[1 - \frac{p_{k-1,m}(\lambda_i) - 1 + \alpha}{p_{k-1,m}(\lambda_i)} \frac{t}{\lambda_i} \right] p_{k-1,m}(t).$$

This polynomial satisfies the conditions $p(0) = 1$ and $p(\lambda_i) = 1-\alpha$. If, in addition, (5.1) is satisfied, the root of the linear term in brackets is less than λ_m, and the polynomial $q(t)$ alternates in sign at k points

$$\lambda_m = t_1, t_2, \ldots, t_k = 1$$

on the interval $[\lambda_m, \lambda_1]$.

Now let p also satisfy the conditions of the theorem. We claim that $|p(t_i)| \geq |q(t_i)|$ for at least one t_i. For if not, the polynomial $p - q$ alternates in sign at the t_i and hence has $k-1$ zeros in the interval $[\lambda_m, \lambda_1]$. But by the conditions of the theorem $p - q$ is zero at zero and λ_i. Hence $p - q$ is a polynomial of degree not greater than k with $k+1$ zeros and must be identically zero — a contradiction.

Because the values of $|p_{k-1,m}(t_i)|$ are equal, the absolute value of $q(t_i)$ is smallest at λ_m, and evaluating $|q(\lambda_m)|$ gives the lower bound (5.2). ∎

Note that when i is large the condition (5.1) is not very restrictive. For in that case, the right-hand side is near λ_i/α, which cannot approach λ_m unless α is very small.

The import of this theorem is that the cost of reducing the residual at the small eigenvalues is to make the polynomial large on an interval containing the initial eigenvalues. This is because the factor λ_m/λ_i in (5.2) becomes enormous for i large. For

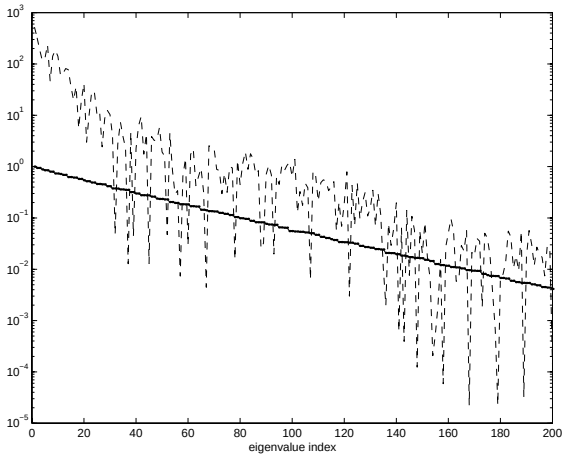


FIG. 6.1. *Eigenvalues (solid line) and components of b (dashed line).*

example, Theorem 5.1 implies that if $m = 8$ and $k = 5$ in our running example, then to get a reduction of .9 ($\alpha = .1$) in the residual at $\lambda_{25} = 6.8\text{e}{-6}$, the absolute value of the residual polynomial will have to be to be greater than 71 at some point in $[\lambda_8, \lambda_1]$. Such a large polynomial is more likely to cause a residual magnification than a residual reduction in $[\lambda_8, \lambda_1]$. This likelihood is increased by the fact that for our example the roots of p_k are spread out over $[\lambda_m, \lambda_1]$, where $m > k$, and the zeros of p_k cannot be near all m of these eigenvalues.

On the other hand, in our example, suppose we wish to reduce the residual at λ_{15} from $\tilde{\beta}_{15}$ to $\tilde{\beta}_{15}(1 - \alpha)$ for $\alpha = 0.11$. The lower bound (5.2) for the residual polynomial is 0.0086. The maximum value of the residual polynomial in $[\lambda_8, \lambda_1]$ is, in fact, 0.037, which is insufficiently greater than 0.0086 to render the reduction at λ_{15} harmless. Once again, the dispersal of the roots of $p_{k,m}$ has a beneficial effect — this time allowing a modest decrease in the residual at the intermediate eigenvalues.

In the last section we observed that MINRES eschews a quick reduction of the residual polynomials over a limited range in favor of expanding the range in which they are moderately small. The theory of this section shows that the deferred reduction has another benefit: it gives the algorithm an opportunity to reduce the residual by a modest amount on the intermediate eigenvalues.

6. Another example. To confirm our analysis of MINRES, we consider a two dimensional blurring operator. Specifically, the original matrix A is the Kronecker product $T \otimes T$, where T is a symmetric Toeplitz matrix whose first row is given by

$$t_{1k} = e^{-0.1(1-k)^2}, \qquad k = 1, \ldots, 32.$$

Figure 6.1 exhibits the first 200 eigenvalues of the matrix and the absolute values components of b in the spectral coordinate system. The error e consists of white noise scaled so that $\|e\|/\|b\| = 0.01$.

Figure 6.2 plots the errors from spectral regularization and MINRES. The contrast is dramatic. The MINRES iteration reduces the error to a minimum of 195 in four iterations. Spectral regularization requires the first 100 components of the solution be computed to reduce the error to a minimum of 192. (However, the plot shows that acceptable accuracy is attained for, say, 30 components.)

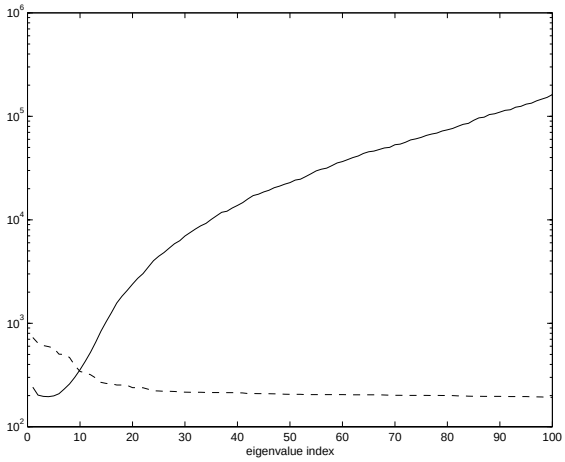


FIG. 6.2. *Errors from spectral regularization (dashed line) and MINRES regularization (solid line).*

The analysis of section 3 suggests that the solution will deteriorate when the residual norm becomes less than the error norm. In fact, the fourth, fifth, and sixth residual norms are 8.66, 7.86, and 7.52. This should be compared with the error norm 7.78.

From the comments in section 4 we would expect that residual polynomials would reduce the error by spreading their roots over a number of eigenvalues rather than by approximating specific eigenvalues. The roots for the first five residual polynomials are given below.

(1)	9.3e−01				
(2)	9.8e−01	7.3e−01			
(3)	9.9e−01	8.1e−01	5.7e−01		
(4)	9.9e−01	8.7e−01	6.9e−01	3.6e−01	
(5)	1.0e+00	9.0e−01	7.4e−01	5.4e−01	2.2e−01

The largest root is converging to one, which in our normalization is an eigenvalue of A. But otherwise, the roots are dispersed. For example, the smallest root of p_4 is 0.36, which corresponds to λ_{34}.

7. Conclusions. In the introduction to this paper we emphasized that this work would not provide a complete analysis of MINRES as a regularization technique. We can analyze MINRES applied to well-posed problems because asymptotic bounds become applicable long before a solution of the desired accuracy is attained. On the other hand, for ill-posed problems in our examples, MINRES does not have a chance to settle down before the optimal solution is obtained. In effect, we are analyzing the early transient behavior of the algorithm — always a difficult problem.

In spite of not concluding with a general theorem, our analysis sheds considerable light on the behavior of MINRES as a regularizer. Theorem 3.1 is a very general result about ill-posed problems. It suggests that attempting to reduce the residual norm below the error norm will result in a worsened solution. As we have already noted, the theorem can be taken as a justification of the Morozov discrepancy principle.

We have mentioned that the bound (4.4) on the residual norm is too crude to be used in practical applications. Nonetheless, it is good enough to suggest the strategy that MINRES uses to obtain rapid reduction of the residual norm. The free parameter m in the bound is essentially a measure of the spread of the zeros of the residual polynomial. As we have seen, the bound is optimized by keeping m greater than the iteration number k. This suggests that MINRES will attempt to minimize the residual by spreading the roots of its residual polynomials over a larger number of eigenvalues. Our numerical examples support this conjecture.

Finally, Theorem 5.1 shows us why MINRES does not tend to reduce the residual at a small eigenvalue while homing in on the regularized solution. Essentially, a small value of the residual polynomial at a small eigenvalue causes the polynomial to be large on the larger eigenvalues — just where it is most desirable to obtain a decrease in the residuals.

For definiteness, we have restricted ourselves to MINRES applied to a positive definite system. Obviously, some of our results apply to other systems and methods — e.g., least squares regularization of overdetermined systems. Such generalizations, however, will be subject for future work.

8. Appendix. *Proof of Theorem* 4.1. Let $c_k(\tau)$ denote the Chebychev polynomial of degree k and define $s_m(t)$ and $p_{k,m}(t)$ as in (4.2) and (4.3), respectively. Now $p_{k,m}$ is a polynomial of degree k satisfying $p_{k,m}(0) = 1$, and therefore from (4.1), we have

$$\|r^{(k)}\|_2^2 \leq \sum_{i=1}^{n} p_{k,m}^2(\lambda_i)\tilde{\beta}_i = \|p_{k,m}(\Lambda)\tilde{b}\|_2^2. \tag{8.1}$$

When $i \leq m$, $c_k^2(s_m(\lambda_i)) \leq 1$. Moreover, one can show that the denominator of $p_{k,m}$, which is independent of i, is bounded above [10]:

$$\left| \frac{1}{c_k(\frac{\lambda_1+\lambda_m}{\lambda_1-\lambda_m})} \right| \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k. \tag{8.2}$$

Thus, it remains to determine an upper bound for the values $c_k(s_m(\lambda_i)),\ i > m$.

For notational convenience, we fix i, let $d = \frac{\lambda_i}{\lambda_1}$, and note that since $\kappa = \frac{\lambda_1}{\lambda_m}$, $\frac{\lambda_i}{\lambda_m} = d\kappa$. Using this notation, we write $s_{m,i} \equiv s_m(\lambda_i) = (\frac{\kappa+1-2d\kappa}{\kappa-1})$.

Now $c_k(s) = \cosh(k \cosh^{-1} s)$. We use the formulas

$$\cosh x = \frac{e^x + e^{-x}}{2} \tag{8.3}$$

and

$$\ln z = \cosh^{-1}\left(\frac{z + z^{-1}}{2}\right) \text{ when } x = \ln z \tag{8.4}$$

to determine our upper bound as outlined below. We note that this type of proof technique was used at least as early as the 1960s (see [3]).

(1) Find a z such that $(\frac{z+z^{-1}}{2}) = s_{m,i}, i > m$.
(2) Use z and (8.3) to determine $c_k(s_{m,i}), i > m$.
(3) Find an upper bound for $c_k^2(s_{m,i}), i > m$.
(4) Use this bound and (8.2) to deduce an upper bound for $p_{k,m}^2(s_{m,i})$ $i > m$.
(5) Use the above information to obtain the upper bound on the right-hand side of (8.1).

Step 1. We first set

$$\frac{z + z^{-1}}{2} = s_{m,i}$$

to obtain

$$z^2 - 2s_{m,i}z + 1 = 0.$$

Since this is a quadratic equation, z is given by

$$\begin{aligned} z &= \frac{2s_{m,i} \pm \sqrt{4s_{m,i}^2 - 4}}{2} \\ &= \left(\frac{\kappa + 1 - 2d\kappa}{\kappa - 1}\right) \pm \sqrt{\left(\frac{\kappa + 1 - 2d\kappa}{\kappa - 1}\right)^2 - 1} \\ &= \frac{1}{\kappa - 1}\left(\kappa + 1 - 2d\kappa \pm 2\sqrt{(\kappa - \kappa d)(1 - \kappa d)}\right) \\ &= \frac{(\sqrt{\kappa - \kappa d} \pm \sqrt{1 - \kappa d})^2}{\kappa - 1}. \end{aligned}$$

(8.5)

For convenience, we take z to be

$$z = \frac{(\sqrt{\kappa - \kappa d} + \sqrt{1 - \kappa d})^2}{\kappa - 1}. \tag{8.6}$$

Step 2. Now $c_k(s_{m,i}) = \cosh(k \ln z) = \cosh(\ln z^k)$. Using (8.3), we obtain

$$\begin{aligned} c_k(s_{m,i}) &= \frac{z^k + z^{-k}}{2} \\ &= \frac{z^{2k} + 1}{2z^k}. \end{aligned}$$

(8.7)

Step 3. Therefore,

$$c_k^2(s_{m,i}) = \frac{1}{4}\left(z^{2k} + \frac{1}{z^{2k}} + 2\right).$$

(8.8)

To get an upper bound on $c_k^2(s_{m,i})$, let us first try to get an upper bound on $\frac{1}{z^{2k}}$ by getting a lower bound on z. Observe that the numerator of z in (8.6) is bounded below by $\kappa - d\kappa$, so our lower bound for z is $\frac{\kappa - d\kappa}{\kappa - 1}$. Since $d\kappa \leq 1$, we therefore obtain

$$\frac{1}{z} \leq \frac{\kappa - 1}{\kappa - d\kappa} \leq 1,$$

and so $\frac{1}{z^{2k}} \leq 1$.

Next, we have the following expression for z^{2k}:

$$z^{2k} = \left(\frac{\kappa + 1 + 2\left((\kappa - d\kappa)(1 - d\kappa)\right)^{1/2} - 2d\kappa}{\kappa - 1} \right)^{2k}.$$

Therefore, we have the following bound on $c_k^2(s_{m,i})$:

$$c_k^2(s_{m,i}) \leq \frac{1}{4} \left(3 + \left(\frac{\kappa + 1 + 2\left((\kappa - d\kappa)(1 - d\kappa)\right)^{1/2} - 2d\kappa}{\kappa - 1} \right)^{2k} \right). \tag{8.9}$$

Step 4. Now we may use (8.2) and (8.9) to determine an upper bound on $p_{k,m}^2(s_{m,i})$ for $i > m$:

$$p_{k,m}^2(s_{m,i}) \leq \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^{2k} \left(3 + \left(\frac{\kappa + 1 + 2\left((\kappa - d\kappa)(1 - d\kappa)\right)^{1/2} - 2d\kappa}{\kappa - 1} \right)^{2k} \right).$$

Using $\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} = \frac{\kappa-1}{\kappa+2\sqrt{\kappa}+1}$, we cancel like terms on the right-hand side above and factor out $\sqrt{\kappa}$ from the radical to obtain

$$p_{k,m}^2(s_{m,i}) \leq 3 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^{2k} + \left(\frac{\kappa + 1 + 2\sqrt{\kappa}\left((1 - \frac{\lambda_i}{\lambda_1})(1 - \frac{\lambda_i}{\lambda_m}) \right)^{1/2} - 2\frac{\lambda_i}{\lambda_m}}{\kappa + 1 + 2\sqrt{\kappa}} \right)^{2k}.$$

The second term of the right-hand side above can be bounded above by

$$\left(1 - \frac{2\frac{\lambda_i}{\lambda_m}}{\kappa + 1 + 2\sqrt{\kappa}} \right)^{2k} = \left(1 - 2\frac{\lambda_i}{\lambda_1 + \lambda_m + 2\sqrt{\lambda_1 \lambda_m}} \right)^{2k}.$$

Thus,

$$p_{k,m}^2(s_{m,i}) \leq 3 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^{2k} + \left(1 - 2\frac{\lambda_i}{\lambda_1 + \lambda_m + 2\sqrt{\lambda_1 \lambda_m}} \right)^{2k}. \tag{8.10}$$

Step 5. From (8.2) when $i \leq m$,

$$p_{k,m}^2(s_{m,i}) \leq \left(\frac{l}{c_k(s_{m,i})} \right)^2 \leq 4 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^{2k}.$$

Breaking up the sum over the first m and last $n - m$ terms from (8.1) and (8.10), we observe

$$\|r_k\|_2^2 \leq 4 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^{2k} \|\tilde{b}_m\|_2^2 + \sum_{i=m+1}^{n} \left(1 - 2\frac{\lambda_i}{(\sqrt{\lambda_1} + \sqrt{\lambda_m})^2} \right)^{2k} \tilde{\beta}_i^{\,2}.$$

9. Acknowledgments. We wish to thank Per Christian Hansen and Dianne O'Leary for their useful comments on the first draft of this paper.

REFERENCES

[1] K. E. Atkinson, *An Introduction to Numerical Analysis*, John Wiley, New York, 1978.
[2] A. Björck and L. Eldén, *Methods in Numerical Algebra for Ill-Posed Problems*, Technical Report LiTH-MAT-R-1979-33, Department of Mathematics, Linköping University, Linköping, Sweden, 1979.
[3] J. W. Daniel, *The conjugate gradient method for linear and nonlinear operator equations*, SIAM J. Numer. Anal., 4 (1967), pp. 10–26.
[4] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.
[5] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, 1997.
[6] M. Hanke, private communication.
[7] P. C. Hansen, *The truncated SVD as a method for regularization*, BIT, 27 (1987), pp. 534–553.
[8] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems*, SIAM, Philadelphia, 1998.
[9] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.
[10] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, 2nd ed., Springer-Verlag, New York, 1993.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 629–635
© 1999 Society for Industrial and Applied Mathematics

THE MOORE–PENROSE GENERALIZED INVERSE FOR SUMS OF MATRICES*

JAMES ALLEN FILL† AND DONNIELL E. FISHKIND‡

Abstract. In this paper we exhibit, under suitable conditions, a neat relationship between the Moore–Penrose generalized inverse of a sum of two matrices and the Moore–Penrose generalized inverses of the individual terms. We include an application to the parallel sum of matrices.

Key words. Moore–Penrose generalized inverse, Sherman–Morrison–Woodbury formula, singular value decomposition, rank additivity, parallel sum

AMS subject classifications. Primary, 15A09; Secondary, 15A18

PII. S0895479897329692

1. Background and main result. In the late 1940s and the 1950s Sherman and Morrison [11], [12], Woodbury [13], Bartlett [2], and Bodewig [4] discovered the following result. As in [7], $M_{m,n}$ denotes the space of complex-valued $m \times n$ matrices and, when $m = n$, this is shortened to M_n.

THEOREM 1 (Sherman–Morrison–Woodbury). *For $s \leq n$, let $A \in M_n$ and $G \in M_s$ both be invertible, and let $Y, Z \in M_{n,s}$. Then $A + YGZ^*$ is invertible if and only if $G^{-1} + Z^*A^{-1}Y$ is invertible, in which case*

$$(A + YGZ^*)^{-1} = A^{-1} - A^{-1}Y(G^{-1} + Z^*A^{-1}Y)^{-1}Z^*A^{-1}.$$

The Sherman–Morrison–Woodbury (SMW) formula and related formulas are reviewed in Henderson and Searle [6]. The SMW formula has been used in a wide variety of applications; an excellent review by Hager [5] describes some of the applications to statistics, networks, structural analysis, asymptotic analysis, optimization, and partial differential equations.

In 1992, Riedel [10] proved an analogous formula (Theorem 2) for some cases where A is singular. All matrices, including singular and even nonsquare matrices, have a *Moore–Penrose generalized inverse. Given a matrix $A \in M_{m,n}$, the Moore–Penrose generalized inverse* of A, denoted $A^\dagger$, is the unique matrix in $M_{n,m}$ satisfying the conditions

$$AA^\dagger A = A, \tag{1}$$

$$A^\dagger AA^\dagger = A^\dagger, \tag{2}$$

$$AA^\dagger \text{ is Hermitian, and} \tag{3}$$

$$A^\dagger A \text{ is Hermitian.} \tag{4}$$

*Received by the editors November 6, 1997; accepted for publication (in revised form) by G. Styan November 1, 1998; published electronically December 28, 1999.
http://www.siam.org/journals/simax/21-2/32969.html

†Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, MD 21218-2682 (jimfill@jhu.edu). The research of this author was supported by NSF grants DMS–96–26756 and DMS–98–03780 and by the Acheson J. Duncan Fund for the Advancement of Research in Statistics.

‡Department of Mathematics and Statistics, University of Southern Maine, Portland, ME 04104 (fishkind@payson.usm.maine.edu). This author carried out research for this work while affiliated with The Johns Hopkins University and was supported by U.S. Department of Education GAANN grant P200A40303.

In particular, if $A = U\Sigma V^*$ is a singular value decomposition of A (that is, if $U \in M_m$ and $V \in M_n$ are unitary and $\Sigma \in M_{m,n}$ has $\Sigma_{i,i} \geq 0$ for $1 \leq i \leq \min(m,n)$ and $\Sigma_{i,j} = 0$ otherwise), then it may be verified (by checking (1)–(4)) that $A^\dagger = V\Sigma^\dagger U^*$, where $\Sigma^\dagger$ is defined by

$$\Sigma^\dagger_{i,j} := \begin{cases} \frac{1}{\Sigma_{i,i}} & \text{if } i = j \text{ and } \Sigma_{i,i} \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Classical references on generalized inverses are [3] and [9].

THEOREM 2 (Riedel). *Let s and n be positive integers with $s \leq n$; $A \in M_n$; $G \in M_s$; $Y, Y_p \in M_{n,s}$; $Z, Z_p \in M_{n,s}$. Assume $R(Y) \subseteq R(A)$, $R(Y_p) \perp R(A)$, $R(Z) \subseteq R(A^*)$, $R(Z_p) \perp R(A^*)$, G is invertible, Y_p is of full rank, and Z_p is of full rank. Assume also that $R(Y_p) = R(Z_p)$. Then*

$$(A + (Y + Y_p)G(Z + Z_p)^*)^\dagger = A^\dagger - DZ^*A^\dagger - A^\dagger YC^* + D(G^{-1} + Z^*A^\dagger Y)C^*,$$

where $C := Y_p(Y_p^*Y_p)^{-1}$ and $D := Z_p(Z_p^*Z_p)^{-1}$.

The matrices $(Y + Y_p)G(Z + Z_p)^*$ in Theorem 2 and YGZ^* in Theorem 1 are referred to as the *update matrices* to the *initial matrix* A. A version of Riedel's theorem (Theorem 2) for the special case where we seek the Moore–Penrose generalized inverse of a rank-one update to the initial matrix can be found in [9].

Riedel verifies Theorem 2 by checking conditions (1)–(4). It must, however, be noted that the hypothesis $R(Y_p) = R(Z_p)$ of Theorem 2 is nowhere used in the verification of Theorem 2, and thus *Theorem 2 is true without this part of the hypothesis.* It is this key observation that allows us to make use of Theorem 2 in this paper. *When we refer to Theorem 2 henceforth, we will be referring to this theorem without the aforementioned unnecessary hypothesis.*

The purpose of this paper, given matrices A and B and suitable conditions, is to relate $(A + B)^\dagger$ cleanly to $A^\dagger$ and $B^\dagger$. This is done in Theorem 3, our main result, using Riedel's theorem. (For a subspace Ω we denote by P_Ω the orthogonal projection onto Ω.)

THEOREM 3. *Let $A, B \in M_n$ with $\operatorname{rank}(A + B) = \operatorname{rank} A + \operatorname{rank} B$. Then*

$$(A + B)^\dagger = (I - S)\, A^\dagger\, (I - T) + SB^\dagger T, \tag{5}$$

where

$$S := (P_{R(B^*)}P_{R(A^*)^\perp})^\dagger \qquad \text{and}$$
$$T := (P_{R(A)^\perp}P_{R(B)})^\dagger.$$

EXAMPLE 4. Without the rank-additivity hypothesis [$\operatorname{rank}(A + B) = \operatorname{rank} A + \operatorname{rank} B$], the conclusion of Theorem 3 is (generally) false. For example, let A and B be 1×1 matrices with 1 as their only entry. In the notation of Theorem 3, we compute $S = ([1][0])^\dagger = [0]$ and $T = ([0][1])^\dagger = [0]$. Hence,

$$(I - S)A^\dagger(I - T) + SB^\dagger T = [1][1]^\dagger[1] + [0][1]^\dagger[0] = [1]$$

while $(A + B)^\dagger = [\frac{1}{2}]$, contrary to the assertion of Theorem 3. But note $\operatorname{rank}[2] = 1 \neq 2 = \operatorname{rank}[1] + \operatorname{rank}[1]$.

Is it possible, however, that the rank-additivity hypothesis in the statement of Theorem 3 can be eliminated in favor of a weaker condition? We show in Proposition 5

that the rank-additivity hypothesis cannot be avoided in any proof of Theorem 3 which employs Riedel's theorem (Theorem 2), since rank additivity is shown to be implied by the hypotheses of Riedel's theorem. (As mentioned, our proof of Theorem 3 relies on Theorem 2.)

For conditions when $\operatorname{rank}(A+B) = \operatorname{rank} A + \operatorname{rank} B$, see [8].

Remark. The matrices S and T appearing in (5) are far from determined by (5). For example, let x and y be orthonormal vectors in $\mathbf{C}^n$ with $n \geq 3$, and let

$$A := xx^*, \quad B := yy^*.$$

Applying Theorem 3, we obtain

$$(A+B)^\dagger = (I-yy^*)A(I-yy^*) + (yy^*)B(yy^*),$$

which simplifies to

$$(xx^* + yy^*)^\dagger = xx^* + yy^*. \tag{6}$$

But applying Theorem 3 with the roles of A and B reversed, we obtain the different formula

$$(A+B)^\dagger = (xx^*)A(xx^*) + (I-xx^*)B(I-xx^*)$$

which, however, also simplifies to (6).

1.1. Derivation of main result (Theorem 3). Our proof of Theorem 3 is based on the following proposition.

PROPOSITION 5. *Let s and n be positive integers with $s \leq n$; $A \in M_n$; $G \in M_s$; $Y, Y_p \in M_{n,s}$; $Z, Z_p \in M_{n,s}$. Assume $R(Y) \subseteq R(A)$, $R(Y_p) \perp R(A)$, $R(Z) \subseteq R(A^*)$, and $R(Z_p) \perp R(A^*)$.*

Of the following statements, 1 *implies* 2. *Conversely,* 2 *and* 3 *imply* 1.

1. *Y_p and Z_p are of full rank.*
2. $\operatorname{rank}[A + (Y+Y_p)G(Z+Z_p)^*] = \operatorname{rank} A + \operatorname{rank}[(Y+Y_p)G(Z+Z_p)^*]$.
3. $\operatorname{rank}[(Y+Y_p)G(Z+Z_p)^*] = s$.

Proposition 5 is used in proving Theorem 3, but it also demonstrates that rank additivity (of the initial matrix and the update matrix) is implied by the hypotheses of Theorem 2; since our proof of Theorem 3 relies on Theorem 2, the rank-additivity hypothesis of Theorem 3 is, for us, unavoidable.

Proof of Proposition 5. Using the assumption $R(Y) \subseteq R(A)$, we find

$$R[A + (Y+Y_p)G(Z+Z_p)^*] \subseteq R(A) + R(Y) + R(Y_p) = R(A) + R(Y_p).$$

Thus, if statements 2 and 3 hold, then

$$\operatorname{rank} A + s = \operatorname{rank}[A + (Y+Y_p)G(Z+Z_p)^*] \leq \operatorname{rank} A + \operatorname{rank} Y_p,$$

from which we conclude that $\operatorname{rank} Y_p \geq s$, that is, that Y_p (and similarly Z_p) is of full rank (statement 1).

Conversely, suppose Y_p and Z_p are of full rank (statement 1). We have

$$\operatorname{rank} Y_p = s \geq \operatorname{rank} G \geq \operatorname{rank}\left[(Y+Y_p)G(Z+Z_p)^*\right]. \tag{7}$$

In [10], Riedel points out that (when Y_p and Z_p are of full rank)

$$[A + (Y+Y_p)G(Z+Z_p)^*]\,[A + (Y+Y_p)G(Z+Z_p)^*]^\dagger = AA^\dagger + Y_pY_p^\dagger.$$

By the orthogonality of $R(A)$ and $R(Y_p)$, we have $\operatorname{rank}(AA^\dagger + Y_pY_p^\dagger) = \operatorname{rank}(AA^\dagger) + \operatorname{rank}(Y_pY_p^\dagger)$. (Without loss of generality, $AA^\dagger$ and $Y_pY_p^\dagger$ share the same unitary matrices in their singular value decompositions because of this orthogonality.) Thus,

$$\begin{aligned}\operatorname{rank}[A + (Y + Y_p)G(Z + Z_p)^*] &= \operatorname{rank}(AA^\dagger + Y_pY_p^\dagger)\\ &= \operatorname{rank}(AA^\dagger) + \operatorname{rank}(Y_pY_p^\dagger)\\ &= \operatorname{rank} A + \operatorname{rank} Y_p\\ &\geq \operatorname{rank} A + \operatorname{rank}[(Y + Y_p)G(Z + Z_p)^*],\end{aligned}$$

the last inequality holding by (7). Because (trivially)

$$\operatorname{rank}[A + (Y + Y_p)G(Z + Z_p)^*] \leq \operatorname{rank} A + \operatorname{rank}[(Y + Y_p)G(Z + Z_p)^*],$$

we conclude

$$\operatorname{rank}[A + (Y + Y_p)G(Z + Z_p)^*] = \operatorname{rank} A + \operatorname{rank}[(Y + Y_p)G(Z + Z_p)^*],$$

that is, statement 2 holds. □

In proving Theorem 3 we will need also the following three facts about the Moore–Penrose generalized inverse that can be verified directly from (1)–(4). For positive integers t and n such that $t \leq n$, let $L_{n,t}$ denote a matrix of size $n \times t$ with ones on the diagonal and zeros elsewhere. Let r, s, p, and q be positive integers with $s \leq p$ and $r \leq q$, and let $A \in M_{r,s}$, $U \in M_r$, and $V \in M_s$ with U and V unitary. Then

$$(L_{q,r}AL_{p,s}^*)^\dagger = L_{p,s}A^\dagger L_{q,r}^* \tag{8}$$

and

$$(UAV^*)^\dagger = VA^\dagger U^*. \tag{9}$$

If A is of full rank with $r \geq s$, then

$$A^\dagger = (A^*A)^{-1}A^*. \tag{10}$$

Proof of Theorem 3. To simplify notation, and since n is fixed, we shorten $L_{n,t}$ to L_t for $t \leq n$. Let A and B have respective singular value decompositions $U_A\Sigma_AV_A^*$ and $U_B\Sigma_BV_B^*$, where, without loss of generality, exactly the first s diagonal entries of Σ_B are nonzero and exactly the first r diagonal entries of Σ_A are zero.

Note that

$$A + B = A + U_BL_sL_s^*\Sigma_BL_sL_s^*V_B^* = A + (Y + Y_p)G(Z + Z_p)^*, \tag{11}$$

where we define

$$\begin{aligned}G &:= L_s^*\Sigma_BL_s,\\ Y &:= P_{R(A)}U_BL_s = [U_A(I - L_rL_r^*)U_A^*]U_BL_s,\\ Y_p &:= P_{R(A)^\perp}U_BL_s = [U_AL_rL_r^*U_A^*]U_BL_s,\\ Z &:= P_{R(A^*)}V_BL_s = [V_A(I - L_rL_r^*)V_A^*]V_BL_s,\\ Z_p &:= P_{R(A^*)^\perp}V_BL_s = [V_AL_rL_r^*V_A^*]V_BL_s.\end{aligned}$$

Note that G, Y, Y_p, Z, and Z_p satisfy all of the hypotheses of Theorem 2 since Y_p and Z_p are of full rank by Proposition 5 (because $\operatorname{rank} B = s$ and $\operatorname{rank}(A + B) = \operatorname{rank} A + \operatorname{rank} B$).

We next observe that (with D and C defined as in Theorem 2)

$$\begin{aligned} DG^{-1}C^* &= DL_s^*\Sigma_B^\dagger L_s C^* \\ &= DL_s^*V_B^*V_B\Sigma_B^\dagger U_B^*U_BL_sC^* \\ &= DL_s^*V_B^*B^\dagger U_BL_sC^* \\ &= D(Z^*+Z_p^*)B^\dagger(Y+Y_p)C^*, \end{aligned} \tag{12}$$

and thus by Theorem 2 and (12) we have that

$$(A+B)^\dagger = (I-DZ^*)A^\dagger(I-YC^*)+(DZ^*+DZ_p^*)B^\dagger(YC^*+Y_pC^*). \tag{13}$$

This is the basic form of $(A+B)^\dagger$ that we seek, and we proceed to compute DZ^*, YC^*, DZ_p^*, and Y_pC^*.

Because

$$n \geq \operatorname{rank}(A+B) = \operatorname{rank}A + \operatorname{rank}B = n-r+s,$$

we have $r \geq s$. By this, the fact that projection matrices are Hermitian and idempotent, and (8)–(10), we get

$$\begin{aligned} YC^* &= Y(Y_p^*Y_p)^{-1}Y_p^* \\ &= P_{R(A)}U_BL_s(L_s^*U_B^*P_{R(A)^\perp}^*P_{R(A)^\perp}U_BL_s)^{-1}L_s^*U_B^*P_{R(A)^\perp}^* \\ &= P_{R(A)}U_BL_s(L_s^*U_B^*P_{R(A)^\perp}U_BL_s)^{-1}L_s^*U_B^*P_{R(A)^\perp} \\ &= P_{R(A)}U_BL_s(L_s^*U_B^*U_AL_rL_r^*U_A^*U_BL_s)^{-1}L_s^*U_B^*U_AL_rL_r^*U_A^* \\ &= P_{R(A)}U_BL_s[(L_r^*U_A^*U_BL_s)^*(L_r^*U_A^*U_BL_s)]^{-1}(L_r^*U_A^*U_BL_s)^*L_r^*U_A^* \\ &= P_{R(A)}U_BL_s(L_r^*U_A^*U_BL_s)^\dagger L_r^*U_A^* \\ &= P_{R(A)}U_B(L_rL_r^*U_A^*U_BL_sL_s^*)^\dagger U_A^* \\ &= P_{R(A)}(U_AL_rL_r^*U_A^*U_BL_sL_s^*U_B^*)^\dagger \\ &= P_{R(A)}(P_{R(A)^\perp}P_{R(B)})^\dagger \\ &= P_{R(A)}T \end{aligned} \tag{14}$$

and also

$$\begin{aligned} DZ^* &= Z_p(Z_p^*Z_p)^{-1}Z^* \\ &= P_{R(A^*)^\perp}V_BL_s(L_s^*V_B^*P_{R(A^*)^\perp}^*P_{R(A^*)^\perp}V_BL_s)^{-1}L_s^*V_B^*P_{R(A^*)}^* \\ &= P_{R(A^*)^\perp}V_BL_s(L_s^*V_B^*P_{R(A^*)^\perp}V_BL_s)^{-1}L_s^*V_B^*P_{R(A^*)} \\ &= V_AL_rL_r^*V_A^*V_BL_s[L_s^*V_B^*V_AL_rL_r^*V_A^*V_BL_s]^{-1}L_s^*V_B^*P_{R(A^*)} \\ &= V_AL_r(L_r^*V_A^*V_BL_s)[(L_r^*V_A^*V_BL_s)^*(L_r^*V_A^*V_BL_s)]^{-1}L_s^*V_B^*P_{R(A^*)} \\ &= V_AL_r(L_r^*V_A^*V_BL_s)^{*\dagger}L_s^*V_B^*P_{R(A^*)} \\ &= V_AL_r(L_s^*V_B^*V_AL_r)^\dagger L_s^*V_B^*P_{R(A^*)} \\ &= V_A(L_sL_s^*V_B^*V_AL_rL_r^*)^\dagger V_B^*P_{R(A^*)} \\ &= (V_BL_sL_s^*V_B^*V_AL_rL_r^*V_A^*)^\dagger P_{R(A^*)} \\ &= (P_{R(B^*)}P_{R(A^*)^\perp})^\dagger P_{R(A^*)} \\ &= SP_{R(A^*)}. \end{aligned} \tag{15}$$

Similarly, we get

$$Y_p C^* = P_{R(A)^\perp} T \qquad \text{and}$$
$$DZ_p^* = S P_{R(A^*)^\perp}. \tag{16}$$

By plugging (14)–(16) into (13), and noting that $P_{R(A^*)} A^\dagger = A^\dagger$ and $A^\dagger P_{R(A)} = A^\dagger$, the assertion of Theorem 3 follows. □

2. Application to the parallel sum. It is well known in elementary electronics that if two resistors with resistances r_1 and r_2 are placed in parallel, then the cumulative resistance r is computed by the formula

$$r = r_1 (r_1 + r_2)^{-1} r_2 = \left(\frac{1}{r_1} + \frac{1}{r_2} \right)^{-1}. \tag{17}$$

With the idea of generalizing this notion to matrices, Anderson and Duffin [1] define, for $A, B \in M_n$, the *parallel sum* of A and B as

$$A : B := A(A + B)^\dagger B, \tag{18}$$

which, in the case that A and B are (scalar) resistances, is exactly the formula in (17). An alternative definition for the parallel sum of A and B can be found in Rao and Mitra [9], where it is defined as

$$A \| B := \left(A^\dagger + B^\dagger \right)^\dagger, \tag{19}$$

which, in the case that A and B are (scalar) resistances, is again exactly the formula in (17). Given some assumptions on A and B, [9] presents necessary and sufficient conditions for the two definitions of parallel sum to agree.

The following result uses Theorem 3 to provide, under certain conditions, a neat equation relating $A \| B$ to A and B.

COROLLARY 6. *Let* $A, B \in M_n$ *with* $\operatorname{rank}(A \| B) = \operatorname{rank} A + \operatorname{rank} B$. *Then*

$$A \| B = (I - R) A (I - W) + RBW,$$

where

$$R := \left(P_{R(B)} P_{R(A)^\perp} \right)^\dagger \qquad \textit{and}$$
$$W := \left(P_{R(A^*)^\perp} P_{R(B^*)} \right)^\dagger.$$

Corollary 6 is an immediate corollary of Theorem 3, where $A^\dagger$ and $B^\dagger$ of Theorem 6 play the roles of A and B in Theorem 3.

Acknowledgments. The authors thank Adi Ben-Israel, Shih-Ping Han, Roger Horn, and Kurt Riedel for valuable discussions and also thank two anonymous referees for their helpful suggestions.

REFERENCES

[1] W. N. ANDERSON, JR., AND R. J. DUFFIN, *Series and parallel addition of matrices*, J. Math. Anal. Appl., 26 (1969), pp. 576–594.

[2] M. S. Bartlett, *An inverse matrix adjustment arising in discriminant analysis*, Ann. Math. Statist., 22 (1951), pp. 107–111.
[3] A. Ben-Israel and T. Greville, *Generalized Inverses*: *Theory and Applications*, Wiley-Interscience, New York, 1974.
[4] E. Bodewig, *Matrix Calculus*, North-Holland, Amsterdam, 1959.
[5] W. W. Hager, *Updating the inverse of a matrix*, SIAM Rev., 31 (1989), pp. 221–239.
[6] H. V. Henderson and S. R. Searle, *On deriving the inverse of a sum of matrices*, SIAM Rev., 23 (1981), pp. 53–60.
[7] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.
[8] G. Marsaglia and G. P. H. Styan, *When does* $\operatorname{rank}(A + B) = \operatorname{rank} A + \operatorname{rank} B$*?*, Canad. Math. Bull., 15 (1972), pp. 451–452.
[9] C. R. Rao and S. K. Mitra, *Generalized Inverse of Matrices and its Applications*, Wiley-Interscience, New York, 1971.
[10] K. S. Riedel, *A Sherman–Morrison–Woodbury identity for rank augmenting matrices with application to centering*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 659–662.
[11] J. Sherman and W. J. Morrison, *Adjustment of an inverse matrix corresponding to changes in the elements of a given column or given row of the original matrix*, Ann. Math. Statist., 20 (1949), p. 621.
[12] J. Sherman and W. J. Morrison, *Adjustment of an inverse matrix corresponding to a change in one element of a given matrix*, Ann. Math. Statist., 21 (1950), pp. 124–127.
[13] M. A. Woodbury, *Inverting Modified Matrices*, Technical Report 42, Statistical Research Group, Princeton University, Princeton, NJ, 1950.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 636–641
© 1999 Society for Industrial and Applied Mathematics

A NOTE ON P_1- AND LIPSCHITZIAN MATRICES*

G. S. R. MURTHY†, S. K. NEOGY‡, AND F. THUIJSMAN§

Abstract. The linear complementarity problem (q, A) with data $A \in R^{n\times n}$ and $q \in R^n$ involves finding a nonnegative $z \in R^n$ such that $Az+q \geq 0$ and $z^t(Az+q) = 0$. Cottle and Stone introduced the class of P_1-matrices and showed that if A is in $P_1\backslash Q$, then $K(A)$ (the set of all q for which (q, A) has a solution) is a half-space and (q, A) has a unique solution for every q in the interior of $K(A)$. Extending the results of Murthy, Parthasarathy, and Sriparna [*Ann. Dynamic Games*, to appear], we present a number of equivalent characterizations of $P_1\backslash Q$. Also, we present yet another characterization of P-matrices. This widens the range of matrix classes for which a conjecture raised by Murthy, Parthasarathy, and Sriparna [*SIAM J. Matrix Anal. Appl.*, 19 (1998), pp. 898–905] characterizing the class of Lipschitzian matrices is true.

Key words. linear complementarity problem, matrix classes, Lipschitz property

AMS subject classification. 90C33

PII. S0895479897329898

1. Introduction.

The linear complementarity problem (LCP) with data $A \in \boldsymbol{R}^{n\times n}$ and $q \in \boldsymbol{R}^n$ involves finding a vector $z \in \boldsymbol{R}^n$ such that

$$Az + q \geq 0, \quad z \geq 0, \quad \text{and } z^t(Az + q) = 0.$$

Many have tackled a number of problems on existence and solvability of LCP by studying the properties of A; this has lead to the evolution of a wide range of matrix classes. Good background material on the subject can be found in [2, 13].

We shall briefly introduce the concepts and notation required for presentation of the results of this paper. For $A \in \boldsymbol{R}^{n\times n}$, $q \in \boldsymbol{R}^n$, let $F(q, A) = \{z \in \boldsymbol{R}^n_+ : Az+q \geq 0\}$, $S(q, A) = \{z \in F(q, A) : (Az + q)^t z = 0\}$, and $K(A) = \{p \in \boldsymbol{R}^n : S(p, A) \neq \emptyset\}$. For most of the other notation, we shall follow that used in [2]. Let $\bar{n} = \{1, 2, \ldots, n\}$ and for $\alpha, \beta \subseteq \bar{n}$, let $A_{\alpha\beta}$ denote the submatrix of A obtained by retaining only rows corresponding to α and columns corresponding to β. A similar notation is used for vectors. The complement of $\alpha \subseteq \bar{n}$ with respect to $\bar{n}$ will be denoted by $\bar{\alpha}$. The vector e will denote the vector of ones (of appropriate order).

We define a number of matrix classes studied extensively in the literature of LCPs. Say that $A \in \boldsymbol{Q}$ if $S(q, A) \neq \phi$ for all $q \in \boldsymbol{R}^n$; $A \in \boldsymbol{Q}_0$ if for all q, $F(q, A) \neq \phi \Rightarrow S(q, A) \neq \phi$; $A \in \boldsymbol{P}_0(\boldsymbol{P})$ (A is nondegenerate) if $\det A_{\alpha\alpha} \geq 0 (> 0)(\neq 0)$ for all $\alpha \subseteq \bar{n}$; $A \in \boldsymbol{P}_1$ if $A \in \boldsymbol{P}_0$ and $\det A_{\alpha\alpha} = 0$ for some unique α; $A \in \boldsymbol{INS}_k$ if $|S(q, A)| = k$ for all $q \in \text{int } K(A)$; $A \in \boldsymbol{R}_0$ if $|S(0, A)| = 1$; $A \in \boldsymbol{E}_0(\boldsymbol{E})$ if $|S(q, A)| = 1$ for all $q > 0\, (q \geq 0)$; and $A \in \boldsymbol{E'}$ if $|S(q, A)| = 1$ for all $q \geq 0$ and $|S(0, A)| \geq 2$. The class $\boldsymbol{INS}_1$ is denoted by $\boldsymbol{U}$. Next, say that $A \in \boldsymbol{Q}$ if $A_{\alpha\alpha} \in \boldsymbol{Q}$ for all $\alpha \subseteq \bar{n}$. Similar notation is used for other classes.

For $A \in \boldsymbol{R}^{n\times n}$ with $\det A_{\alpha\alpha} \neq 0$ for an α, the principal pivotal transform (PPT) of A with respect to α, denoted by $\wp_\alpha(A)$, is defined as $M = \wp_\alpha(A)$, where $M_{\alpha\alpha} =$

*Received by the editors November 10, 1997; accepted for publication (in revised form) by R. Cottle April 6, 1998; published electronically December 28, 1999.
http://www.siam.org/journals/simax/21-2/32989.html

†Indian Statistical Institute, Street No. 8, Habsiguda, Hyderabad 500 007, India (murthygsr@hotmail.com).
‡Indian Statistical Institute, 7 SJS Sansanwal Marg, New Delhi 110 016, India (skn@isid.ac.in).
§University of Maastricht, P.O. Box 616, 6200 MD Maastricht, The Netherlands (frank@math.unimaas.nl).

$(A_{\alpha\alpha})^{-1}$, $M_{\alpha\bar{\alpha}} = -(A_{\alpha\alpha})^{-1}A_{\alpha\bar{\alpha}}$, $M_{\bar{\alpha}\alpha} = A_{\bar{\alpha}\alpha}(A_{\alpha\alpha})^{-1}$, and $M_{\bar{\alpha}\bar{\alpha}} = A_{\bar{\alpha}\bar{\alpha}} - M_{\bar{\alpha}\alpha}A_{\alpha\bar{\alpha}}$. Whenever we refer to PPTs, we mean only those which are well defined. It is a well-established result that if A belongs to $\boldsymbol{Q}$, $\boldsymbol{Q}_0$, $\boldsymbol{P}_0$, $\boldsymbol{P}$, $\boldsymbol{P}_1$, or $\boldsymbol{INS}_k$, then every PPT of A is also in the respective class.

For any $A \in \boldsymbol{R}^{m\times n}$, the associated two-person zero-sum matrix game is described as follows. There are two players: row player R and column player C. R chooses a row i, C chooses a column j, and R pays C an amount equal to a_{ij}. A mixed strategy for $R(C)$ is a probability vector $p \in \boldsymbol{R}^m (q \in \boldsymbol{R}^n)$. The fundamental theorem of von Neumann asserts that there exist strategies $\bar{p}$ and $\bar{q}$, called optimal strategies for R and C, respectively, and a real number v, called the value of the game (denoted by $v(A)$), such that

$$\bar{p}^t A \le ve \quad \text{and} \quad A\bar{q} \ge ve.$$

A strategy is said to be completely mixed if all its coordinates are positive. A game is said to be completely mixed if all the optimal strategies are completely mixed. We shall refer to the game by the matrix itself.

Cottle and Stone [3] showed that if $A \in \boldsymbol{P}_1\backslash\boldsymbol{Q}$, then $A \in \boldsymbol{U}$ and that $K(A)$ is a half-space. They further showed that if, in addition, $\det A = 0$, then the support of the half-space $K(A)$ can be chosen to be a positive vector. Murthy, Parthasarathy, and Sriparna [11] showed that if A is a completely mixed game with $v(A) = 0$, then $A \in \boldsymbol{P}_1\backslash\boldsymbol{Q}$ if and only if $A \in \boldsymbol{U}$. In section 2, we show that $A \in \boldsymbol{P}_1\backslash\boldsymbol{Q}$ if and only if $A \in \boldsymbol{U}$ and $K(A)$ is a half-space with a positive support. Indeed, we give a number of equivalent conditions of the statement that $A \in \boldsymbol{P}_1\backslash\boldsymbol{Q}$ and $\det A = 0$.

In section 3, we provide some results relating to the conjecture raised by Murthy, Parthasarathy, and Sriparna [12] concerning a constructive characterization of Lipschitzian matrices. In particular, we show that a certain (constructive) property is necessarily satisfied by Lipschitzian matrices as well as nondegenerate ***INS***-matrices and that the converse is true with an additional assumption.

2. Results on $\boldsymbol{P}_1$-matrices. For the sake of completeness, we state relevant portions of some known theorems.

THEOREM 1 (see Cottle and Stone [3]). *Suppose $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{P}_1$. The following statements hold:*

(i) *There exists a unique α such that $\wp_\alpha(A) \in \boldsymbol{P}_1$ and $\det \wp_\alpha(A) = 0$.*
(ii) *A is in $\boldsymbol{Q}_0$.*
(iii) *If $A \notin \boldsymbol{Q}$ and $\det A = 0$, then $A \in \boldsymbol{U}$ and $K(A) = \{q : \pi^t q \ge 0\}$ for some $\pi > 0$.*

THEOREM 2 (see Kaplansky [7]). *Suppose $A \in \boldsymbol{R}^{m\times n}$. The following statements hold for the matrix game A:*

(i) *If the row player R has a completely mixed strategy, then $Aq = v(A)e$ for all optimal strategies q of the column player C.*
(ii) *If $v(A) = 0$, then the game is completely mixed if and only if [$m = n$, $\operatorname{rank}(A) = n - 1$, and all cofactors of A are different from zero and have the same sign].*
(iii) *If the matrix game A is completely mixed, then $v(A) = \det A/$(sum of all cofactors). If, in addition, $v(A) = 0$, then $K(A) = \{q : \pi^t q \ge 0\}$ for some $\pi > 0$.*

THEOREM 3 (see Murthy, Parthasarathy, and Sriparna [11]). *Suppose $A \in \boldsymbol{R}^{m\times n}$. If the matrix game A is completely mixed with $v(A) = 0$, then $A \in \boldsymbol{Q}_0$ and $K(A) = \{q : \pi^t q \ge 0\}$ for some $\pi > 0$.*

Remark 4. It is easy to see that if A is in $\boldsymbol{P}_1$ and $\det A = 0$, then A is a completely mixed game with $v(A) = 0$. However, the converse is not true. This is not true even with an additional assumption that A is a $\boldsymbol{P}_0$-matrix. Consider the example

$$A = \begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{bmatrix}.$$

Note that A is a completely mixed game with $A \in \boldsymbol{P}_0$ but $A \notin \boldsymbol{P}_1$.

LEMMA 5. *Suppose $A \in \boldsymbol{R}^{n\times n}$, $n \geq 2$. If $A \in \boldsymbol{E'}$, then for every proper subset α of $\bar{n}$ such that* $\det A_{\alpha\alpha} = 0$, *$A_{\alpha\alpha} \notin \boldsymbol{P}_1$.*

Proof. Suppose for some proper subset α, $A_{\alpha\alpha} \in \boldsymbol{P}_1$ and $\det A_{\alpha\alpha} = 0$. Then $A_{\alpha\alpha}$ is a completely mixed game with value 0. So there exists a positive vector z_α such that $A_{\alpha\alpha}z_\alpha = 0$. Let $q_{\bar{\alpha}}$ be such that $A_{\bar{\alpha}\alpha}z_\alpha + q_{\bar{\alpha}} > 0$ and $q_{\bar{\alpha}} > 0$. Then z is a nontrivial solution of (q, A), where $z^t = (z_\alpha^t, 0_{\bar{\alpha}}^t)$ and $q = (0_\alpha^t, q_{\bar{\alpha}}^t)$. Since q is a nonzero nonnegative vector, this contradicts the assumption that $A \in \boldsymbol{E'}$. This completes the proof. □

THEOREM 6. *Suppose $A \in \boldsymbol{R}^{n\times n}$, $n \geq 2$. The following statements are equivalent:*

(i) $A \in \boldsymbol{P}_1\backslash Q$ *and* $\det A = 0$.
(ii) $A \in \boldsymbol{U}$ *and* $K(A) = \{q : \pi^t q \geq 0\}$ *for some positive vector* π.
(iii) $A \in \boldsymbol{E'} \cap \boldsymbol{P}_1$.
(iv) $A \in \boldsymbol{E'} \cap \bar{U}$.
(v) $A \in \boldsymbol{U}$ *and* A *is completely mixed with value zero.*
(vi) A^t *satisfies any one of the above conditions* (i)–(v).
(vii) A^t *satisfies all of the above conditions* (i)–(v).

Proof. Cottle and Stone [3] have already proved that (i) implies (ii).

Suppose (ii) holds. Then, as $K(A)$ is a half-space, it is convex and hence $A \in \boldsymbol{Q}_0\backslash\boldsymbol{Q}$ (see [5]). So $A \in \boldsymbol{P}_0$ (see Theorem 4 of [3]). Since $\boldsymbol{P} \subseteq \boldsymbol{Q}$, $\det A_{\alpha\alpha} = 0$ for at least one α. If possible, let α be a proper subset of $\bar{n}$ such that $\det A_{\alpha\alpha} = 0$. There exists $d_\alpha \neq 0$ such that $A_{\alpha\alpha}d_\alpha = 0$. We may assume that $e_\alpha + d_\alpha > 0$. Let $q_\alpha = -A_{\alpha\alpha}e_\alpha$. Choose $q_{\bar{\alpha}}$ to be any positive vector satisfying (*a*) $A_{\bar{\alpha}\alpha}e_\alpha + q_{\bar{\alpha}} > 0$, (*b*) $A_{\bar{\alpha}\alpha}(e_\alpha + d_\alpha) + q_{\bar{\alpha}} > 0$, and (*c*) $\pi_\alpha^t q_\alpha + \pi_{\bar{\alpha}}^t q_{\bar{\alpha}} > 0$. Let $q^t = (q_\alpha^t, q_{\bar{\alpha}}^t)$, $z^t = (z_\alpha^t, 0_{\bar{\alpha}}^t)$, and $d^t = (d_\alpha^t, 0_{\bar{\alpha}}^t)$. It is easy to check that z and $z + d$ are two distinct solutions of (q, A). However, since $\pi^t q > 0$, $q \in \operatorname{int} K(A)$, this contradicts that $A \in \boldsymbol{U}$. It follows that $\det A = 0$ and $A \in \boldsymbol{P}_1$. From the corollary of Lemma 1 of [1], $(0, A)$ has at least two solutions. Now it is easy to see that $A \in \boldsymbol{E'}$ (also see Danao [4]).

Suppose (iii) holds. Since $A \in \boldsymbol{E'}$, $A \notin \boldsymbol{R}_0$. So $A \in \boldsymbol{P}_1\backslash\boldsymbol{Q}$. Let z be a nontrivial solution of $(0, A)$. From Lemma 5, it follows that $z > 0$ and hence $Az = 0$. From Theorem 1(iii), $A \in U$. As $A \in \boldsymbol{P}_1$ and $\det A = 0$, all proper principal submatrices of A are in $\boldsymbol{P}$ and hence in $\boldsymbol{U}$. Hence $A \in \bar{\boldsymbol{U}}$.

Suppose (iv) holds. Since $A \in \boldsymbol{E'}$ and $n \geq 2$, all diagonal entries of A are positive (if $a_{11} < 0$, then $(1, 0, \ldots, 0)^t$ is a nontrivial solution of (q, A) for some nonzero nonnegative q). Now, if for some α all proper principal submatrices of $A_{\alpha\alpha}$ are all $\boldsymbol{P}$-matrices and $\det A_{\alpha\alpha} < 0$, then $(A_{\alpha\alpha})^{-1}$ is an $\boldsymbol{N}$-matrix and hence a $\boldsymbol{Q}_0$-matrix (if $A_{\alpha\alpha} \leq 0$, then obviously $(A_{\alpha\alpha})^{-1} \in \boldsymbol{Q}_0$, and if $A_{\alpha\alpha} \not\leq 0$, then from Theorem 6.6.4 of [2] $(A_{\alpha\alpha})^{-1} \in \boldsymbol{Q}$). Since $\boldsymbol{U} \cap \boldsymbol{Q}_0 \subseteq \boldsymbol{P}_0$, this would contradict the hypothesis. So $\det A_{\alpha\alpha} \geq 0$. From Lemma 5, it follows that $\det A_{\alpha\alpha} > 0$ for all proper subsets α. Since $A \notin \boldsymbol{R}_0$, it follows that $A \in \boldsymbol{P}_1$, $\det A = 0$, and, from Remark 4, A is completely mixed.

Suppose (v) holds. From Theorem 2, $\det A = 0$. From Theorem 3, $A \in \boldsymbol{Q}_0$ and $K(A) = \{q : \pi^t q \geq 0\}$ for some positive vector π. Since $A \in \boldsymbol{U} \cap \boldsymbol{Q}_0$, $A \in \boldsymbol{P}_0$. Suppose $\det A_{\alpha\alpha} = 0$ for some α with $|\alpha| < n$. Then there exists $d_\alpha \neq 0$ such that $A_{\alpha\alpha} d_\alpha = 0$. Let $q_\alpha = -A_{\alpha\alpha} e_\alpha$. Then, for $q_{\bar{\alpha}}$ positive and sufficiently large, we observe $q \in \text{int } K(A)$ and (q, A) has two solutions, that is, $(e^t_\alpha, 0^t_{\bar{\alpha}})$ and $(e^t_\alpha + \lambda d^t_\alpha, 0^t_{\bar{\alpha}})$ for some λ sufficiently small but positive. This contradicts that $A \in \boldsymbol{U}$, and thus all proper principal minors of A are positive. Since $K(A)$ is a half-space, $A \notin \boldsymbol{Q}$.

To complete the rest of the proof, we will show that (i) holds if and only if (i) holds for A^t. It suffices to show that (i) implies $A^t \notin \boldsymbol{Q}$. Suppose (i) holds. From the proof of Lemma 5 it follows that there exists $z > 0$ such that $Az = 0$. Then $z^t A^t = 0$. Note that this implies $S(q, A) = \emptyset$ for all $q < 0$. Hence $A \notin \boldsymbol{Q}$. □

3. Block property. The study of stability of solution sets in LCPs leads to the class of Lipschitzian matrices. A matrix $A \in \boldsymbol{R}^{n\times n}$ is said to be Lipschitzian if there exists a real number $\lambda > 0$ such that for every $p, q \in K(A)$ and every $z \in S(q, A)$ there exists an $x \in S(p, A)$ such that $\|x - z\| \leq \lambda \|p - q\|$. Mangasarian and Shiau [8] showed that $\boldsymbol{P}$-matrices are Lipschitzian. Gowda [6] showed that if A is Lipschitzian and (q, A) has a unique solution for some nondegenerate q (q is said to be nondegenerate if $z + Az + q > 0$ for all $z \in S(q, A)$), then A is in $\boldsymbol{P}$. Later, Murthy, Parthasarathy, and Sabatini [9] showed that a $\boldsymbol{Q}$-matrix is Lipschitzian if it is a $\boldsymbol{P}$-matrix. Stone [15] showed that Lipschitzian matrices are nondegenerate ***INS***-matrices and conjectured that the converse is also true. Murthy, Parthasarathy, and Sriparna [12] introduced the block property and showed that Lipschitzian as well as nondegenerate ***INS***-matrices satisfy the block property; they further conjectured that if a matrix has block property, then it must be Lipschitzian. In this article, we modify the definition of block property and show that the revised property is also a necessary condition for both Lipschitzian and nondegenerate ***INS***-matrices. Furthermore, we show that if a matrix A has the revised block property and (q, A) has a unique solution for some nondegenerate q, then A is Lipschitzian.

The main significance of block property (as well as the revised one) is that it can be checked in finite number of steps by examining the sign structures of all PPTs. It must be mentioned here that no finite characterization is known for checking whether a given matrix is a Lipschitzian matrix or not. Thus, if the conjecture is true, one would have a finite condition for checking the Lipschitzian property.

DEFINITION 7. *Say that $A \in \boldsymbol{R}^{n\times n}$ has block property $(B^\star)$ if for every PPT M of A there exists a partition $\alpha_1, \alpha_2, \ldots, \alpha_k$ of $\{1, 2, \ldots, n\}$ such that the following hold:*

(i) *$M_{\alpha_i\alpha_i}$, are negative $\boldsymbol{N}$-matrices for $i = 1, 2, \ldots, k-1$, $M_{\alpha_i\alpha_j} = 0$ for $i \neq j$, $i, j = 1, 2, \ldots, k-1$, and the diagonal entries of $M_{\alpha_k\alpha_k}$ are positive.*

(ii) *For any index sets α and β such that $\alpha \Delta \beta = \{i\}$ for some i, $\det M_{\alpha\alpha} \det M_{\beta\beta} > 0$ provided $\{x \in \boldsymbol{R}^{n-1}_+ : [-M_{.\alpha} I_{.\bar{\alpha}}]_{.(i)} x > 0\} \neq \emptyset$.*

THEOREM 8. *Suppose $A \in \boldsymbol{R}^{n\times n}$ is a nondegenerate ***INS***-matrix. Then A has property $(B^\star)$.*

Proof. In [12] it was shown that A satisfies condition (i) of Definition 7. From Theorem 6.6.20 of [2], it follows that M does not have any reflecting facets (see also Definition 6.2.10 and Definition 6.6.14 of [2]). From this, the second condition of the definition follows. □

COROLLARY 9. *Suppose $A \in \boldsymbol{R}^{n\times n}$ is a Lipschitzian matrix. Then A has property $(B^\star)$.*

Proof. The proof follows from the fact that Lipschitzian matrices are nondegenerate ***INS***-matrices. □

THEOREM 10. *Suppose $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{E}_0$. Assume that A has property $(B^\star)$. Then $A \in \boldsymbol{P}$.*

Proof. We will prove this by induction. Since $A \in \boldsymbol{E}_0$, $a_{ii} > 0$ for all i. Assume that $A_{\alpha\alpha} \in \boldsymbol{P}$ for all α such that $|\alpha| = k-1$, $k \geq 2$. Let β be any arbitrary index set such that $|\beta| = k$. Suppose $\det A_{\beta\beta} < 0$. Then the inverse of $A_{\beta\beta}$ is an $\boldsymbol{N}$-matrix. From (i) of property $(B^\star)$, $(A_{\beta\beta})^{-1} \leq 0$ and all its diagonal entries are negative. This would imply that $v((A_{\beta\beta})^{-1}) < 0$ and hence $v(A_{\beta\beta}) < 0$. However, this is not possible, as $A \in \boldsymbol{E}_0$ (see Definition 3.9.4 and Corollary 3.9.7 of [2]). □

THEOREM 11. *Suppose $A \in \boldsymbol{R}^{n\times n}$ has property $(B^\star)$. Assume that (q, A) has a unique solution for some nondegenerate q. Then $A \in \boldsymbol{P}$.*

Proof. We may assume without loss of generality that $q > 0$. If (p, A) has a unique solution for every positive vector p, then $A \in \boldsymbol{E}_0$ and hence $A \in \boldsymbol{P}$. Suppose, to the contrary, that (p, A) has a nontrivial solution for some $p > 0$. This means there exists a facet in the positive orthant and an open ball B of positive vectors such that for every $q' \in B \cap H^+$, (q', A) has a unique solution, and for every $q' \in B \cap H^-$, (q', A) has at least two solutions, where H^+ and H^- are the half-spaces containing the facets. Let α be the index set corresponding to the facet. Then $\alpha \neq \emptyset$. Since the facet intersects the nonnegative orthant in the interior, the system defined in condition (ii) of Definition 7 is nonempty. Condition (ii) of the definition implies for $i \notin \alpha$, $-A_{.i}$ and $I_{.i}$ are on the opposing sides of the facet in question. But then it would mean that (q', A) has at least two solutions for some $q' \in B \cap H^+$—a contradiction. It follows that (p, A) has a unique solution for every positive vector p. Hence $A \in \boldsymbol{E}_0$ and the rest follows from Theorem 10. □

From the above we have a characterization of $\boldsymbol{P}$-matrices (see [2] for characterizations of $\boldsymbol{P}$-matrices), namely, $A \in \boldsymbol{P}$ if and only if A has property $(B^\star)$ and (q, A) has a unique solution for some nondegenerate q.

Acknowledgments. The authors wish to thank Prof. O. J. Vrieze for some useful discussions and for inviting the first two authors to visit the Department of Mathematics, University of Maastricht. The first author offers special thanks to Prof. K. G. Murty, University of Michigan at Ann Arbor, for partially supporting his foreign tour.

REFERENCES

[1] M. AGANAGIC AND R. W. COTTLE, *A note on Q-matrices*, Math. Programming, 16 (1979), pp. 347–377.
[2] R. W. COTTLE, J. S. PANG, AND R. E. STONE, *The Linear Complementarity Problem*, Academic Press, Boston, 1992.
[3] R. W. COTTLE AND R. E. STONE, *On the uniqueness of solutions to linear complementarity problems*, Math. Programming, 27 (1983), pp. 191–231.
[4] R. A. DANAO, *A note on E′-matrices*, Linear Algebra Appl., 259 (1997), pp. 299–305.
[5] B. C. EAVES, *The linear complementarity problem*, Management Sci., 17 (1971), pp. 612–634.
[6] M. S. GOWDA, *On the continuity of the solution map in linear complementarity problems*, SIAM J. Optim., 2 (1992), pp. 619–634.
[7] I. KAPLANSKY, *A contribution to von Neumann's theory of games*, Ann. of Math. (2), 46 (1945), pp. 474–479.
[8] O. L. MANGASARIAN AND T.-H. SHIAU, *Lipschitz continuity of solutions of linear inequalities, programs and complementarity problems*, SIAM J. Control Optim., 25 (1987), pp. 583–595.
[9] G. S. R. MURTHY, T. PARTHASARATHY, AND M. SABATINI, *Lipschitzian Q-matrices are P-matrices*, Math. Programming, 74 (1996), pp. 55–58.

[10] G. S. R. Murthy, T. Parthasarathy, and B. Sriparna, *Constructive characterization of Lipschitzian Q_0-matrices*, Linear Algebra Appl., 252 (1997), pp. 323–337.

[11] G. S. R. Murthy, T. Parthasarathy, and B. Sriparna, *Linear complementarity problems in static and dynamic games*, Ann. Dynamic Games, to appear.

[12] G. S. R. Murthy, T. Parthasarathy, and B. Sriparna, *Some recent results on the linear complementarity problem*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 898–905.

[13] K. G. Murty, *Linear Complementarity, Linear and Nonlinear Programming*, Heldermann-Verlag, Berlin, 1988.

[14] T. D. Parsons, *Applications of principal pivoting*, in Proceeding of the Princeton Symposium on Mathematical Programming, H. W. Kuhn, ed., Princeton University Press, Princeton, NJ, 1970, pp. 561–581.

[15] R. E. Stone, *Lipschitzian matrices are nondegenerate INS-matrices*, in Complementarity and Variational Problems, C. Ferris and J. S. Pang, eds., SIAM, Philadelphia, PA, 1997, pp. 440–451.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 642–645
© 1999 Society for Industrial and Applied Mathematics

REAL SOLUTIONS OF THE EQUATION $\Phi^t(A) = \frac{1}{n}J_n$*

ZHANG XIAN†, YANG ZHONGPENG‡, AND CAO CHONGGUANG‡

Abstract. For nonsingular $n \times n$ matrices A, $\Phi(A) = A \circ A^{-T}$ ($\circ$ denotes the Hadamard product and A^{-T} the inverse transpose, $(A^{-1})^T$, of A) arises in mathematical control theory associated with chemical engineering design problems and in a matrix theoretic problem involving the relation between the diagonal entries and eigenvalues. For any positive integer t, we show that the equation

$$\Phi^t(A) = \frac{1}{n}J_n$$

has at least 2^{t-1} and 2^t mutually Φ-distinct real solutions for the case $n = 4$ and $n > 4$, respectively, and that these solutions can be determined by an inverted iteration. We also prove that the above equation has only one Φ-distinct real solution $\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ for $t = 1$ and has no real solutions for $t \geq 2$ when $n = 2$. The above results answer the problem that has been proposed by Johnson and Shapiro in [*SIAM J. Algebraic Discrete Methods*, 7 (1986), pp. 627–644]. When does the equation $\Phi(A) = A \circ A^{-T} = \frac{1}{n}J_n$ have a real solution?

Key words. Hadamard product, Φ-distinct, H–matrix

AMS subject classifications. 15A24, 93A99, 65F99

PII. S0895479896313334

1. Introduction.

The Hadamard product of two matrices $B = (b_{ij})$ and $C = (c_{ij})$ is defined by $B \circ C = (b_{ij}c_{ij})$. For a nonsingular $n \times n$ matrix A, we investigate the map

$$\Phi(A) = A \circ A^{-T},$$

where A^{-T} means the inverse transpose, $\left(A^{-1}\right)^T$, of A. The matrix $\Phi(A)$ arises in mathematical control theory in chemical engineering design problems, where it is known as the relative gain array, and also in a matrix theoretic problem involving the relation between the diagonal entries and eigenvalues. A basic question about the map Φ is to determine its range.

Let $M_n(\mathbf{R})$ and $\sigma(A)$ be the set of $n \times n$ matrices over real number field $\mathbf{R}$ and eigenvalues of matrix A, respectively. We denote by I_n and J_n the $n \times n$ identity matrix and all-one matrix, respectively.

From the definition of Φ, it is easy to see that every matrix in the range of Φ has row and column sums of 1. Johnson and Shapiro [1, p. 634], however, showed that not every matrix with row and column sums 1 lying in the range of Φ is true by the equation $\Phi(A) = \frac{1}{3}J_3$. Thus it has no real solution and they proposed the open problem: When does the equation

$$(\star) \qquad \Phi(A) = \frac{1}{n}J_n$$

*Received by the editors December 6, 1996; accepted for publication (in revised form) by R. Brualdi February 12, 1999; published electronically December 28, 1999. This research was supported by NSF of China and NSF of Heilongjiang Province.
http://www.siam.org/journals/simax/21-2/31333.html

†Department of Mathematics, Heilongjiang University, Harbin 150080, People's Republic of China.
‡Department of Mathematics, Jilin Teacher's College, Jilin 132013, People's Republic of China.

have a real solution? In this paper we study the more general problem: the existence of real solutions of the equation

$$(\star\star) \qquad \Phi^t(A) = \frac{1}{n}J_n$$

for any positive integer t, where Φ^t is the map Φ applied t times. Obviously, the above open problem is only a special case when $t = 1$.

Johnson and Shapiro showed that $(\star)$ has no real solutions when $n = 3$ (see [1, Example 2]), and hence $(\star\star)$ also has no real solutions. Hence we consider the case $n \neq 3$.

The main results are given in sections 2 and 3 for the case $n \geq 4$ and $n = 2$, respectively. An open problem is proposed in section 4.

2. The case $n \geq 4$. In this section, we assume that $n \geq 4$.

LEMMA 1 [1, Observations 2 and 4]. *For every nonsigular $A \in M_n(\mathbf{R})$,*

(i) *if D and E are $n \times n$ diagonal, nonsingular matrices, then $\Phi(DAE) = \Phi(A)$;*

(ii) *if P and Q are $n \times n$ permutation matrices, then $\Phi(PAQ) = P\Phi(A)Q$.*

DEFINITION 1. *Suppose A and B in $M_n(\mathbf{R})$. If there exist diagonal, nonsingular matrices D and E such that $A = DBE$, then A and B are said to be diagonally equivalent.*

LEMMA 2. *For $\lambda \neq 0$, let $A(\lambda) = \frac{1}{\lambda}[(\lambda + n)I_n - J_n]$.*

(i) *The matrix $A(\lambda)$ is nonsingular if $\lambda \neq -n$. When $\lambda = -n$ we get $A(-n) = \frac{1}{n}J_n$.*

(ii) *If $\lambda \neq -n$, we have $\Phi(A(\lambda)) = A(\mu)$, where $\mu = \lambda(\lambda + n)$.*

(iii) *If $\alpha \neq \beta$, then $A(\alpha)$ and $A(\beta)$ are not diagonally equivalent.*

Proof. (i) Let $e = (1, \ldots, 1)^T$. Then $\sigma(J_n) = \{n, 0, \ldots, 0\}$ because ee^T and e^Te have the same nonzero eigenvalues (see [2, Theorem 1.3.20]), and hence $\sigma(A(\lambda)) = \{1, 1 + \frac{n}{\lambda}, \ldots, 1 + \frac{n}{\lambda}\}$. Thus $A(\lambda)$ is nonsingular if $\lambda \neq -n$. It is easy to see that $A(-n) = \frac{1}{n}J_n$ from the definition of $A(\lambda)$.

(ii) Since $\lambda \neq -n$, $A(\lambda)$ is nonsingular from part (i) of the lemma. Thus

$$\Phi(A(\lambda)) = A(\lambda) \circ A(\lambda)^{-T} = \left\{\frac{1}{\lambda}[(\lambda + n)I_n - J_n]\right\} \circ \left\{\lambda[\frac{1}{\lambda + n}I_n + \frac{1}{\lambda(\lambda + n)}J_n]\right\}.$$

By a direct computation, we obtain $\Phi(A(\lambda)) = A(\mu)$, where $\mu = \lambda(\lambda + n)$.

(iii) Suppose there exist nonsingular matrices $D = \operatorname{diag}(d_1, \ldots, d_n)$ and $E = \operatorname{diag}(e_1, \ldots, e_n)$ such that $A(\alpha) = DA(\beta)E$. Then $d_ie_j = \frac{\beta}{\alpha}$ for any $i \neq j$ and $d_ie_i = \frac{\beta(\alpha+n-1)}{\alpha(\beta+n-1)}$ for any i. This is impossible.

The lemma follows.

LEMMA 3. *Let $f(\lambda) = \lambda(\lambda + n)$, and let t be a positive integer. If $n > 4$, in the interval $(-\frac{n^2}{4}, 0]$ there are 2^t distinct, real solutions to the equation $f^t(\lambda) = -n$, where f^t is the map f applied t times. If $n = 4$, in the interval $(-\frac{n^2}{4}, 0]$ there are 2^{t-1} distinct, real solutions to the equation $f^t(\lambda) = -n$.*

Proof. We only prove the lemma by induction on t when $n > 4$ (because the proof is similar when $n = 4$). $t = 1$ being obvious, we may assume the result for $t - 1$ $(t \geq 2)$.

Let $\mu = f(\lambda)$. Then $f^t(\lambda) = f^{t-1}(\mu)$. By the inductive hypothesis, in the interval $(-\frac{n^2}{4}, 0]$ there are 2^{t-1} real, distinct solutions to the equation $f^{t-1}(\mu) = -n$ (i.e., there are mutually distinct $\mu_1, \mu_2, \ldots, \mu_{2^{t-1}} \in (-\frac{n^2}{4}, 0]$ such that $f^{t-1}(\mu_i) = -n$).

For $i = 1, 2, \ldots, 2^{t-1}$, the solutions to the equation $f(\lambda) = \mu_i$ are

$$\lambda_{i,1} = \frac{-n + \sqrt{n^2 + 4\mu_i}}{2} \quad \text{and} \quad \lambda_{i,-1} = \frac{-n - \sqrt{n^2 + 4\mu_i}}{2}. \tag{1}$$

Obviously, $\lambda_{1,1}, \lambda_{1,-1}, \lambda_{2,1}, \lambda_{2,-1}, \ldots, \lambda_{2^{t-1},1}$ and $\lambda_{2^{t-1},-1}$ are 2^t real solutions, in the interval $(-\frac{n^2}{4}, 0]$, to the equation $f^t(\lambda) = -n$. Now we will show that $\lambda_{1,1}$, $\lambda_{1,-1}$, $\lambda_{2,1}$, $\lambda_{2,-1}, \ldots, \lambda_{2^{t-1},1}$ and $\lambda_{2^{t-1},-1}$ are mutually distinct.

Case 1. Suppose there exist $p, q \in \{1, -1\}$ and distinct $i, j \in \{1, 2, \ldots, 2^{t-1}\}$ such that $\lambda_{i,p} = \lambda_{j,q}$. Then $\mu_i = \mu_j$ from (1), which contradicts $\mu_i \neq \mu_j$.

Case 2. Suppose $\lambda_{i,1} = \lambda_{i,-1}$ for some $i \in \{1, 2, \ldots, 2^{t-1}\}$. Then $\mu_i = -\frac{n^2}{4}$ from (1), which contradicts $\mu_i \in (-\frac{n^2}{4}, 0]$.

Summarizing, $\lambda_{1,1}, \lambda_{1,-1}, \lambda_{2,1}, \lambda_{2,-1}, \ldots, \lambda_{2^{t-1},1}$ and $\lambda_{2^{t-1},-1}$, in the interval $(-\frac{n^2}{4}, 0]$, are 2^t distinct, real solutions to the equation $f^t(\lambda) = -n$. The lemma follows.

THEOREM 1. *Let t be a positive integer. Then if $n > 4$, there are 2^t distinct, real values of λ such that $\Phi^t(A(\lambda)) = \frac{1}{n}J_n$ and hence the equation $\Phi^t(A) = \frac{1}{n}J_n$ has at least 2^t nondiagonally equivalent solutions. When $n = 4$, there are 2^{t-1} distinct, real values of λ such that $\Phi^t(A(\lambda)) = \frac{1}{4}J_4$ and hence the equation $\Phi^t(A) = \frac{1}{4}J_4$ has at least 2^{t-1} nondiagonally equivalent solutions.*

Proof. If $\lambda \neq -n$, part (ii) of Lemma 2 tells us that $\Phi(A(\lambda)) = A(f(\lambda))$. Thus, $\Phi^t(A(\lambda)) = A(f^t(\lambda))$ and the result follows from Lemma 3 and parts (i) and (iii) of Lemma 2.

Remark 1. *We can find 2^t and 2^{t-1} mutually nondiagonally equivalent real solutions $A(\lambda_0)$ to the equation $\Phi^t(A) = \frac{1}{n}J_n$ for the case $n > 4$ and $n = 4$, respectively, where λ_0 satisfies the following inverted iteration:*

$$\begin{cases} \lambda_t = -n, \\ \lambda_k = \frac{-n \pm \sqrt{n^2 + 4\lambda_{k+1}}}{2}, \quad k = 0, 1, \ldots, t-1. \end{cases} \tag{2}$$

In particular, the equation $\Phi(A) = \frac{1}{n}J_n$ has at least two nondiagonally equivalent real solutions $\frac{2}{-n+\epsilon\sqrt{n^2-4n}}(\frac{-n+\epsilon\sqrt{n^2-4n}}{2}I_n - J_n)$ if $n > 4$, where $\epsilon = \pm 1$. When $n = 4$, the equation $\Phi(A) = \frac{1}{4}J_4$ has at least one nondiagonally equivalent real solution $-I_4 + \frac{1}{2}J_4$.

Proof. We consider only the case $n > 4$. (The proof is similar for $n = 4$.) From the proof of Lemma 3, it is easy to see that the equation $f^t(\lambda) = -n$ has 2^t distinct real solutions and these solutions can be obtained from the inverted iteration (2). Again applying $\Phi^t(A(\lambda)) = A(f^t(\lambda))$ and parts (i) and (iii) of Lemma 2, we can obtain 2^t mutually distinct real solutions $A(\lambda_0)$ to the equation $\Phi^t(A) = \frac{1}{n}J_n$, where λ_0 follows from (2). In particular, if $t = 1$, $A(\lambda_0) = -I_4 + \frac{1}{2}J_4$.

3. The case $n = 2$. In this section, we assume that $n = 2$.

THEOREM 2. *Let $A = (a_{ij}) \in M_2(\mathbf{R})$. Then A is a real solution to the equation $\Phi(A) = \frac{1}{2}J_2$ if and only if A and $\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ are diagonally equivalent. Hence, there is only one nondiagonally equivalent solution to the equation $\Phi(A) = \frac{1}{2}J_2$.*

Proof. A calculation shows that if $A \in M_2(\mathbf{R})$, then $\Phi(A) = \frac{1}{2}J_2$ if and only if $a_{11}a_{22} = -a_{12}a_{21} \neq 0$. Let $D = \operatorname{diag}(1, \frac{a_{11}}{a_{21}})$ and $E = \operatorname{diag}(\frac{1}{a_{11}}, \frac{1}{a_{12}})$. Then $DAE = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$; i.e., A and $\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ are diagonally equivalent.

The converse is evident by applying part (i) of Lemma 1.

THEOREM 3. *Suppose $t \geq 2$. Then the equation $\Phi^t(A) = \frac{1}{2}J_2$ has no real solutions.*

Proof. Suppose $A = (a_{ij}) \in M_2(\mathbf{R})$ is a real solution of the equation $\Phi^t(A) = \frac{1}{2}J_2$. Then $\Phi(A)$ is nonsingular, and hence $a_{11}a_{22} \neq -a_{12}a_{21}$. Let $B = \Phi(A)$.

Case 1. If $|a_{11}a_{22}| > |a_{12}a_{21}|$, then B is an H-matrix. By applying Theorem 2 in [1], we have $\lim_{k\to\infty} \Phi^k(A) = \lim_{k\to\infty} \Phi^{k-1}(B) = I_2$. This contradicts $\Phi^t(A) = \frac{1}{2}J_2$.

Case 2. If $|a_{11}a_{22}| < |a_{12}a_{21}|$, then BP is an H-matrix, where $P = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. Hence $\lim_{k\to\infty} \Phi^k(BP) = I_2$ from Theorem 2 in [1]. Again applying part (ii) of Lemma 1, we obtain $\lim_{k\to\infty} \Phi^k(B) = P$. Further, $\lim_{k\to\infty} \Phi^k(A) = \lim_{k\to\infty} \Phi^{k-1}(B) = P$. This contradicts $\Phi^t(A) = \frac{1}{2}J_2$.

Remark 2. From Theorems 1 and 2, we obtain that the equation $\Phi(A) = \frac{1}{n}J_n$ has at least one real solution for $n \neq 3$. This answers the open problem that has been proposed by Johnson and Shapiro [1].

4. An open problem. Are there real solutions to the equation $\Phi^t(A) = \frac{1}{n}J_n$, which are not diagonally equivalent to those found by the authors when $n \geq 4$?

Acknowledgment. The author wishes to thank the referee for providing helpful reports.

REFERENCES

[1] C. R. JOHNSON AND H. M. SHAPIRO, *Mathematical aspects of the relative gain array* ($A \circ A^{-T}$), SIAM J. Algebraic Discrete Methods, 7 (1986), pp. 627–644.

[2] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, London, New York, 1985.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 646–667
© 2000 Society for Industrial and Applied Mathematics

MATRIX ALGEBRAS AND DISPLACEMENT DECOMPOSITIONS*

CARMINE DI FIORE†

Abstract. A class ξ of algebras of symmetric $n \times n$ matrices, related to Toeplitz-plus-Hankel structures and including the well-known algebra $\mathcal{H}$ diagonalized by the Hartley transform, is investigated. The algebras of ξ are then exploited in a general displacement decomposition of an arbitrary $n \times n$ matrix A. Any algebra of ξ is a 1-*space*, i.e., it is spanned by n matrices having as first rows the vectors of the canonical basis. The notion of 1-space (which generalizes the previous notions of $\mathcal{L}_1$ space [Bevilacqua and Zellini, *Linear and Multilinear Algebra*, 25 (1989), pp. 1–25] and Hessenberg algebra [Di Fiore and Zellini, *Linear Algebra Appl.*, 229 (1995), pp. 49–99]) finally leads to the identification in ξ of three new (non-Hessenberg) matrix algebras close to $\mathcal{H}$, which are shown to be associated with fast *Hartley-type* transforms. These algebras are also involved in new efficient centrosymmetric Toeplitz-plus-Hankel inversion formulas.

Key words. matrix algebras, displacement rank, Toeplitz-plus-Hankel matrices, inversion formulas, discrete Fourier transform, discrete Hartley transform

AMS subject classifications. 15A09, 65F05, 65T20

PII. S0895479896300747

1. Introduction.

It is well known that the inverse of any nonsingular Toeplitz matrix $T = (t_{i-j})_{i,j=1}^{n}$ can be represented using lower and upper triangular Toeplitz matrices L_m, U_m via the Gohberg–Semencul formula $T^{-1} = L_1U_1 + L_2U_2$ [23]. Kailath, Kung, and Morf [28] extended this result by showing that any $n \times n$ matrix A can be decomposed as

$$A = \sum_{m=1}^{\alpha} L_m U_m \tag{1.1}$$

with α equal to the *displacement rank* of A, i.e., $\alpha = \operatorname{rank}(A - ZAZ^T)$, where $Z = (\delta_{i,\,j+1})_{i,\,j=1}^{n}$. On the basis of the ideas introduced in [28], different fast algorithms for the inversion or the factorization of structured matrices such as Toeplitz-like [27, 31, 33], Cauchy-like [19, 24], and polynomial Vandermonde-like matrices [29, 30] have been developed (see also [7, 25, 31]).

Besides the triangular Toeplitz used in [23, 28], other algebras have been exploited in *displacement formulas* of type (1.1), for example, ε-circulant [1, 18, 20], τ algebra [6, 16, 32], and algebras of dimension greater than n [8, 9]. In [16], most of these algebras appear as special instances of Hessenberg algebras, which allows one to regain the known displacement formulas in a more general context and to obtain new decompositions of high efficiency (especially if A is the inverse of a Toeplitz-plus-Hankel matrix) [16, 10, 17].

If A is a *Toeplitz-like* matrix, that is, A has a *small* displacement rank α, then the known displacement formulas let one compute the matrix-vector product $A\mathbf{f}$, $\mathbf{f} \in \mathbb{C}^n$, by means of a small number of fast discrete transforms (assuming preprocessing on A). These transforms are discrete Fourier transforms (DFT) in cases of formulas

*Received by the editors March 18, 1996; accepted for publication (in revised form) by R. Freund May 17, 1998; published electronically January 19, 2000. This work was partially supported by the Istituto Nazionale di Alta Matematica, Cittá Universitaria, Rome, Italy.

http://www.siam.org/journals/simax/21-2/30074.html

†Dipartimento di Matematica, Università di Roma "Tor Vergata," Via della Ricerca Scientifica, 00133 Rome, Italy.

involving triangular Toeplitz or ε-circulant matrices [1, 9, 17, 21, 22] and are sine or cosine transforms in cases of formulas involving τ or $\tau_{\varepsilon,\varphi}$ matrices [6, 10, 16, 17, 32], and therefore they are all associated with Hessenberg algebras [16].

In this paper we further extend the results of [6, 9, 10, 16, 17, 18, 20, 21, 22, 28, 32] in the sense that we introduce a new class of matrix algebras $\mathbb{L}$, including Hessenberg and other algebras of matrices diagonalized by means of Hartley [11, 12] or *Hartley-type* transforms, which have not been yet considered in displacement literature. This extension requires the study of matrix algebras containing the matrix $T^{\beta,\beta}_{\varepsilon,\varphi}$ displayed at the beginning of section 2. Notice that the algebra $\mathcal{H}$ of the matrices diagonalized by the Hartley transform (see [5]) contains the matrix $T^{1,1}_{0,0}$. The appropriate mechanism for capturing algebras $\mathbb{L}$ such that $\mathbb{L} \supset T^{\beta,\beta}_{\varepsilon,\varphi}$, which are generally not Hessenberg, is the notion of 1-*space* (which is an extension of the notions of $\mathcal{L}_1$ space [4] and Hessenberg algebra [16]).

A 1-space is a space of $n \times n$ matrices A spanned by n matrices J_k having as first rows the vectors of the canonical basis of $\mathbb{C}^n$. If $[z_1 z_2 \cdots z_n]$ is the first row of A, then each a_{ij} is a linear combination in $\mathbb{C}$ of $z_1, z_2, \ldots, z_n$. Any space of matrices simultaneously diagonalized by a nonsingular matrix M whose first row has all nonzero entries can be easily checked to be a 1-space. This is the main reason why the introduction of 1-spaces allows one to extend the range of algebras which could be used, in principle, in (possibly) efficient displacement formulas. In particular, the algebra $\mathcal{H}$ diagonalized by the Hartley transform [5] is a 1-space even though it is not a Hessenberg algebra.

The results of this paper are now described in detail.

In section 2 we state some properties of commutative 1-spaces used throughout the paper. Then we define a class of symmetric 1-spaces $\xi(\varphi, \beta, \mathbf{p})$, $\varphi, \beta \in \mathbb{C}$, $\mathbf{p} \in \mathbb{C}^n - 1$ in terms of matrices of different dimensions from the algebra τ (τ is the algebra generated by $T^{0,0}_{0,0}$). The main result of section 2 is Theorem 2.5, where the symmetric 1-algebras (closed 1-spaces), including the matrix $T^{\beta,\beta}_{\varepsilon,\varphi}$, are shown to be the spaces $\xi(\varphi, \beta, \mathbf{p})$ with $\mathbf{p}$ running among the solutions of a linear system with coefficients depending upon φ, β, and ε.

In section 3 a *general* displacement formula for a matrix A in terms of 2α matrices from two arbitrary symmetric 1-algebras $\mathbb{L} \supset T^{\beta,\beta}_{\varepsilon,\varphi}$ and $\mathbb{L}' \supset T^{\beta',\beta'}_{\varepsilon',\varphi'}$ is obtained under the assumption that the rank of $AT^{\beta,\beta}_{\varepsilon,\varphi} - T^{\beta,\beta}_{\varepsilon,\varphi}A$ is α (see Theorem 3.2). This formula extends some formulas of [10] to the case of non-Hessenberg algebras.

In sections 4 and 5 the results of Theorems 2.5 and 3.2 are investigated and specialized. In particular it is shown that the Hartley algebra $\mathcal{H}$ introduced in [5] is an element of the class of 1-algebras ξ characterized in Theorem 2.5 and that there are at least three other algebras of ξ, called η, μ, and $\mathcal{K}$, which are associated with fast Hartley-type discrete transforms (see Theorem 5.2 and the following remark). Moreover, new decompositions of the inverse of an arbitrary centrosymmetric Toeplitz-plus-Hankel matrix $T + H = (t_{i-j} + h_{i+j} - 2)^n_{i,\,j} = 1$ in terms of matrices from $\mathcal{H}$, $\mathcal{K}$, η, and μ are obtained. In particular it is shown that there exist $\mathbf{a}, \mathbf{b} \in \mathbb{C}^n$ such that

$$(T + H)^{-1} = [\mu(\mathbf{a}) + I]\eta(\mathbf{b}) - \mu(\mathbf{b})[\eta(\mathbf{a}) - I]. \tag{1.2}$$

(Here $\mathbb{L}(\mathbf{z})$ denotes the matrix of $\mathbb{L}$ whose first row is $\mathbf{z}^T$.) Under the assumption that the vectors $\mathbf{a}$ and $\mathbf{b}$ are known, formula (1.2) lets one calculate the matrix-vector product $(T+H)^{-1}\mathbf{f}$, $\mathbf{f} \in \mathbb{C}^n$, by means of 10 fast discrete transforms reducible to 8 in case $H = 0$, $[T^{-1}]_{11} \neq 0$, matching both best limits known so far [1, 10, 16]. In any case, the number of transforms reduces to 6 (as in [1, 10, 16, 21, 22]) if the transforms

of vectors not depending upon $\mathbf{f}$ are included in the preprocessing stage, where $\mathbf{a}$ and $\mathbf{b}$ are computed.

2. A class of algebras of symmetric matrices. The main result of this section (Theorem 2.5) is a characterization of all spaces $\mathbb{L}$ of $n \times n$ matrices containing the matrix

$$T_{\varepsilon,\varphi}^{\beta,\beta} = \begin{pmatrix} \varepsilon & 1 & 0 & \cdot & \cdot & 0 & \beta \\ 1 & 0 & 1 & . & & & 0 \\ 0 & 1 & . & . & . & & \cdot \\ . & . & . & . & . & . & \cdot \\ . & & . & . & . & 1 & 0 \\ 0 & & & . & 1 & 0 & 1 \\ \beta & 0 & \cdot & \cdot & 0 & 1 & \varphi \end{pmatrix}, \quad \varepsilon, \varphi, \beta \in \mathbb{C} \tag{2.1}$$

and satisfying the following three properties: $A = A^T$, $\forall\, A \in \mathbb{L}$; $AB \in \mathbb{L}$, $\forall\, A,\, B \in \mathbb{L}$; $\mathbb{L}$ is a 1-space (see Definition 2.1). Notice that the properties of symmetry and closure imply the commutativity of $\mathbb{L}$. Moreover, requiring $\mathbb{L}$ to be a 1-space essentially means that any matrix of $\mathbb{L}$ is determined once its first row is given.

The interest of matrix algebras including $T_{\varepsilon,\varphi}^{\beta,\beta}$ and of possible *displacement* decompositions involving them (see sections 3 and 4) is in the fact that for a Toeplitz-plus-Hankel matrix $T + H$, $[T + H]_{ij} = t_{i-j} + h_{i+j-2}$, $i, j = 1, \ldots, n$, the rank of $(T + H)T_{\varepsilon,\varphi}^{\beta,\beta} - T_{\varepsilon,\varphi}^{\beta,\beta}(T + H)$ is 4 for all values of ε, φ, β (see [26] for the case $\varepsilon = \varphi = \beta = 0$). In section 5, this fact finally leads to efficient inversion formulas for $T + H$ involving Hartley-type matrix algebras. The appropriate mechanism with which to capture algebras including $T_{\varepsilon,\varphi}^{\beta,\beta}$ is the notion of 1-space introduced below.

Let $M_n(\mathbb{C})$ be the space of $n \times n$ matrices with entries in the complex field $\mathbb{C}$ and let $\mathbf{e}_k$, $k = 1, \ldots, n$, be the vectors of $\mathbb{C}^n$ $\mathbf{e}_k = [0 \cdots 0 \underset{k}{1} 0 \cdots 0]^T$.

DEFINITION 2.1. *A subset $\mathbb{L}$ of $M_n(\mathbb{C})$ is a 1-space if there exist n $n \times n$ matrices $J_{\mathbf{k}} \in \mathbb{L}$, $k = 1, \ldots, n$, such that $\mathbb{L} = \{\sum_{k=1}^{n} a_k J_k : a_k \in \mathbb{C}\}$ and*

$$\mathbf{e}_1^T J_k = \mathbf{e}_k^T, \quad k = 1, \ldots, n.$$

Closed (under matrix multiplication) 1-spaces are also called 1-algebras.

Many significant classes of spaces of matrices have 1-space structure. Some examples are the group (or, more generally, hypergroup) matrix algebras [18, 3] and the intersection algebras of the association schemes [2, pp. 52–57]; a simple example is the space of all symmetric Toeplitz matrices (which is not a matrix algebra).

Moreover, every space $H_X = \{\sum_{k=1}^{n} a_k X^{k-1} : a_k \in \mathbb{C}\}$, where X is an $n \times n$ lower Hessenberg matrix, is a 1-space if the entries $[X]_{i,\, i+1}$ are all nonzero. In this case we also have that $H_X = \{A \in M_n(\mathbb{C}) : AX = XA\}$ because X is nonderogatory. In [16] H_X is called Hessenberg algebra (HA) and, for $\mathbf{z} = [z_1 \cdots z_n]^T \in \mathbb{C}^n$, $H_X(\mathbf{z})$ denotes the matrix of H_X whose first row is $\mathbf{z}^T$. For our purposes it is useful to recall the HAs corresponding to the choices $X = T_{\varepsilon,\varphi}$ and $X = P_\beta$, where

$$T_{\varepsilon,\varphi} = \begin{pmatrix} \varepsilon & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ 1 & 0 & 1 & . & & & . \\ 0 & 1 & . & . & . & & . \\ \cdot & \cdot & . & . & . & . & . \\ . & & . & . & . & 1 & 0 \\ & & & . & 1 & 0 & 1 \\ 0 & \cdot & \cdot & \cdot & 0 & 1 & \varphi \end{pmatrix} \quad \text{and} \quad P_\beta = \begin{pmatrix} 0 & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ \cdot & 0 & 1 & . & & & . \\ \cdot & & \cdot & \cdot & \cdot & & . \\ \cdot & & & \cdot & \cdot & . & . \\ \cdot & & & & . & 1 & 0 \\ 0 & & & & & 0 & 1 \\ \beta & 0 & \cdot & \cdot & \cdot & \cdot & 0 \end{pmatrix}. \tag{2.2}$$

These HAs are denoted, respectively, by $\tau_{\varepsilon,\varphi}$ and C_β in conformity with [10, 16, 17, 20]. In fact the (non-Hessenberg) algebras containing $T^{\beta,\beta}_{\varepsilon,\varphi}$ studied in Theorem 2.5 and in section 4 are defined in terms of matrices from $\tau_{\varepsilon,\varphi}$ and C_β. Notice that the matrices of $\tau_{\varepsilon,\varphi}$ and of C_β are, respectively, symmetric and persymmetric, in particular $C_\beta(\mathbf{z}) = \sum_{i=1}^n z_i P_\beta^{i-1}$. C_β is the space of β-circulant matrices, and $C = C_1$ is the well-known space of circulant matrices [14].

Finally, observe that any space $\mathbb{L}$ defined as the set of all matrices diagonalized by a nonsingular matrix M is a 1-space if $[M]_{1,i} \neq 0\ \forall i$, because, in this case, $\mathbb{L} = \{Md(M^T\mathbf{z})d(M^T\mathbf{e}_1)^{-1}M^{-1} : \mathbf{z} \in \mathbb{C}^n\}$, where for $\mathbf{z} \in \mathbb{C}^n$ $d(\mathbf{z}) = \mathrm{diag}(z_i, i = 1, \ldots, n)$. As a consequence, the algebra $\mathcal{H}$ diagonalized by the Hartley transform (see [5]) is a 1-space even though it is not an HA. Recall that matrices from $\mathcal{H}$ are symmetric and that $\mathcal{H}$ contains the matrix $T^{1,1}_{0,0}$. Thus $\mathcal{H}$ is an example of a symmetric 1-algebra including $T^{\beta,\beta}_{\varepsilon,\varphi}$ for $\beta \neq 0$.

Following the notation used for HAs, if $\mathbb{L}$ is a 1-space and $\mathbf{z} \in \mathbb{C}^n$, $\mathbb{L}(\mathbf{z})$ denotes the matrix of $\mathbb{L}$ whose first row is $\mathbf{z}^T$, i.e., $\mathbb{L}(\mathbf{z}) = \sum_{i=1}^n z_i J_i$, where J_i are the matrices in Definition 2.1. Notice that $A \in \mathbb{L}$ iff $A = \mathbb{L}(A^T\mathbf{e}_1)$.

PROPOSITION 2.2. *Let $\mathbb{L}$ be a commutative 1-space. Then*

(i) *$\mathbb{L}$ is closed under matrix multiplication and $I \in \mathbb{L}$;*

(ii) $\mathbf{x}^T\mathbb{L}(\mathbf{y}) = \mathbf{y}^T\mathbb{L}(\mathbf{x}) \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{C}^n$;

(iii) $\mathbb{L}(\mathbb{L}(\mathbf{x})^T\mathbf{y}) = \mathbb{L}(\mathbf{y})\mathbb{L}(\mathbf{x}) \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{C}^n$.

Proof. As $J_kJ_s = J_sJ_k\ \forall s, k$, we have that $\mathbf{e}_k^T J_s = \mathbf{e}_s^T J_k\ \forall s, k$. Consequently, $J_1 \equiv \mathbb{L}(\mathbf{e}_1)$ is the identity matrix I. Moreover, for all i, j, $\mathbf{e}_i^T(\sum_{r=1}^n [J_s]_{kr} J_r)\mathbf{e}_j = \sum_{r=1}^n [J_s]_{kr}[J_r]_{ij} = \sum_{r=1}^n [J_s]_{kr}[J_i]_{rj} = [J_sJ_i]_{kj} = [J_iJ_s]_{kj} = [J_kJ_s]_{ij}$ and thus

$$J_kJ_s = \sum_{r=1}^n [J_s]_{kr} J_r \quad \forall\, s, k,$$

that is, assertion (i) holds. For (iii) observe that, by (i), both $\mathbb{L}(\mathbf{y})\mathbb{L}(\mathbf{x})$ and $\mathbb{L}(\mathbb{L}(\mathbf{x})^T\mathbf{y})$ are in $\mathbb{L}$ and have $\mathbf{y}^T\mathbb{L}(\mathbf{x})$ as first row. Finally, for (ii) use (iii) and the commutativity of $\mathbb{L}$. ∎

Proposition 2.2 and the following notation are used throughout the paper. The symbol I^i_j, $1 \leq i,\ j \leq n$, denotes the $(|j - i| + 1) \times n$ $(0, 1)$ matrix, which maps a vector $\mathbf{z} = [z_1 \cdots z_n]^T \in \mathbb{C}^n$ into the vector $I^i_j\mathbf{z} = [z_i \cdots z_j]^T \in \mathbb{C}^{|j-i|+1}$. Thus $I = I^1_n$ and $J = I^n_1$ are, respectively, the $n \times n$ identity and the reversion matrix. I and J also denote, respectively, identity and reversion matrices of dimensions different from n. Also, set $e_k = I^1_{n-1}\mathbf{e}_k$, $k = 1, \ldots, n-1$, and $\hat{\mathbf{z}} = [z_k \cdots z_1]^T = J\mathbf{z}$ if $\mathbf{z} \in \mathbb{C}^k$.

Now we state Theorem 2.5, where the symmetric closed 1-spaces containing $T^{\beta,\beta}_{\varepsilon,\varphi}$ are shown to be the spaces $\xi(\varphi, \beta, \mathbf{p})$ in Definition 2.4 obtained by choosing as $\mathbf{p}$ the solutions of (2.6). As a consequence (see section 4) for given ε, φ, β, there are as many symmetric 1-algebras including $T^{\beta,\beta}_{\varepsilon,\varphi}$ as the solutions of equation (2.6), i.e., none, an infinite number, or only one, depending upon the values of ε, φ, β. A preliminary Lemma 2.3 follows.

LEMMA 2.3. (i) *Let A be an $n \times n$ matrix and $\mathbf{x}_m$ and $\mathbf{y}_m$, $m = 1, \ldots, \alpha$, vectors of $\mathbb{C}^n$ such that $AT_{\varepsilon,0} - T_{\varepsilon,0}A = \sum_{m=1}^\alpha \mathbf{x}_m\mathbf{y}_m^T$. Then*

$$A = \sum_{m=1}^{\alpha} \begin{pmatrix} 0 & \cdots\cdots & 0 \\ \vdots & & \\ \vdots & \tau(I^2_n\mathbf{x}_m) & \\ 0 & & \end{pmatrix} \Omega_\varepsilon(\mathbf{y}_m) + \Omega_\varepsilon(A^T\mathbf{e}_1), \tag{2.3}$$

where $\tau = \tau_{0,0}$ and $\Omega_\varepsilon = \tau_{\varepsilon,0}$.

(ii) *In particular, for* $\mathbf{z} \in \mathbb{C}^n$,

$$\Omega_\varepsilon(\mathbf{z}) = \tau(\mathbf{z}) - \varepsilon \begin{pmatrix} 0 & \cdot & \cdot & \cdot & 0 \\ \vdots & & & & \\ \vdots & & \tau(I_n^2\mathbf{z}) & & \\ 0 & & & & \end{pmatrix}. \tag{2.4}$$

Proof. For (i) see [16]. (ii) follows from the identities

$$\tau(\mathbf{z})T_{\varepsilon,0} - T_{\varepsilon,0}\tau(\mathbf{z}) = \tau(\mathbf{z})T_{0,0} - T_{0,0}\tau(\mathbf{z}) + \varepsilon[\tau(\mathbf{z})\mathbf{e}_1\mathbf{e}_1^T - \mathbf{e}_1\mathbf{e}_1^T\tau(\mathbf{z})] = \varepsilon(\mathbf{z}\mathbf{e}_1^T - \mathbf{e}_1\mathbf{z}^T)$$

and from assertion (i) for $A = \tau(\mathbf{z})$. □

DEFINITION 2.4. *For* φ, $\beta \in \mathbb{C}$, $\mathbf{p} \in \mathbb{C}^{n-1}$, *define the space of* $n \times n$ *matrices*

$$\begin{aligned} \xi &\equiv \xi(\varphi, \beta, \mathbf{p}) \\ &= \left\{ \tau(\mathbf{z}) - \begin{pmatrix} 0 & \cdot & \cdot & \cdot & 0 \\ & \vdots\, \tau(I_{n-1}^2\mathbf{z})\, \vdots & & & \\ 0 & \cdot & \cdot & \cdot & 0 \end{pmatrix} (\varphi I + \beta J) + \begin{pmatrix} 0 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ \vdots & & & & & & \\ 0 & & & J\Omega_\varphi(I_2^n\mathbf{z})\Omega_\varphi(\mathbf{p})\, J & & & \end{pmatrix} : \mathbf{z} \in \mathbb{C}^n \right\} \end{aligned} \tag{2.5}$$

and denote by $\xi(\mathbf{z})$ *the matrix of* ξ *whose first row is* $\mathbf{z}^T$.

THEOREM 2.5. *If* $\mathbb{L}$ *is a symmetric closed* 1*-space containing the matrix* $T_{\varepsilon,\varphi}^{\beta,\beta}$ *for some* ε, φ, $\beta \in \mathbb{C}$, *then* $\mathbb{L} = \xi(\varphi, \beta, \mathbf{p})$ *with* $\mathbf{p}$ *such that*

$$\Omega_\varphi(\beta e_1 + e_{n-1})\mathbf{p} = (\varphi - \varepsilon)e_1. \tag{2.6}$$

Conversely, every space of matrices $\xi(\varphi, \beta, \mathbf{p})$ *with* $\mathbf{p}$ *solving* (2.6) *for some* $\varepsilon \in \mathbb{C}$ *is a symmetric closed* 1*-space containing the matrix* $T_{\varepsilon,\varphi}^{\beta,\beta}$*; moreover,* $\xi(\varphi, \beta, \mathbf{p}) = \{A \in M_n(\mathbb{C}) : AT_{\varepsilon,\varphi}^{\beta,\beta} = T_{\varepsilon,\varphi}^{\beta,\beta}A$ *and* $A\xi(\mathbf{e}_n) = \xi(\mathbf{e}_n)A\}$.

Proof. Let $\mathbb{L}$ be a symmetric closed 1-space containing the matrix $T_{\varepsilon,\varphi}^{\beta,\beta}$ and let A be an arbitrary element of $\mathbb{L}$. Notice that $AT_{\varepsilon,\varphi}^{\beta,\beta} = T_{\varepsilon,\varphi}^{\beta,\beta}A$ and therefore $I_n^2A\mathbf{e}_1(e_1 + \beta e_{n-1})^T + BT_{0,\varphi}^{()} = (e_1 + \beta e_{n-1})(I_n^2A\mathbf{e}_1)^T + T_{0,\varphi}^{()}B$, where B and $T_{0,\varphi}^{()}$ are the $(n-1)\times(n-1)$ lower-right submatrices of A and $T_{\varepsilon,\varphi}^{\beta,\beta}$, respectively. Right- and left-multiply this equality by the matrix J to obtain

$$JBJT_{\varphi,0}^{()} - T_{\varphi,0}^{()}JBJ = (\beta e_1 + e_{n-1})(I_2^nA\mathbf{e}_1)^T - (I_2^nA\mathbf{e}_1)(\beta e_1 + e_{n-1})^T \tag{2.7}$$

($T_{\varphi,0}^{()} = JT_{0,\varphi}^{()}J$). The identity (2.7) and Lemma 2.3(i) (with n replaced by $n-1$) yield

$$JBJ = \begin{pmatrix} 0 & \cdot\cdot & 0 \\ \vdots & & \\ 0 & J & \end{pmatrix}\Omega_\varphi(I_2^nA\mathbf{e}_1) - \begin{pmatrix} 0 & \cdot & \cdot & \cdot & 0 \\ \vdots & & & & \\ 0 & \tau(I_2^{n-1}A\mathbf{e}_1) & & & \end{pmatrix} \Omega_\varphi(\beta e_1 + e_{n-1}) + \Omega_\varphi(I_2^nA\mathbf{e}_n).$$

Therefore,

$$B = \begin{pmatrix} & J & 0 \\ & & \vdots \\ 0 & \cdot\cdot & 0 \end{pmatrix} J\Omega_\varphi(I_2^nA\mathbf{e}_1)J - \begin{pmatrix} \tau(I_2^{n-1}A\mathbf{e}_1) & 0 \\ & \vdots \\ 0 & \cdot\cdot\cdot\cdot\ 0 \end{pmatrix} J\Omega_\varphi(\beta e_1 + e_{n-1})J + J\Omega_\varphi(I_2^nA\mathbf{e}_n)J$$

and, by the equality (2.4),

$$B = \begin{pmatrix} & J & 0 \\ & & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \tau(I_2^n A\mathbf{e}_1) - \begin{pmatrix} \tau(I_2^{n-1}A\mathbf{e}_1)\, 0 \\ \vdots \\ 0 \quad \cdots \quad 0 \end{pmatrix} J$$

$$- \beta \begin{pmatrix} \tau(I_2^{n-1}A\mathbf{e}_1)\, 0 \\ \vdots \\ 0 \quad \cdots \quad 0 \end{pmatrix} + J\Omega_\varphi(I_2^n A\mathbf{e}_n)J. \tag{2.8}$$

As a consequence of (2.8) we have

$$\begin{pmatrix} 0 & (I_n^2 A\mathbf{e}_1)^T \\ \vdots & B \\ 0 & \end{pmatrix} = J \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & & \\ 0 & \tau(I_2^n A\mathbf{e}_1) & \end{pmatrix} - \beta \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \tau(I_2^{n-1}A\mathbf{e}_1) & \vdots \\ 0 & \cdots & 0 \end{pmatrix}$$

$$- \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \tau(I_2^{n-1}A\mathbf{e}_1) & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & & J \\ 0 & & \end{pmatrix}$$

$$+ \begin{pmatrix} 0 & \cdots & \cdots & 0 \\ \vdots & & & J\Omega_\varphi(I_2^n A\mathbf{e}_n)\, J \\ 0 & & & \end{pmatrix}. \tag{2.9}$$

As $A = \mathbb{L}(A\mathbf{e}_1)$, by Proposition 2.2(ii), $A\mathbf{e}_n = \mathbb{L}(\mathbf{e}_n)A\mathbf{e}_1$, i.e., $A\mathbf{e}_n = (J + \left(\begin{smallmatrix} 0 & \cdots & 0 \\ \vdots & & \\ 0 & Q & \end{smallmatrix}\right))A\mathbf{e}_1$, for a certain $(n-1)\times(n-1)$ matrix Q not depending upon A. Thus $I_2^n A\mathbf{e}_n = I_{n-1}^1 A\mathbf{e}_1 + JQJI_2^n A\mathbf{e}_1$ and (2.9) becomes

$$\begin{pmatrix} 0 & (I_n^2 A\mathbf{e}_1)^T \\ \vdots & B \\ 0 & \end{pmatrix} = J \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & & \\ 0 & \tau(I_2^n A\mathbf{e}_1) & \end{pmatrix} - \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \tau(I_2^{n-1}A\mathbf{e}_1) & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & & J \\ 0 & & \end{pmatrix}$$

$$+ \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & & \\ 0 & \tau(I_{n-1}^1 A\mathbf{e}_1) & \end{pmatrix}$$

$$- \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \tau(I_{n-1}^2 A\mathbf{e}_1) & \vdots \\ 0 & \cdots & 0 \end{pmatrix} (\varphi I + \beta J) + \begin{pmatrix} 0 & \cdots & \cdots & 0 \\ \vdots & & & J\Omega_\varphi(JQJI_2^n A\mathbf{e}_1)J \\ 0 & & & \end{pmatrix}. \tag{2.10}$$

Notice that the sum of the first three matrices on the right-hand side of (2.10) plus $A\mathbf{e}_1\mathbf{e}_1^T$ is the matrix $\tau(A\mathbf{e}_1)$. In fact the identity $\tau(A\mathbf{e}_1)T_{0,0} = T_{0,0}\tau(A\mathbf{e}_1)$ implies that (2.7) holds for $\varphi = \beta = 0$ and for B ($T_{0,0}^{()}$) the $(n-1)\times(n-1)$ lower-right submatrix of $\tau(A\mathbf{e}_1)$ ($T_{0,0}$); the thesis follows from (2.10), which then holds for $\varphi = \beta = 0$ and $Q = 0$. Thus we have an explicit expression of $A \in \mathbb{L}$:

$$A = \tau(A\mathbf{e}_1) - \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \tau(I_{n-1}^2 A\mathbf{e}_1) & \vdots \\ 0 & \cdots & 0 \end{pmatrix} (\varphi I + \beta J) + \begin{pmatrix} 0 & \cdots & \cdots & 0 \\ \vdots & & & J\Omega_\varphi(JQJI_2^n A\mathbf{e}_1)\, J \\ 0 & & & \end{pmatrix}.$$

By exploiting it for $A = \mathbb{L}(\mathbf{e}_n) = J + \left(\begin{smallmatrix} 0 & \cdots & 0 \\ \vdots & & \\ 0 & Q & \end{smallmatrix}\right)$, we realize that $JQJ = \Omega_\varphi(JQJe_1)$ or, equivalently, that $JQJ = \Omega_\varphi(\mathbf{p})$ for some $\mathbf{p} \in \mathbb{C}^{n-1}$ not depending upon A.

Therefore, by Proposition 2.2(iii), the generic matrix A of a symmetric closed 1-space containing $T^{\beta,\beta}_{\varepsilon,\varphi}$ has the expression

$$A = \tau(A\mathbf{e}_1) - \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \tau(I^2_{n-1}A\mathbf{e}_1) & \vdots \\ 0 & \cdots & 0 \end{pmatrix}(\varphi I + \beta J) + \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & J\Omega_\varphi(I^n_2 A\mathbf{e}_1)\Omega_\varphi(\mathbf{p})\, J & \\ 0 & & \end{pmatrix} \tag{2.11}$$

for some $\mathbf{p} \in \mathbb{C}^{n-1}$. In particular, (2.11) must be verified for $A = T^{\beta,\beta}_{\varepsilon,\varphi}$ and thus $\mathbf{p}$ must verify (2.6).

Now let us prove the second part of Theorem 2.5. Consider the space $\xi = \xi(\varphi, \beta, \mathbf{p})$ in Definition 2.4 and assume that $\mathbf{p}$ solves equation $\Omega_\varphi(\beta e_1 + e_{n-1})\mathbf{p} = (\varphi - \varepsilon)e_1$ for some $\varepsilon \in \mathbb{C}$. The matrix $T^{\beta,\beta}_{\varepsilon,\varphi}$ is an element of ξ; in fact, by Proposition 2.2(iii), $\xi(\varepsilon\mathbf{e}_1 + \mathbf{e}_2 + \beta\mathbf{e}_n) = T^{\beta,\beta}_{\varepsilon,\varphi}$. Obviously, ξ is a symmetric 1-space. Thus we have to prove only that ξ is equal to the space $\mathbb{A}$ defined as

$$\mathbb{A} = \left\{A \in M_n(\mathbb{C}) : AT^{\beta,\beta}_{\varepsilon,\varphi} = T^{\beta,\beta}_{\varepsilon,\varphi}A \quad \text{and} \quad A\xi(\mathbf{e}_n) = \xi(\mathbf{e}_n)A\right\} \tag{2.12}$$

since the closure of ξ follows from the closure of $\mathbb{A}$. Observe that $\mathbb{A}$ is a linear space whose dimension is not greater than n. In fact, let A_i, $i = 1, \ldots, k$, be k linearly independent matrices of $\mathbb{A}$. If $k > n$, then there exist k elements of $\mathbb{C}$, z_i, $i = 1, \ldots, k$, not all null and such that $\sum_{i=1}^k z_i \mathbf{e}_1^T A_i = \mathbf{0}^T$. The matrix $\sum_{i=1}^k z_i A_i$ is an element of $\mathbb{A}$ and $\mathbf{e}_1^T(\sum_{i=1}^k z_i A_i) = \mathbf{0}^T$. However, if a matrix $A \in \mathbb{A}$, then it satisfies the identities

$$\begin{aligned} &\mathbf{e}_1^T AT^{\beta,\beta}_{\varepsilon,\varphi} = \varepsilon\mathbf{e}_1^T A + \mathbf{e}_2^T A + \beta\mathbf{e}_n^T A, \qquad \mathbf{e}_1^T A\xi(\mathbf{e}_n) = \mathbf{e}_n^T A, \\ &\mathbf{e}_i^T AT^{\beta,\beta}_{\varepsilon,\varphi} = \mathbf{e}_{i-1}^T A + \mathbf{e}_{i+1}^T A, \quad i = 2, \ldots, n-1. \end{aligned} \tag{2.13}$$

If, moreover, $\mathbf{e}_1^T A = \mathbf{0}^T$ from (2.13), it follows that $A = 0$. Thus the matrix $\sum_{i=1}^k z_i A_i$ above must be null and the A_i's are linearly dependent, that is, a contradiction. Now we show that $\xi \subset \mathbb{A}$. As a consequence of this fact and of the inequalities $\dim \xi = n$ and $\dim \mathbb{A} \le n$, we have that $\xi = \mathbb{A}$.

For $\mathbf{z} \in \mathbb{C}^n$, set

$$M(\mathbf{z}) = \tau(\mathbf{z}) - \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \tau(I^2_{n-1}\mathbf{z}) & \vdots \\ 0 & \cdots & 0 \end{pmatrix}(\varphi I + \beta J), \quad N(\mathbf{z}) = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & J\Omega_\varphi(I^n_2\mathbf{z})\Omega_\varphi(\mathbf{p})\, J & \\ 0 & & \end{pmatrix}$$

and notice that $\xi(\mathbf{z}) = M(\mathbf{z}) + N(\mathbf{z})$. By exploiting the equality $T^{\beta,\beta}_{\varepsilon,\varphi} = T^{\beta,\beta}_{\varphi,\varphi} + (\varepsilon - \varphi)\mathbf{e}_1\mathbf{e}_1^T$, as well as the fact that the first row and the first column of $N(\mathbf{z})$ are null, and the equality $M(\mathbf{z})T^{\beta,\beta}_{\varphi,\varphi} = T^{\beta,\beta}_{\varphi,\varphi}M(\mathbf{z})$ (the proof of this identity is obvious and mechanical and thus is omitted), we have

$$\xi(\mathbf{z})T^{\beta,\beta}_{\varepsilon,\varphi} - T^{\beta,\beta}_{\varepsilon,\varphi}\xi(\mathbf{z}) = (\varepsilon - \varphi)(\mathbf{z}\mathbf{e}_1^T - \mathbf{e}_1\mathbf{z}^T) + N(\mathbf{z})T^{\beta,\beta}_{\varphi,\varphi} - T^{\beta,\beta}_{\varphi,\varphi}N(\mathbf{z}).$$

As

$$N(\mathbf{z})T^{\beta,\beta}_{\varphi,\varphi} - T^{\beta,\beta}_{\varphi,\varphi}N(\mathbf{z}) = \begin{pmatrix} 0 & -\mathbf{p}^T\Omega_\varphi(\beta e_1 + e_{n-1})\Omega_\varphi(I^n_2\mathbf{z})J \\ J\Omega_\varphi(I^n_2\mathbf{z})\Omega_\varphi(\beta e_1 + e_{n-1})\mathbf{p} & O \end{pmatrix}$$

the assumption $\Omega_\varphi(\beta e_1 + e_{n-1})\mathbf{p} = (\varphi - \varepsilon)e_1$ yields

$$N(\mathbf{z})T^{\beta,\beta}_{\varphi,\varphi} - T^{\beta,\beta}_{\varphi,\varphi}N(\mathbf{z}) = (\varphi - \varepsilon)(\mathbf{z}\mathbf{e}_1^T - \mathbf{e}_1\mathbf{z}^T),$$

and therefore $\xi(\mathbf{z})T^{\beta,\beta}_{\varepsilon,\varphi} = T^{\beta,\beta}_{\varepsilon,\varphi}\xi(\mathbf{z})\ \ \forall \mathbf{z} \in \mathbb{C}^n$.

Now set $Q = \xi(\mathbf{z})\xi(\mathbf{e}_n) - \xi(\mathbf{e}_n)\xi(\mathbf{z})$. Notice that $\mathbf{e}_1^T Q = \mathbf{z}^T N(\mathbf{e}_n) - \mathbf{e}_n^T N(\mathbf{z}) = \mathbf{0}^T$. Therefore, as $Q^T = -Q$, the first row and the first column of Q are null. Moreover, $QT^{\beta,\beta}_{\varepsilon,\varphi} = T^{\beta,\beta}_{\varepsilon,\varphi}Q$, which implies

$$Q = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & & \\ 0 & & \tau_{0,\varphi}(\mathbf{x}) \end{pmatrix}$$

for some $\mathbf{x} \in \mathbb{C}^{n-1}$. Thus Q is simultaneously symmetric and skewsymmetric; therefore, $Q = \xi(\mathbf{z})\xi(\mathbf{e}_n) - \xi(\mathbf{e}_n)\xi(\mathbf{z}) = 0\ \forall \mathbf{z} \in \mathbb{C}^n$. □

3. 1-algebras and displacement formulas. The algebras characterized in Theorem 2.5 are now involved in a *general* decomposition formula (see Theorem 3.2 below) which leads, in the next section, to new significant displacement decompositions corresponding to special choices of these matrix algebras. A preliminary Lemma 3.1 generalizing related results on HAs [16, 10, 17] follows below. The role of this lemma in the proof of Theorem 3.2 is analogous to the role of *orthogonality relations* in the proof of displacement decompositions involving group matrices [18]. In Lemma 3.1 and Theorem 3.2 A denotes an arbitrary $n \times n$ matrix.

LEMMA 3.1. *Let $\mathbb{L}$ be a commutative 1-space and let $X \in \mathbb{L}$. If $\mathbf{x}_m$, $\mathbf{y}_m \in \mathbb{C}^n$, $m = 1, \ldots, \alpha$, are such that $AX - XA = \sum_{m=1}^{\alpha} \mathbf{x}_m\mathbf{y}_m^T$, then $\sum_{m=1}^{\alpha} \mathbf{x}_m^T\mathbb{L}(\mathbf{y}_m)^T = \mathbf{0}^T$.*

Proof. By Proposition 2.2(ii), for $r = 1, \ldots, n$,

$$\sum_{m=1}^{\alpha} \mathbf{x}_m^T\mathbb{L}(\mathbf{y}_m)^T\mathbf{e}_r = \sum_{m=1}^{\alpha} \mathbf{x}_m^T J_r^T\mathbf{y}_m = \sum_{m=1}^{\alpha}\sum_{i,j=1}^{n} [\mathbf{x}_m]_i[\mathbf{y}_m]_j[J_r^T]_i$$

$$= \sum_{i,j=1}^{n} [AX - XA]_{ij}[J_r]_{ji} = \sum_{i=1}^{n}[(AX - XA)J_r]_{ii} = \sum_{i=1}^{n}[(AJ_r)X - X(AJ_r)]_{ii} = 0. \quad \square$$

THEOREM 3.2. *Let $\mathbb{L}$ and $\mathbb{L}'$ be two symmetric closed 1-spaces containing the matrices $T^{\beta,\beta}_{\varepsilon,\varphi}$ and $T^{\beta',\beta'}_{\varepsilon',\varphi'}$, respectively. If $AT^{\beta,\beta}_{\varepsilon,\varphi} - T^{\beta,\beta}_{\varepsilon,\varphi}A = \sum_{m=1}^{\alpha} \mathbf{x}_m\mathbf{y}_m^T$, then*

$$\begin{aligned} &(\varepsilon - \varepsilon')A + (\beta - \beta')(A\mathbb{L}(\mathbf{e}_n) + \mathbb{L}'(\mathbf{e}_n)A) + (\varphi - \varphi')\mathbb{L}'(\mathbf{e}_n)A\mathbb{L}(\mathbf{e}_n) \\ &\quad = \sum_{m=1}^{\alpha} \mathbb{L}'(\mathbf{x}_m)\mathbb{L}(\mathbf{y}_m) + \mathbb{L}'((\varepsilon - \varepsilon')\mathbf{e}_1 + (\beta - \beta')\mathbf{e}_n)\mathbb{L}(A^T\mathbf{e}_1) \\ &\qquad + \mathbb{L}'((\beta - \beta')\mathbf{e}_1 + (\varphi - \varphi')\mathbf{e}_n)\mathbb{L}(A^T\mathbf{e}_n). \end{aligned} \tag{3.1}$$

Proof. Let X be a symmetric $n \times n$ matrix such that if $AX = XA$ and $\mathbf{e}_1^T A = \mathbf{e}_n^T A = \mathbf{0}^T$, then $A = 0$. Set $[X]_{1n} = [X]_{n1} = \beta$, $[X]_{11} = \varepsilon$, and $[X]_{nn} = \varphi$ and let X' be the $n \times n$ matrix defined by $X = X' + (\varepsilon - \varepsilon')\mathbf{e}_1\mathbf{e}_1^T + (\beta - \beta')(\mathbf{e}_1\mathbf{e}_n^T + \mathbf{e}_n\mathbf{e}_1^T) + (\varphi - \varphi')\mathbf{e}_n\mathbf{e}_n^T$. The assertion of Theorem 3.2 is now shown for X and X' instead of for $T^{\beta,\beta}_{\varepsilon,\varphi}$ and $T^{\beta',\beta'}_{\varepsilon',\varphi'}$, respectively. The thesis will follow because $T^{\beta,\beta}_{\varepsilon,\varphi}$ and $T^{\beta',\beta'}_{\varepsilon',\varphi'}$ satisfy the hypotheses on X and X'. (The simple proof of this fact is left to the reader.)

Let M and N be the matrices on the left-hand side and on the right-hand side in equality (3.1), respectively. We shall prove that if $AX - XA = \sum_{m=1}^{\alpha} \mathbf{x}_m\mathbf{y}_m^T$, then

$(M-N)X = X(M-N)$ and $\mathbf{e}_1^T(M-N) = \mathbf{e}_n^T(M-N) = \mathbf{0}^T$, and therefore, by the hypothesis on X, $M = N$.

The equality $\mathbf{e}_1^T M = \mathbf{e}_1^T N$ is easily verifiable by exploiting Lemma 3.1. The equalities $(M-N)X = X(M-N)$ and $\mathbf{e}_n^T M = \mathbf{e}_n^T N$ are equivalent to the equalities

$$\begin{aligned}[(\beta-\beta')\mathbf{e}_1 + (\varphi-\varphi')\mathbf{e}_n]&\left\{\sum_{m=1}^{\alpha}\mathbf{x}_m^T\mathbb{L}'(\mathbf{e}_n)\mathbb{L}(\mathbf{y}_m)\right\}\\ &= [(\beta-\beta')\mathbf{e}_1 + (\varphi-\varphi')\mathbf{e}_n]\{(\varepsilon-\varepsilon')[\mathbf{e}_n^T A - \mathbf{e}_1^T A\mathbb{L}(\mathbf{e}_n)]\\ &\quad + \mathbf{e}_n^T\mathbb{L}'(\mathbf{e}_n)[(\beta-\beta')(A-\mathbb{L}(A^T\mathbf{e}_1)) + (\varphi-\varphi')(A\mathbb{L}(\mathbf{e}_n)-\mathbb{L}(A^T\mathbf{e}_n))]\}\end{aligned}$$

and

$$\begin{aligned}\sum_{m=1}^{\alpha}\mathbf{x}_m^T\mathbb{L}'(\mathbf{e}_n)\mathbb{L}(\mathbf{y}_m) &= (\varepsilon-\varepsilon')[\mathbf{e}_n^T A - \mathbf{e}_1^T A\mathbb{L}(\mathbf{e}_n)]\\ &\quad + \mathbf{e}_n^T\mathbb{L}'(\mathbf{e}_n)[(\beta-\beta')(A-\mathbb{L}(A^T\mathbf{e}_1)) + (\varphi-\varphi')(A\mathbb{L}(\mathbf{e}_n)-\mathbb{L}(A^T\mathbf{e}_n))],\end{aligned} \tag{3.2}$$

respectively. The proof of the second equivalence is simple. Let us prove the first one.

$$\begin{aligned}NX-XN = \sum_{m=1}^{\alpha}&[\mathbb{L}'(\mathbf{x}_m)X - X\mathbb{L}'(\mathbf{x}_m)]\mathbb{L}(\mathbf{y}_m) + (\beta-\beta')[\mathbb{L}'(\mathbf{e}_n)X - X\mathbb{L}'(\mathbf{e}_n)]\mathbb{L}(A^T\mathbf{e}_1)\\ &+ (\varphi-\varphi')[\mathbb{L}'(\mathbf{e}_n)X - X\mathbb{L}'(\mathbf{e}_n)]\mathbb{L}(A^T\mathbf{e}_n).\end{aligned}$$

For the sake of simplicity, set $Q = \mathbb{L}'(\mathbf{e}_n)X - X\mathbb{L}'(\mathbf{e}_n)$ and then exploit the equality $X = X' + (\varepsilon-\varepsilon')\mathbf{e}_1\mathbf{e}_1^T + (\beta-\beta')(\mathbf{e}_1\mathbf{e}_n^T + \mathbf{e}_n\mathbf{e}_1^T) + (\varphi-\varphi')\mathbf{e}_n\mathbf{e}_n^T$ to obtain

$$\begin{aligned}NX - XN &= \sum_{m=1}^{\alpha}\{(\varepsilon-\varepsilon')(\mathbf{x}_m\mathbf{e}_1^T - \mathbf{e}_1\mathbf{x}_m^T) + (\beta-\beta')\\ &\qquad\times[\mathbf{x}_m\mathbf{e}_n^T + \mathbb{L}'(\mathbf{e}_n)\mathbf{x}_m\mathbf{e}_1^T - \mathbf{e}_1\mathbf{x}_m^T\mathbb{L}'(\mathbf{e}_n) - \mathbf{e}_n\mathbf{x}_m^T]\\ &\qquad+(\varphi-\varphi')[\mathbb{L}'(\mathbf{e}_n)\mathbf{x}_m\mathbf{e}_n^T - \mathbf{e}_n\mathbf{x}_m^T\mathbb{L}'(\mathbf{e}_n)]\}\mathbb{L}(\mathbf{y}_m)\\ &\qquad+(\beta-\beta')Q\mathbb{L}(A^T\mathbf{e}_1) + (\varphi-\varphi')Q\mathbb{L}(A^T\mathbf{e}_n)\\ &= \sum_{m=1}^{\alpha}\{(\varepsilon-\varepsilon')[\mathbf{x}_m\mathbf{y}_m^T - \mathbf{e}_1\mathbf{x}_m^T\mathbb{L}(\mathbf{y}_m)]\\ &\qquad+(\beta-\beta')[\mathbf{x}_m\mathbf{y}_m^T\mathbb{L}(\mathbf{e}_n) + \mathbb{L}'(\mathbf{e}_n)\mathbf{x}_m\mathbf{y}_m^T\\ &\qquad\quad-\mathbf{e}_1\mathbf{x}_m^T\mathbb{L}'(\mathbf{e}_n)\mathbb{L}(\mathbf{y}_m) - \mathbf{e}_n\mathbf{x}_m^T\mathbb{L}(\mathbf{y}_m)]\\ &\qquad+(\varphi-\varphi')[\mathbb{L}'(\mathbf{e}_n)\mathbf{x}_m\mathbf{y}_m^T\mathbb{L}(\mathbf{e}_n) - \mathbf{e}_n\mathbf{x}_m^T\mathbb{L}'(\mathbf{e}_n)\mathbb{L}(\mathbf{y}_m)]\}\\ &\qquad+(\beta-\beta')Q\mathbb{L}(A^T\mathbf{e}_1) + (\varphi-\varphi')Q\mathbb{L}(A^T\mathbf{e}_n).\end{aligned}$$

By exploiting the assumption $AX - XA = \sum_{m=1}^{\alpha}\mathbf{x}_m\mathbf{y}_m^T$ and Lemma 3.1, the last expression becomes

$$\begin{aligned}&(\varepsilon-\varepsilon')(AX-XA) + (\beta-\beta')\\ &\times\left[(AX-XA)\mathbb{L}(\mathbf{e}_n) + \mathbb{L}'(\mathbf{e}_n)(AX-XA) - \mathbf{e}_1\sum_{m=1}^{\alpha}\mathbf{x}_m^T\mathbb{L}'(\mathbf{e}_n)\mathbb{L}(\mathbf{y}_m)\right]\end{aligned}$$

$$
\begin{aligned}
&+ (\varphi-\varphi')\left[\mathbb{L}'(\mathbf{e}_n)(AX-XA)\mathbb{L}(\mathbf{e}_n) - \mathbf{e}_n\sum_{m=1}^{\alpha}\mathbf{x}_m^T\mathbb{L}'(\mathbf{e}_n)\mathbb{L}(\mathbf{y}_m)\right]\\
&+ (\beta-\beta')Q\mathbb{L}(A^T\mathbf{e}_1) + (\varphi-\varphi')Q\mathbb{L}(A^T\mathbf{e}_n)\\
=\ & (\varepsilon-\varepsilon')(AX-XA) + (\beta-\beta')\\
&\times\left[A\mathbb{L}(\mathbf{e}_n)X - XA\mathbb{L}(\mathbf{e}_n) + \mathbb{L}'(\mathbf{e}_n)AX - X\mathbb{L}'(\mathbf{e}_n)A - QA - \mathbf{e}_1\sum_{m=1}^{\alpha}\mathbf{x}_m^T\mathbb{L}'(\mathbf{e}_n)\mathbb{L}(\mathbf{y}_m)\right]\\
&+ (\varphi-\varphi')\left[\mathbb{L}'(\mathbf{e}_n)A\mathbb{L}(\mathbf{e}_n)X - X\mathbb{L}'(\mathbf{e}_n)A\mathbb{L}(\mathbf{e}_n) - QA\mathbb{L}(\mathbf{e}_n)\right.\\
&-\ \left.\mathbf{e}_n\sum_{m=1}^{\alpha}\mathbf{x}_m^T\mathbb{L}'(\mathbf{e}_n)\mathbb{L}(\mathbf{y}_m)\right] + (\beta-\beta')Q\mathbb{L}(A^T\mathbf{e}_1) + (\varphi-\varphi')Q\mathbb{L}(A^T\mathbf{e}_n)\\
=\ & MX - XM + (\beta-\beta')Q[\mathbb{L}(A^T\mathbf{e}_1) - A] + (\varphi-\varphi')Q[\mathbb{L}(A^T\mathbf{e}_n) - A\mathbb{L}(\mathbf{e}_n)]\\
&- [(\beta-\beta')\mathbf{e}_1 + (\varphi-\varphi')\mathbf{e}_n]\sum_{m=1}^{\alpha}\mathbf{x}_m^T\mathbb{L}'(\mathbf{e}_n)\mathbb{L}(\mathbf{y}_m).
\end{aligned}
$$

By replacing $\mathbf{x}_m$ with $\mathbf{e}_n$ in the expression of $\mathbb{L}'(\mathbf{x}_m)X - X\mathbb{L}'(\mathbf{x}_m)$ obtained above, we have $Q = (\varepsilon-\varepsilon')(\mathbf{e}_n\mathbf{e}_1^T - \mathbf{e}_1\mathbf{e}_n^T) + (\beta-\beta')[\mathbb{L}'(\mathbf{e}_n)\mathbf{e}_n\mathbf{e}_1^T - \mathbf{e}_1\mathbf{e}_n^T\mathbb{L}'(\mathbf{e}_n)] + (\varphi-\varphi')[\mathbb{L}'(\mathbf{e}_n)\mathbf{e}_n\mathbf{e}_n^T - \mathbf{e}_n\mathbf{e}_n^T\mathbb{L}'(\mathbf{e}_n)]$. Thus

$$
\begin{aligned}
NX-XN = MX-XM &+ (\beta-\beta')\{[(\beta-\beta')\mathbf{e}_1 + (\varphi-\varphi')\mathbf{e}_n]\mathbf{e}_n^T\mathbb{L}'(\mathbf{e}_n)[A - \mathbb{L}(A^T\mathbf{e}_1)]\\
&+ [(\varepsilon-\varepsilon')\mathbf{e}_1 - (\varphi-\varphi')\mathbb{L}'(\mathbf{e}_n)\mathbf{e}_n][\mathbf{e}_n^TA - \mathbf{e}_1^TA\mathbb{L}(\mathbf{e}_n)]\}\\
&+ (\varphi-\varphi')\{[(\beta-\beta')\mathbf{e}_1 + (\varphi-\varphi')\mathbf{e}_n]\mathbf{e}_n^T\mathbb{L}'(\mathbf{e}_n)[A\mathbb{L}(\mathbf{e}_n) - \mathbb{L}(A^T\mathbf{e}_n)]\\
&+ [(\varepsilon-\varepsilon')\mathbf{e}_n + (\beta-\beta')\mathbb{L}'(\mathbf{e}_n)\mathbf{e}_n][\mathbf{e}_n^TA - \mathbf{e}_1^TA\mathbb{L}(\mathbf{e}_n)]\}\\
&- [(\beta-\beta')\mathbf{e}_1 + (\varphi-\varphi')\mathbf{e}_n]\sum_{m=1}^{\alpha}\mathbf{x}_m^T\mathbb{L}'(\mathbf{e}_n)\mathbb{L}(\mathbf{y}_m)\\
= MX - XM &+ [(\beta-\beta')\mathbf{e}_1 + (\varphi-\varphi')\mathbf{e}_n]\{(\varepsilon-\varepsilon')[\mathbf{e}_n^TA - \mathbf{e}_1^TA\mathbb{L}(\mathbf{e}_n)]\\
&+ \mathbf{e}_n^T\mathbb{L}'(\mathbf{e}_n)[(\beta-\beta')(A - \mathbb{L}(A^T\mathbf{e}_1)) + (\varphi-\varphi')(A\mathbb{L}(\mathbf{e}_n) - \mathbb{L}(A^T\mathbf{e}_n))]\}\\
&- [(\beta-\beta')\mathbf{e}_1 + (\varphi-\varphi')\mathbf{e}_n]\left\{\sum_{m=1}^{\alpha}\mathbf{x}_m^T\mathbb{L}'(\mathbf{e}_n)\mathbb{L}(\mathbf{y}_m)\right\}
\end{aligned}
$$

and the first equivalence is proved. Now we have to prove (3.2) and the proof of Theorem 3.2 will be complete, because then the equality preceding (3.2)—identical to (3.2), but a factor—is satisfied. For $s = 1, \ldots, n$

$$
\begin{aligned}
&\sum_{m=1}^{\alpha}\mathbf{x}_m^T\mathbb{L}'(\mathbf{e}_n)\mathbb{L}(\mathbf{y}_m)\mathbf{e}_s = \sum_{m=1}^{\alpha}\mathbf{x}_m^T\mathbb{L}'(\mathbf{e}_n)\mathbb{L}(\mathbf{e}_s)\mathbf{y}_m = \sum_{m=1}^{\alpha}\sum_{i,\,j=1}^{n}[\mathbf{x}_m]_i[\mathbf{y}_m]_j[\mathbb{L}'(\mathbf{e}_n)\mathbb{L}(\mathbf{e}_s)]_{ij}\\
&= \sum_{i,\,j=1}^{n}[AX-XA]_{ij}[\mathbb{L}(\mathbf{e}_s)\mathbb{L}'(\mathbf{e}_n)]_{ji} = \sum_{i=1}^{n}[A\mathbb{L}(\mathbf{e}_s)X\mathbb{L}'(\mathbf{e}_n) - XA\mathbb{L}(\mathbf{e}_s)\mathbb{L}'(\mathbf{e}_n)]_{ii}
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{n}[-A\mathbb{L}(\mathbf{e}_s)(\mathbb{L}'(\mathbf{e}_n)X - X\mathbb{L}'(\mathbf{e}_n))]_{ii}\\
&= \sum_{i=1}^{n}[-A\mathbb{L}(\mathbf{e}_s)\{(\varepsilon-\varepsilon')(\mathbf{e}_n\mathbf{e}_1^T - \mathbf{e}_1\mathbf{e}_n^T) + (\beta-\beta')[\mathbb{L}'(\mathbf{e}_n)\mathbf{e}_n\mathbf{e}_1^T - \mathbf{e}_1\mathbf{e}_n^T\mathbb{L}'(\mathbf{e}_n)]\\
&\quad + (\varphi-\varphi')[\mathbb{L}'(\mathbf{e}_n)\mathbf{e}_n\mathbf{e}_n^T - \mathbf{e}_n\mathbf{e}_n^T\mathbb{L}'(\mathbf{e}_n)]\}]_{ii}\\
&= -(\varepsilon-\varepsilon')\sum_{i=1}^{n}\mathbf{e}_1^T[A\mathbb{L}(\mathbf{e}_n)\mathbf{e}_s\mathbf{e}_1^T - A\mathbf{e}_s\mathbf{e}_n^T]\mathbf{e}_i - (\beta-\beta')\sum_{i=1}^{n}\mathbf{e}_i^T[A\mathbb{L}(\mathbf{e}_s)\mathbb{L}'(\mathbf{e}_n)\mathbf{e}_n\mathbf{e}_1^T\\
&\quad - A\mathbf{e}_s\mathbf{e}_n^T\mathbb{L}'(\mathbf{e}_n)]\mathbf{e}_i - (\varphi-\varphi')\sum_{i=1}^{n}\mathbf{e}_i^T[A\mathbb{L}(\mathbf{e}_s)\mathbb{L}'(\mathbf{e}_n)\mathbf{e}_n\mathbf{e}_n^T - A\mathbb{L}(\mathbf{e}_n)\mathbf{e}_s\mathbf{e}_n^T\mathbb{L}'(\mathbf{e}_n)]\mathbf{e}_i\\
&= (\varepsilon-\varepsilon')[\mathbf{e}_n^TA\mathbf{e}_s - \mathbf{e}_1^TA\mathbb{L}(\mathbf{e}_n)\mathbf{e}_s] + (\beta-\beta')[\mathbf{e}_n^T\mathbb{L}'(\mathbf{e}_n)A\mathbf{e}_s - \mathbf{e}_n^T\mathbb{L}'(\mathbf{e}_n)\mathbb{L}(\mathbf{e}_s)A^T\mathbf{e}_1]\\
&\quad + (\varphi-\varphi')[\mathbf{e}_n^T\mathbb{L}'(\mathbf{e}_n)A\mathbb{L}(\mathbf{e}_n)\mathbf{e}_s - \mathbf{e}_n^T\mathbb{L}'(\mathbf{e}_n)\mathbb{L}(\mathbf{e}_s)A^T\mathbf{e}_n],
\end{aligned}
$$

that is, (3.2) holds. □

Remark. It is clear that Theorem 3.2 holds unchanged if $T_{\varepsilon,\varphi}^{\beta,\beta} = T_{\varepsilon,\varphi} + \beta(\mathbf{e}_1\mathbf{e}_n^T + \mathbf{e}_n\mathbf{e}_1^T)$ is replaced by $M = Y + \beta(\mathbf{e}_1\mathbf{e}_n^T + \mathbf{e}_n\mathbf{e}_1^T)$, where Y is a generic symmetric tridiagonal matrix having at least $n-2$ nonzero entries $[Y]_{i,\,i+1}$, and $T_{\varepsilon',\varphi'}^{\beta',\beta'}$ is replaced by $M' = Y' + \beta'(\mathbf{e}_1\mathbf{e}_n^T + \mathbf{e}_n\mathbf{e}_1^T)$, where $Y' = Y + (\varepsilon'-\varepsilon)\mathbf{e}_1\mathbf{e}_1^T + (\varphi'-\varphi)\mathbf{e}_n\mathbf{e}_n^T$ (set $[Y]_{11} = \varepsilon$, $[Y]_{nn} = \varphi$). (In fact M and M' satisfy the assumptions on X and X' at the beginning of the proof.) This result includes Theorems 3.2 ($i = 1$, $i = n$) and 3.4 of [10].

For $\beta = \beta' = 0$ the result stated in Theorem 3.2 leads to displacement decompositions exploiting symmetric HAs that include, as special instances, some of the most significant formulas stated in [10] (see Corollaries 4.1 and 4.2 in [10]). In the next section it is shown that formula (3.1) also leads to significant decompositions exploiting 1-spaces which are not HAs. More specifically, in these last decompositions some Hartley-type matrix algebras will be involved.

4. The algebras $\boldsymbol{\eta, \mu, \mathcal{H}, \mathcal{K}}$. Theorem 3.2 can be exploited in order to obtain—as special cases of formula (3.1)—effective displacement decompositions of a generic matrix A. To this end we need to know if, and under what assumptions, there exist symmetric 1-algebras containing matrices of the form $T_{\varepsilon,\varphi}^{\beta,\beta}$. Theorem 2.5 relates the existence of such spaces $\xi(\varphi,\beta,\mathbf{p})$ to the existence of vectors $\mathbf{p}$ solving the linear system

$$I_{\beta\varphi}\mathbf{p} = (\varphi-\varepsilon)e_1, \tag{4.1}$$

where $I_{\beta\varphi}$ is the $(n-1)\times(n-1)$ matrix

$$I_{\beta\varphi} = \Omega_\varphi(\beta e_1 + e_{n-1}) = \begin{pmatrix} \beta & & & & \\ & \ddots & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \beta \end{pmatrix} + \begin{pmatrix} & & & & 1 \\ & & & 1 & -\varphi \\ & & \iddots & \iddots & \\ & 1 & \iddots & & \\ 1 & -\varphi & & & \end{pmatrix}. \tag{4.2}$$

Equation (4.1) may have no solution, infinite solutions, or only one solution; these three cases and the corresponding 1-spaces $\xi(\varphi,\beta,\mathbf{p})$ are studied in Proposition 4.2. For the sake of simplicity, for U, V $n\times n$ matrices, set $\mathcal{C}_V(U) = UV - VU$.

LEMMA 4.1. *$I_{\beta\varphi}$ is nonsingular iff $\exists\ \mathbf{z} \in \mathbb{C}^{n-1}$ and $\delta \in \mathbb{C}$, $\delta \neq 0$, such that $\mathbf{z}^T I_{\beta\varphi} = \delta e_1^T$. In this case $I_{\beta,\varphi}^{-1} = \delta^{-1}\Omega_\varphi(\mathbf{z})$.*

Proof. The assertion holds for any matrix A of a commutative 1-space $\mathbb{L}$; in fact, if $\mathbf{z}^T A = \delta e_1^T$, then $e_i^T \mathbb{L}(\mathbf{z})A = \mathbf{z}^T \mathbb{L}(e_i)A = \mathbf{z}^T A \mathbb{L}(e_i) = \delta e_i^T$, $i = 1, \ldots, n-1$, that is, $\delta^{-1}\mathbb{L}(\mathbf{z})A = I$. ∎

PROPOSITION 4.2. *We have the following three cases.*

(i) *$I_{\beta\varphi}$ singular and $\varepsilon \neq \varphi$: There is no symmetric 1-algebra containing $T_{\varepsilon,\varphi}^{\beta,\beta}$.*

(ii) *$I_{\beta\varphi}$ singular and $\varepsilon = \varphi$: There are infinite symmetric 1-algebras containing $T_{\varphi,\varphi}^{\beta,\beta}$ and therefore $T_{\varphi,\varphi}^{\beta,\beta}$ is derogatory. More specifically, these spaces are the $\xi(\varphi, \beta, \mathbf{p})$ (in (2.5)) where $\mathbf{p}$ is such that $I_{\beta\varphi}\mathbf{p} = \mathbf{0}$, and they can be represented as*

$$\xi(\varphi, \beta, \mathbf{p}) = \left\{ A \in M_n(\mathbb{C}) : \mathcal{C}_A\left(T_{\varphi,\varphi}^{\beta,\beta}\right) = \mathcal{C}_A(\xi(\mathbf{e}_n)) = 0 \right\}. \tag{4.3}$$

Only one of them is also persymmetric, and we call it $\tau_{\varphi,\varphi}^{\beta,\beta}$. We have

$$\tau_{\varphi,\varphi}^{\beta,\beta} = \xi(\varphi, \beta, \mathbf{0}) = \left\{ A \in M_n(\mathbb{C}) : \mathcal{C}_A\left(T_{\varphi,\varphi}^{\beta,\beta}\right) = \mathcal{C}_A(J) = 0 \right\}. \tag{4.4}$$

(iii) *$I_{\beta\varphi}$ nonsingular: For any $\varepsilon \in \mathbb{C}$ there exists a unique symmetric 1-algebra containing $T_{\varepsilon,\varphi}^{\beta,\beta}$. Moreover, if $\tau_{\varepsilon,\varphi}^{\beta,\beta}$ denotes such a space, we have*

$$\tau_{\varepsilon,\varphi}^{\beta,\beta} = \xi\left(\varphi, \beta, (\varphi - \varepsilon) I_{\beta\varphi}^{-1} e_1\right) = \left\{ A \in M_n(\mathbb{C}) : \mathcal{C}_A\left(T_{\varepsilon,\varphi}^{\beta,\beta}\right) = 0 \right\}. \tag{4.5}$$

Therefore, $T_{\varepsilon,\varphi}^{\beta,\beta}$ is nonderogatory and $\tau_{\varepsilon,\varphi}^{\beta,\beta}$ is the set of all polynomials in $T_{\varepsilon,\varphi}^{\beta,\beta}$.

Proof of Proposition 4.2(i). Assume that $I_{\beta\varphi}$ is singular and that $\varepsilon \neq \varphi$. By the first part of Theorem 2.5, a symmetric 1-algebra containing $T_{\varepsilon,\varphi}^{\beta,\beta}$ is equal to $\xi(\varphi, \beta, \mathbf{p})$, where $\mathbf{p}$ is such that $I_{\beta\varphi}\mathbf{p} = (\varphi - \varepsilon)e_1$. Then, by Lemma 4.1, $I_{\beta\varphi}$ is invertible, that is, a contradiction.

Proof of Proposition 4.2(ii). Assume that $I_{\beta\varphi}$ is singular and that $\varepsilon = \varphi$. Then the vectors $\mathbf{p} \in \mathbb{C}^{n-1}$ satisfying the equality $I_{\beta\varphi}\mathbf{p} = (\varphi - \varepsilon)e_1 = \mathbf{0}$ are infinite and, by the second part of Theorem 2.5, every space $\xi(\varphi, \beta, \mathbf{p})$ is a symmetric 1-algebra containing $T_{\varphi,\varphi}^{\beta,\beta}$, and it can be represented as in (4.3). The matrix $T_{\varphi,\varphi}^{\beta,\beta}$ is derogatory, because otherwise the set of all polynomials in $T_{\varphi,\varphi}^{\beta,\beta}$ should be an n-dimensional subspace of each $\xi(\varphi, \beta, \mathbf{p})$, which is absurd. Finally, among the $\xi(\varphi, \beta, \mathbf{p})$'s, there is only one containing the matrix J (or, equivalently, for which $\xi(\mathbf{e}_n) = J$), that is, $\xi(\varphi, \beta, \mathbf{0})$.

Proof of Proposition 4.2(iii). Assume that $I_{\beta\varphi}$ is nonsingular. By the second part of Theorem 2.5, $\xi(\varphi, \beta, (\varphi - \varepsilon)I_{\beta\varphi}^{-1}e_1)$ is a symmetric 1-algebra containing $T_{\varepsilon,\varphi}^{\beta,\beta}$. By the first part of Theorem 2.5, there is no other symmetric 1-algebra containing $T_{\varepsilon,\varphi}^{\beta,\beta}$. As regards the identity (4.5), notice that $\tau_{\varepsilon,\varphi}^{\beta,\beta} \subset \{A \in M_n(\mathbb{C}) : \mathcal{C}_A(T_{\varepsilon,\varphi}^{\beta,\beta}) = 0\}$. Conversely, let A be a matrix commuting with $T_{\varepsilon,\varphi}^{\beta,\beta}$ and consider the space $\tau_{\varepsilon',\varphi}^{\beta,\beta} = \xi(\varphi, \beta, (\varphi - \varepsilon')I_{\beta\varphi}^{-1}e_1)$, $\varepsilon' \neq \varepsilon$. Then apply Theorem 3.2 for $\mathbb{L} = \tau_{\varepsilon,\varphi}^{\beta,\beta}$ and $\mathbb{L}' = \tau_{\varepsilon',\varphi}^{\beta,\beta}$ to the matrix A to obtain $(\varepsilon - \varepsilon')A = (\varepsilon - \varepsilon')\tau_{\varepsilon,\varphi}^{\beta,\beta}(A^T\mathbf{e}_1)$. ∎

Now two interesting classes of matrix algebras $\mathcal{S}$ and $\mathcal{R}$, both corresponding to case (ii) in Proposition 4.2, are investigated. These algebras are also exploited to state, as special instances of formula (3.1), new efficient decompositions of a generic centrosymmetric matrix A (Theorem 4.3). Notice that the algebra $\mathcal{H}$, studied in [5] and related to the *Hartley transform*, is a particular element of $\mathcal{S}$.

The class $\mathcal{S}$. Let $\varphi = \varepsilon = 0$ and $\beta = 1$ in (4.1)–(4.2). As $\Omega_0(e_1 + e_{n-1}) = \tau(e_1 + e_{n-1}) = I + J$ is singular, by Proposition 4.2(ii) there are infinite symmetric 1-algebras containing the matrix $T^{1,1}_{0,0}$, i.e., the spaces $\xi(0, 1, \mathbf{p}^{\mathrm{SK}})$, where $\mathbf{p}^{\mathrm{SK}}$ is an arbitrary skewsymmetric vector ($\hat{\mathbf{p}}^{\mathrm{SK}} = -\mathbf{p}^{\mathrm{SK}}$). These spaces are denoted by $\mathcal{S}(\cdot; \mathbf{p}^{\mathrm{SK}})$ and can be represented as

$$\mathcal{S}(\cdot; \mathbf{p}^{\mathrm{SK}}) = \xi(0, 1, \mathbf{p}^{\mathrm{SK}}) = \{A \in M_n(\mathbb{C}) : \mathcal{C}_A(T^{1,1}_{0,0}) = \mathcal{C}_A(\mathcal{S}(\mathbf{e}_n; \mathbf{p}^{\mathrm{SK}})) = 0\}. \tag{4.6}$$

Each algebra $\mathcal{S}(\cdot; \mathbf{p}^{\mathrm{SK}})$ contains the algebra C^{S} of all $n \times n$ symmetric circulant matrices; therefore, by the identity $\{A : \mathcal{C}_A(T^{1,1}_{0,0}) = 0\} = C + JC$ (found in [9]), $\mathcal{S}(\cdot; \mathbf{p}^{\mathrm{SK}})$ must be equal to $C^{\mathrm{S}} + J\tilde{C}$ for some subset $\tilde{C}$ (depending upon $\mathbf{p}^{\mathrm{SK}}$) of the space C of circulant matrices.

Algebra η. If $\mathbf{p}^{\mathrm{SK}} = \mathbf{0}$ we have the space

$$\eta = \mathcal{S}(\cdot; \mathbf{0}) = \xi(0, 1, \mathbf{0}) = \tau^{1,1}_{0,0} = \{A \in M_n(\mathbb{C}) : \mathcal{C}_A(T^{1,1}_{0,0}) = \mathcal{C}_A(J) = 0\}. \tag{4.7}$$

Notice that $\eta = C^{\mathrm{S}} + JC^{\mathrm{S}}$; in fact $C^{\mathrm{S}} + JC^{\mathrm{S}} \subset \eta$ and

$$\begin{aligned}\dim(C^{\mathrm{S}} + JC^{\mathrm{S}}) &= \dim C^{\mathrm{S}} + \dim JC^{\mathrm{S}} - \dim C^{\mathrm{S}} \cap JC^{\mathrm{S}} = 2\dim C^{\mathrm{S}} - \dim C^{\mathrm{S}} \cap JC^{\mathrm{S}} \\ &= \begin{cases} 2\left(\dfrac{n}{2} + 1\right) - 2 & \text{if } n \text{ is even,} \\ 2\left(\dfrac{n+1}{2}\right) - 1 & \text{if } n \text{ is odd,} \end{cases}\end{aligned}$$

that is, $\dim(C^{\mathrm{S}} + JC^{\mathrm{S}}) = n$.

Algebra $\mathcal{H}$. If $\mathbf{p}^{\mathrm{SK}} = \frac{1}{2}(e_2 - e_{n-2})$, we have the space

$$\mathcal{H} = \mathcal{S}\left(\cdot; \frac{1}{2}(e_2 - e_{n-2})\right) = \xi\left(0, 1, \frac{1}{2}(e_2 - e_{n-2})\right). \tag{4.8}$$

Notice that $\mathcal{H} = C^{\mathrm{S}} + JPC^{\mathrm{SK}}$, where P is the circulant matrix whose first row is $\mathbf{e}_2^T$ ($P = P_1 = C(\mathbf{e}_2)$) and C^{SK} is the set of all $n \times n$ skewsymmetric circulant matrices (a matrix A is skewsymmetric if $A^T = -A$). To prove this fact, first observe that $C^{\mathrm{S}} + JPC^{\mathrm{SK}}$ is commutative and that the matrices $T^{1,1}_{0,0}$ and

$$\mathcal{H}(\mathbf{e}_n) = \mathcal{S}\left(\mathbf{e}_n; \frac{1}{2}(e_2 - e_{n-2})\right) = J + \frac{1}{2}\begin{pmatrix} 0 & \cdots & 0 \\ \vdots & & \\ 0 & \tau(e_2 - e_{n-2}) & \end{pmatrix} \tag{4.9}$$

are elements of $C^{\mathrm{S}} + JPC^{\mathrm{SK}}$. The commutativity follows from the commutativity of the space C. Moreover, the matrices $\frac{1}{2}T^{1,1}_{0,0}$ and $JP\left(-\frac{1}{2}(P - P^T)\right)$ are elements of C^{S} and JPC^{SK}, respectively, and their sum is the matrix in (4.9). Thus $C^{\mathrm{S}} + JPC^{\mathrm{SK}} \subset \mathcal{H}$. But

$$\begin{aligned}\dim(C^{\mathrm{S}} + JPC^{\mathrm{SK}}) &= \dim C^{\mathrm{S}} + \dim JPC^{\mathrm{SK}} - \dim C^{\mathrm{S}} \cap JPC^{\mathrm{SK}} \\ &= \begin{cases} \left(\dfrac{n}{2} + 1\right) + \left(\dfrac{n}{2} - 1\right) - 0 & \text{if } n \text{ is even,} \\ \left(\dfrac{n+1}{2}\right) + \left(\dfrac{n-1}{2}\right) - 0 & \text{if } n \text{ is odd,} \end{cases}\end{aligned}$$

that is, $\dim(C^{\mathrm{S}} + JPC^{\mathrm{SK}}) = n$, and the identity $\mathcal{H} = C^{\mathrm{S}} + JPC^{\mathrm{SK}}$ is proved. In [5] it is shown that the matrices of $\mathcal{H}$ are simultaneously diagonalized by a similarity

transformation known as Hartley transform (see also Theorem 5.2 in the next section). A greater attention has been devoted to this particular *real* transform since Bracewell [11, 12] introduced the fast Hartley transform (FHT).

Observe that the proper inclusion $\mathcal{H} \supset C^{\mathrm{S}}$ is exploited in [5] to determine a new preconditioner of symmetric Toeplitz systems, competitive with the more usual circulant preconditioners (see also [13]). All algebras $\mathcal{S}(\cdot;\mathbf{p}^{\mathrm{SK}})$ include C^{S} and, besides $\mathcal{H}$, there may be other algebras $\mathcal{S}(\cdot;\mathbf{p}^{\mathrm{SK}})$ whose matrices are simultaneously diagonalized by a fast transform (this is, the case of $\eta = \mathcal{S}(\cdot;\mathbf{0})$; see Theorem 5.2). As it will be shown in a forthcoming paper, some of the algebras $\mathcal{S}(\cdot;\mathbf{p}^{\mathrm{SK}})$ (together with some other $\mathcal{R}(\cdot;\mathbf{p}^{\mathrm{S}})$ algebras described below) can lead to other efficient preconditioners of Toeplitz systems.

The class $\mathcal{R}$. The choice $\varphi = \varepsilon = 0$, $\beta = -1$ leads to symmetric 1-algebras—containing $T_{0,0}^{-1,-1}$—naturally related to those of the class $\mathcal{S}$. These are the following:

$$\mathcal{R}(\cdot;\mathbf{p}^{\mathrm{S}}) = \xi(0,-1,\mathbf{p}^{\mathrm{S}}) = \left\{A \in M_n(\mathbb{C}) : \mathcal{C}_A\left(T_{0,0}^{-1,-1}\right) = \mathcal{C}_A(\mathcal{R}(\mathbf{e}_n;\mathbf{p}^{\mathrm{S}})) = 0\right\}, \tag{4.10}$$

where $\mathbf{p}^{\mathrm{S}}$ is an arbitrary symmetric vector ($\hat{\mathbf{p}}^{\mathrm{S}} = \mathbf{p}^{\mathrm{S}}$). Each algebra $\mathcal{R}(\cdot;\mathbf{p}^{\mathrm{S}})$ contains the algebra C_{-1}^{S} of all $n \times n$ symmetric (-1)-circulant matrices; therefore, by the identity $\{A : \mathcal{C}_A(T_{0,0}^{-1,-1}) = 0\} = C_{-1} + JC_{-1}$ (found in [9]), $\mathcal{R}(\cdot;\mathbf{p}^{\mathrm{S}}) = C_{-1}^{\mathrm{S}} + J\tilde{C}_{-1}$ for some subset $\tilde{C}_{-1}$ (depending on $\mathbf{p}^{\mathrm{S}}$) of the space C_{-1} of (-1)-circulant matrices.

Algebra μ. If $\mathbf{p}^{\mathrm{S}} = \mathbf{0}$, we have the space

$$\mu = \mathcal{R}(\cdot;\mathbf{0}) = \xi(0,-1,\mathbf{0}) = \tau_{0,0}^{-1,-1} = \{A \in M_n(\mathbb{C}) : \mathcal{C}_A(T_{0,0}^{-1,-1}) = \mathcal{C}_A(J) = 0\} \tag{4.11}$$

naturally related to η. Notice that $\mu = C_{-1}^{\mathrm{S}} + JC_{-1}^{\mathrm{S}}$; in fact $C_{-1}^{\mathrm{S}} + JC_{-1}^{\mathrm{S}} \subset \mu$ and

$$\begin{aligned}
\dim(C_{-1}^{\mathrm{S}} + JC_{-1}^{\mathrm{S}}) &= \dim C_{-1}^{\mathrm{S}} + \dim JC_{-1}^{\mathrm{S}} - \dim C_{-1}^{\mathrm{S}} \cap JC_{-1}^{\mathrm{S}} \\
&= 2\dim C_{-1}^{\mathrm{S}} - \dim C_{-1}^{\mathrm{S}} \cap JC_{-1}^{\mathrm{S}} \\
&= \begin{cases} 2\left(\dfrac{n}{2}\right) - 0 & \text{if } n \text{ is even,} \\ 2\left(\dfrac{n+1}{2}\right) - 1 & \text{if } n \text{ is odd,} \end{cases}
\end{aligned}$$

that is, $\dim(C_{-1}^{\mathrm{S}} + JC_{-1}^{\mathrm{S}}) = n$.

Algebra $\mathcal{K}$. If $\mathbf{p}^{\mathrm{S}} = -\frac{1}{2}(e_2 + e_{n-2})$, we have the space

$$\mathcal{K} = \mathcal{R}\left(\cdot; -\frac{1}{2}(e_2 + e_{n-2})\right) = \xi\left(0,-1,-\frac{1}{2}(e_2 + e_{n-2})\right) \tag{4.12}$$

naturally related to $\mathcal{H}$. Notice that $\mathcal{K} = C_{-1}^{\mathrm{S}} + JP_{-1}C_{-1}^{\mathrm{SK}}$, where $P_{-1} = C_{-1}(\mathbf{e}_2)$ and C_{-1}^{SK} is the set of all $n \times n$ skewsymmetric (-1)-circulant matrices. In order to prove this fact, first show (by proceeding as for $\mathcal{H}$) the inclusion $C_{-1}^{\mathrm{S}} + JP_{-1}C_{-1}^{\mathrm{SK}} \subset \mathcal{K}$, and then use the identity

$$\begin{aligned}
\dim(C_{-1}^{\mathrm{S}} + JP_{-1}C_{-1}^{\mathrm{SK}}) &= \dim C_{-1}^{\mathrm{S}} + \dim JP_{-1}C_{-1}^{\mathrm{SK}} - \dim C_{-1}^{\mathrm{S}} \cap JP_{-1}C_{-1}^{\mathrm{SK}} \\
&= \begin{cases} \left(\dfrac{n}{2}\right) + \left(\dfrac{n}{2}\right) - 0 & \text{if } n \text{ is even,} \\ \left(\dfrac{n+1}{2}\right) + \left(\dfrac{n-1}{2}\right) - 0 & \text{if } n \text{ is odd,} \end{cases} = n.
\end{aligned}$$

The matrices of $\mathcal{K}$ are simultaneously diagonalized by a similarity transformation analogue to the Hartley transform (skew-Hartley transform). Also the algebra μ is associated with a fast discrete transform. (See Theorem 5.2 and the following remark.) In Theorem 4.3 the most significant displacement decompositions are stated in terms of the algebras η, μ, $\mathcal{H}$, and $\mathcal{K}$.

THEOREM 4.3. *If* $AT_{0,0}^{1,1} - T_{0,0}^{1,1}A = \sum_{m=1}^{\alpha} \mathbf{x}_m \mathbf{y}_m^T$, *then*

$$A\mathcal{S}(\mathbf{e}_n;\mathbf{p}^{\mathrm{SK}}) + \mathcal{R}(\mathbf{e}_n;\mathbf{p}^{s})A = \frac{1}{2}\sum_{m=1}^{\alpha} \mathcal{R}(\mathbf{x}_m;\mathbf{p}^{\mathrm{S}})\mathcal{S}(\mathbf{y}_m;\mathbf{p}^{\mathrm{SK}}) + \mathcal{S}(A^T\mathbf{e}_n;\mathbf{p}^{\mathrm{SK}}) + \mathcal{R}(\mathbf{e}_n;\mathbf{p}^{s})\mathcal{S}(A^T\mathbf{e}_1;\mathbf{p}^{\mathrm{SK}}) \tag{4.13}$$

and, in particular,

$$AJ + JA = \frac{1}{2}\sum_{m=1}^{\alpha} \mu(\mathbf{x}_m)\eta(\mathbf{y}_m) + \eta((AJ+JA)^T\mathbf{e}_1), \tag{4.14}$$

$$AJ + \mathcal{K}(\mathbf{e}_n)A = \frac{1}{2}\sum_{m=1}^{\alpha} \mathcal{K}(\mathbf{x}_m)\eta(\mathbf{y}_m) + \eta(A^T\mathbf{e}_n) + \mathcal{K}(\mathbf{e}_n)\eta(A^T\mathbf{e}_1), \tag{4.15}$$

$$A\mathcal{H}(\mathbf{e}_n) + JA = \frac{1}{2}\sum_{m=1}^{\alpha} \mu(\mathbf{x}_m)\mathcal{H}(\mathbf{y}_m) + \mathcal{H}(A^T\mathbf{e}_n) + J\mathcal{H}(A^T\mathbf{e}_1), \tag{4.16}$$

$$A\mathcal{H}(\mathbf{e}_n) + \mathcal{K}(\mathbf{e}_n)A = \frac{1}{2}\sum_{m=1}^{\alpha} \mathcal{K}(\mathbf{x}_m)\mathcal{H}(\mathbf{y}_m) + \mathcal{H}(A^T\mathbf{e}_n) + \mathcal{K}(\mathbf{e}_n)\mathcal{H}(A^T\mathbf{e}_1). \tag{4.17}$$

Proof. For (4.13) set $\varepsilon = \varphi = \varepsilon' = \varphi' = 0$, $\beta = 1$, $\beta' = -1$ in Theorem 3.2. The particular cases (4.14)–(4.17) correspond, respectively, to the choices $\mathbf{p}^{\mathrm{S}} = \mathbf{p}^{\mathrm{SK}} = \mathbf{0}$, $\mathbf{p}^{\mathrm{S}} = -\frac{1}{2}(e_2 + e_{n-2})$ and $\mathbf{p}^{\mathrm{SK}} = \mathbf{0}$, $\mathbf{p}^{\mathrm{S}} = \mathbf{0}$ and $\mathbf{p}^{\mathrm{SK}} = \frac{1}{2}(e_2 - e_{n-2})$, and $\mathbf{p}^{\mathrm{S}} = -\frac{1}{2}(e_2 + e_{n-2})$ and $\mathbf{p}^{\mathrm{SK}} = \frac{1}{2}(e_2 - e_{n-2})$. □

If the matrix A is *centrosymmetric* (i.e., $AJ = JA$) the formulas (4.14)–(4.16) give explicit representations of A in terms of the algebras μ, η, $\mathcal{H}$, and $\mathcal{K}$. In fact the matrices $2J$, $J + \mathcal{K}(\mathbf{e}_n)$, and $J + \mathcal{H}(\mathbf{e}_n)$ are invertible. (It can be shown that $\det(J + \mathcal{H}(\mathbf{e}_n)) = \det(J + \mathcal{K}(\mathbf{e}_n)) = (-1)^{(n-1)/2}2n$ if n is odd; $\det(J + \mathcal{H}(\mathbf{e}_n)) = (-1)^{n/2}n^2$, $\det(J + \mathcal{K}(\mathbf{e}_n)) = (-1)^{n/2}4$ if n is even.) Notice that by Proposition 4.2(i) a symmetric 1-algebra containing $T_{\varepsilon,0}^{\beta,\beta}$, where $\beta = 1$ or $\beta = -1$, may exist only if $\varepsilon = 0$. As a (nonobvious) consequence of this fact, Theorem 3.2 cannot yield effective representations of a *generic* matrix A including algebras $\mathcal{S}(\cdot;\mathbf{p}^{\mathrm{SK}})$ or $\mathcal{R}(\cdot;\mathbf{p}^{\mathrm{S}})$. However, Theorem 3.2 yields such generic formulas, also in terms of non-Hessenberg algebras, if we let both $\mathbb{L}$ and $\mathbb{L}'$ be matrix algebras of the type considered in Proposition 4.2(iii). An example is easily obtained by choosing $\varphi' = \varphi$, $\beta' = \beta$ (in Theorem 3.2) and then—in order to ensure the existence of symmetric 1-algebras $\mathbb{L} \supset T_{\varepsilon,\varphi}^{\beta,\beta}$ and $\mathbb{L}' \supset T_{\varepsilon',\varphi}^{\beta,\beta}$ for $\varepsilon \neq \varepsilon'$—by requiring $I_{\beta\varphi}$ in (4.2) to be nonsingular (see Proposition 4.2). For the sake of brevity we mention only some values of β and φ for which $I_{\beta\varphi}$ is nonsingular and $I_{\beta\varphi}^{-1}$ is known (in the sense of Lemma 4.1) for any value of n: φ arbitrary, $\beta = 0$ [10]; $\varphi = 0$, $\beta^2 \neq 1$; $\varphi = 2$, $\beta = 1$; $\varphi = -2$, $\beta = -1$.

Formula (4.14) is exploited in section 5 to state a simple espression of the inverse of a centrosymmetric Toeplitz-plus-Hankel matrix $T + H$. This expression allows us to calculate $(T+H)^{-1}\mathbf{f}$, $\mathbf{f} \in \mathbb{C}^n$, by performing essentially 10 DFTs reducible to 8 in the case $H = 0$, $[T^{-1}]_{11} \neq 0$, matching both best limits known so far.

5. Toeplitz-plus-Hankel inversion formulas. Theorem 3.2, the results of the previous section, and the fact that the rank of $\mathcal{C}_{T_{\varepsilon,\varphi}^{\beta,\beta}}((T+H)^{-1})$ is 4 for all values of ε, φ, β (see [26] for the case $\varepsilon=\varphi=\beta=0$) yield new representations of the inverse of a Toeplitz-plus-Hankel matrix $T+H$ (or, more generally, of $(T+H)$-like matrices, that is, structured matrices A for which $\operatorname{rank}\mathcal{C}_{T_{\varepsilon,\varphi}^{\beta,\beta}}(A)$ is small with respect to n). These are similar to other formulas found in [1, 6, 9, 10, 16, 17, 20, 23, 32], but they involve new n-dimensional matrix algebras different from HAs. The formulas so obtained can be used to solve a linear system $(T+H)\mathbf{x}=\mathbf{f}$, $\mathbf{f}=\mathbb{C}^n$, in $O(n\log n)$ arithmetic operations (via the computation of $(T+H)^{-1}\mathbf{f}$), provided the 8 vectors defining $\mathcal{C}_{T_{\varepsilon,\varphi}^{\beta,\beta}}((T+H)^{-1})$ are known. Here only the centrosymmetric case is considered in detail.

This approach (compared to a direct triangular factorization of $T+H$ [33, 27]) is significant especially in case a distinction is emphasized between a preprocessing stage—where only operations on elements of $T+H$ are performed—and a successive stage of complexity $O(n\log n)$, where the linear system $(T+H)\mathbf{x}=\mathbf{f}$, $\mathbf{f}\in\mathbb{C}^n$, is solved. This distinction is justified when many different linear systems $(T+H)\mathbf{x}=\mathbf{f}_i$ have to be solved. The same point of view is assumed by Gohberg and Olshevsky in [21, 22], where the complexity of the computation of $A\mathbf{f}$ with preprocessing on A is studied for different types of structured matrices A, including the case $A=T^{-1}$ for a generic Toeplitz T. (Some results on the complexity of the preprocessing stage are also given in [21, 22].) In particular, they show that the application of T^{-1} to the vector $\mathbf{f}$ can be accomplished with a cost of 6 DFTs of order n and thus generalize the analogous result obtained by Ammar and Gader in the Hermitian case [1]. We mention the fact that if T is symmetric, the above limit can be reduced to 11 DFTs of order $\frac{n}{2}$ by using a formula for T^{-1} involving circulant and (-1)-circulant matrices of order $\frac{n}{2}$ (see [15, 17]). Moreover, it is known [10, 16] that 6 discrete transforms are also enough to compute the product $(T+H)^{-1}\mathbf{f}$, where $T+H$ is a centrosymmetric Toeplitz-plus-Hankel matrix. This fact is also shown in the present paper by using a decomposition of $(T+H)^{-1}$ in terms of Hartley-type matrix algebras (see the remarks after Theorems 5.1 and 5.2).

Let T, $[T]_{ij}=t_{i-j}$, and H, $[H]_{ij}=h_{i+j-2}$, $i,j=1,\dots,n$, be, respectively, a symmetric Toeplitz and a persymmetric Hankel matrix with complex elements, and assume that $T+H$ is nonsingular. Then [26]

$$\begin{aligned}(T+H)^{-1}T_{\varphi,\varphi}^{\beta,\beta}-T_{\varphi,\varphi}^{\beta,\beta}(T+H)^{-1}&=(\mathbf{x}_1-\varphi\mathbf{e}_1-\beta\mathbf{e}_n)\mathbf{w}_1^T+(\hat{\mathbf{x}}_1-\varphi\mathbf{e}_n-\beta\mathbf{e}_1)\hat{\mathbf{w}}_1^T\\&\quad-\mathbf{w}_1(\mathbf{x}_1-\varphi\mathbf{e}_1-\beta\mathbf{e}_n)^T-\hat{\mathbf{w}}_1(\hat{\mathbf{x}}_1-\varphi\mathbf{e}_n-\beta\mathbf{e}_1)^T,\end{aligned}\tag{5.1}$$

where $\mathbf{w}_1$ and $\mathbf{x}_1$ are such that

$$(T+H)\mathbf{w}_1=\mathbf{e}_1\quad\text{and}\quad(T+H)\mathbf{x}_1=[t_1+h_{-1}\ t_2+h_0\cdots t_n+h_{n-2}]^T, h_{-1},t_n\in\mathbb{C}$$

(see also [16, 10]). Equality (5.1) for $\beta=0$, $\varphi=1$ and Theorem 3.2 for $\varepsilon=\varphi=1$, $\varepsilon'=\varphi'=-1$, $\beta=\beta'=0$ let us regain the decomposition of $(T+H)^{-1}$

$$2(T+H)^{-1}=\tau_{-1,-1}(\mathbf{x}_1+\mathbf{e}_1)\tau_{1,1}(\mathbf{w}_1)-\tau_{-1,-1}(\mathbf{w}_1)\tau_{1,1}(\mathbf{x}_1-\mathbf{e}_1)\tag{5.2}$$

found in [10]. Moreover, Theorem 3.2 (via Theorem 4.3) yields new decompositions of $(T+H)^{-1}$ in terms of the matrix algebras η, μ, $\mathcal{H}$, and $\mathcal{K}$ studied in section 4.

THEOREM 5.1.

$$(T+H)^{-1}=\frac{1}{2}\{\mu(\hat{\mathbf{x}}_1+\mathbf{e}_1)\eta(\mathbf{w}_1)-\mu(\mathbf{w}_1)\eta(\hat{\mathbf{x}}_1-\mathbf{e}_1)\},\tag{5.3}$$

$$(T+H)^{-1} = \frac{1}{2}(J+\mathcal{K}(\mathbf{e}_n))^{-1}\{[\mathcal{K}(\mathbf{x}_1+\mathbf{e}_n)+\mathcal{K}(\hat{\mathbf{x}}_1+\mathbf{e}_1)J]\eta(\mathbf{w}_1)$$

$$-[\mathcal{K}(\mathbf{w}_1)+\mathcal{K}(\hat{\mathbf{w}}_1)J]\eta(\mathbf{x}_1-\mathbf{e}_n)\}, \tag{5.4}$$

$$(T+H)^{-1} = \{\mu(\mathbf{x}_1+\mathbf{e}_n)[\mathcal{H}(\mathbf{w}_1)+J\mathcal{H}(\hat{\mathbf{w}}_1)]$$

$$-\mu(\mathbf{w}_1)[\mathcal{H}(\mathbf{x}_1-\mathbf{e}_n)+J\mathcal{H}(\hat{\mathbf{x}}_1-\mathbf{e}_1)]\}\frac{1}{2}(J+\mathcal{H}(\mathbf{e}_n))^{-1}. \tag{5.5}$$

Proof. Exploit (5.1) for $\varphi = 0$, $\beta = 1$ and formulas (4.14), (4.15), and (4.16) of Theorem 4.3, respectively. ∎

Formulas (5.2)–(5.5) can be used to compute $(T+H)^{-1}\mathbf{f}$ by means of a constant number of DFTs, Hartley-type transforms, trigonometric transforms, or mixed-type transforms all computable in $O(n\log n)$ arithmetic operations (see [5, 11, 10, 34], Theorem 5.2, and the following remark). In particular, formula (5.3) is competitive with the formulas found in [16, 10]. In fact, as an immediate consequence of Theorem 5.2, the matrix by vector product $(T+H)^{-1}\mathbf{f}$, $\mathbf{f}\in\mathbb{C}^n$, can be calculated by performing essentially 10 order n DFTs if $(T+H)^{-1}$ is replaced by its expression in (5.3) and if $\mathbf{x}_1$ and $\mathbf{w}_1$ are assumed as known. Moreover, we shall see that, for $H=0$ and $w_{11}=[T^{-1}]_{11}\neq 0$, the number of DFTs can be reduced to 8. The limits 10 and 8 are identical to those obtained in [10] with (5.2); however, here the limit 8 is obtained without the further assumption that the entries of T are real, and the coefficient of n in the surplus of $O(n)$ operations is smaller. Recall that the limit 8 has been obtained for the first time by Ammar and Gader in [1]. Both in [1, 16, 10] and in (5.3) the number of discrete transforms is in any case 6 if the transforms of vectors not depending upon $\mathbf{f}$ are included in the preprocessing stage. Moreover, notice that Rost [32] obtains a simple representation for the "classical" Hankel Bezoutian (and therefore for H^{-1}) in terms of $\tau_{0,0}$ and $\tau_{0,1}$ matrices and refers to a future work concerning with the Toeplitz-plus-Hankel case and with the study of computational properties of these representations.

In the next theorem, $d(\mathbf{z})$, $\mathbf{z}\in\mathbb{C}^n$, denotes the $n\times n$ diagonal matrix whose (k,k) element is z_k, $k=1,\dots,n$, and $\mathbf{i}$ is the imaginary unit. Moreover, if A is an $n\times n$ matrix with complex entries, then A^H denotes the transposed conjugate of A.

THEOREM 5.2. *Set* $\rho=\exp(-\mathbf{i}\pi/n)$, $\bar{\rho}=\rho^{-1}$, $\omega=\rho^2$, $[F]_{ij}=\frac{1}{\sqrt{n}}\,\omega^{(i-1)(j-1)}$, $i,j=1,\dots,n$, $D_\rho=\mathrm{diag}(\rho^{i-1}, i=1,\dots,n)$, *and* $D_\omega=D_\rho^2$. *Then, for all* $\mathbf{z}\in\mathbb{C}^n$,

$$\eta(\mathbf{z}) = M_\eta\Lambda(M_\eta^T\mathbf{z})M_\eta^H, \qquad \Lambda(M_\eta^T\mathbf{z}) = d(M_\eta^T\mathbf{z})d(M_\eta^T\mathbf{e}_1)^{-1}, \tag{5.6}$$

$$\mu(\mathbf{z}) = M_\mu\Lambda(M_\mu^T\mathbf{z})M_\mu^H, \qquad \Lambda(M_\mu^T\mathbf{z}) = d(M_\mu^T\mathbf{z})d(M_\mu^T\mathbf{e}_1)^{-1}, \tag{5.7}$$

where M_η *and* M_μ *are the unitary matrices:*

$$M_\eta = \frac{1}{\sqrt{2}}F\begin{pmatrix} \sqrt{2} & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 1 & & & & & -\omega & \\ \vdots & & \ddots & & & \iddots & & \vdots \\ \vdots & & & 1 & 0 & -\omega^{\frac{n}{2}-1} & & \vdots \\ \vdots & & & 0 & \sqrt{2} & 0 & & \vdots \\ \vdots & & & \omega^{\frac{n}{2}+1} & 0 & 1 & & \vdots \\ & \iddots & & & & & \ddots & \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & \omega^{n-1} & 1 \end{pmatrix}, \tag{5.8}$$

$$M_\mu = \frac{1}{\sqrt{2}} D_\rho F \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & -\overline{\rho}^{n-1} \\ 0 & 1 & \ddots & & & & & \iddots & \iddots & 0 \\ \vdots & \ddots & \ddots & \ddots & & & \iddots & \iddots & \iddots & \vdots \\ \vdots & \iddots & \ddots & \ddots & 0 & 0 & -\overline{\rho}^{3} & \iddots & \iddots & \vdots \\ \vdots & & \iddots & 0 & 1 & -\overline{\rho} & 0 & & & \vdots \\ \vdots & & & 0 & -\overline{\rho}^{n-1} & 1 & 0 & & & \vdots \\ \vdots & & \iddots & \iddots & 0 & 0 & 1 & \ddots & \iddots & \vdots \\ \vdots & \iddots & \iddots & \iddots & & & \ddots & \ddots & \ddots & \vdots \\ 0 & -\overline{\rho}^{3} & \iddots & & & & & \ddots & \ddots & 0 \\ -\overline{\rho} & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 1 \end{pmatrix} \tag{5.9}$$

for n even, and

$$M_\eta = \frac{1}{\sqrt{2}} F \begin{pmatrix} \sqrt{2} & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 1 & \ddots & & & & & & \iddots & -\omega \\ \vdots & \ddots & \ddots & \ddots & & & & \iddots & \iddots & 0 \\ \vdots & \iddots & \ddots & \ddots & 0 & 0 & \iddots & \iddots & \iddots & \vdots \\ \vdots & & \iddots & 0 & 1 & -\omega^{\frac{n-1}{2}} & 0 & \iddots & \iddots & \vdots \\ \vdots & & & \iddots & \omega^{\frac{n+1}{2}} & 1 & \iddots & \iddots & \iddots & \vdots \\ \vdots & & 0 & \iddots & \ddots & \ddots & \ddots & 0 & \iddots & \vdots \\ \vdots & \iddots & \iddots & \iddots & 0 & 0 & \iddots & \iddots & \iddots & \vdots \\ \vdots & \iddots & & \iddots & \iddots & \iddots & \iddots & \iddots & 0 & \vdots \\ 0 & \omega^{n-1} & 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 & 0 \end{pmatrix}, \tag{5.10}$$

$$M_\mu = \frac{1}{\sqrt{2}} D_\rho F \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & -\omega^{\frac{n+1}{2}} \\ 0 & \ddots & \ddots & & & & \iddots & \iddots & 0 \\ \vdots & \ddots & \ddots & \ddots & & & \iddots & \iddots & \vdots \\ \vdots & \iddots & 1 & 0 & -\omega^{n-1} & \iddots & \iddots & \iddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & \iddots & \iddots & \vdots \\ \vdots & & 0 & \sqrt{2} & 0 & \iddots & \iddots & \iddots & \vdots \\ \vdots & \iddots & \iddots & \ddots & \ddots & \ddots & \iddots & \iddots & \vdots \\ \vdots & \iddots & \iddots & \omega & 0 & 1 & \iddots & \iddots & \vdots \\ 0 & \ddots & & \ddots & & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & & \ddots & & \ddots & \ddots & \vdots \\ \omega^{\frac{n-1}{2}} & 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 & 0 \end{pmatrix} \tag{5.11}$$

for n odd. Moreover, for all $\mathbf{z} \in \mathbb{C}^n$,

$$\mathcal{H}(\mathbf{z}) = \sqrt{n}\ H_+ d(H_+\mathbf{z}) H_+ = \sqrt{n}\ H_- d(H_-\mathbf{z}) H_-, \tag{5.12}$$

$$\mathcal{K}(\mathbf{z}) = \sqrt{n}\ K_+ d(K_+^T\mathbf{z}) K_+^T = \sqrt{n}\ K_- d(K_-^T\mathbf{z}) K_-^T, \tag{5.13}$$

where H_+, H_-, K_+, and K_- are the orthonormal matrices defined by

$$[H_\pm]_{ij} = (1/\sqrt{n}\,) \left(\cos \frac{2\pi(i-1)(j-1)}{n} \pm \sin \frac{2\pi(i-1)(j-1)}{n} \right), \tag{5.14}$$

$$[K_\pm]_{ij} = (1/\sqrt{n}\,) \left(\cos \frac{\pi(i-1)(2j-1)}{n} \pm \sin \frac{\pi(i-1)(2j-1)}{n} \right), \tag{5.15}$$

$i, j = 1, \ldots, n$.

Proof. The equalities (5.6) and (5.7) are shown only in the case n even ($n = 2m$). In the case n odd, the proof is similar. Notice that in order to find the matrices in (5.8), (5.10) and (5.9), (5.11), we had to look for a matrix diagonalizing J among the matrices diagonalizing $T_{0,0}^{1,1}$ and $T_{0,0}^{-1,-1}$, respectively.

Let us prove (5.6). Set $\mathbf{c}_i = \sqrt{n}\,F\mathbf{e}_i$, $i = 1, \ldots, n$. By using the identities $\hat{\mathbf{c}}_i = \omega^{n-i+1}\mathbf{c}_{n-i+2}$, $i = 2, \ldots, m$ (recall that, for a vector $\mathbf{z}$, $\hat{\mathbf{z}} = J\mathbf{z}$), one can easily show that

$$M_\eta = \frac{1}{\sqrt{2n}}\,[\sqrt{2}\mathbf{c}_1\ \mathbf{c}_2 + \hat{\mathbf{c}}_2 \cdots \mathbf{c}_m + \hat{\mathbf{c}}_m\ \sqrt{2}\mathbf{c}_{m+1}\ \ \mathbf{c}_{m+2} - \hat{\mathbf{c}}_{m+2} \cdots \mathbf{c}_n - \hat{\mathbf{c}}_n]. \tag{5.16}$$

Moreover, as $T_{0,0}^{1,1} = P_1 + P_1^H$ and $P_1 = FD_\omega F^H$, $T_{0,0}^{1,1}F = F(D_\omega + D_\omega^H)$, i.e.,

$$T_{0,0}^{1,1}[\mathbf{c}_1\ \mathbf{c}_2 \cdots \mathbf{c}_n] = [\mathbf{c}_1\ \mathbf{c}_2 \cdots \mathbf{c}_n]\,\mathrm{diag}(2\cos\frac{2\pi(j-1)}{n}, \quad j = 1, \ldots, n). \tag{5.17}$$

By the centrosymmetry of $T_{0,0}^{1,1}$ (besides $\mathbf{c}_j$) also $\hat{\mathbf{c}}_j$ is an eigenvector of $T_{0,0}^{1,1}$ with associated eigenvalue $2\cos\frac{2\pi(j-1)}{n}$. This remark and equalities (5.17) and (5.16) allow us to say that

$$T_{0,0}^{1,1}M_\eta = M_\eta\,\mathrm{diag}\left(2\cos\frac{2\pi(j-1)}{n},\ j = 1, \ldots, n\right). \tag{5.18}$$

From (5.16) it also follows that

$$\eta(\mathbf{e}_n)M_\eta = JM_\eta = M_\eta \begin{pmatrix} I & O \\ O & -I \end{pmatrix}, \tag{5.19}$$

where the I in (5.19) is the $m \times m$ identity matrix ($\hat{\mathbf{c}}_1 = \mathbf{c}_1$, $\hat{\mathbf{c}}_{m+1} = -\mathbf{c}_{m+1}$). By exploiting, respectively, (5.18) and (5.19), we have that the matrix $M_\eta d(M_\eta^T\mathbf{z})\,d(M_\eta^T\mathbf{e}_1)^{-1}M_\eta^H$ commutes with the matrices $T_{0,0}^{1,1}$ and J $\forall\mathbf{z} \in \mathbb{C}^n$. Moreover, as $M_\eta M_\eta^H = I$, its first row is $\mathbf{z}^T$, and therefore, by (4.7), we have $\eta(\mathbf{z}) = M_\eta d(M_\eta^T\mathbf{z})\,d(M_\eta^T\mathbf{e}_1)^{-1}M_\eta^H$.

Let us prove (5.7). Set $\mathbf{c}_i = \sqrt{n}D_\rho F\mathbf{e}_i$, $i = 1, \ldots, n$. The identities $\hat{\mathbf{c}}_i = -\bar{\rho}^{2i-1}\mathbf{c}_{n+1-i}$, $i = 1, \ldots, m$, yield

$$M_\mu = \frac{1}{\sqrt{2n}}\,[\mathbf{c}_1 + \hat{\mathbf{c}}_1 \cdots \mathbf{c}_m + \hat{\mathbf{c}}_m\ \ \mathbf{c}_{m+1} - \hat{\mathbf{c}}_{m+1} \cdots \mathbf{c}_n - \hat{\mathbf{c}}_n]. \tag{5.20}$$

Moreover, as $T_{0,0}^{-1,-1} = P_{-1} + P_{-1}^H$ and $P_{-1} = D_\rho F\rho D_\omega F^H D_\rho^H$, we have $T_{0,0}^{-1,-1}D_\rho F = D_\rho F(\rho D_\omega + \bar{\rho}D_\omega^H)$, i.e.,

$$T_{0,0}^{-1,-1}[\mathbf{c}_1\ \mathbf{c}_2 \cdots \mathbf{c}_n] = [\mathbf{c}_1\ \mathbf{c}_2 \cdots \mathbf{c}_n]\,\mathrm{diag}\left(2\cos\frac{\pi(2j-1)}{n}, \quad j = 1, \ldots, n\right). \tag{5.21}$$

As in the case of (5.6), the equalities (5.21) and (5.20) yield

$$T_{0,0}^{-1,-1}M_\mu = M_\mu\,\mathrm{diag}\left(2\cos\frac{\pi(2j-1)}{n}, \quad j = 1, \ldots, n\right),$$

$$\mu(\mathbf{e}_n)M_\mu = JM_\mu = M_\mu \begin{pmatrix} I & O \\ O & -I \end{pmatrix},$$

where I is the $m \times m$ identity matrix. Thus the matrix $M_\mu d(M_\mu^T \mathbf{z}) d(M_\mu^T \mathbf{e}_1)^{-1} M_\mu^H$ commutes with the matrices $T_{0,0}^{-1,-1}$ and J $\forall \mathbf{z} \in \mathbb{C}^n$. Moreover, as $M_\mu M_\mu^H = I$, its first row is $\mathbf{z}^T$ and therefore, by (4.11), we have $\mu(\mathbf{z}) = M_\mu d(M_\mu^T \mathbf{z}) d(M_\mu^T \mathbf{e}_1)^{-1} M_\mu^H$.

Finally, let us prove (5.13). This proof is analogous to the proof of the first equality in (5.12), which is in [5]. Notice that $D_\rho F = \frac{1}{\sqrt{n}} (M - \mathbf{i}N)$, where $[M]_{ij} = \cos \frac{\pi(i-1)(2j-1)}{n}$ and $[N]_{ij} = \sin \frac{\pi(i-1)(2j-1)}{n}$, $i, j = 1, \ldots, n$. Moreover, from the identities $(D_\rho F)^H D_\rho F = I$ and $(D_\rho F)^T D_\rho F = J$, we have

$$M^T M + N^T N = nI \quad \text{and} \quad M^T N + N^T M = 0,$$

respectively. Observe that $K_+ = \frac{1}{\sqrt{n}} (M + N)$ $\left[K_- = \frac{1}{\sqrt{n}} (M - N)\right]$. Thus, by the above equalities, $K_+^T K_+ = I$ $[K_-^T K_- = I]$. Moreover $M = MJ = -JP_{-1}M$ and $-N = NJ = -JP_{-1}N$; therefore, $K_+ J = -JP_{-1}K_+$ $[K_- J = -JP_{-1}K_-]$.

Let A be a generic (-1)-circulant matrix. We know that $(D_\rho F)^H A D_\rho F = D_A$, where D_A is a diagonal matrix and thus

$$\text{(5.22)} \qquad \mathrm{Re}D_A = \frac{1}{n}(M^T A M + N^T A N), \quad \mathrm{Im}D_A = \frac{1}{n} (N^T A M - M^T A N).$$

From (5.22) it follows that if A is a (-1)-circulant matrix, then

$$K_+^T A K_+ = \mathrm{Re}D_A - J\mathrm{Im}D_A \quad [K_-^T A K_- = \mathrm{Re}D_A + J\mathrm{Im}D_A].$$

Now let E be a generic element of $\mathcal{K} = C_{-1}^{\mathrm{S}} + JP_{-1}C_{-1}^{\mathrm{SK}}$ and assume that the entries of E are real, i.e., $E = E_{-1}^{\mathrm{S}} + JP_{-1}E_{-1}^{\mathrm{SK}}$, where E_{-1}^{S} is a real symmetric (-1)-circulant matrix and E_{-1}^{SK} is a real skewsymmetric (-1)-circulant matrix. Observe that the eigenvalues of E_{-1}^{S} is a real skewsymmetric (-1)-circulant matrix. Observe that the eigenvalues of E_{-1}^{S} and E_{-1}^{SK} are, respectively, real and purely imaginary. Thus

$$\begin{aligned} K_+^T E K_+ &= K_+^T E_{-1}^{\mathrm{S}} K_+ + K_+^T J P_{-1} E_{-1}^{\mathrm{SK}} K_+ = K_+^T E_{-1}^{\mathrm{S}} K_+ - J K_+^T E_{-1}^{\mathrm{SK}} K_+ \\ &= \mathrm{Re}D_{E_{-1}^{\mathrm{S}}} + \mathrm{Im}D_{E_{-1}^{\mathrm{SK}}} \\ &\left[K_-^T E K_- = \mathrm{Re}D_{E_{-1}^{\mathrm{S}}} - \mathrm{Im}D_{E_{-1}^{\mathrm{SK}}}\right]. \end{aligned}$$

We have proved that $K_\pm^T E K_\pm = d(\mathbf{z}_E^\pm)$ for some $\mathbf{z}_E^\pm \in \mathbb{R}^n$. The thesis, in the real case, follows from the equalities $\mathbf{e}_1^T E K_\pm = \mathbf{e}_1^T K_\pm d(\mathbf{z}_E^\pm) = \mathbf{z}_E^{\pm T} d(K_\pm^T \mathbf{e}_1) = \frac{1}{\sqrt{n}}\, \mathbf{z}_E^{\pm T}$. For the complex case, simply observe that if $\mathbf{z} \in \mathbb{C}^n$, then $\mathbf{z} = \mathbf{z}_1 + \mathbf{i}\mathbf{z}_2$, where $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^n$, and that $\mathcal{K}(\mathbf{z}) = \mathcal{K}(\mathbf{z}_1 + \mathbf{i}\mathbf{z}_2) = \mathcal{K}(\mathbf{z}_1) + \mathbf{i}\mathcal{K}(\mathbf{z}_2)$. ∎

Remark. If n is an integer power of 2, then the *skew-Hartley* transform $\sqrt{n}\, K_\pm \mathbf{z}$ $(\sqrt{n}\, K_\pm^T \mathbf{z})$, $\mathbf{z} \in \mathbb{R}^n$, can be computed in at most $\frac{3}{2} n \log_2 n$ additions and $n \log_2 n$ multiplications of real numbers, i.e., with the same cost of the Hartley transform $\sqrt{n}\, H_\pm \mathbf{z}$. (For this last transform, see [5] and the references cited therein.) In fact, for $K_\pm^{(n)} = K_\pm$ we have

$$K_\pm^{(n)} = \frac{1}{\sqrt{2}} Q \begin{pmatrix} K_\pm^{\left(\frac{n}{2}\right)} & K_\pm^{\left(\frac{n}{2}\right)} \\ K_\pm^{\left(\frac{n}{2}\right)} R_\pm & -K_\pm^{\left(\frac{n}{2}\right)} R_\pm \end{pmatrix},$$

where $R_\pm = \mathrm{diag}(\cos \frac{(2j-1)\pi}{n}, j = 1, \ldots, \frac{n}{2}) \pm J\, \mathrm{diag}(\sin \frac{(2j-1)\pi}{n}, j = 1, \ldots, \frac{n}{2})$ and Q is the permutation matrix $Q\mathbf{e}_j = \mathbf{e}_{2j-1}$, $Q\mathbf{e}_{n-j+1} = \mathbf{e}_{n-2j+2}$, $j = 1, \ldots, \frac{n}{2}$. (For

$H_{\pm}^{(n)} = H_{\pm}$ an analogous identity holds, where $R_{\pm} = \operatorname{diag}(\cos \frac{2\pi(j-1)}{n},\ j = 1, \ldots, \frac{n}{2}) \pm JP_{\beta} \operatorname{diag}(\sin \frac{2\pi(j-1)}{n},\ j = 1, \ldots, \frac{n}{2})$.

If $H = 0$ and $w_{11} = [T^{-1}]_{11} \neq 0$, then $\mathbf{x}_1 = -(1/w_{11})P_0\mathbf{w}_1$ [25] (see also [16]). By exploiting this fact and the identities (5.6) and (5.7) in Theorem 5.2, formula (5.3) becomes

$$T^{-1} = \frac{1}{2w_{11}}\{\mu(\mathbf{w}_1)\eta(JP_1\mathbf{w}_1) - \mu(JP_{-1}\mathbf{w}_1)\eta(\mathbf{w}_1)\} \tag{5.23}$$

$$= \frac{1}{2w_{11}} M_{\mu}\{\Lambda(M_{\mu}^T\mathbf{w}_1)M_{\mu}^H M_{\eta}\Lambda(M_{\eta}^T JP_1\mathbf{w}_1)$$

$$- \Lambda(M_{\mu}^T JP_{-1}\mathbf{w}_1)M_{\mu}^H M_{\eta}\Lambda(M_{\eta}^T\mathbf{w}_1)\}M_{\eta}^H.$$

Observe that the vectors $\mathbf{z}$ in the four matrices $\Lambda(\mathbf{z})$ appearing in this last formula can be computed in $O(n)$ arithmetic operations once that $F\mathbf{w}_1$ and $FD_{\rho}\mathbf{w}_1$ are calculated (use the identities $F(JP_1)\mathbf{w}_1 = (JP_1)F\mathbf{w}_1$ and $FD_{\rho}(JP_{-1})\mathbf{w}_1 = -JFD_{\rho}\mathbf{w}_1$). Thus, if $\mathbf{w}_1$ is known, the vector $T^{-1}\mathbf{f}$, $\mathbf{f} \in \mathbb{C}^n$, can be computed by performing eight DFTs plus $\mathrm{O}(n)$ arithmetic operations.

In [1] Ammar and Gader obtain the same result by exploiting the representation in terms of circulant and (-1)-circulant matrices

$$T^{-1} = \frac{1}{2w_{11}}\{C_{-1}(\mathbf{w}_1)C(\mathbf{w}_1)^T + C_{-1}(\mathbf{w}_1)^T C(\mathbf{w}_1)\}, \tag{5.24}$$

which is a consequence of the following formula, holding for a *generic* nonsingular Toeplitz matrix $T = (t_{i-j})_{i,j=1}^n$,

$$T^{-1} = \frac{1}{2}\{C_{-1}(\hat{\mathbf{w}}_n)C(\mathbf{e}_1 - \hat{\mathbf{x}}_1) + C_{-1}(\mathbf{e}_1 + \hat{\mathbf{x}}_1)C(\hat{\mathbf{w}}_n)\}, \tag{5.25}$$

where $\mathbf{w}_n = T^{-1}\mathbf{e}_n$ and $T\mathbf{x}_1 = [t_1\ t_2 \cdots t_n]^T$, $t_n \in \mathbb{C}$ (see also [16]). Formulas of type (5.25), generalizing the Ammar–Gader formula (5.24), were first derived by Gohberg and Olshevsky in [20, 22]. Notice that, by using formula (5.25) or the analogous formulas in [20, 22], the product $T^{-1}\mathbf{f}$ for a generic T can be calculated with essentially 10 order n DFTs [21, 22], i.e., with the same amount of computation required to compute $(T+H)^{-1}\mathbf{f}$ for $T = T^T$ and $H = JHJ$ via (5.3). Both in (5.24), (5.25) and in (5.3), (5.23) the number of discrete transforms is 6 if the transforms of vectors not depending upon $\mathbf{f}$ are included in the preprocessing stage. Thus formulas (5.3) and (5.23) seem to be the analogues of the Ammar–Gader–Gohberg–Olshevsky-type formulas for the centrosymmetric Toeplitz-plus-Hankel case.

Acknowledgments. I wish to thank Professor Paolo Zellini for proposing the problem and for his comments and suggestions. I would also like to thank the referees for their useful comments.

REFERENCES

[1] G. Ammar and P. D. Gader, *New decompositions of the inverse of a Toeplitz matrix*, in Signal Processing, Scattering and Operator Theory, and Numerical Methods, in Proc. Int. Symp. MTNS-89, Vol. III, M. A. Kaashoek, J. H. van Schuppen, and A. C. N. Ran, eds., Birkhäuser, Boston, MA, 1990, pp. 421–428.

[2] E. Bannai and T. Ito, *Algebraic Combinatorics.* I. *Association Schemes*, Benjamin/Cummings, Menlo Park, CA, 1984.

[3] R. B. Bapat and V. S. Sunder, *On hypergroups of matrices*, Linear and Multilinear Algebra, 29 (1991), pp. 125–140.
[4] R. Bevilacqua and P. Zellini, *Closure, commutativity and minimal complexity of some spaces of matrices*, Linear and Multilinear Algebra, 25 (1989), pp. 1–25.
[5] D. Bini and P. Favati, *On a matrix algebra related to the discrete Hartley transform*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 500–507.
[6] D. Bini and V. Pan, *Improved parallel computations with Toeplitz-like and Hankel-like matrices*, Linear Algebra Appl., 188/189 (1993), pp. 3–29.
[7] D. Bini and V. Pan, *Polynomial and Matrix Computations,* Vol. 1. *Fundamental Algorithms*, Birkhäuser, Boston, MA, 1994.
[8] E. Bozzo, *Matrix Algebras and Discrete Transforms*, Ph.D. thesis, TD-1/94, University of Pisa, Pisa, Italy, 1994.
[9] E. Bozzo, *Algebras of higher dimension for displacement decompositions and computations with Toeplitz plus Hankel matrices*, Linear Algebra Appl., 230 (1995), pp. 127–150.
[10] E. Bozzo and C. Di Fiore, *On the use of certain matrix algebras associated with discrete trigonometric transforms in matrix displacement decomposition*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 312–326.
[11] R. N. Bracewell, *The fast Hartley transform*, in Proc. IEEE, 72 (1984), pp. 1010–1018.
[12] R. N. Bracewell, *The Hartley Transform*, Clarendon Press, Oxford, 1986.
[13] R. H. Chan and M. K. Ng, *Conjugate gradient methods for Toeplitz systems*, SIAM Rev., 38 (1996), pp. 427–482.
[14] P. J. Davis, *Circulant Matrices*, John Wiley, New York, 1979.
[15] C. Di Fiore, *Decomposizioni di Matrici e Classi di Algebre di Matrici Commutative*, Ph.D. thesis, University "La Sapienza," Rome, Italy, 1994–95.
[16] C. Di Fiore and P. Zellini, *Matrix decompositions using displacement rank and classes of commutative matrix algebras*, Linear Algebra Appl., 229 (1995), pp. 49–99.
[17] C. Di Fiore and P. Zellini, *Matrix displacement decompositions and applications to Toeplitz linear systems*, Linear Algebra Appl., 268 (1998), pp. 197–225.
[18] P. D. Gader, *Displacement operator based decompositions of matrices using circulants or other group matrices*, Linear Algebra Appl., 139 (1990), pp. 111–131.
[19] I. Gohberg, T. Kailath, and V. Olshevsky, *Fast Gaussian elimination with partial pivoting for matrices with displacement structure*, Math. Comp., 64 (1995), pp. 1557–1576.
[20] I. Gohberg and V. Olshevsky, *Circulants, displacements and decompositions of matrices*, Integral Equations Operator Theory, 15 (1992), pp. 730–743.
[21] I. Gohberg and V. Olshevsky, *Fast algorithms with preprocessing for matrix-vector multiplication problems*, J. Complexity, 10 (1994), pp. 411–427.
[22] I. Gohberg and V. Olshevsky, *Complexity of multiplication with vectors for structured matrices*, Linear Algebra Appl., 202 (1994), pp. 163–192.
[23] I. Gohberg and A. Semencul, *On the inversion of finite Toeplitz matrices and their continuous analogs*, Mat. Issled., 7(2) (1972), pp. 201–233 (in Russian).
[24] G. Heinig, *Inversion of generalized Cauchy matrices and other classes of structured matrices*, in Linear Algebra for Signal Processing, IMA Vol. Math. Appl. 69, A. Bojanczyk and G. Cybenko, eds., Springer, New York, 1994, pp. 63–81.
[25] G. Heinig and K. Rost, *Algebraic Methods for Toeplitz-like Matrices and Operators*, Akademie–Verlag, Berlin, 1984.
[26] G. Heinig and K. Rost, *On the inverses of Toeplitz-plus-Hankel matrices*, Linear Algebra Appl., 106 (1988), pp. 39–52.
[27] T. Kailath and J. Chun, *Generalized displacement structure for block-Toeplitz, Toeplitz-block, and Toeplitz-derived matrices*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 114–128.
[28] T. Kailath, S. Kung, and M. Morf, *Displacement ranks of matrices and linear equations*, J. Math. Anal. Appl., 68 (1979), pp. 395–407.
[29] T. Kailath and V. Olshevsky, *Displacement structure approach to Chebyshev-Vandermonde and related matrices*, Integral Equations Operator Theory, 22 (1995), pp. 65–92.
[30] T. Kailath and V. Olshevsky, *Displacement-structure approach to polynomial Vandermonde and related matrices*, Linear Algebra Appl., 261 (1997), pp. 49–90.
[31] T. Kailath and A. H. Sayed, *Displacement structure: Theory and applications*, SIAM Rev. 37 (1995), pp. 297–386.
[32] K. Rost, *Generalized companion matrices and matrix representations for generalized Bezoutians*, Linear Algebra Appl., 193 (1993), pp. 151–172.
[33] A. H. Sayed, H. Lev-Ari, and T. Kailath, *Fast triangular factorization of the sum of quasi-Toeplitz and quasi-Hankel matrices*, Linear Algebra Appl., 191 (1993), pp. 77–106.
[34] Z. Wang, *Fast algorithms for the discrete W transform and for the discrete Fourier transform*, IEEE Trans. Acoust. Speech Signal Process., 32 (1984), pp. 803–816.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 668–678
© 2000 Society for Industrial and Applied Mathematics

CONVEXITY OF THE JOINT NUMERICAL RANGE*

CHI-KWONG LI† AND YIU-TUNG POON‡

Dedicated to Professor Yik-Hoi Au-Yeung on the occasion of his retirement.

Abstract. Let $A = (A_1, \ldots, A_m)$ be an m-tuple of $n \times n$ Hermitian matrices. For $1 \le k \le n$, the kth *joint numerical range* of A is defined by

$$W_k(A) = \{(\mathrm{tr}\,(X^*A_1X), \ldots, \mathrm{tr}\,(X^*A_mX)) : X \in \mathbf{C}^{n\times k}, X^*X = I_k\}.$$

We consider linearly independent families of Hermitian matrices $\{A_1, \ldots, A_m\}$ so that $W_k(A)$ is convex. It is shown that m can reach the upper bound $2k(n-k)+1$. A key idea in our study is relating the convexity of $W_k(A)$ to the problem of constructing rank k orthogonal projections under linear constraints determined by A. The techniques are extended to study the convexity of other generalized numerical ranges and the corresponding matrix construction problems.

Key words. convexity, numerical range, matrix constriction

AMS subject classifications. 15A60, 15A57, 15A63

PII. S0895479898343516

1. Introduction.

Let $A = (A_1, \ldots, A_m)$ be an m-tuple of $n \times n$ Hermitian matrices. For $1 \le k \le n$, the kth *joint numerical range* of A is defined as

$$W_k(A) = \{(\mathrm{tr}\,(X^*A_1X), \ldots, \mathrm{tr}\,(X^*A_mX)) : X \in \mathbf{C}^{n\times k},\ X^*X = I_k\}.$$

When $k = 1$, it reduces to the usual joint numerical range of A that is useful in the study of various pure and applied subjects (see [2, 3, 4, 5, 7]). In particular, in the study of structured singular values arising in robust stability (see [6, 7, 15]), it is important that $W_1(A)$ is convex. Unfortunately, $W_1(A)$ is not always convex if $m > 3$ (see, e.g., [1]). We modify the example in [1] to show that the same conclusion holds for $W_k(A)$ in the following example.

Example 1.1. Let $A_1 = \left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right) \oplus 0_{n-2}$, $A_2 = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right) \oplus 0_{n-2}$, $A_3 = \left(\begin{smallmatrix} 0 & i \\ -i & 0 \end{smallmatrix}\right) \oplus 0_{n-2}$, $A_4 = I_2 \oplus 2I_{k-1} \oplus 0_{n-k-1}$, $A_j = I_n$ for all other $j \le m$. Then $W_k(A)$ is not convex.

Proof. Suppose $W_k(A)$ is convex. Note that $2k-1$ is the sum of the k largest eigenvalues of A_4. If $X \in \mathbf{C}^{n\times k}$ with $X^*X = I_k$ such that $\mathrm{tr}\,(X^*A_4X) = 2k-1$, then (see, e.g., [10]) the column space of X must be spanned by k eigenvectors of A_4 corresponding to the k largest eigenvalues. Thus, there exists a $k \times k$ unitary matrix V such that XV has columns $\alpha e_1 + \beta e_2, e_3, \ldots, e_{k+1}$, where $|\alpha|^2 + |\beta|^2 = 1$. Then one can check that the subset

$$S = \{(a_1, \ldots, a_m) \in W_k(A) : a_4 = 2k-1\} = \{(a, b, c, 2k-1, k, \ldots, k) : a^2+b^2+c^2 = 1\}$$

of $W_k(A)$ is not convex, which is a contradiction. ∎

Apart from these negative results, one would still hope to get the convexity conclusion if A has some special structure. In this paper, we consider linearly independent

*Received by the editors August 12, 1998; accepted for publication (in revised form) by R. Bhatia December 4, 1998; published electronically January 19, 2000.

http://www.siam.org/journals/simax/21-2/34351.html

†Department of Mathematics, College of William and Mary, P.O. Box 8795, Williamsburg, VA 23187-8795 (ckli@math.wm.edu). The research of this author was partially supported by an NSF grant. This paper was done while he was visiting the University of Toronto in the academic year 1998–1999 and was supported by a Faculty Research grant of the College of William and Mary.

‡Department of Mathematics, Iowa State University, Ames, IA 50011 (poon@math.iastate.edu).

families $\{A_1, \ldots, A_m\}$ so that $W_k(A)$ is convex. In particular, we show that the maximum value of m is $2k(n-k)+1$, which is much larger than 3 as ensured by the general theorem (see [2]).

A key idea in our study is to view $W_k(A)$ as the image of the set of all rank k projections under the linear map

$$\phi(P) = (\operatorname{tr} A_1 P, \ldots, \operatorname{tr} A_m P).$$

To make this claim precise, denote by

$$\mathcal{U}(C) = \{U^* C U : U^* U = I_n\}$$

the *unitary similarity orbit* of a given Hermitian matrix C. If

$$J_k = I_k \oplus 0_{n-k},$$

then $\mathcal{U}(J_k)$ is the set of rank k orthogonal projections. Since a matrix P belongs to $\mathcal{U}(J_k)$ if and only if $P = XX^*$ for some $X \in \mathbf{C}^{n\times k}$ satisfying $X^*X = I_k$, we have

$$W_k(A) = \{(\operatorname{tr} A_1 P, \ldots, \operatorname{tr} A_m P) : P \in \mathcal{U}(J_k)\}.$$

Thus, $W_k(A)$ is just the image of $\mathcal{U}(J_k)$ under the linear projection onto the linear space spanned by $\{A_1, \ldots, A_m\}$. With this new formulation of $W_k(A)$, its convex hull can be written as

$$\operatorname{conv} W_k(A) = \{(\operatorname{tr} A_1 X, \ldots, \operatorname{tr} A_m X) : X \in \operatorname{conv} \mathcal{U}(J_k)\}.$$

Consequently, we have the following theorem.

THEOREM 1.2. *Let $A = (A_1, \ldots, A_m)$ be an m-tuple of $n \times n$ Hermitian matrices. Then $W_k(A)$ is convex if and only if for any $X \in \operatorname{conv} \mathcal{U}(J_k)$ there exists $P \in \mathcal{U}(J_k)$ so that $\operatorname{tr} A_j P = \operatorname{tr} A_j X$ for all $j = 1, \ldots, m$.*

This rather simple observation turns out to be very useful in our study. Moreover, if $A_1, \ldots, A_m$ are chosen from the standard basis for $n \times n$ Hermitian matrices, then the constraints $\operatorname{tr} A_j P = \operatorname{tr} A_j X$ for all $j = 1, \ldots, m$, are just specifications of certain entries of P. Thus, the problem reduces to construction of a rank k orthogonal projection P with some specified entries.

We shall use $\{E_{11}, E_{12}, \ldots, E_{nn}\}$ to denote the standard basis for $\mathbf{C}^{n\times n}$ in our discussion. Moreover, the following observations will be used frequently:

1. $W_k(A) = W_k(V^* A_1 V, \ldots, V^* A_m V)$ for any unitary V.
2. $W_k(A)$ is convex if and only if $W_k(I, A_1, \ldots, A_m)$ is convex.
3. Let $\{B_1, \ldots, B_s\}$ be a basis for $\operatorname{span}\{A_1, \ldots, A_m\}$. Then $W_k(A)$ is convex if and only if $W_k(B_1, \ldots, B_s)$ is convex.
4. Suppose $W_k(A)$ is convex. If $B_j \in \operatorname{span}\{I, A_1, \ldots, A_m\}$ for $1 \le j \le s$, then $W_k(B_1, \ldots, B_s)$ is convex.

2. The *k*th numerical range and orthogonal projections. In this section, we study linearly independent families $\{A_1, \ldots, A_m\}$ for which $W_k(A_1, \ldots, A_m)$ is convex. We call such a family a *linearly independent convex family* for the kth numerical range. In particular, we would like to study maximal (in the sense of set inclusion) linearly independent convex families.

We begin with the following result on the completion of a certain partial Hermitian matrix to a rank k orthogonal projection.

THEOREM 2.1. *Let $1 \le k < n$ and $X \in \mathbf{C}^{k\times(n-k)}$. There exists a rank k orthogonal projection of the form $\left(\begin{smallmatrix} * & X \\ X^* & * \end{smallmatrix}\right)$ if and only if $\|X\| \le 1/2$.*

Proof. Suppose X is an off-diagonal submatrix of a rank k orthogonal projection. Then $\|X\| \le 1/2$ by the result in [12]. For the converse, let $m = \min\{k, n-k\}$. Suppose $2X$ has singular value decomposition UDV, where U and V are unitary and

$$D_{rs} = \begin{cases} \sin t_r & \text{if } r = s \le m, \\ 0 & \text{otherwise,} \end{cases}$$

with $\pi/2 \ge t_1 \ge \cdots \ge t_m \ge 0$. Set

$$P = \begin{cases} U\,\mathrm{diag}\,(1+\cos t_1, \ldots, 1+\cos t_k)U^*/2 & \text{if } k \le n/2, \\ U(\mathrm{diag}\,(1+\cos t_1, \ldots, 1+\cos t_m) \oplus I_{2k-n})U^*/2 & \text{otherwise,} \end{cases}$$

and

$$Q = \begin{cases} V^*(\mathrm{diag}\,(1-\cos t_1, \ldots, 1-\cos t_k) \oplus 0_{n-2k})V/2 & \text{if } k \le n/2, \\ V^*\mathrm{diag}\,(1-\cos t_1, \ldots, 1-\cos t_m)V/2 & \text{otherwise.} \end{cases}$$

One easily checks that $\left(\begin{smallmatrix} P & X \\ X^* & Q \end{smallmatrix}\right)$ is a rank k orthogonal projection. ∎

Clearly, the set of $k \times (n-k)$ matrices X satisfying $\|X\| \le 1/2$ is convex. By observations 1 and 2, we have the following result.

THEOREM 2.2. *Let $A = (A_1, \ldots, A_m)$ be an m-tuple of $n \times n$ Hermitian matrices. Suppose $1 \le k < n$ and*

$$\mathcal{S} = \left\{ \alpha I + \begin{pmatrix} 0 & X \\ X^* & 0 \end{pmatrix} : \alpha \in \mathbf{R},\ X \in \mathbf{C}^{k\times(n-k)} \right\}.$$

If there exists a unitary U such that $U^*A_jU \in \mathcal{S}$ for all j, then $W_k(A)$ is convex.

Let $1 \le k < n$. Since all matrices in $\mathcal{U}(J_k)$ have Frobenius norm equal to $\sqrt{k}$, the set $\mathcal{U}(J_k)$ is highly nonconvex in the sense that no three points in $\mathcal{U}(J_k)$ are collinear [18]. It is somewhat surprising that the projection of $\mathcal{U}(J_k)$ to a subspace of dimension $2k(n-k)+1$ can be convex. In particular, if $n = 2k$, then $2k(n-k)+1 = n^2/2+1$, which is more than half of the dimension of the space of $n \times n$ Hermitian matrices! In any event, we have the following result showing that $2k(n-k)+1$ is indeed the upper limit.

THEOREM 2.3. *The unitary orbit $\mathcal{U}(J_k)$ is a homogeneous manifold, and the tangent space at J_k equals*

$$\{i(HJ_k - J_kH) : H = H^*\} = \left\{ \begin{pmatrix} 0 & X \\ X^* & 0 \end{pmatrix} : X \in \mathbf{C}^{k\times(n-k)} \right\},$$

which has dimension $2k(n-k)$. Consequently, if $A = (A_1, \ldots, A_m)$ is such that

$$\dim\,(\mathrm{span}\,\{I, A_1, \ldots, A_m\}) > 2k(n-k)+1,$$

then $W_k(A)$ is not convex.

Proof. The set $\mathcal{U}(J_k)$ is the orbit of J_k under the group action $(U, C) \mapsto U^*CU$ for any unitary U and Hermitian C. Thus $\mathcal{U}(J_k)$ is a homogeneous manifold.

Since every unitary matrix admits a representation of the form e^{iH} for some Hermitian H, every smooth path of unitary matrices $U(t)$, $-1 \le t \le 1$, $U(0) = I$, is

of the form $U(t) = e^{iH(t)}$, where $H(t)$, $-1 \le t \le 1$, is a smooth path of Hermitian matrices such that $H(0) = 0$. Thus we have

$$\left.\frac{d(e^{iH(t)}J_k e^{-iH(t)})}{dt}\right|_{t=0} = i(H'(0)J_k - J_k H'(0))$$

and the tangent space of $\mathcal{U}(J_k)$ at J_k is equal to

$$\{i(HJ_k - J_k H) : H = H^*\} = \left\{ \begin{pmatrix} 0 & X \\ X^* & 0 \end{pmatrix} : X \in \mathbf{C}^{k\times(n-k)} \right\},$$

which has dimension $2k(n-k)$.

Let $A = (A_1, \ldots, A_m)$ and $\mathcal{S} = \operatorname{span}\{I, A_1, \ldots, A_m\}$ with $\dim \mathcal{S} > 2k(n-k)+1$. We prove that $W_k(A)$ is not convex. Suppose $W_k(A)$ is indeed convex. By observation 4, we can find a linearly independent subset $\{I, B_1, \ldots, B_s\}$ of $\{I, A_1, \ldots, A_m\}$ with $s = 2k(n-k)+1$ such that $W_k(B_1, \ldots, B_s)$ is also convex.

We claim that $W_k(B_1, \ldots, B_s)$ has a nonempty interior in $\mathbf{R}^s$. If this isn't true, then the convex set $W_k(B_1, \ldots, B_s)$ must lie in a certain hyperplane

$$\mathcal{P} = \{v \in \mathbf{R}^s : v^t w = d\},$$

for some unit vector $w = (w_1, \ldots, w_s)^t \in \mathbf{R}^s$ and $d \in \mathbf{R}$. But then we have

$$\operatorname{tr}\left(\left(\sum_{j=1}^{s} w_j B_j\right) X\right) = d$$

for all $X \in \mathcal{U}(J_k)$, i.e., $W_k(\sum_{j=1}^s w_j B_j) = \{d\}$. It follows that (see [10]) $\sum_{j=1}^s w_j B_j = dI/k$, contradicting the fact that $\{I, B_1, \ldots, B_s\}$ is linearly independent.

Now $\mathcal{U}(J_k)$ is a homogeneous manifold of (real) dimension $2k(n-k)$ and $\phi(\mathcal{U}(J_k)) = W_k(B_1, \ldots, B_s)$ with

$$\phi(P) = (\operatorname{tr} B_1 P, \ldots, \operatorname{tr} B_s P).$$

The set $W_k(B_1, \ldots, B_s) \subseteq \mathbf{R}^s$ with $s = 2k(n-k)+1$ cannot have a nonempty interior. So, the assumption that $\dim \mathcal{S} > 2k(n-k)+1$ cannot be true. □

By the last two theorems, we see that a basis for the tangent space of $\mathcal{U}(J_k)$ at any point, together with the identity matrix, form a maximal linearly independent convex family for the kth generalized numerical range.

Next we turn to other linearly independent families $\{A_1, \ldots, A_m\}$ so that $W_k(A)$ is convex. By the result in [14], we have the following theorem.

THEOREM 2.4. *Let $1 \le k \le n/2$ and $P \in \mathbf{C}^{k\times k}$ be Hermitian. Then there exists a rank k orthogonal projection of the form $\begin{pmatrix} P & * \\ * & * \end{pmatrix}$ if and only if all eigenvalues of P lie in $[0,1]$.*

Clearly, the set of matrices $P \in \mathbf{C}^{k\times k}$ with eigenvalues lying in $[0,1]$ is convex. Using Theorem 1.2, observations 1 and 2, we obtain the following theorem.

THEOREM 2.5. *Let $A = (A_1, \ldots, A_m)$ be an m-tuple of Hermitian matrices. Suppose*

$$\mathcal{S} = \{P \oplus \alpha I_{n-k} : \alpha \in \mathbf{R},\ P \text{ is } k \times k \text{ Hermitian }\}.$$

If there exists a unitary U such that $U^* A_j U \in \mathcal{S}$ for all j, then $W_k(A)$ is convex.

3. Convex families for the first joint numerical range. We first identify a maximal linearly independent convex family for the first joint numerical range which is different from those constructed in the previous section.

THEOREM 3.1. *Let $A = (A_1, \ldots, A_m)$ be an m-tuple of Hermitian matrices. Suppose*

$$\mathcal{S} = \text{span}\,(\{E_{jj} : 1 \le j \le n\} \cup \{E_{j,j+1} + E_{j+1,j} : 1 \le j \le n-1\}).$$

If there exists a unitary U such that $U^* A_j U \in \mathcal{S}$ for all j, then $W_1(A)$ is convex.

Proof. Let $A = (A_1, \ldots, A_{2n-1})$ so that $A_j = E_{jj}$ for $1 \le j \le n$ and $A_{n+j} = (E_{j,j+1} + E_{j+1,j})/2$ for $1 \le j \le n-1$. Then $(x_1, \ldots, x_{2n-1}) \in W_1(A)$ if and only if there is a unit vector $v = (v_1, \ldots, v_n)^t \in \mathbf{C}^n$ such that

$$\begin{aligned} x_j &= |v_j|^2, && \text{if } 1 \le j \le n, \\ x_{n+j} &= (v_j \bar{v}_{j+1} + \bar{v}_j v_{j+1})/2, && \text{if } 1 \le j < n. \end{aligned}$$

These conditions hold if and only if $(x_1, \ldots, x_{2n-1})$ satisfies

$$\sum_{j=1}^{n} x_j = 1, \quad x_j \ge 0 \ \text{ for } 1 \le j \le n, \ \text{ and}$$

$$|x_{n+j}|^2 \le x_j x_{j+1} \text{ for } 1 \le j \le n-1.$$

Now suppose that $x = (x_1, \ldots, x_{2n-1})$, $y = (y_1, \ldots, y_{2n-1}) \in W_1(A)$, and $z = (z_1, \ldots, z_{2n-1})$ equals $(x+y)/2$. Clearly, we have

$$\sum_{j=1}^{n} z_j = 1, \quad z_j \ge 0 \ \text{ for } \ j = 1, \ldots, n.$$

Moreover, for $j = 1, \ldots, n-1$,

$$4|z_{n+j}|^2 = |x_{n+j} + y_{n+j}|^2 \le (\sqrt{x_j x_{j+1}} + \sqrt{y_j y_{j+1}})^2 \le (x_j + y_j)(x_{j+1} + y_{j+1}).$$

Hence $z \in W_1(A)$. Since $W_1(A)$ is closed, we conclude that it is a convex set. ∎

A result of Horn [8] implies that $W_1(E_{11}, E_{22}, \ldots, E_{nn})$ is convex. Theorem 3.1 strengthens this statement.

Note that a maximal linearly independent convex family of the first joint numerical range may not have $2(n-1)+1$ elements as shown in the following example.

Example 3.2. Suppose $(k, n) = (1, 3)$. Then $\{I_3, A_1, A_2, A_3\}$ with

$$A_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \ A_2 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \ A_3 = \begin{pmatrix} 0 & i & 0 \\ -i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

is a maximal linearly independent convex family for $W_1(A)$.

Proof. Suppose there exists A_4 such that $\{I, A_1, \ldots, A_4\}$ is linearly independent and $W_1(I_3, A_1, \ldots, A_4)$ is convex. One may replace A_4 by a matrix of the form $A_4 - (a_0 I_3 + a_1 A_1 + a_2 A_2 + a_3 A_3)$ so that the leading 2×2 principal submatrix is zero. Since $W_1(I_3, A_1, \ldots, A_4)$ is convex if and only if $W_1(A)$ is convex with $A = (A_1, \ldots, A_4)$, we can focus on $W_1(A) \subseteq \mathbf{R}^4$. Using the unit vectors $(1, 0, 0)^t$ and $(0, 1, 0)^t$, we see

that $(1,0,0,0),(-1,0,0,0) \in W_1(A)$. By convexity, we see that $(0,0,0,0) \in W_1(A)$. One easily checks that a unit vector giving rise to the point $(0,0,0,0) \in W_1(A)$ must be of the form $(0,0,\mu) \in \mathbf{C}^3$. As a result, the $(3,3)$ entry of A_4 must be 0. Let U be a unitary matrix of the form $U = \hat{U} \oplus [1]$ so that $\hat{A}_4 = U^* A_4 U = E_{13} + E_{31}$. Then

$$\operatorname{span}\{A_1, A_2, A_3, \hat{A}_4\} = \operatorname{span}\{U^* A_j U : 1 \le j \le 4\}.$$

Thus $W_1(A)$ is convex if and only if $W_1(\hat{A})$ is convex, where $\hat{A} = (A_1, A_2, A_3, \hat{A}_4)$. Note that

$$W_1(\hat{A}) = \left\{ \begin{array}{c} (|v_1|^2 - |v_2|^2, v_1\bar{v}_2 + v_2\bar{v}_1, i(v_2\bar{v}_1 - v_1\bar{v}_2), v_1\bar{v}_3 - v_3\bar{v}_1) : \\ v_1, v_2, v_3 \in \mathbf{C}, \sum_{j=1}^3 |v_j|^2 = 1 \end{array} \right\}.$$

Using the unit vectors $(1/2, \pm 1/2, 1/\sqrt{2})^t$, we see that $(0, \pm 1/2, 0, 1/\sqrt{2}) \in W_1(\hat{A})$. However, their midpoint $(0,0,0,1/\sqrt{2}) \notin W_1(\hat{A})$, which is a contradiction. ∎

In connection with the above discussion, it would be interesting to solve the following problems.

Problem 3.3. Characterize maximal linearly independent convex families for the first (or the kth) joint numerical range.

Problem 3.4. Determine maximal linearly independent convex families with the smallest number of elements.

4. A theorem on nonconvexity. Let $n \ge 3$. Then there exists a 4-tuple A of $n \times n$ Hermitian matrices such that $W_1(A)$ is not convex as mentioned in the introduction. Suppose A_1, A_2, A_3 are $n \times n$ Hermitian matrices such that $\{I, A_1, A_2, A_3\}$ is linearly dependent, say, $\operatorname{span}\{I, A_1, A_2, A_3\} = \operatorname{span}\{I, A_1, A_2\}$. Then for any Hermitian matrix A_4, $W(A_1, A_2, A_4)$ is convex, and so is $W(I, A_1, A_2, A_3, A_4)$ by observations 2 and 3.

In the following, we show that as long as $\{I, A_1, A_2, A_3\}$ is linearly independent, one can find A_4 so that $W_1(A_1, A_2, A_3, A_4)$ is not convex. We first establish the following auxiliary result, which is of independent interest.

THEOREM 4.1. *Let A_1, A_2, A_3 be $n \times n$ Hermitian matrices. If $\{I, A_1, A_2, A_3\}$ is linearly independent, then there exists $X \in \mathbf{C}^{n\times 2}$ with $X^*X = I_2$ such that $\{I_2, X^*A_1X, X^*A_2X, X^*A_3X\}$ is linearly independent.*

Proof. We assume that $n > 2$ to avoid trivial consideration. Since I, A_1, A_2 are linearly independent, the complex matrix $A_1 + iA_2$ cannot be written as $\alpha H + \beta I$ for any $\alpha, \beta \in \mathbf{C}$ and Hermitian matrix H. By [14, Theorem 3.5], there exists $X \in \mathbf{C}^{n\times 2}$ such that $X^*(A_1 + iA_2)X$ is not normal. By [9, Lemma 1.3.1], we may assume that $X^*(A_1 + iA_2)X$ has equal diagonal entries $\gamma_1 + i\gamma_2$ with $\gamma_1, \gamma_2 \in \mathbf{R}$. Replace A_1 and A_2 by suitable linear combinations of $A_1 - \gamma_1 I$ and $A_2 - \gamma_2 I$; we may assume that

$$X^*A_1X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad X^*A_2X = \begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix}.$$

If $\{I_2, X^*A_1X, X^*A_2X, X^*A_3X\}$ is linearly independent, then we are done. If not, we may replace A_3 by a suitable linear combination of I, A_1, A_2, A_3 so that $X^*A_3X = 0$ and $\{I, A_1, A_2, A_3\}$ is still linearly independent. Since $A_3 \ne 0$, there exists $Y = [X|y] \in \mathbf{C}^{n\times 3}$ such that $Y^*Y = I_3$ and

$$\hat{A}_1 = Y^*A_1Y = \begin{pmatrix} 0 & 1 & a_1 \\ 1 & 0 & a_2 \\ \bar{a}_1 & \bar{a}_2 & a_3 \end{pmatrix}, \quad \hat{A}_2 = Y^*A_2Y = \begin{pmatrix} 0 & i & b_1 \\ -i & 0 & b_2 \\ \bar{b}_1 & \bar{b}_2 & b_3 \end{pmatrix},$$

$$\text{and} \qquad \hat{A}_3 = Y^* A_3 Y = \begin{pmatrix} 0 & 0 & c_1 \\ 0 & 0 & c_2 \\ \bar{c}_1 & \bar{c}_2 & c_3 \end{pmatrix} \neq 0.$$

We may assume that $c_2 \geq |c_1|$; otherwise, we may replace $\hat{A}_3$ by $D^* \hat{A}_3 D$ with $D = \text{diag}(1, \mu, 1)$ or $D = \left(\begin{smallmatrix} 0 & \mu \\ 1 & 0 \end{smallmatrix}\right) \oplus [1]$ for some suitable $\mu \in \mathbf{C}$ with $|\mu| = 1$. One can then replace $\hat{A}_1$ and $\hat{A}_2$ by suitable linear combinations of $D^* \hat{A}_1 D$ and $D^* \hat{A}_2 D$, and assume that $\hat{A}_1$ and $\hat{A}_2$ are still of the form in the display.

Now consider $P = \left(\begin{smallmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \end{smallmatrix}\right)^t$ with $\theta \in (0, \pi/2)$. Then

$$P^* A_1 P = \sin\theta \begin{pmatrix} 0 & \cot\theta + a_1 \\ \cot\theta + \bar{a}_1 & (a_2 + \bar{a}_2)\cos\theta + a_3 \sin\theta \end{pmatrix},$$

$$P^* A_2 P = \sin\theta \begin{pmatrix} 0 & i\cot\theta + b_1 \\ -i\cot\theta + \bar{b}_1 & (b_2 + \bar{b}_2)\cos\theta + b_3 \sin\theta \end{pmatrix},$$

$$P^* A_3 P = \sin\theta \begin{pmatrix} 0 & c_1 \\ \bar{c}_1 & 2c_2 \cos\theta + c_3 \sin\theta \end{pmatrix}.$$

We claim that there exists $\theta \in (0, \pi/2)$ such that $\{I_2, P^* \hat{A}_1 P, P^* \hat{A}_2 P, P^* \hat{A}_3 P\}$ is linearly independent, and hence the conclusion of the theorem follows. Since all the $(1,1)$ entries of $P^* \hat{A}_1 P$, $P^* \hat{A}_2 P$, and $P^* \hat{A}_3 P$ equal 0, I_2 is not a linear combination of these three matrices. To establish our claim, we need only show that there exists $\theta \in (0, \pi/2)$ such that $P^* \hat{A}_1 P$, $P^* \hat{A}_2 P$, and $P^* \hat{A}_3 P$ are linearly independent. To this end, construct the following matrix B, whose rows contain the real and imaginary parts of the entries of $P^* \hat{A}_j P$:

$$B = \sin\theta \begin{pmatrix} \cot\theta + (a_1 + \bar{a}_1)/2 & i(\bar{a}_1 - a_1)/2 & (a_2 + \bar{a}_2)\cos\theta + a_3 \sin\theta \\ (b_1 + \bar{b}_1)/2 & \cot\theta + i(\bar{b}_1 - b_1)/2 & (b_2 + \bar{b}_2)\cos\theta + b_3 \sin\theta \\ (c_1 + \bar{c}_1)/2 & i(\bar{c}_1 - c_1)/2 & 2c_2 \cos\theta + c_3 \sin\theta \end{pmatrix}.$$

Now,

$$\det(B)/\sin^4\theta = \det \begin{pmatrix} \cot\theta + (a_1 + \bar{a}_1)/2 & i(\bar{a}_1 - a_1)/2 & (a_2 + \bar{a}_2)\cot\theta + a_3 \\ (b_1 + \bar{b}_1)/2 & \cot\theta + i(\bar{b}_1 - b_1)/2 & (b_2 + \bar{b}_2)\cot\theta + b_3 \\ (c_1 + \bar{c}_1)/2 & i(\bar{c}_1 - c_1)/2 & 2c_2 \cot\theta + c_3 \end{pmatrix}$$

can be viewed as $p(\cot\theta)$ for some real polynomial p of degree 3. If $\det(B) = 0$ for all $\theta \in (0, \pi/2)$, then the coefficient of $\cot^3\theta$ in $p(\cot\theta)$ is 0, which is just $2c_2$ by expanding $\det(B)/\sin^4\theta$. Since $c_2 \geq |c_1|$, we see that $c_1 = 0$ as well. Now, consider the coefficient of $\cot^2\theta$ in $p(\cot\theta)$, which is just c_3. Again, it has to be 0. It follows that $\hat{A}_3 = 0$, which is a contradiction. ∎

Now we are ready to state the nonconvexity result of the joint numerical range.

THEOREM 4.2. *Suppose A_1, A_2, A_3 are $n \times n$ Hermitian matrices such that $\{I, A_1, A_2, A_3\}$ is linearly independent. Then there exists an $n \times n$ Hermitian matrix A_4 such that $W_1(A_1, A_2, A_3, A_4)$ is not convex.*

Proof. By the previous theorem, there exists $X \in \mathbf{C}^{n \times 2}$ such that $X^* X = I_2$ and $\{I_2, X^* A_1 X, X^* A_2 X, X^* A_3 X\}$ is linearly independent. Let $A_4 = XX^*$. Then by [11, Lemma 4.1] we have $(a, b, c, 1) \in W_1(A_1, A_2, A_3, A_4)$ if and only if $(a, b, c) \in W_1(X^* A_1 X, X^* A_2 X, X^* A_3 X)$, which is not convex [1]. The result follows. ∎

We remark that Theorem 4.1 and its proof can be easily modified to deal with infinite dimensional self-adjoint operators A_1, A_2, A_3. In general, it is interesting to solve the following problem.

Problem 4.3. If a linearly independent family $\{A_1, \ldots, A_m\}$ of (finite- or infinite-dimensional) self-adjoint operators is given, where $(r-1)^2 < m \le r^2$, does there exist an X such that $X^*X = I_r$ and $\{X^*A_jX : j = 1, \ldots, m\}$ is linearly independent?

By private communication, Doug Farenick pointed out that this problem can be studied in the context of unital completely positive maps on C^*-algebras.

5. Related results and questions. There are many variations of our problems. We mention a few of them in what follows.

5.1. Real symmetric matrices. In applications, one often has to consider real symmetric matrices instead of complex Hermitian matrices. One can modify the results and proofs on complex Hermitian matrices and get the following analogues for real symmetric matrices.

THEOREM 5.1. *Let $A = (A_1, \ldots, A_m)$ be an m-tuple of real symmetric matrices. Suppose $1 \le k \le n/2$ and*

$$\mathcal{S} = \left\{ \alpha I + \begin{pmatrix} 0_k & X \\ X^t & 0_{n-k} \end{pmatrix} : \alpha \in \mathbf{R},\ X \in \mathbf{R}^{k\times(n-k)} \right\},$$

or

$$\mathcal{S} = \{ P \oplus \alpha I_{n-k} : \alpha \in \mathbf{R},\ P \text{ is } k \times k \text{ real symmetric} \}.$$

If there is a real orthogonal Q such that $Q^tA_jQ \in \mathcal{S}$, then $W_k(A)$ is convex.

THEOREM 5.2. *The orthogonal similarity orbit $\mathcal{O}(J_k)$ is a homogeneous manifold, and the tangent space at J_k equals*

$$\{(KJ_k - J_kK) : K = -K^t\} = \left\{ \begin{pmatrix} 0 & X \\ X^t & 0 \end{pmatrix} : X \in \mathbf{R}^{k\times(n-k)} \right\},$$

which has dimension $k(n-k)$. Consequently, if $A = (A_1, \ldots, A_m)$ such that

$$\dim(\operatorname{span}\{I, A_1, \ldots, A_m\}) > k(n-k) + 1,$$

then $W_k(A)$ is not convex.

By this theorem and a result of Horn [8], we have the following corollary.

COROLLARY 5.3. *The elements of $A = (E_{11}, E_{22}, \ldots, E_{nn})$ form a maximal linearly independent convex family for $W_1(A)$.*

5.2. Rectangular matrices. Suppose $A = (A_1, \ldots, A_m)$, where $A_1, \ldots, A_m$ are $n \times r$ matrices over $\mathbf{F} = \mathbf{R}$ or $\mathbf{C}$. To be specific, we assume that $n \le r$ in our discussion. Otherwise, consider the transposes of $A_1, \ldots, A_m$. For $1 \le k \le n$, define

$$V_k(A) = \left\{ \left(\sum_{j=1}^{k} x_j^* A_i y_j \right)_{i=1}^{m} : \begin{matrix} \{x_1, \ldots, x_k\} \text{ is an orthonormal set of } \mathbf{F}^n \\ \{y_1, \ldots, y_k\} \text{ is an orthonormal set of } \mathbf{F}^r \end{matrix} \right\},$$

which is a subset of $\mathbf{F}^m$. Let $\mathcal{V}_k$ be the collection of $n \times r$ matrices X such that X^*X is a rank k orthogonal projection. It is not hard to see that

$$V_k(A) = \{(\operatorname{tr} X^*A_1, \ldots, \operatorname{tr} X^*A_m) : X \in \mathcal{V}_k\}.$$

It was shown in [13, Theorems 14 and 41] that $V_k(A)$ is not convex in general if $m > 2$. Again, if $A_1, \ldots, A_m$ belong to a certain subset, then one can get convexity for a much larger m as shown in the following result.

THEOREM 5.4. *Let $A = (A_1, \ldots, A_m)$ be an m-tuple of $n \times r$ matrices. Suppose*

$$\mathcal{S} = \left\{ \begin{pmatrix} X & 0 \\ 0 & 0 \end{pmatrix} \in \mathbf{F}^{n\times r} : X \in \mathbf{F}^{p\times q} \right\},$$

where $\min\{p, q\} \le k \le n + r - p - q$. If there exist square matrices U and V with $U^*U = I_n$ and $V^*V = I_r$ such that $UA_jV \in \mathcal{S}$ for all j, then $V_k(A)$ is convex.

The proof of this theorem depends on the following theorem.

THEOREM 5.5. *Let $(p, q) \le (n, r)$ in the entrywise sense, and $\min\{p, q\} \le k \le n + r - p - q$. Then a $p \times q$ matrix X can be embedded in an $n \times r$ matrix Y so that Y^*Y is a rank k orthogonal projection if and only if $\|X\| \le 1$. Hence, the collection of all such $X \in \mathbf{F}^{p\times q}$ is a convex set.*

Proof. By a result of Thompson [17] and our assumptions on the positive integers n, r, p, q, k, we see that a $p \times q$ matrix X with singular values $s_1 \ge \cdots \ge s_k$ can be embedded in an $n \times r$ matrix with the k largest singular values equal to 1 and the rest equal to 0 if and only if $s_i \le 1$. The result follows. □

Let $R_k = E_{11} + E_{22} + \cdots + E_{kk} \in \mathbf{F}^{n\times r}$. In the complex case, the tangent space of the manifold $\mathcal{V}_k$ at R_k is

$$\begin{aligned} \mathcal{T}(R_k) &= \{i(R_kH + GR_k) : H = H^*,\ G = G^*\} \\ &= \left\{ \begin{pmatrix} iX & Y \\ Z & 0_{(n-k)\times(r-k)} \end{pmatrix} : X = X^* \in \mathbf{C}^{k\times k} \right\} \end{aligned}$$

and has real dimension $k^2 + 2k(n + r - 2k)$. In the real case, the tangent space of the manifold $\mathcal{V}_k$ at R_k is

$$\begin{aligned} \mathcal{T}(R_k) &= \left\{R_kH + GR_k : H = -H^t,\ G = -G^t\right\} \\ &= \left\{ \begin{pmatrix} X & Y \\ Z & 0_{(n-k)\times(r-k)} \end{pmatrix} : X = -X^t \in \mathbf{R}^{k\times k} \right\} \end{aligned}$$

and has real dimension $k(k-1)/2 + k(n + r - 2k)$. It would be interesting to see whether one can construct a maximal linearly independent convex family for $V_k(A)$ of this size.

5.3. Complex symmetric matrices. Let $A = (A_1, \ldots, A_m)$, where $A_1, \ldots, A_m$ are $n \times n$ complex symmetric matrices. For $1 \le k \le n$, let $J_k = I_k \oplus 0_{n-k}$, and define

$$W_k^t(A) = \{(\operatorname{tr} J_kU^tA_1U, \ldots, \operatorname{tr} J_kU^tA_mU) : U^*U = I\}.$$

Then Theorem 5.1 is valid for the complex case. However, the tangent space of the manifold at J_k is

$$\begin{aligned} \mathcal{U}^t(J_k) &= \left\{i(J_kH + H^tJ_k) : H = H^*\right\} \\ &= \left\{ i \begin{pmatrix} X & Y \\ Y^t & 0_{n-k} \end{pmatrix} : X = X^t \in \mathbf{R}^{k\times k},\ Y \in \mathbf{C}^{k\times(n-k)} \right\} \end{aligned}$$

and has real dimension $k(k+1)/2 + 2k(n-k)$. It is interesting to know whether one can find a maximal linearly independent convex family for $W_k^t(A)$ of this size.

Note that one may consider $W_k^t(A)$ for an m-tuple of complex matrices $A = (A_1, \ldots, A_m)$. However, it is easy to check that $W_k^t(A)$ is the same as $W_k^t(\hat{A}_1, \ldots, \hat{A}_m)$, where $\hat{A}_j$ is the symmetric part of A_j for all j. For skew-symmetric matrices, one needs a different treatment, as shown in the next subsection.

5.4. Skew-symmetric matrices. Suppose $A_1, \ldots, A_m$ are skew-symmetric matrices over $\mathbf{F} = \mathbf{R}$ or $\mathbf{C}$. For $1 \le k \le n/2$, let $T_k = \left(\begin{smallmatrix} 0_k & I_k \\ -I_k & 0_k \end{smallmatrix}\right) \oplus 0_{n-2k}$, and define

$$W_k^t(A) = \{(\operatorname{tr} T_k U^t A_1 U, \ldots, \operatorname{tr} T_k U^t A_m U) : U^* U = I\}.$$

We have the following convexity theorem.

THEOREM 5.6. *Let $A = (A_1, \ldots, A_m)$ be an m-tuple of $n \times n$ skew-symmetric matrices. Suppose*

$$\mathcal{S} = \{E_{j,k+j} : j = 1, \ldots, k\}, \qquad \text{or}$$

$$\mathcal{S} = \left\{ \begin{pmatrix} 0_p & 0 & X \\ 0 & 0 & 0 \\ -X^t & 0 & 0_q \end{pmatrix} : X \in \mathbf{F}^{p \times q},\ \|X\| \le 1 \right\}$$

for some positive integers p and q such that $n - (p+q) \ge k \ge \min\{p, q\}$. If there exists a square matrix U with $U^*U = I_n$ such that $U^t A_j U \in \mathcal{S}$ for all j, then $W_k^t(A)$ is convex.

The proof of this theorem depends on the following two results. First we have the following (see [16] and its references).

THEOREM 5.7. *Let $(d_1, \ldots, d_k) \in \mathbf{F}^k$. There exists U with $U^*U = I_n$ such that the $(1, k+1), (2, k+2), \ldots, (k, 2k)$ entries of $U^t T_k U$ equal $d_1, \ldots, d_k$ if and only if $|d_j| \le 1$ for all j. Hence, the collection of such $(d_1, \ldots, d_k)$ forms a convex set.*

Note that the singular values of the matrix $\left(\begin{smallmatrix} 0_k & X \\ -X^t & 0_{n-k} \end{smallmatrix}\right)$ are just two copies of those of X. Applying the result on rectangular matrices to the $k \times (n-k)$ right top corner of a skew-symmetric matrix, we have the following result.

THEOREM 5.8. *Let p and q be positive integers with $n - (p+q) \ge k \ge \min\{p, q\}$. Then a $p \times q$ matrix X can be embedded in the right top corner of an $n \times n$ skew-symmetric matrix with the $2k$ largest singular values equal to 1 and the rest equal to 0 if $\|X\| \le 1$. Hence, the collection of all such X is a convex set.*

Acknowledgment. The first author would like to express his thanks to Professor M. D. Choi for making the visit to the University of Toronto possible and to the staff of the University of Toronto for their warm hospitality.

Note added in proof. Professor M. D. Choi pointed out that the answer of Problem 4.3 is negative. For instance, the 10×10 Hermitian matrices $A_j = E_{1j} + E_{j1}$, $j = 2, \ldots, 10$, are linearly independent, but it is impossible to have a 10×3 matrix V with $V^*V = I_3$ so that $\{V^* A_j V : j = 2, \ldots, 10\}$ is linearly independent. A modified problem would be the following: Given linearly independent Hermitian operators $A_1, \ldots, A_m$, find the smallest positive integer r so that there exists V with $V^*V = I_r$ for which the set $\{V^* A_j V : j = 1, \ldots, m\}$ is still linearly independent.

REFERENCES

[1] Y. H. Au-Yeung and Y. T. Poon, *A remark on the convexity and positive definiteness concerning Hermitian matrices*, Southeast Asian Bull. Math., 3 (1979), pp. 85–92.

[2] Y. H. Au-Yeung and N. K. Tsing, *An extension of the Hausdorff-Toeplitz theorem on the numerical range*, Proc. Amer. Math. Soc., 89 (1983), pp. 215–218.

[3] P. Binding, D. Farenick, and C. K. Li, *A dilation and norm in several variable operator theory*, Canad. J. Math., 47 (1995), pp. 449–461.

[4] P. Binding and C. K. Li, *Joint ranges of Hermitian matrices and simultaneous diagonalization*, Linear Algebra Appl., 151 (1991), pp. 157–168.

[5] M. Cho and M. Takaguchi, *Some classes of commuting m-tuples of operators*, Studia Math., 80 (1984), pp. 245–259.

[6] J. C. Doyle, *Analysis of feedback systems with structured uncertainties*, Proc. IEE-D, 129 (1982), pp. 242–250.

[7] M. Fan and A. Tits, *m-form numerical range and the computation of the structured singular values*, IEEE Trans. Automat. Control, AC-33 (1988), pp. 284–289.

[8] A. Horn, *Doubly stochastic matrices and the diagonal of a rotation matrix*, Amer. J. Math., 76 (1954), pp. 620–630.

[9] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[10] C. K. Li, *The C-convex matrices*, Linear and Multilinear Algebra, 21 (1987), pp. 303–312.

[11] C. K. Li, *Matrices with some extremal properties*, Linear Algebra Appl., 101 (1988), pp. 255–267.

[12] C. K. Li and R. Mathias, *Inequalities on the singular values of an off-diagonal block of a Hermitian matrix*, to appear in J. Inequalities and Applications; preprint available at http://www.math.wm.edu/~ckli/pub.html.

[13] C. K. Li and T. Y. Tam, *Numerical ranges arising from simple Lie algebras—convexity and inclusion relation*, preprint available at http://www.auburn.edu/~tamtiny/pub.html.

[14] C. K. Li and N. K. Tsing, *On the k^{th} matrix numerical range*, Linear and Multilinear Algebra, 28 (1991), pp. 229–239.

[15] Y. T. Poon, *On the convex hull of the multiform numerical range*, Linear and Multilinear Algebra, 37 (1994), pp. 221–224.

[16] T. Y. Tam, *Partial superdiagonal elements and singular values of a complex skew-symmetric matrix*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 737–754.

[17] R. C. Thompson, *Principal submatrices* IX: *Interlacing inequalities for singular values of submatrices*, Linear Algebra Appl., 5 (1972), pp. 1–12.

[18] R. C. Thompson, *Research problem: The matrix numerical range*, Linear and Multilinear Algebra, 21 (1987), pp. 321–323.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 679–693
© 2000 Society for Industrial and Applied Mathematics

CONDITIONING OF RECTANGULAR VANDERMONDE MATRICES WITH NODES IN THE UNIT DISK*

FERMÍN S. V. BAZÁN†

Abstract. Let $W_N = W_N(z_1, z_2, \ldots, z_n)$ be a rectangular Vandermonde matrix of order $n \times N$, $N \geq n$, with distinct nodes z_j in the unit disk and z_j^{k-1} as its (j,k) entry. Matrices of this type often arise in frequency estimation and system identification problems. In this paper, the conditioning of W_N is analyzed and bounds for the spectral condition number $\kappa_2(W_N)$ are derived. The bounds depend on n, N, and the separation of the nodes. By analyzing the behavior of the bounds as functions of N, we conclude that these matrices may become well conditioned, provided the nodes are close to the unit circle but not extremely close to each other and provided the number of columns of W_N is large enough. The asymptotic behavior of both the conditioning itself and the bounds is analyzed and the theoretical results arising from this analysis verified by numerical examples.

Key words. Vandermonde matrices, singular values, almost normal matrices, exponential modeling

AMS subject classifications. 15A12, 65F35, 93A30

PII. S0895479898336021

1. Introduction.

Let $W_N = W_N(z_1, z_2, \ldots, z_n)$ be a Vandermonde matrix of order $n \times N$, $N \geq n$, with distinct nonzero complex nodes in the unit disk and z_j^{k-1} as its (j,k) entry, that is,

$$W_N = W_N(z_1, z_2, \ldots, z_n) = \begin{bmatrix} 1 & z_1 & z_1^2 & \cdots & z_1^{N-1} \\ 1 & z_2 & z_2^2 & \cdots & z_2^{N-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & z_n & z_n^2 & \cdots & z_n^{N-1} \end{bmatrix}, \tag{1.1}$$

and let $\kappa_2(W_N) = \|W_N\|\|W_N^\dagger\|$ be its 2-norm condition number, where † stands for Moore–Penrose pseudoinverse. These matrices arise in approximation theory and polynomial interpolation, but their most frequent use is in the framework of signal processing.

The purpose of this work is to derive bounds for $\kappa_2(W_N)$ and to analyze their behavior for fixed $n, z_1, \ldots, z_n$ and increasing N. Motivation for this analysis can be encountered in problems related to frequency estimation, parameter identification, and quantitative analysis of time domain nuclear magnetic resonance (NMR) data, among others; see, for instance, Van Huffel [17], Van Huffel et al. [18], and Bazán and Bavastri [1]. For illustration purposes, suppose we have a set of data h_l ($l = 0, 1, \ldots M-1$) involving $2n$ unknown parameters $r_j, z_j \in \mathbb{C}$ through the model

$$h_l = r_1 z_1^l + r_2 z_2^l + \cdots + r_n z_n^l,$$

*Received by the editors March 20, 1998; accepted for publication (in revised form) by N. J. Higham April 15, 1999; published electronically January 25, 2000. This research was performed while the author was a visitor at FUNDP, Namur, Belgium, and was supported by CNPq, Brazil, by grant 201299/96-8(NV).

http://www.siam.org/journals/simax/21-2/33602.html

†Departamento de Matemática, Universidade Federal de Santa Catarina, Florianópolis, Santa Catarina 88040-900, Brazil (fermin@mtm.ufsc.br).

and suppose we wish to determine these parameters from the available data. This is a separable problem that can be solved as follows. First, note that if all z_j are available then the r_j can be estimated by solving the least squares problem

$$\min \|W_M^T r - h\|, \tag{1.2}$$

where T denotes the transpose of a matrix, W_M is as before, and $h \in \mathbb{C}^M$ contains the data h_l. The sensitivity of the r_j to perturbations in the data is governed by $\kappa_2(W_M)$. If both r_j and z_j are unknown instead, then one starts by finding first the z_j. For instance, these z_j can be found from the roots of a polynomial $p_N(z) = a_0 + a_1 z + \cdots + a_{N-1} z^{N-1} + z^N$, whose coefficients are computed by solving a least squares problem of type

$$\min \|Ha + b\|, \tag{1.3}$$

where H is an $L \times N$ Hankel matrix with h_{i+j-2} as its (i, j) entry, $n \le L, N \le M - n$, and $b = [h_N, h_{N+1}, \ldots, h_{N+L-1}]^T$ (Wei and Majda [19]). Once the z_j are found, the r_j are readily computed by solving (1.2). In this case, one can prove that the sensitivity of the roots z_j to perturbations in the coefficients depends on $\|W_N^\dagger\|$ (see, for instance, Bazán, Toint, and Zambaldi [4]), and once more $\kappa_2(W_N)$ is important.

Although there are many other applications in signal processing where $\kappa_2(W_N)$ plays a crucial role, little is known about the behavior of $\kappa_2(W_N)$ for $N > n$, which is in contrast with the well-known fact that square Vandermonde matrices are in general ill conditioned (see, e.g., Gautschi [7], Gautschi and Inglese [8], Cordoba, Gautschi, and Ruscheweyh [5], and Tyrtyshnikov [15]). This justifies our interest to analyze this condition number.

Rectangular Vandermonde matrices also appear in trigonometric polynomial interpolation problems, but we shall restrict ourselves to analyzing the conditioning of those Vandermonde matrices that appear implicitly in frequency estimation and parameter identification. The term *implicitly* is used because in these applications the nodes z_j are known only approximately: the z_j appear after preliminary computations involving data h_l that is noise corrupted. Thus, what we pursue is a theoretical prediction of the sensitivity of these problems to perturbations in the data rather than an effective computation of $\kappa_2(W_N)$.

An interesting feature often numerically observed in the above problems is that $\kappa_2(W_N)$ is substantially better than the bad condition number $\kappa_2(W_n)$ when N is large enough, but a mathematical explanation for this phenomenon is still lacking. In this paper, we intend to substantiate this observation by exhibiting lower and upper bounds for $\kappa_2(W_N)$ which say much about the condition number itself. The bounds depend on n, z_j, and N, and their quality essentially depends on the separation of the nodes in the unit disk. In particular, we show that the numbers $\kappa_2(W_N)$ become rather small for large N, provided the nodes are not extremely close to each other and of modulus not much smaller than one. On the other hand, to evaluate the quality of the bounds as N grows, we analyze the asymptotic value of both the bounds and the condition number itself as N grows infinitely. As a by-product, we obtain that if all $|z_j| = 1$, then the asymptotic value of the bounds is equal to 1.

The paper is organized as follows. In section 2 we describe preliminary background information. The main results are presented in section 3 where we derive our bounds and perform an analyzis about the behavior of them as N grows infinitely. In section 4 we describe some numerical results involving Vandermonde matrices related to signal processing applications. Finally, we present some conclusions in section 5.

2. An estimate for $\|\mathbf{W}_\mathbf{N}^\dagger\|$ and other preliminaries. In this section we perform a preliminary analysis of the behavior of $\|W_N^\dagger\|$ as a function of N and present a few basic results. In what follows the singular values of a matrix $A \in \mathbb{C}^{M\times N}$ are denoted by $\sigma_i(A)$ and arranged in increasing order, i.e., $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_q(A)$, where $q = \min(M, N)$. The spectrum of $A \in \mathbb{C}^{N\times N}$ is denoted by $\lambda(A)$. Also, $\|A\|$ and $\|A\|_F$ denote the spectral and the Frobenius norm of A, respectively; A^* denotes the transpose conjugate of A.

The following result, whose proof follows from the Rayleigh characterization of $\sigma_n^2(W_N)$ as an eigenvalue of $W_N W_N^*$, provides information about the behavior of $\|W_N^\dagger\|$ as a function of N.

THEOREM 1. *Let W_N be a Vandermonde matrix as in* (1.1) *but with distinct nodes in the complex plane. Then for $N > n$, $\|W_N^\dagger\|$ decreases monotonically with N and this decrease is strict along the subsequence $\{n, 2n, \ldots\}$. Moreover, if $N = p\cdot n$, $p \geq 1$, then*

$$\|W_{p\cdot n}^\dagger\| \leq \frac{\|W_n^{-1}\|}{\sqrt{1+\beta^{2n}+\beta^{4n}+\cdots+\beta^{2(p-1)n}}}, \tag{2.1}$$

where $\beta = \min|z_j|$, $j = 1, \ldots, n$.

This preliminary result suggests that, since $\|W_N^\dagger\|$ decreases monotonically with N, it is not unreasonable to expect improvements of $\kappa_2(W_N)$ as N increases, though the behavior of $\kappa_2(W_N)$ as a function of N may be not always monotonic, as illustrated in section 4. We also observe that the bound (2.1) as a function of N may be very pessimistic if W_n is very ill conditioned. In spite of this, it can be very useful for theoretical purposes as we illustrate below.

THEOREM 2. *Let $\widehat{f}_N$ be the minimum 2-norm solution of the underdetermined system*

$$W_N f = Z^N e,$$

where $Z = \operatorname{diag}(z_1, \ldots, z_n)$ and e is the vector in $\mathbb{R}^n$ of all ones. Suppose that all nodes z_j satisfy either $|z_j| = 1$ or $|z_j| < 1$; then, whenever $N \to \infty$ we have

$$\|\widehat{f}_N\| \to 0.$$

Proof. It is sufficient to note that $\|\widehat{f}_N\| = \|W_N^\dagger Z^N e\| \leq \|W_N^\dagger\|\sqrt{n}\alpha^N$, where $\alpha = \max_j|z_j|$, $j = 1, 2, \ldots, n$, and then to use Theorem 1. ∎

The vector $\widehat{f}_N$ we have just described appears in connection with linear prediction problems and autoregressive modeling of time series. It is currently referred to as *vector of predictor parameters*; see Bazán and Bezerra [2] or Cybenko [6] for details. In our context, it predicts the last column of W_{N+1}. This fact yields the following interesting relationship:

$$ZW_N = W_N\mathcal{F}_N, \tag{2.2}$$

where $\mathcal{F}_N$ is an $N \times N$ companion matrix of the form

$$\mathcal{F}_N = [e_2, e_3, \ldots, e_{N-1}, \widehat{f}_N],$$

in which e_i denotes the ith canonical vector in $\mathbb{R}^N$.

Define F_N by

$$F_N = V_N^* \mathcal{F}_N V_N, \tag{2.3}$$

where

$$V_N = W_N^*(W_N W_N^*)^{-1/2}.$$

Note that F_N is well defined for $W_N W_N^*$ is positive definite. Then the following result is obtained.

THEOREM 3. *The matrix F_N introduced in* (2.3) *has an eigenvalue decomposition given by*

$$F_N = Q_N Z Q_N^{-1}, \tag{2.4}$$

where $Q_N = (W_N W_N^*)^{-1/2}$. Furthermore, this eigenvector matrix satisfies $\kappa_2(Q_N) = \kappa_2(W_N)$.

Proof. The proof is straightforward. □

Much of our results concerning bounds for $\kappa_2(W_N)$ depend on the eigenvalue and singular value spectra of F_N. The following theorem characterizes the singular spectrum of F_N.

THEOREM 4. *Let F_N be as in* (2.3). *Then its singular spectrum is described by*

$$\begin{aligned}
\sigma_1^2(F_N) &= \frac{2 + \|\widehat{f}_N\|^2 - \|p_1\|^2 + \sqrt{(\|\widehat{f}_N\|^2 + \|p_1\|^2)^2 - 4|f_1|^2}}{2}, \\
\sigma_j^2(F_N) &= 1, \quad j = 2, n-1, \\
\sigma_n^2(F_N) &= \frac{2 + \|\widehat{f}_N\|^2 - \|p_1\|^2 - \sqrt{(\|\widehat{f}_N\|^2 + \|p_1\|^2)^2 - 4|f_1|^2}}{2},
\end{aligned} \tag{2.5}$$

where p_1 is the first column of $\mathcal{P}_N$, the orthogonal projector onto $\mathcal{R}(W_N^*)$, and f_1 the first component of the vector $\widehat{f}_N$ introduced in Theorem 2.

Proof. Using (2.4) we have that

$$F_N F_N^* = Q_N Z Q_N^{-2} Z^* Q_N = Q_N^* Z W_N W_N^* Z^* Q_N = V_N^* \mathcal{F}_N \mathcal{F}_N^* V_N,$$

where the last equality is because of (2.2). Since $\mathcal{F}_N$ is a companion matrix, this can be now rewritten as

$$F_N F_N^* = I + xx^* - yy^*, \tag{2.6}$$

where $x = V_N^* \widehat{f}_N$ and $y = V_N^* e_1$. Hence, it can be proved (Bazán and Toint [3]) that $\lambda(F_N F_N^*)$ is formed by $n-2$ eigenvalues equal to 1 and the remaining ones are given by

$$\begin{aligned}
\gamma_1 &= \frac{2 + \|x\|^2 - \|y\|^2 + \sqrt{(\|x\|^2 + \|y\|^2)^2 - 4(x^*y)^2}}{2}, \\
\gamma_2 &= \frac{2 + \|x\|^2 - \|y\|^2 - \sqrt{(\|x\|^2 + \|y\|^2)^2 - 4(x^*y)^2}}{2}.
\end{aligned}$$

On the other hand, note that the Rayleigh quotient associated with (2.6) ensures that

$$\gamma_2 = \min\{\lambda(F_N F_N^*)\} \leq 1 \leq \max\{\lambda(F_N F_N^*)\} = \gamma_1, \tag{2.7}$$

and these inequalities lead to $\gamma_1 = \sigma_1^2(F_N)$ and $\gamma_2 = \sigma_n^2(F_N)$. To conclude the proof, we first note that the columns of V_N form an orthonormal basis of $\mathcal{R}(W_N^*)$ and therefore $\mathcal{P}_N = V_N V_N^*$. Hence we have that

$$\|y\|^2 = y^*y = e_1^* V_N V_N^* e_1 = e_1^* \mathcal{P}_N^* \mathcal{P}_N e_1 = \|\mathcal{P}_N e_1\|^2,$$

since $\mathcal{P}_N$ is a projector, and that

$$\begin{aligned} \|x\|^2 &= x^*x = \widehat{f}_N^* V_N V_N^* \widehat{f}_N = \widehat{f}_N^* \widehat{f}_N = \|\widehat{f}_N\|^2, \\ |x^*y| &= |\widehat{f}_N^* V_N V_N^* e_1| = |\widehat{f}_N^* e_1| = |f_1|, \end{aligned}$$

since $\widehat{f}_N \in \mathcal{R}(W_N^*)$. This concludes the proof. ∎

3. Bounds for $\kappa_2(W_N)$. We first recall an important result concerning the sensitivity of eigenvalues to matrix perturbations, which tells us that, for given $A \in \mathbb{C}^{n\times n}$, if one knows a measure of closeness between its singular values and the absolute values of its eigenvalues, then one can always estimate the conditioning of the related eigenvalue problem as a function of this closeness (see, for instance, Smith [13]). A conclusion obtained along these lines is that the closeness of *all* singular values to the absolute values of the corresponding eigenvalues implies a well-conditioned eigenvalue problem (Ruhe [12]). Intuitively, this is because in that case, the spectrum of A behaves *almost* as the spectrum of a normal matrix whose eigenvalue problem is perfectly conditioned.

The bounds that we derive rely then on the fact that the conditioning of the eigenvalue problem related to F_N is governed by $\kappa_2(W_N)$ (Theorem 3) and also on the observation that we can measure the above-mentioned closeness exactly (Theorems 3 and 4). To show this we need to recall some measures for the sensitivity of matrix eigenvalue problems. In fact, let u_j and v_j be left and right eigenvectors of A related to the eigenvalue λ_j and define

$$s_j = \frac{u_j^* v_j}{\|u_j\|\|v_j\|}. \tag{3.1}$$

It is well known that when λ_j is simple, the number $|s_j|^{-1}$ is finite and uniquely determined. This number measures the sensitivity of λ_j to perturbations in A and it is known as *the condition number* of λ_j; see, e.g., Wilkinson [20, p. 314]. The theorem below states some relations involving these condition numbers and an overall condition number for the eigenvalue problem measured by the Frobenius norm.

THEOREM 5. *Suppose $A \in \mathbb{C}^{n\times n}$ has an eigenvalue decomposition $A = X\Lambda X^{-1}$ and assume all eigenvalues λ_j of A are simple. Then the following inequalities hold:*

$$1 \le |s_j^{-1}| \le \left[1 + \frac{D^2(A)}{(n-1)\delta_j^2}\right]^{\frac{n-1}{2}}, \quad 1 \le j \le n, \quad \text{where } \delta_j = \min_{\substack{1\le k\le n \\ k\ne j}} |\lambda_j - \lambda_k|, \tag{3.2}$$

and $D^2(A) = \|A\|_F^2 - \sum_{j=1}^n |\lambda_j|^2$. Furthermore, the condition number $\kappa_F(X) = \|X\|_F\|X^{-1}\|_F$ satisfies

$$\sum_{j=1}^n \frac{1}{|s_j|} \le \kappa_F(X) \le n \cdot \max_j \frac{1}{|s_j|} \cdot \frac{\max_j \|e_j^* X^{-1}\|}{\min_j \|e_j^* X^{-1}\|}. \tag{3.3}$$

Proof. Inequalities (3.2) and the left inequality of (3.3) are consequences of Theorems 5 and 3 by Smith [13], respectively. In order to prove the remaining inequality, we first introduce $u_j^* = e_j^* X^{-1}$ and $v_j = Xe_j$. Since these vectors are left and right eigenvectors related to λ_j, using the definition of $\kappa_F(X)$ and (3.1), it follows that

$$k_F^2(X) = \sum_{j=1}^{n} \|Xe_j\|^2 \sum_{j=1}^{n} \|e_j^* X^{-1}\|^2 = \sum_{j=1}^{n} \frac{1}{|s_j|^2} \cdot \frac{1}{\|e_j^* X^{-1}\|^2} \sum_{j=1}^{n} \|e_j^* X^{-1}\|^2.$$

The desired inequality follows then by bounding the terms inside the sums. □

Note that if A is normal the bounds for $|s_j^{-1}|$ in (3.2) are sharp since $D^2(A) = 0$. Bounds (3.3) will become sharp if besides normality all right eigenvectors have the same 2-norm. Number $D(A)$ is referred to as *departure from normality* of A (see Henrici [10], or Golub and Van Loan [9, p. 314]). Our results concerning bounds for $\kappa_2(W_N)$ are stated in the following main theorem.

THEOREM 6. *Let W_N be the $n \times N$ Vandermonde matrix with nodes z_j in the unit disk. Define $\alpha = \max_j |z_j|$, $\beta = \min_j |z_j|$, and $\delta = \min_{j,k} |z_j - z_k|$, $j \neq k$. Also, let D_N be the departure from normality of matrix F_N defined in* (2.3)*, that is, $D_N^2 = D^2(F_N) = \|F_N\|_F^2 - (|z_1|^2 + \cdots + |z_n|^2)$. Then, for $N > n \geq 2$, the 2-norm condition number of W_N satisfies*

$$\frac{\sigma_1(F_N)}{\alpha} \leq \kappa_2(W_N) \leq \frac{1}{2}\left(\eta + \sqrt{\eta^2 - 4}\right), \tag{3.4}$$

where $\eta = \rho - n + 2$,

$$\rho = n\left[1 + \frac{D_N^2}{(n-1)\delta^2}\right]^{\frac{n-1}{2}} \phi_N(\alpha, \beta), \tag{3.5}$$

and

$$\phi_N(\alpha, \beta) = \sqrt{\frac{1 + \alpha^2 + \alpha^4 + \cdots + \alpha^{2(N-1)}}{1 + \beta^2 + \beta^4 + \cdots + \beta^{2(N-1)}}}. \tag{3.6}$$

Proof. The left inequality follows after taking 2-norm in both sides of (2.4) and using Theorem 3. To prove the right inequality we first bound $\kappa_F(W_N)$. It suffices to note that $\|e_j^* Q_N^{-1}\| = \|e_j^* (W_N W_N^*)^{1/2}\| = \|e_j^* W_N\|$ and then apply (3.3) and (3.2). This yields

$$\kappa_F(W_N) \leq \rho, \tag{3.7}$$

with ρ as in (3.5). Now, we recall a well-known result involving two condition numbers for a nonsingular $X \in \mathbb{C}^{n\times n}$ (see Smith [13, Theorem 1]) which states that

$$n - 2 + \kappa_2(X) + \kappa_2^{-1}(X) \leq \kappa_F(X).$$

Solving this inequality for $\kappa_2(X)$ and adapting the result to our problem, we obtain

$$\kappa_2(W_N) \leq \frac{1}{2}\left[\kappa_F(W_N) - n + 2 + \sqrt{(\kappa_F(W_N) - n + 2)^2 - 4}\right].$$

The proof concludes by using (3.7) in this inequality. □

We now analyze the quality of the upper bound (3.4) as a function of N. Note that the bound depends on three factors: the separation of the nodes z_j in the unit disk, the departure from normality D_N, and $\phi_N(\alpha,\beta)$. As the contribution of $\phi_N(\alpha,\beta)$ is only modest because the nodes lie in the unit disk, it is clear that the quality of the bound mostly depends on the ratio $D_N^2/(n-1)\delta^2$. Thus, moderate bounds will be obtained provided D_N^2 is of the same order of magnitude as $(n-1)\delta^2$ and n is not very large. Therefore, what remains to do is to analyze the behavior of D_N as a function of N. As the following lemma shows, this is possible, since by Theorem 4, we know the singular values of F_N (and hence $\|F_N\|_F$) exactly.

LEMMA 7. *Let D_N be the departure from normality of matrix F_N. Then, for each $N \geq n$,*

$$(n-1)+\frac{\prod_{j=1}^{n}|z_j|^2}{1+\|\widehat{f}_N\|^2}-\sum_{j=1}^{n}|z_j|^2 \leq D_N^2 \leq (n-1)+\|\widehat{f}_N\|^2+\prod_{j=1}^{n}|z_j|^2-\sum_{j=1}^{n}|z_i|^2, \tag{3.8}$$

and therefore

$$D_\infty^2 = \lim_{N\to\infty} D_N^2 = (n-1)+\prod_{j=1}^{n}|z_j|^2-\sum_{j=1}^{n}|z_j|^2.$$

Proof. From the fact that the product of the singular values of a square matrix is equal to the product of the absolute values of the corresponding eigenvalues, we have, by Theorem 4,

$$\sigma_1^2(F_N)\sigma_n^2(F_N) = \prod_{j=1}^{n}|z_j|^2.$$

Hence, since $\|F_N\|_F^2 = \sigma_1^2(F_N)+\cdots+\sigma_n^2(F_N)$, using the definition of D_N^2, we obtain

$$D_N^2 = (n-2)+\sigma_1^2(F_N)+\frac{\prod_{j=1}^{n}|z_j|^2}{\sigma_1^2(F_N)}-\sum_{j=1}^{n}|z_j|^2. \tag{3.9}$$

Since

$$1 \leq \sigma_1^2(F_N) \leq 1+\|\widehat{f}_N\|^2, \tag{3.10}$$

by Theorem 4, inequalities (3.8) follow by applying (3.10) into (3.9). The last part follows after taking the limit as N is going to infinity in both sides of (3.8) and using Theorem 2. ∎

Lemma 7 shows that, while the behavior of D_N^2 as a function of N depends on the speed at which $\|\widehat{f}_N\|^2$ converges to zero as N increases, the size of D_N^2 for N large will ultimately depend on the size of the nodes themselves. We therefore conclude that whenever N is large enough and $|z_j| \approx 1$, the number D_N^2 will be quite small (though D_N may be large for small N). If in addition we assume that the number n is not very large, then W_N should be well conditioned unless the nodes are extremely close to each other.

A comment is needed regarding the speed at which $\|\widehat{f}_N\|^2$ converges to zero. This speed is difficult to estimate because it depends to some extent on the behavior of $\|W_N^\dagger\|$ as a function of N, which ultimately depends on the nodes themselves.

Despite this, the author's experience is that in many applications where the nodes are not much smaller than 1, e.g., when analyzing slow-decaying signals for frequency extraction, moderate values of N are sufficient to ensure small values of $\|\widehat{f}_N\|^2$ (see the examples discussed in [3]). As Vandermonde matrices related to signal processing contain a number of nodes that is typically not very large (genuine applications, in NMR and modal analysis, for example, point out n ranging from 2 to 16; see [17], [18], and [1]), we conclude those matrices should be well conditioned even for moderate values of N.

The upper bound in (3.4) also applies to the case of real nodes. However, we are aware that in this case, the bound could behave quite differently than for complex nodes, since for real nodes close to each other the condition $D_N^2 \approx (n-1)\delta^2$ is difficult to satisfy. All these observations are illustrated by numerical examples in section 4.

We now consider other consequences of Lemma 7. Suppose first that we substitute the upper bound (3.8) in (3.7). This yields the bound

$$\kappa_F(W_N) \leq n\left[1+\frac{(n-1)+\|\widehat{f}_N\|^2+\prod_{i=1}^n|z_i|^2-\sum_{i=1}^n|z_i|^2}{(n-1)\delta^2}\right]^{\frac{n-1}{2}}\phi_N(\alpha,\beta), \tag{3.11}$$

where the factor between brackets decreases with N since $\|\widehat{f}_N\|^2$ does also (see [3, Lemma 6]). Consequently, whenever this factor decreases faster than $\phi_N(\alpha,\beta)$ increases, the bound decreases too. This suggests that in these conditions, it is not unreasonable to expect improvements of the conditioning itself as N increases. We believe this is the most reliable explanation for what one often observes in practice, so far without theoretical explanation (see, e.g., [16] or [4]).

Another consequence of Lemma 7 is that the asymptotic value D_∞^2 now allows us to obtain the asymptotic value of the bounds, at least for two *frequent configurations* of the nodes occurring in practical applications: nodes inside the unit disk and nodes on the unit disk. Before this, however, we prove that the following limit

$$\kappa_{2,W_\infty} = \lim_{N\to\infty}\kappa_2(W_N) \tag{3.12}$$

exists and that it is computable for our two configurations of the nodes. This can be useful if one wishes to evaluate the quality of the bounds.

First, we consider the case where the nodes satisfy $|z_j|<1$, $j=1\ldots,n$. Note that in this case, if we let $A_N = W_N W_N^*$ and denote the (j,k) entry of this matrix by $(A_N)_{j,k}$, then

$$(A_N)_{j,k} = 1+z_j\bar{z}_k+(z_j\bar{z}_k)^2+\cdots+(z_j\bar{z}_k)^{N-1} = \frac{1-(z_j\bar{z}_k)^N}{1-z_j\bar{z}_k},\quad 1\leq j,k\leq n, \tag{3.13}$$

where the bar denotes complex conjugation. Let $C = A_N + B_N$, where B_N is an $n\times n$ matrix whose (j,k) entry is

$$(B_N)_{j,k} = \frac{(z_j\bar{z}_k)^N}{1-z_j\bar{z}_k}.$$

It then follows that A_N, B_N, and C are all positive definite, and thus we have that

$$\sigma_j(C)\geq\sigma_j(A_N),\qquad j=1,\ldots,n. \tag{3.14}$$

We also note that as $B_N = Z^N C Z^{N*}$ we have $\|B_N\| \le \alpha^{2N}\|C\|$. Using the definition of C, (3.14), and the last inequality, it is not difficult to prove that

$$(1-\alpha^{2N})\kappa_2(C) \le \kappa_2^2(W_N) \le \frac{\kappa_2(C)}{1-\alpha^{2N}\kappa_2(C)} \quad \text{if} \quad \alpha^{2N}\kappa_2(C) < 1. \tag{3.15}$$

These inequalities show that $\kappa_{2,W_\infty} = \sqrt{\kappa_2(C)}$, which can be computed, e.g., via an eigenvalue solver. Although the estimates (3.15) become sharp as N goes to infinity, they are only partially useful in the finite case: (3.15) holds only for

$$N > -\frac{1}{2}\frac{\ln(\kappa_2(C))}{\ln(\alpha)}, \tag{3.16}$$

which may be very large when $\alpha \approx 1$. If the nodes lie on the unit circle, say, $z_j = \exp(\imath t_j)$, with $\imath = \sqrt{-1}$, observe from (3.13) that $(A_N)_{j,j} = N$, $j = 1, \ldots, n$. Using this observation and the fact that

$$|(A_N)_{j,k}| = \left|\frac{\sin(N(t_j - t_k)/2)}{\sin((t_j - t_k)/2)}\right| \le \frac{2}{|z_j - z_k|} \quad \text{for } j \ne k, \tag{3.17}$$

which follows from (3.13), by applying the Gerschgorin theorem to A_N it is easy to prove that

$$1 \le \kappa_2^2(W_N) \le \frac{1+\frac{2n-2}{\delta N}}{1-\frac{2n-2}{\delta N}} \quad \text{if} \quad N > 2(n-1)/\delta. \tag{3.18}$$

Despite the fact that these estimates could hold for potentially large values of N when δ is very small, they are sufficiently informative to allow us to conclude that $k_{2,W_\infty} = 1$. Note that the same result can be derived from (3.4) because in this case $\phi_N = 1$ and $D_\infty = 0$. Thus we have proved the following lemma.

LEMMA 8. *Suppose the nodes z_j $(j = 1{:}n)$ lie either on or inside the unit disk. Then*

$$\kappa_{2,W_\infty} = \begin{cases} 1 & \text{if } |z_j| = 1, \\ \sqrt{\kappa_2(C)} & \text{if } |z_j| < 1, \end{cases}$$

where C is as above.

We are now ready to describe the asymptotic behavior of the bounds in the case where all nodes lie inside the unit disk. This is done in the following corollary.

COROLLARY 9. *Suppose W_N as before and assume all nodes z_j lie inside the unit disk. Then*

$$\frac{1}{\alpha} \le \kappa_{2,W_\infty} \le \frac{1}{2}\left(\breve{\eta} + \sqrt{\breve{\eta}^2 - 4}\right), \tag{3.19}$$

where

$$\breve{\eta} = n\left[1 + \frac{n-1+\prod_{j=1}^n |z_j|^2 - \sum_{j=1}^n |z_j|^2}{(n-1)\delta^2}\right]^{\frac{n-1}{2}} \sqrt{\frac{1-\beta^2}{1-\alpha^2}} - n + 2 \tag{3.20}$$

with α, β, and δ as before.

Proof. It suffices to take the limit when N is going to infinity in both sides of (3.4) and take into account inequalities (3.10) and Lemma 7. ∎

If we now suppose that the nodes are complex with the same magnitude, and if, in addition, we assume that the separation of the nodes is not smaller than the distance of the nodes to the unit circle, then the following simplified version of the upper bound of Corollary 9 is obtained.

COROLLARY 10. *Suppose* $|z_1| = |z_2| = \cdots = |z_n| < 1$. *Also, suppose that* $1-\beta^2 \leq \delta^2$. *Then*

$$\kappa_{2,W_\infty} \leq n\, 2^{\frac{n-1}{2}} - n + 2.$$

Proof. First, notice in (3.6) that $\phi_N(\alpha, \beta) = 1$ for all N since $\alpha = \beta$. Also, since $\beta = |z_i| < 1$, observe that $D_\infty^2 = (n-1)(1-\beta^2) + \beta^{2n} - \beta^2 \leq (n-1)(1-\beta^2)$, for $\beta^{2n} < \beta^2$. Using these observations the upper bound of Corollary 9 reduces to

$$\kappa_{2,W_\infty} \leq n \left(1 + \frac{(1-\beta^2)}{\delta^2}\right)^{\frac{n-1}{2}} - n + 2,$$

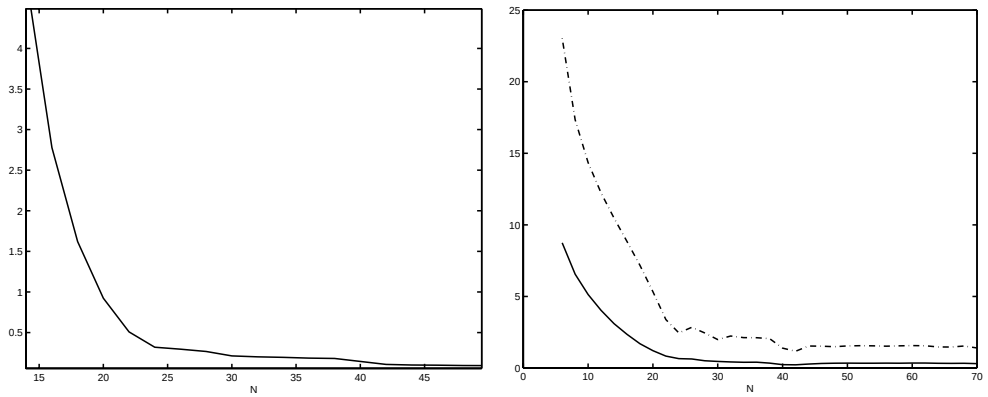
which ensures the statement of the corollary since by assumption $(1-\beta^2) \leq \delta^2$. ∎

As our theoretical results suggest that the numbers $\kappa_2(W_N)$ should be rather small under conditions that are reasonable in several signal processing applications, namely, n small and $|z_j|$ not much smaller than 1, it is natural to ask what one may expect about the behavior of $\kappa_2(W_N)$ if the Vandermonde matrix contains a large number of nodes. This question is very difficult to answer for two reasons. First, as we do not know the speed at which $\|\widehat{f}_N\|^2$ converges to zero, extremely large values of N may be required to achieve "reasonably" small values of D_N^2. Second, even if D_N^2 is small, in the absence of a priori information on the configuration of the nodes, the size of the ratio $D_N^2/(n-1)\delta^2$ (which was shown to govern the quality of upper bound in (3.4)), seems impossible to predict. In spite of this, motivated by the results obtained for the case where the nodes lie on the unit circle, we conjecture that if we assume the nodes close to the unit circle and uniformly spaced, then W_N could be reasonably conditioned, though this could hold for excessively large values of N, as suggested in (3.18).

4. Numerical examples. This section numerically illustrates the behavior of the bounds of Theorem 6 and the remarks made throughout the last section. We analyze Vandermonde matrices arising from frequency estimation and system identification problems where the nodes often satisfy $|z_j| \approx 1$, since they are of the form $z_j = e^{(d_j + \imath\omega_j)\Delta t}$, where $\imath = \sqrt{-1}$, $d_j \leq 0$, and $d_j \Delta t \approx 0$. For each example we show the behavior of the upper bound in (3.4) and the conditioning itself as functions of N, illustrating the role of the departure from normality D_N on that behavior. The departure from normality D_N is computed by using (3.9). The asymptotic values of the upper and lower bounds of Corollary 9, which we denote here by L_∞ and U_∞, respectively, are also shown. All numerical experiments were performed in MATLAB.

We also tried to compare estimates (3.15) with those obtained from Theorem 6, but no illustrative comparison was possible. Whereas the bounds provided by this theorem were sufficiently informative for moderate values of N, this no longer occurred with the estimates (3.15) which required large values of N for its realization, because of the restriction (3.16).

4.1. Example 1: Conditioning of a Vandermonde matrix whose nodes are the poles of a stable time invariant linear system. In this example we



(a) (b)

FIG. 1. (a) *Behavior of* $\|\widehat{f}_N\|^2$. (b) *Upper bound of Theorem* 6 *and* $\kappa_2(W_N)$ *on a logarithmic scale* ($\kappa_2(W_N)$*: solid line, upper bound: dashed-dotted line).*

TABLE 1
Nodes of Vandermonde matrix, their absolute values, and corresponding separations.

j	z_j	$\lvert z_j\rvert$	δ_j^2
1	$0.9856 \pm 0.1628\imath$	0.9990	0.0794
2	$0.8976 \pm 0.4305\imath$	0.9955	0.0264
3	$0.8127 \pm 0.5690\imath$	0.9921	0.0264

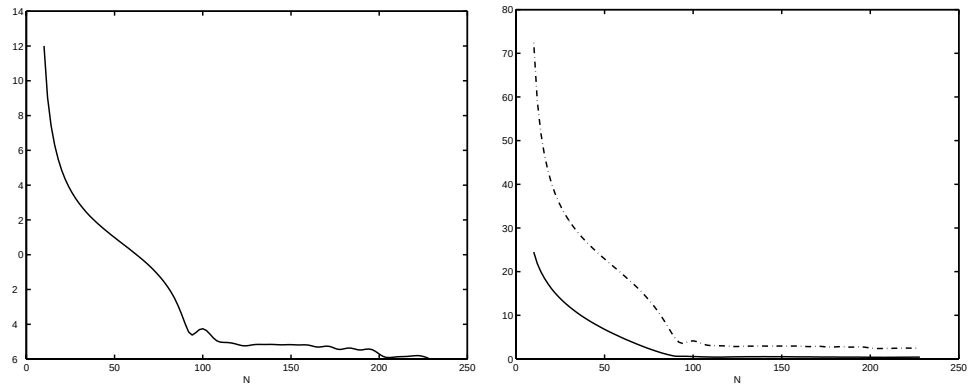
analyze a Vandermonde matrix W_N of order $6 \times N$ (i.e., $n = 6$), whose nodes are the eigenvalues of the matrix

$$A = \begin{pmatrix} 0.9856 & 0.1628 & 0 & 0 & 0 & 0 \\ -0.1628 & 0.9856 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.8976 & 0.4305 & 0 & 0 \\ 0 & 0 & -0.4305 & 0.8976 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.8127 & 0.5690 \\ 0 & 0 & 0 & 0 & -0.5690 & 0.8127 \end{pmatrix}.$$

This matrix concentrates dynamical properties of a simulated discrete time invariant linear system. It is often used in testing system identification algorithms and was taken from [11]. The nodes of W_N, as well as their absolute values and separations δ_j, are illustrated in Table 1.

As in this case all $|z_j|$ are close to 1 and the nodes are not excessively close to each other, the number D_N^2 should be quite small because the fast decrease of $\|\hat{f}_N\|^2$ towards 0, as illustrated in Figure 1(a), even for values of N near to 25. This justifies the relatively small upper bounds for $\kappa_2(W_N)$ for $N \geq 30$ illustrated in Figure 1(b). The logarithmic scale in that figure is because for N near to 6, both $\kappa_2(W_N)$ and the corresponding upper bounds are excessively large. For illustration, while some values of the conditioning itself are $\kappa_2(W_6) = 6.2996 \times 10^3$, $\kappa_2(W_8) = 702.1543$, $\kappa_2(W_{20}) = 3.3559$, for the same values of N the bound decreases from 1.0143×10^{10} down to 206.9487. The situation is even more favorable for $N = 30$, where we obtain $\kappa_2(W_{30}) = 1.5899$ and 7.2880 for the bound.

For this example the limiting values of both the bounds and the conditioning



(a) (b)

FIG. 2. (a) *Behavior of* D_N^2 *on a logarithmic scale.* (b) *Upper bound of Theorem* 6 *and* $\kappa_2(W_N)$ *on a logarithmic scale* ($\kappa_2(W_N)$*: solid line, upper bound: dashed-dotted line.)*

TABLE 2
Nodes of Vandermonde matrix, their absolute values, and corresponding separations.

j	z_j	$\lvert z_j\rvert$	δ_j^2
1	$0.9699 \pm 0.2249\imath$	0.9956	0.0042
2	$0.9532 \pm 0.2931\imath$	0.9972	0.0049
3	$0.9844 \pm 0.1619\imath$	0.9976	0.0031
4	$0.9921 \pm 0.1065\imath$	0.9977	0.0031
5	$0.9972 \pm 0.0485\imath$	0.9989	0.0034

itself are

$$L_\infty = 1.0010, \qquad U_\infty = 12.7454, \qquad \text{and} \quad \kappa_{2,W_\infty} = 2.7927.$$

A comparison of the behavior of bounds of Theorem 6 with that of estimates (3.15) was not possible because the latter hold only for $N \geq 983$.

4.2. Example 2: Conditioning of a Vandermonde matrix arising from a practical application. This example considers a Vandermonde matrix of order $10 \times N$ (i.e., $n = 10$), whose nodes z_j contain information regarding frequencies and decay factors of a vibrating structure. In this case, because the nodes z_j were synthesized from a slow-decaying discrete real signal (see [4] for details), they come in complex conjugate pairs and are close to the unit circle. This is illustrated in Table 2, where the separations of the nodes δ_j are also included.

For this matrix we deduce that the asymptotic values of both the bounds and κ_{2,W_∞} are

$$L_\infty = 1.0016, \qquad U_\infty = 11.8223, \qquad \text{and} \quad \kappa_{2,W_\infty} = 1.7470.$$

Although these asymptotic values are very apparent, it is also important to observe the behavior of the upper bound and the number $\kappa_2(W_N)$ itself as N increases. This is illustrated in Figure 2(b) on a logarithmic scale, from which we see that the upper bound starts to approximate well its asymptotic value $\ln(11.8223) \approx 2.4698$ for values of N near to 90 and that the quality of this approximation is rather good for $N = 200$. We also see that the behavior of the bound for $10 \leq N \leq 90$ is remarkably

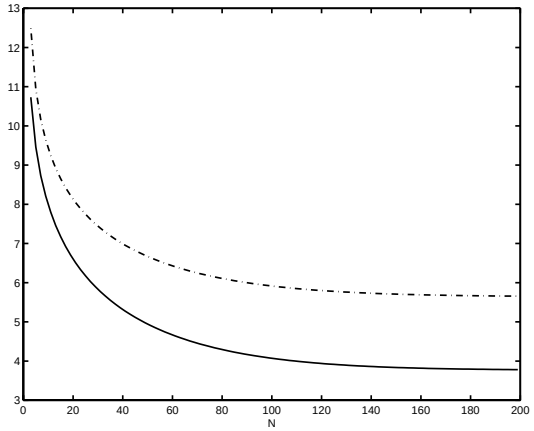


FIG. 3. *Upper bound of Theorem* 6 *(dashed-dotted line) and* $\kappa_2(W_N)$ *(solid line) on a logarithmic scale.*

pessimistic. These two well-distinguished forms of behavior of the bound depend on the behavior of the ratio $D_N^2/(n-1)\delta^2$ in (3.5): while for $10 \leq N \leq 89$, we obtain $0.0327 \leq D_N^2 \leq 1.6348\times 10^5$, which indicates that D_N^2 is larger than $(n-1)\delta^2 = 0.0282$, this no longer occurs for $N = 90$, where we obtain $D_N^2 = 0.0189$. The situation is even more favorable when $N = 200$, where $D_N^2 = 0.0036$, a value very close to that of $D_\infty^2 = 0.0012$. The number D_N^2 is displayed as a function of N in Figure 2(a). The behavior of the conditioning itself as a function of N is still more striking: while for $10 \leq N \leq 50$, we obtain $1.3505 \times 10^3 \leq \kappa_2(W_N) \leq 4.3183 \times 10^{10}$, indicating that W_N is ill conditioned for those values of N, that behavior starts to change drastically for $N \geq 50$. A few values illustrate this change: $\kappa_2(W_{50}) = 880.6591$, $\kappa_2(W_{56}) = 264.2212$, and $\kappa_2(W_{60}) = 125.3186$, and for N large, say, $N = 250$, $\kappa_2(W_{250}) = 1.5322$.

Estimates (3.15) hold for $N \geq 383$, and therefore no useful comparison with bounds (3.4) is possible.

4.3. Example 3: Conditioning of a Vandermonde matrix with real nodes. In this example we consider a Vandermonde matrix with real positive numbers as nodes. The nodes arise from an exponential model that simulates fluorescence decay data: $h_k = 2z_1^k + 3z_2^k + 0.1z_3^k$, where $z_1 = e^{-0.05} = 0.9512$, $z_2 = e^{-0.03} = 0.9704$, and $z_3 = e^{-0.02} = 0.9802$ (see [14]). For this case, $(n-1)\delta^2 = 1.9025 \times 10^{-4}$ and $D_\infty^2 = 0.0113$. Hence, as commented in section 4, because the quantity $(n-1)\delta^2$ is always larger than D_N^2, the behavior of the upper bound is rather different from that observed in the case where nodes are complex. The following asympotic values illustrate this observation:

$$L_\infty = 1.02, \qquad U_\infty = 283.2361, \qquad \text{and} \quad \kappa_{2,W_\infty} = 43.2820.$$

The behavior of both the conditioning itself and the corresponding upper bound are illustrated in Figure 3. Estimates (3.15) hold in this case for $N \geq 189$.

5. Conclusions and future work. We have derived lower and upper bounds for the conditioning of rectangular Vandermonde matrices, both depending on the dimension of the problem. The behavior of these bounds for large N allows us to predict

that the conditioning itself may be good, under conditions that are reasonable in many applications. Specifically, we have proved that these bounds can be small, provided the dimension of the problem is large enough, with the effect strengthened when the nodes are close to one in absolute value but not excessively close to each other. This was illustrated with numerical examples where we considered Vandermonde matrices with nodes arising from simulated and real applications linked to frequency estimation and parameter identification, in which case the above mentioned conditions are very often satisfied. However, we have not yet investigated the rate at which $\|W_N^\dagger\|$ itself decreases to its limit value.

Perhaps the most important contribution of this work is the qualitative prediction of the improvement of the conditioning of rectangular Vandermonde matrices that is obtained when the dimension is large enough. This information may play a crucial role in the solution of several signal processing applications, in which taking *N large enough* is not a severe restriction. Moreover, given that rank-deficient Hankel matrices are directly related to rectangular Vandermonde matrices, one can determine bounds for the singular values of the former by using the results presented in this paper. This may be of interest in signal subspace approaches for solving frequency estimation problems where the size of the smallest nonzero singular value of Hankel matrices plays a crucial role. This is the object of current research.

Acknowledgments. The author wishes to thank B. Colson, E. Cornelis, and the Department of Mathematics at FUNDP (Namur, Belgium) for providing a cordial environment. Special thanks are due to Ph. L. Toint for helpful discussions and comments regarding the presentation of this paper and to the referees for their suggestions and constructive criticism. Thanks also go to Nick Higham for many suggestions that have improved the presentation of this paper. The author is particularly grateful to one referee for providing estimates (3.15) and (3.18).

REFERENCES

[1] F. S. V. Bazán and C. Bavastri, *An optimized pseudo-inverse algorithm (OPIA) for multi-input multi-output modal parameter identification*, Mechanical Systems and Signal Processings, 10 (1996), pp. 365–380.

[2] F. S. V. Bazán and L. H. Bezerra, *On zero location of predictor polynomials*, Numer. Linear Algebra Appl., 4 (1997), pp. 459–468.

[3] F. S. V. Bazán and Ph. L. Toint, *Singular Values of Predictor Matrices and Signal Eigenvalue Bounds*, Technical report 5, Department of Mathematics, FUNDP, Namur, Belgium, 1998.

[4] F. S. V. Bazán, Ph. L. Toint, and M. C. Zambaldi, *A Conjugate-Gradients Based Method for Harmonic Retrieval Problems that Does Not Use Explicit Signal Subspace Computation*, Technical report 16, Department of Mathematics, FUNDP, Namur, Belgium, 1997.

[5] A. Cordoba, W. Gautschi, and S. Ruscheweyh, *Vandermonde matrices on the circle: Spectral properties and conditioning*, Numer. Math., 57 (1990), pp. 577–591.

[6] G. Cybenko, *Restrictions of normal operators, Padé approximation and autoregressive time series*, SIAM J. Math. Anal., 15 (1984), pp. 753–767.

[7] W. Gautschi, *Norm estimates for inverses of Vandermonde matrices*, Numer. Math., 23 (1975), pp. 337–347.

[8] W. Gautschi and G. Inglese, *Lower bounds for the condition number of Vandermonde matrices*, Numer. Math., 52 (1988), pp. 241–250.

[9] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.

[10] P. Henrici, *Bounds for iterates, inverses, spectral variation and fields of values of non-normal matrices*, Numer. Math., 4 (1962), pp. 320–329.

[11] J. N. Juang, M. Phan, L. G. Horta, and R. W. Longman, *Identification of observer/Kalman filter Markov parameters: Theory and experiments*, J. Guidance Control Dynam., 16 (1993), pp. 28–36.

[12] A. Ruhe, *On the closeness of eigenvalues and singular values for almost normal matrices*, Linear Algebra Appl., 11 (1975), pp. 87–94.
[13] R. A. Smith, *The condition numbers of the matrix eigenvalue problem*, Numer. Math., 10 (1967), pp. 232–240.
[14] B. Soni and J. Eisenfeld, *System identification of models exhibiting exponential, harmonic and resonant modes*, in Applied Nonlinear Analysis, V. Lakshmikantham, ed., Academic Press, New York, 1979, pp. 555–568.
[15] E. E. Tyrtyshnikov, *How bad are Hankel matrices?*, Numer. Math., 67 (1994), pp. 261–269.
[16] A. Van Der Veen, E. F. Deprettere, and A. Lee Swindlehurst, *Subspace-based signal analysis using singular value decomposition*, Proceedings of the IEEE, 81 (9), 1993, pp. 1277–1309.
[17] S. Van Huffel , *Enhanced resolution based on minimum variance estimation and exponential data modeling*, Signal Processing, 33 (1993), pp. 333–355.
[18] S. Van Huffel, H. Chen, C. Decanniere, and P. Van Hecke, *Algorithm for time-domain NMR data fitting based on total least squares*, J. Magnetic Resonance, A 110 (1994), pp. 228–237.
[19] M. Wei and G. Majda, *A new theoretical approach for Prony's method*, Linear Algebra Appl., 136 (1990), pp. 119–132.
[20] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, UK, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 694–698
© 2000 Society for Industrial and Applied Mathematics

ON A NEWTON-LIKE METHOD FOR SOLVING ALGEBRAIC RICCATI EQUATIONS*

CHUN-HUA GUO† AND ALAN J. LAUB‡

Abstract. An exact line search method has been introduced by Benner and Byers [*IEEE Trans. Automat. Control*, 43 (1998), pp. 101–107] for solving continuous algebraic Riccati equations. The method is a modification of Newton's method. A convergence theory is established in that paper for the Newton-like method under the strong hypothesis of controllability, while the original Newton's method needs only the weaker hypothesis of stabilizability for its convergence theory. It is conjectured there that the controllability condition can be weakened to the stabilizability condition. In this article we prove that conjecture.

Key words. continuous algebraic Riccati equations, Newton's method, maximal symmetric solution, convergence rate, line search

AMS subject classifications. 15A24, 65F30, 93B40

PII. S0895479898348519

1. Introduction.

Consider the continuous algebraic Riccati equation (CARE)

$$\mathcal{R}(X) = XDX - XA - A^TX - C = 0, \tag{1.1}$$

where $A, D, C \in \mathbb{R}^{n\times n}$ and $D^T = D$, $C^T = C$. Matrix equations of this form occur in many important applications (see [11], [9]). We limit our attention to real equations, since the complex analog of these equations can be studied in a similar manner. All matrices in this note are real matrices.

For $A, H \in \mathbb{R}^{n\times n}$, the pair (A, H) is said to be controllable if

$$\operatorname{rank}(H\ AH\ A^2H \cdots A^{n-1}H) = n.$$

The pair (A, H) is called stabilizable if there is a $K \in \mathbb{R}^{n\times n}$ such that $A - HK$ is stable, i.e., all its eigenvalues are in the open left half-plane. Note that the matrix K can be chosen to be symmetric if the matrix H is symmetric (see [9]). It is well known that controllability implies stabilizability (see [9]). The order relation on the set of symmetric matrices is the usual one: $X \geq Y$ if $X - Y$ is positive semidefinite. A symmetric solution X_+ of (1.1) is called maximal if $X_+ \geq X$ for every symmetric solution X.

Let $\mathcal{S}$ be the set of symmetric matrices in $\mathbb{R}^{n\times n}$. The Riccati function $\mathcal{R}$ is clearly a mapping from $\mathcal{S}$ into itself. Moreover, the first Fréchet derivative of $\mathcal{R}$ at a matrix X is a linear map $\mathcal{R}'_X : \mathcal{S} \to \mathcal{S}$ given by

$$\mathcal{R}'_X(S) = -\{(A - DX)^TS + S(A - DX)\}.$$

When (A, D) is stabilizable, then a Newton procedure for the solution of the Riccati equation (1.1) is as follows.

*Received by the editors November 12, 1998; accepted for publication (in revised form) by V. Mehrmann March 19, 1999; published electronically January 25, 2000. This research was supported in part by National Science Foundation grant ECS-9633326.

http://www.siam.org/journals/simax/21-2/34851.html

†Department of Computer Science, University of California, Davis, One Shields Avenue, Davis, CA 95616-8562 (chguo@math.uregina.ca).

‡College of Engineering, University of California, Davis, One Shields Avenue, Davis, CA 95616-5294 (laub@ucdavis.edu).

ALGORITHM 1.1. NEWTON'S METHOD FOR THE CARE.

1. *Choose a symmetric matrix X_0 for which $A - DX_0$ is stable.*
2. *For $k = 0, 1, \ldots$ do:*
   Solve the Lyapunov equation $(A - DX_k)^T N_k + N_k(A - DX_k) = \mathcal{R}(X_k)$;
   compute $X_{k+1} = X_k + N_k$.

Note that the matrix X_0 in the above algorithm can be found by the method described in [12]. The Lyapunov equation has a unique solution whenever the matrix $A - DX_k$ is stable (see [9], for example). The unique solution N_k can be found efficiently by the methods described in [1].

Concerning the convergence of Algorithm 1.1, we have the following result (see [8], [4], [5], and [9]).

THEOREM 1.2. *Assume that $D \geq 0$, $C^T = C$, (A, D) is stabilizable, and there exists a symmetric solution of the CARE* (1.1). *Then there is a maximal symmetric solution X_+ of* (1.1) *and all the eigenvalues of $A - DX_+$ are in the closed left half-plane. Moreover, Algorithm* 1.1 *determines a sequence of symmetric matrices $\{X_k\}_{k=1}^{\infty}$ for which $A - DX_k$ is stable for $k = 1, 2, \ldots$, $X_1 \geq X_2 \geq \cdots$, and $\lim_{k\to\infty} X_k = X_+$. The convergence is quadratic if $A - DX_+$ has no eigenvalues on the imaginary axis.*

Remark 1.1. The situation where $A - DX_+$ has eigenvalues on the imaginary axis has been studied recently in [6]. It is shown in [6] that the convergence of Algorithm 1.1 is either quadratic or linear with rate 1/2 provided that all eigenvalues of $A - DX_+$ on the imaginary axis are semisimple. Whether quadratic convergence is indeed possible remains an open problem. If the convergence is linear, the performance of Algorithm 1.1 can be improved dramatically by using a simple modification strategy (see [6] for details). When $A - DX_+$ has nonsemisimple eigenvalues on the imaginary axis, the convergence behavior of Algorithm 1.1 also remains an open problem.

In this note we limit our attention to the case where $A - DX_+$ has no eigenvalues on the imaginary axis. In this case X_+ is called a stabilizing solution.

It is important to note that $X_0 \geq X_1$ is not true in general. As an example (see [7]), consider the CARE (1.1) with $A = 0, D = I, C = I$. In this case $X_+ = I$, but if $X_0 = \epsilon I$ (ϵ is a small positive number), we have $X_1 = \frac{1+\epsilon^2}{2\epsilon} I$. Thus $\|X_1 - X_+\|_\infty$ can be arbitrarily large even though $\|X_0 - X_+\|_\infty < 1$. In [7], some tests are described to determine whether Newton refinement will improve an approximate solution obtained by other methods. In [3], Benner and Byers introduced step size control into Newton's method for the generalized CARE. The generalized CARE they considered can be reduced to the CARE of the form (1.1). Although the reduction is not always numerically feasible, convergence analysis can always be made on the reduced equation. We can therefore limit our attention to the CARE of the form (1.1).

ALGORITHM 1.3. EXACT LINE SEARCH METHOD (see [3]).

1. *Choose a symmetric matrix X_0 for which $A - DX_0$ is stable.*
2. *For $k = 0, 1, \ldots$ do:*
   Solve the Lyapunov equation $(A - DX_k)^T N_k + N_k(A - DX_k) = \mathcal{R}(X_k)$;
   find a minimizer t_k of $f_k(t) = \|\mathcal{R}(X_k + tN_k)\|_F^2 (0 \leq t \leq 2)$;
   compute $X_{k+1} = X_k + t_k N_k$.

In the above algorithm, $\|\cdot\|_F$ is the Frobenius norm. The algorithm effectively solves the problem of a potentially disastrous first Newton step and can be implemented in such a way that the increased amount of computational work per iteration is marginal (see [3] for details).

The following convergence result is proved in [3].

THEOREM 1.4. *Assume that $D \geq 0$, $C^T = C$, (A, D) is controllable, and there is a stabilizing maximal solution X_+ of the CARE (1.1). If the minimizer t_k of $f_k(t)(0 \leq t \leq 2)$ satisfies $t_k \geq t_L > 0$ for all $k \geq 0$, where t_L is a constant tolerance threshold, then the sequence $\{X_k\}$ in Algorithm 1.3 is well defined and $X_k \to X_+$ quadratically. If $t_k = 0$ for some $k \geq 0$, then $X_k = X_+$.*

In Theorem 1.2, we need only the stabilizability of (A, D) for the convergence of Newton's method. In the above theorem, however, the controllability of (A, D) is required. Since Algorithm 1.3 is designed to improve the performance of Newton's method, it would be desirable to replace the controllability condition in the above theorem by the stabilizability condition. It is conjectured in [3] that this can be done. We will confirm the conjecture in this note.

2. Relaxing the controllability condition. Several interesting results are established in [3] to prove Theorem 1.4. The following two lemmas will be utilized to relax the controllability condition in Theorem 1.4.

LEMMA 2.1 (see [3]). *Assume that $D \geq 0$, $C^T = C$, and $A - DX_+$ is stable. If $A - DX_k$ is stable, then $A - D(X_k + tN_k)$ is also stable for any $t \in [0, 2]$.*

Remark 2.1. The controllability of (A, D) is not needed for this lemma. In the proof (given in [3]) of Theorem 1.4, the controllability of (A, D) is required to prove that the sequence $\{X_k\}$ (already well defined by Lemma 2.1) is bounded. Otherwise the stabilizability of (A, D) is sufficient for the proof of Theorem 1.4.

LEMMA 2.2 (see [3]). *Suppose that $\{X_k\}_{k=1}^{\infty}$ is a sequence of symmetric matrices such that $\{\mathcal{R}(X_k)\}_{k=1}^{\infty}$ is bounded. If (A, D) is controllable, then $\{X_k\}_{k=1}^{\infty}$ is bounded.*

Remark 2.2. In the above lemma, the controllability of (A, D) cannot be replaced by the stabilizability of (A, D). This is illustrated by the CARE (1.1) with

$$D = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad A = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}, \qquad C = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

(see Example 7.9.1 of [9]). Clearly, the pair (A, D) is stabilizable but not controllable. The real symmetric solutions of the CARE can be found to be

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \begin{pmatrix} -1 & b \\ b & -b^2/2 \end{pmatrix}, \quad b \in \mathbb{R}.$$

Note that the solution set of $\mathcal{R}(X) = 0$ is already unbounded. Note however that among these solutions, there is only one stabilizing solution, namely, $X_+ = \mathrm{diag}(1, 0)$. The uniqueness of the stabilizing solution is actually a general result (see Theorem 9.3.1 of [9], for example).

The above observation leads us to formulate the following result (see also Acknowledgments).

LEMMA 2.3. *Suppose that $\{X_k\}_{k=1}^{\infty}$ is a sequence of symmetric matrices such that $\{\mathcal{R}(X_k)\}_{k=1}^{\infty}$ is bounded. If (A, D) is stabilizable and $A - DX_k$ is stable for each $k \geq 1$, then $\{X_k\}_{k=1}^{\infty}$ is bounded.*

Proof. It is clear that we may assume $C = 0$ without loss of generality. Consider the controllable subspace of (A, D):

$$\mathcal{C} = \mathrm{Im}(D\ AD\ A^2D \cdots A^{n-1}D).$$

If $\dim \mathcal{C} = n$, then (A, D) is controllable and the result follows directly from Lemma 2.2. If $\dim\ \mathcal{C} = r < n$, we can consider using the control (or Kalman) normal form

of (A, D) (see [9]). Thus, we take an orthonormal basis for $\mathcal{C}$ and expand it to an orthonormal basis of $\mathbb{R}^n$. Under this basis, the matrices A and D have very simple forms. More precisely, there is an orthogonal matrix U such that

$$A = U^T \begin{pmatrix} A_1 & A_3 \\ 0 & A_2 \end{pmatrix} U, \qquad D = U^T \begin{pmatrix} D_1 & 0 \\ 0 & 0 \end{pmatrix} U,$$

where $A_1, D_1 \in \mathbb{R}^{r\times r}$. Write accordingly

$$X_k = U^T \begin{pmatrix} X_{k,1} & X_{k,3} \\ X_{k,3}^T & X_{k,2} \end{pmatrix} U.$$

Since $A - DX_k$ is stable for all $k \geq 1$, we see immediately that A_2 is stable and $A_1 - D_1X_{k,1}$ is stable for all $k \geq 1$. Moreover, it follows readily from $\dim \mathcal{C} = r$ that

$$\dim \operatorname{Im}(D_1\, A_1D_1 \cdots A_1^{r-1}D_1) = \dim \operatorname{Im}(D_1\, A_1D_1 \cdots A_1^{n-1}D_1) = r.$$

Thus, (A_1, D_1) is controllable.

Since the sequence $\{\mathcal{R}(X_k)\}$ is bounded, direct computation shows that the matrix sequences

$$\{X_{k,1}D_1X_{k,1} - X_{k,1}A_1 - A_1^TX_{k,1}\}, \tag{2.1}$$

$$\{X_{k,1}D_1X_{k,3} - X_{k,1}A_3 - X_{k,3}A_2 - A_1^TX_{k,3}\}, \tag{2.2}$$

and

$$\{X_{k,3}^TD_1X_{k,3} - X_{k,3}^TA_3 - X_{k,2}A_2 - A_3^TX_{k,3} - A_2^TX_{k,2}\} \tag{2.3}$$

are all bounded. Since (2.1) is bounded and (A_1, D_1) is controllable, we know from Lemma 2.2 that $\{X_{k,1}\}$ is bounded. We then know from the boundedness of (2.2) that $\{(A_1 - D_1X_{k,1})^TX_{k,3} + X_{k,3}A_2\}$ is bounded. Since $\{A_1 - D_1X_{k,1}\}$ is bounded, A_2 is stable, and $A_1 - D_1X_{k,1}$ is stable for all $k \geq 1$, it follows from Lemma 2.4 below that $\{X_{k,3}\}$ is bounded. Now that $\{X_{k,3}\}$ has been shown to be bounded, it follows from the boundedness of (2.3) that $\{X_{k,2}A_2 + A_2^TX_{k,2}\}$ is bounded. Since A_2 is stable, $\{X_{k,2}\}$ is also bounded by Lemma 2.4. Therefore, $\{X_k\}$ is bounded. ∎

The following lemma has been used in the proof of Lemma 2.3.

LEMMA 2.4. *Let $\mathcal{A}, \mathcal{B}$, and $\mathcal{C}$ be bounded sets of matrices of size $m \times m$, $n \times n$, and $m \times n$, respectively. Assume that the closure of $\cup_{A\in\mathcal{A}}\sigma(A)$ and the closure of $\cup_{B\in\mathcal{B}}\sigma(B)$ are disjoint. Then the set*

$$\mathcal{G} = \{X \mid AX - XB = C,\ A \in \mathcal{A},\ B \in \mathcal{B},\ C \in \mathcal{C}\}$$

is bounded.

Proof. For $A \in \operatorname{clo}(\mathcal{A})$, $B \in \operatorname{clo}(\mathcal{B})$, and $C \in \operatorname{clo}(\mathcal{C})$, the Sylvester equation $AX - XB = C$ has a unique solution $X = F(A, B, C)$ by the assumption. See, for example, [9] or [10]. It is clear that the function F is continuous on the compact set $\operatorname{clo}(\mathcal{A}) \times \operatorname{clo}(\mathcal{B}) \times \operatorname{clo}(\mathcal{C})$. Therefore, F maps this compact set onto a compact set. As a result, $\mathcal{G}$ is bounded. ∎

We can now relax the controllability condition in Theorem 1.4.

THEOREM 2.5. *The conclusions in Theorem* 1.4 *are valid if the pair (A, D) is stabilizable.*

Proof. This follows readily from Remark 2.1, Lemma 2.1, and Lemma 2.3. Note that $\{\mathcal{R}(X_k)\}_{k=1}^{\infty}$ is bounded by Algorithm 1.3 itself. ∎

Acknowledgments. The authors thank Peter Benner for pointing out to them that Lemma 2.3 was formulated earlier as a conjecture in [2] and for sending them a copy of the reference. Chun-Hua Guo would like to thank Peter Lancaster for introducing him to the study of algebraic Riccati equations several years ago. He also gratefully acknowledges the support of an NSERC postdoctoral fellowship.

REFERENCES

[1] R. H. Bartels and G. W. Stewart, *Solution of the matrix equation $AX + XB = C$*, Comm. ACM, 15 (1972), pp. 820–826.
[2] P. Benner, *Contributions to the Numerical Solution of Algebraic Riccati Equations and Related Eigenvalue Problems*, Logos-Verlag, Berlin, 1997.
[3] P. Benner and R. Byers, *An exact line search method for solving generalized continuous-time algebraic Riccati equations*, IEEE Trans. Automat. Control, 43 (1998), pp. 101–107.
[4] W. A. Coppel, *Matrix quadratic equations*, Bull. Austral. Math. Soc., 10 (1974), pp. 377–401.
[5] I. Gohberg, P. Lancaster, and L. Rodman, *On Hermitian solutions of the symmetric algebraic Riccati equation*, SIAM J. Control Optim., 24 (1986), pp. 1323–1334.
[6] C.-H. Guo and P. Lancaster, *Analysis and modification of Newton's method for algebraic Riccati equations*, Math. Comp., 67 (1998), pp. 1089–1105.
[7] C. Kenney, A. J. Laub, and M. Wette, *Error bounds for Newton refinement of solutions to algebraic Riccati equations*, Math. Control Signals Systems, 3 (1990), pp. 211–224.
[8] D. L. Kleinman, *On an iterative technique for Riccati equation computations*, IEEE Trans. Automat. Control, 13 (1968), pp. 114–115
[9] P. Lancaster and L. Rodman, *Algebraic Riccati Equations*, Clarendon Press, Oxford, 1995.
[10] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, 2nd ed., Academic Press, Orlando, 1985.
[11] V. L. Mehrmann, *The Autonomous Linear Quadratic Control Problem*, Lecture Notes in Control and Inform. Sci. 163, Springer-Verlag, Berlin, 1991.
[12] V. Sima, *An efficient Schur method to solve the stabilizing problem*, IEEE Trans. Automat. Control, 26 (1981), pp. 724–725.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 2, pp. 699–702
© 2000 Society for Industrial and Applied Mathematics

THE p-RELATIVE DISTANCE IS A METRIC*

ANDERS BARRLUND†

Abstract. The conjecture that the p-relative distance, $\varrho_p(\alpha,\tilde{\alpha}) = |\alpha-\tilde{\alpha}|/\sqrt[p]{|\alpha|^p+|\tilde{\alpha}|^p}$, is a metric is proved.

Key words. relative distance, metric

AMS subject classification. 15A45

PII. S0895479898340883

1. Introduction.

The *p-relative distance* between α and $\tilde{\alpha} \in \mathbb{C}$ is defined as

$$\varrho_p(\alpha,\tilde{\alpha}) = \frac{|\alpha-\tilde{\alpha}|}{\sqrt[p]{|\alpha|^p+|\tilde{\alpha}|^p}} \quad \text{for } 1 \le p \le \infty. \tag{1.1}$$

For convenience, we define $0/0 := 0$.

Li [2] presented the conjecture that the p-relative distance is a metric. It is trivial that

$$\begin{aligned} &\varrho_p(\alpha,\beta) \ge 0, \\ &\varrho_p(\alpha,\beta) = 0 \quad \text{if and only if } \alpha = \beta, \\ &\varrho_p(\alpha,\beta) = \varrho_p(\beta,\alpha). \end{aligned}$$

However, it has been an open question whether

$$\varrho_p(\alpha,\beta) \le \varrho_p(\alpha,\gamma) + \varrho_p(\gamma,\beta). \tag{1.2}$$

Li [2] proved (1.2) for $\alpha,\beta,\gamma \in \mathbb{R}$. Day [1] proved (1.2) for $p=\infty$. In this paper we prove (1.2) for $\alpha,\beta,\gamma \in \mathbb{C}$.

The paper is outlined as follows: In section 2 we prove a simple lemma which is used to prove (1.2) in section 3. We will let $\|x\|_2$ denote the Euclidean vector norm.

2. A simple lemma.

LEMMA 1. *Let $1 \le p \le \infty$. If $x \ge 0$ and $y > 0$, then*

$$\sqrt[p]{\frac{1+x^p}{1+y^p}} \ge \min(x, 1/y). \tag{2.1}$$

If $x \ge y \ge 1$, then also

$$\sqrt[p]{\frac{1+x^p}{1+y^p}} > \sqrt{x/y}. \tag{2.2}$$

Proof. If $x \le 1/y$, then $x^p y^p \le 1$, which implies that $1+x^p \ge x^p y^p + x^p = (1+y^p)x^p$, from which (2.1) follows. If $x > 1/y$, then $x^p y^p > 1$, which implies that $1+y^p < x^p y^p + y^p = (1+x^p)y^p$, from which (2.1) follows.

*Received by the editors June 24, 1998; accepted for publication (in revised form) by R. Bhatia September 2, 1998; published electronically January 25, 2000.

http://www.siam.org/journals/simax/21-2/34088.html

†Department of Computing Science, University of Umeå, S-90187 Umeå, Sweden (abrr@cs.umu.se).

If $u \geq v \geq 1$, then $(uv-1)(u-v) \geq 0$, which implies that $v+u^2v \geq u+uv^2$ and

$$\frac{1+u^2}{1+v^2} \geq \frac{u}{v}.$$

This inequality with $u = x^{p/2}$ and $v = y^{p/2}$ gives (2.2). □

3. The proof.

THEOREM 1. *Let $\alpha, \beta, \gamma \in \mathbb{C}$ and $1 \leq p \leq \infty$. Then*

$$\varrho_p(\alpha,\beta) \leq \varrho_p(\alpha,\gamma) + \varrho_p(\gamma,\beta), \tag{3.1}$$

where ϱ_p is defined by (1.1).

Proof. Without loss of generality, we prove (3.1) in the cases with $|\alpha| \leq |\beta|$. The cases with $|\beta| < |\alpha|$ are proved by swapping α and β in the rest of the proof.

In cases with $|\gamma| \leq |\alpha| \leq |\beta|$, we get

$$\begin{aligned}\varrho_p(\alpha,\beta) &= \frac{|\alpha-\beta|}{\sqrt[p]{|\alpha|^p+|\beta|^p}} \leq \frac{|\alpha-\gamma|}{\sqrt[p]{|\alpha|^p+|\beta|^p}} + \frac{|\gamma-\beta|}{\sqrt[p]{|\alpha|^p+|\beta|^p}} \\ &\leq \frac{|\alpha-\gamma|}{\sqrt[p]{|\alpha|^p+|\gamma|^p}} + \frac{|\gamma-\beta|}{\sqrt[p]{|\gamma|^p+|\beta|^p}} = \varrho_p(\alpha,\gamma) + \varrho_p(\gamma,\beta).\end{aligned}$$

In cases with $\alpha = 0$, the inequality (3.1) is trivial.

Next we consider cases with $0 < |\alpha| \leq |\beta| \leq |\gamma|$. In these cases (2.1) gives

$$\frac{\sqrt[p]{|\alpha|^p+|\beta|^p}}{\sqrt[p]{|\alpha|^p+|\gamma|^p}} = \left|\frac{\beta}{\alpha}\right| \frac{\sqrt[p]{1+|\alpha/\beta|^p}}{\sqrt[p]{1+|\gamma/\alpha|^p}} \geq \left|\frac{\beta}{\alpha}\right| \min(|\alpha/\beta|, |\alpha/\gamma|) = \left|\frac{\beta}{\gamma}\right|, \tag{3.2}$$

$$\frac{\sqrt[p]{|\alpha|^p+|\beta|^p}}{\sqrt[p]{|\beta|^p+|\gamma|^p}} = \left|\frac{\alpha}{\beta}\right| \frac{\sqrt[p]{1+|\beta/\alpha|^p}}{\sqrt[p]{1+|\gamma/\beta|^p}} \geq \left|\frac{\alpha}{\beta}\right| \min(|\beta/\alpha|, |\beta/\gamma|) = \left|\frac{\alpha}{\gamma}\right|.$$

We also have

$$\left|\frac{\beta}{\gamma}\right||\alpha-\gamma| + \left|\frac{\alpha}{\gamma}\right||\beta-\gamma| = \left|\frac{\beta\alpha}{\gamma}-\beta\right| + \left|\frac{\beta\alpha}{\gamma}-\alpha\right| \geq |\alpha-\beta|. \tag{3.3}$$

Combining (3.2) and (3.3) gives (3.1).

Finally, we consider cases with $0 < |\alpha| < |\gamma| < |\beta|$. Let $r_1 e^{i\theta_1} = \beta/\alpha$ and $r_2 e^{i\theta_2} = \gamma/\alpha$ be the polar decompositions of β/α and γ/α, respectively. Then (3.1) can be rewritten as

$$\frac{\sqrt{1+r_1^2-2r_1\cos\theta_1}}{\sqrt[p]{1+r_1^p}} \leq \frac{\sqrt{1+r_2^2-2r_2\cos\theta_2}}{\sqrt[p]{1+r_2^p}} + \frac{\sqrt{r_1^2+r_2^2-2r_1r_2\cos(\theta_1-\theta_2)}}{\sqrt[p]{r_1^p+r_2^p}}. \tag{3.4}$$

We now derive some inequalities which can be combined to a proof of (3.4). Since $1 < r_2 < r_1$ and $1 \leq p \leq \infty$, we have

$$\begin{aligned}\frac{\sqrt[p]{r_2^p+r_1^p}-\sqrt[p]{1+r_1^p}}{r_2-1} &= \frac{1}{r_2-1}\int_1^{r_2} \frac{d}{dx}\sqrt[p]{x^p+r_1^p}\,dx = \frac{1}{r_2-1}\int_1^{r_2}\left(\frac{x^p}{x^p+r_2^p}\right)^{1-1/p} dx \\ &< \frac{1}{r_1-r_2}\int_{r_2}^{r_1}\left(\frac{x^p}{x^p+1}\right)^{1-1/p} dx = \frac{\sqrt[p]{r_1^p+1}-\sqrt[p]{r_2^p+1}}{r_1-r_2}.\end{aligned} \tag{3.5}$$

Combining (3.5) and $\sqrt[p]{r_1^p + r_2^p} > \sqrt[p]{1 + r_2^p}$ gives

$$\sqrt[p]{r_1^p + r_2^p}\left(\sqrt[p]{1 + r_1^p} - \sqrt[p]{1 + r_2^p}\right)(r_2 - 1) > \sqrt[p]{1 + r_2^p}\left(\sqrt[p]{r_1^p + r_2^p} - \sqrt[p]{1 + r_1^p}\right)(r_1 - r_2),$$

which can be rewritten as

$$\frac{r_1 - 1}{\sqrt[p]{1 + r_1^p}} < \frac{r_2 - 1}{\sqrt[p]{1 + r_2^p}} + \frac{r_1 - r_2}{\sqrt[p]{r_1^p + r_2^p}}. \tag{3.6}$$

The second inequality (2.2) in Lemma 1 gives

$$\sqrt[p]{\frac{1 + r_1^p}{1 + r_2^p}} > \sqrt{r_1/r_2} \tag{3.7}$$

and

$$\sqrt[p]{\frac{1 + r_1^p}{r_1^p + r_2^p}} = \frac{1}{r_2}\sqrt[p]{\frac{1 + r_1^p}{1 + (r_1/r_2)^p}} > \frac{1}{r_2}\sqrt{r_2} = 1/\sqrt{r_2}. \tag{3.8}$$

From the trigonometric identity $\sin(x + y) = \sin(x)\cos(y) + \sin(y)\cos(x)$, we get

$$|\sin(x + y)| \le |\sin(x)||\cos(y)| + |\sin(y)||\cos(x)| \le |\sin(x)| + |\sin(y)|,$$

which implies that

$$|\sin(\theta_1/2)| \le |\sin(\theta_2/2)| + |\sin((\theta_2 - \theta_1)/2)|. \tag{3.9}$$

From the trigonometric identity $\sin^2(x/2) = \frac{1-\cos(x)}{2}$, we get

$$\sqrt{1 + r_2^2 - 2r_2\cos\theta_2} = \left\| \begin{bmatrix} r_2 - 1 \\ 2\sqrt{r_2}\sin(\theta_2/2) \end{bmatrix} \right\|_2 \tag{3.10}$$

and

$$\sqrt{r_1^2 + r_2^2 - 2r_1r_2\cos(\theta_1 - \theta_2)} = \left\| \begin{bmatrix} r_1 - r_2 \\ 2\sqrt{r_1r_2}\sin((\theta_1 - \theta_2)/2) \end{bmatrix} \right\|_2. \tag{3.11}$$

Combining (3.6), (3.7), (3.8), (3.9), (3.10), and (3.11) gives

$$\begin{aligned}
&\sqrt[p]{1 + r_1^p} \cdot \left(\frac{\sqrt{1 + r_2^2 - 2r_2\cos\theta_2}}{\sqrt[p]{1 + r_2^p}} + \frac{\sqrt{r_1^2 + r_2^2 - 2r_1r_2\cos(\theta_1 - \theta_2)}}{\sqrt[p]{r_1^p + r_2^p}} \right) \\
&\quad = \left\| \begin{bmatrix} \frac{\sqrt[p]{1+r_1^p}}{\sqrt[p]{1+r_2^p}}(r_2 - 1) \\ \frac{2\sqrt[p]{1+r_1^p}}{\sqrt[p]{1+r_2^p}}\sqrt{r_2}\sin(\theta_2/2) \end{bmatrix} \right\|_2 + \left\| \begin{bmatrix} \frac{\sqrt[p]{1+r_1^p}}{\sqrt[p]{r_1^p+r_2^p}}(r_1 - r_2) \\ 2\frac{\sqrt[p]{1+r_1^p}}{\sqrt[p]{r_1^p+r_2^p}}\sqrt{r_1r_2}\sin((\theta_1 - \theta_2)/2) \end{bmatrix} \right\|_2 \\
&\quad \ge \left\| \begin{bmatrix} r_1 - 1 \\ 2\sqrt{r_1}\sin(\theta_1/2) \end{bmatrix} \right\|_2 = \sqrt{1 + r_1^2 - 2r_1\cos\theta_1}.
\end{aligned}$$

That is, we have proved (3.4) and hence also (3.1). □

Acknowledgment. The author would like to thank Ren-Cang Li, who informed him about the paper [1].

REFERENCES

[1] D. Day, *A New Metric on the Complex Numbers*, Sandia Technical Report 98–1754, Sandia National Laboratories, Albuquerque, NM, 1998.
[2] R.-C. Li, *Relative perturbation theory.* I. *Eigenvalue and singular value variations*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 956–982.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 3, pp. 703–723
© 2000 Society for Industrial and Applied Mathematics

GRAPH EMBEDDINGS AND LAPLACIAN EIGENVALUES*

STEPHEN GUATTERY† AND GARY L. MILLER‡

Abstract. Graph embeddings are useful in bounding the smallest nontrivial eigenvalues of Laplacian matrices from below. For an $n \times n$ Laplacian, these embedding methods can be characterized as follows: The lower bound is based on a clique embedding into the underlying graph of the Laplacian. An embedding can be represented by a matrix Γ; the best possible bound based on this embedding is $n/\lambda_{\max}(\Gamma^T\Gamma)$, where $\lambda_{\max}$ indicates the largest eigenvalue of the specified matrix. However, the best bounds produced by embedding techniques are not tight; they can be off by a factor proportional to $\log^2 n$ for some Laplacians.

We show that this gap is a result of the representation of the embedding: By including edge directions in the embedding matrix representation Γ, it is possible to find an embedding such that $\Gamma^T\Gamma$ has eigenvalues that can be put into a one-to-one correspondence with the eigenvalues of the Laplacian. Specifically, if λ is a nonzero eigenvalue of either matrix, then n/λ is an eigenvalue of the other. Simple transformations map the corresponding eigenvectors to each other. The embedding that produces these correspondences has a simple description in electrical terms if the underlying graph of the Laplacian is viewed as a resistive circuit. We also show that a similar technique works for star embeddings when the Laplacian has a zero Dirichlet boundary condition, though the related eigenvalues in this case are reciprocals of each other. In the zero Dirichlet boundary case, the embedding matrix Γ can be used to construct the inverse of the Laplacian. Finally, we connect our results with previous techniques for producing bounds and provide an example.

Key words. Laplacian matrices, graph eigenvalues and eigenvectors, graph embeddings

AMS subject classifications. 05C50, 15A18, 15A42

PII. S0895479897329825

1. Introduction.

In this paper we present an exact relationship between graph embeddings and the eigenvalues and eigenvectors of Laplacian matrices. The study of the connection between Laplacian spectra (particularly with respect to λ_2, the second smallest (and smallest nonzero) eigenvalue of the Laplacian) and properties of the associated graphs dates back to Fiedler's work in the 1970s (see, e.g., [10] and [11]). The Laplacian also has an important role in representing physical problems. It occurs in finite difference, finite element, and control volume representations of problems involving elliptic partial differential equations. These problems often include a Dirichlet boundary condition that specifies that the values at a set of vertices are zero. To represent this condition in the Laplacian, the rows and columns corresponding to the boundary vertices are deleted from the matrix. The resulting matrix is positive definite, and it is the smallest eigenvalue of the matrix that is of interest.

Bounds on the smallest nonzero eigenvalues of both forms of the Laplacian have other important applications. Since the matrices are symmetric, their extreme eigenvalues can be used in computing their condition numbers, which are used in the study

*Received by the editors November 10, 1997; accepted for publication (in revised form) by S. A. Vavasis January 27, 1999; published electronically February 10, 2000.
http://www.siam.org/journals/simax/21-3/32982.html

†Department of Computer Science, Bucknell University, Lewisburg, PA 17837 (guattery@bucknell.edu). The research of this author was supported by the National Aeronautics and Space Administration under NASA contract NAS1-19480 while the author was in residence at the Institute for Computer Applications in Science and Engineering (ICASE), NASA Langley Research Center, Hampton, VA 23681.

‡School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 (glmiller@cs.cmu.edu). The work of this author was supported in part by NSF grant CCR-9505472, DARPA contract N00014-95-1246, and Army contract DAAH04-95-1-0607.

of iterative linear system solvers to estimate rates of convergence [18] and to analyze the quality of preconditioners [4, 13]. Bounds on λ_2 are useful in the analysis of spectral partitioning, both because λ_2 occurs in bounds on cut quality [24] and because they can be used in isolating structural properties of the eigenvectors used in making the cuts [16, 28]. The eigenvalue λ_2 has been related to expansion properties of graphs and can be used in determining if a graph is an expander [1, 2].

One common class of techniques for computing such lower bounds uses properties of graph embeddings [9, 15, 20, 22, 26, 27]. In such methods, a graph H is embedded into the graph G under study; that is, vertices of H are identified with vertices of G, and paths in G are specified to correspond to edges in H. Specific metrics on the embedding such as congestion and dilation are computed; they are then used to calculate lower bounds. H is usually chosen to have some specific structure; most often it is a clique or a star. The bounds produced in general are not tight, however. Kahale [22] has shown that for some graphs, the gap can be a factor of $\log^2 n$, where $n = |V(G)|$, the number of vertices in G.

In this paper, we prove an exact relationship between graph embeddings and Laplacian eigenvalues. In particular, we present a matrix representation of embeddings that differs from the representation used in the methods mentioned above. This representation allows any embedding of H into G to be expressed as an $|E(H)|\times|E(G)|$ matrix Γ, where $|E(H)|$ and $|E(G)|$ are the number of edges in H and G, respectively. We also introduce the current flow embedding, which is based on routing unit currents between pairs of vertices when edge weights are viewed as conductances. We denote the embedding matrix for the current flow matrix as Γ_{cf}. We show that, for a clique embedding based on routing a unit current between each pair of vertices, if the Laplacian $L(G)$ has a nonzero eigenvalue λ with eigenvector $\mathbf{u}$, then $\Gamma_{cf}^T\Gamma_{cf}$ has eigenvalue n/λ; the corresponding eigenvector is $B\mathbf{u}$, where B is the edge-vertex incidence matrix of the Laplacian.

In the zero Dirichlet boundary case, we show that, for a star embedding based on routing a unit current between every vertex and the boundary, if the Laplacian $L(G)$ has an eigenvalue λ with eigenvector $\mathbf{u}$, then $\Gamma_{cf}^T\Gamma_{cf}$ has eigenvalue $1/\lambda$; the corresponding eigenvector is $B\mathbf{u}$, where B is the edge-vertex incidence matrix of the Laplacian. In this case, the Laplacian is positive definite; we show that $\Gamma_{cf}\Gamma_{cf}^T$ is L^{-1}.

In both cases, we show the extension of these results to weighted graphs.

In addition to their utility with respect to the problems listed above, our results are interesting because they provide further illustration of the utility of looking at Laplacian spectra problems with respect to electrical circuits.

The rest of this paper is organized as follows: Section 2 provides an overview of work in this area. Section 3 covers the notation we use and background information on graphs and matrices. Section 4 covers general lemmas about our embedding representation. Section 5 covers the current flow embedding for the clique case. Section 6 covers the current flow embedding for the star case with zero Dirichlet boundary and shows how to construct inverses using embeddings. Section 7 presents related results, including the connection between the results in this paper and previous lower bound techniques.

2. Previous work. The study of the connection between Laplacian spectra (particularly with respect to λ_2) and properties of the associated graphs dates back to Fiedler's work in the 1970s (see, e.g., [10] and [11]). These properties have been used in graph algorithms, particularly algorithms for finding small separators [17, 25, 28].

The relationship between graph embeddings and matrix representations has been

the subject of much interesting research. A large proportion of this work has been aimed at bounding the second largest eigenvalues of time-reversible Markov chains in order to bound the mixing time for random walks. The use of clique embeddings to bound eigenvalues arose in the analysis of mixing times for Markov chains by Jerrum and Sinclair [20, 27]. Further work in this direction was done by Diaconis and Stroock [9] and by Sinclair [26]. Kahale [22] generalized this work in terms of methods assigning lengths to the graph edges and showed that the best bound over all edge length assignments is the largest eigenvalue of the matrix $\Gamma^T\Gamma$, where Γ is a matrix representing the path embedding (Kahale [22] also cited unpublished work along these lines by Fill and Sokal). He also gave a semidefinite programming formulation for a model allowing fractional paths and showed that the bound is off by at most a factor of $\log^2 n$. He showed that this gap is tight; he also noted that the results can be applied to Laplacians with suitable modifications.

Guattery, Leighton, and Miller [15] presented a lower bound technique for λ_2 of a Laplacian. It assigns priorities to paths in the embedding and uses these to compute congestions of edges in the original graph with respect to the embedding. Summing the congestions along the edges in a path gives the path congestion; the lower bound is a function of the reciprocal of the maximum path congestion taken over all paths. For the clique case, they showed that this method is the dual of the method presented in [22]; the best lower bounds produced by these methods are the same. They also showed how to apply their method in the zero Dirichlet boundary case by using star embeddings. In the clique case, they showed that using uniform priorities for any tree T gives a lower bound that is within a factor proportional to the logarithm of the diameter of the tree.

Gremban and Miller [14] have shown how to use embeddings to generate support numbers, which also are bounds on the largest and smallest generalized eigenvalues (and hence the spectral condition number) of preconditioned linear systems involving a generalized definition of Laplacians. They also defined the support tree preconditioner and used the support number bounds to prove properties about the quality of these preconditioners. Gremban, Miller, and Zagha have evaluated the performance of these techniques [13].

The construction of the embedding we use below is related to the effective resistances between pairs of vertices in the graph when the graph is viewed as a network of unit conductances. Chandra et al. [8] have defined the maximum such value taken over all distinct pairs as the electrical resistance of a graph and have used that quantity to bound the commute and cover times of random walks on the graph.

3. Terminology, notation, and background results. We assume that the reader is familiar with the basic definitions of graph theory (in particular, for undirected graphs) and with the basic definitions and results of matrix theory. A graph consists of a set of vertices V and a set of edges E; we denote the vertices (respectively, edges) of a particular graph G as $V(G)$ (respectively, $E(G)$) if there is any ambiguity as to which graph is being referred to. When it is clear which graph we are referring to, we use n to denote $|V|$ and m to denote $|E|$.

We use the term *path graph* for a tree that has exactly two vertices of degree 1 (that is, exactly two vertices with exactly one incident edge). That is, a path graph is a graph consisting of exactly its maximal path. A *star* is a tree with exactly one vertex that is not a leaf. We call the nonleaf vertex the *center* of the star.

3.1. Matrices and matrix notation. We use capital letters to represent matrices and bold lower-case letters for vectors. For a matrix A, a_{ij} or $[A]_{ij}$ represents

the element in row i and column j; for the vector $\mathbf{x}$, x_i or $[\mathbf{x}]_i$ represents the ith entry. The notation $\mathbf{x} = 0$ indicates that all entries of the vector $\mathbf{x}$ are zero; $\mathbf{1}$ indicates the vector that has 1 for every entry. Unless specifically noted otherwise, we index the eigenvalues of an $n \times n$ matrix in nondecreasing order: λ_1 represents the smallest eigenvalue, and λ_n the largest. We use the notation $\lambda_i(A)$ (respectively, $\lambda_i(G)$) to indicate the ith eigenvalue of matrix A (respectively, of the Laplacian of graph G) if there is any ambiguity about which matrix (respectively, graph) the eigenvalue belongs to. The notation $\mathbf{u}_i$ represents the eigenvector corresponding to λ_i.

We use the following notational shorthand: Let D be a diagonal matrix with nonnegative entries on the diagonal. Then $\sqrt{D}$ represents the diagonal matrix with $[\sqrt{D}]_{ii} = \sqrt{d_{ii}}$. If all diagonal entries of D are positive, it is obvious that $\sqrt{D}^{-1} = \sqrt{D^{-1}}$.

I_k denotes the $k \times k$ identity matrix.

For two matrices A and B with the same dimensions, $A \geq B$ if for every entry a_{ij} of A, we have $a_{ij} \geq b_{ij}$.

The following is a standard result (see, e.g., [3] for a version related to Laplacians).

LEMMA 3.1. *Let X be a real matrix. Let $Y = X^T X$ and $Z = XX^T$. If $\mathbf{u}$ is an eigenvector of Y with eigenvalue $\lambda > 0$, then $X\mathbf{u}$ is an eigenvector of Z with eigenvalue λ. Likewise, if $\mathbf{v}$ is an eigenvector of Z with eigenvalue $\mu > 0$, then $X^T\mathbf{v}$ is an eigenvector of Y with eigenvalue μ. Thus Y and Z have the same nonzero eigenvalues.*

Proof. $Z(X\mathbf{u}) = XX^T(X\mathbf{u}) = XY\mathbf{u} = \lambda X\mathbf{u}$, so the first claim holds. Likewise, $Y(X^T\mathbf{v}) = X^T X(X^T\mathbf{v}) = X^T Z\mathbf{v} = \mu X^T\mathbf{v}$, so the second claim holds. Note that Y and Z are positive semidefinite, so any nonzero eigenvalues are positive and the final claim holds. ∎

3.2. The Laplacian matrix representation of a graph. A common matrix representation of graphs is the *Laplacian*. Let D be the matrix with d_{ii} equal to the number of edges incident to vertex v_i for $v_i \in V(G)$ and all off-diagonal entries zero. Let A be the adjacency matrix for G ($[A]_{ij} = 1$ if and only if $(v_i, v_j) \in E(G)$, 0 otherwise). Then the Laplacian of G is the matrix $L = D - A$.

The following are useful facts about the Laplacian:

- The Laplacian is symmetric positive semidefinite, so all its eigenvalues are greater than or equal to 0 (see, e.g., [3, 10]).
- A graph G is connected if and only if 0 is an eigenvalue of the Laplacian of G with multiplicity 1 (see, e.g., [10]).
- The following characterization of λ_2 holds (see, e.g., [10]):

$$\lambda_2 = \min_{\mathbf{x}^T\mathbf{1}=0} \frac{\mathbf{x}^T L\mathbf{x}}{\mathbf{x}^T\mathbf{x}}. \tag{3.1}$$

- For any vector $\mathbf{x}$ and Laplacian matrix L of the graph G, we have (see, e.g., [19]) the following:

$$\mathbf{x}^T L\mathbf{x} = \sum_{(v_i,v_j)\in E(G)} (x_i - x_j)^2. \tag{3.2}$$

An *edge-weighted* graph is a graph for which a real, nonzero weight w_{ij} is associated with each edge (v_i, v_j). Fiedler extended the notion of the Laplacian to graphs with positive edge weights [11]; he referred to this representation as the *generalized*

Laplacian. Let w_{ij} be the (positive) weight of edge (i, j) in graph G. Then the entries of the generalized Laplacian L of G are defined as follows: l_{ii} is the sum of the weights of the edges incident to vertex v_i; for $i \neq j$ and $(v_i, v_j) \in E(G)$, $l_{ij} = -w_{ij}$, and $l_{ij} = 0$ otherwise.

With the exception of (3.2), the properties listed above also apply to generalized Laplacians. A slightly modified version of (3.2) holds for the generalized Laplacian L:

$$\mathbf{x}^T L\mathbf{x} = \sum_{(v_i,v_j)\in E(G)} w_{ij}(x_i - x_j)^2.$$

We usually use L to denote the Laplacian of a graph G, and K to represent the Laplacian of the clique K (whether K refers to the graph or the Laplacian will be clear from context).

3.3. Graph subspaces with respect to the Laplacian. It is well known (see, e.g., [6, 7]) that the Laplacian of a graph G can be expressed as the product $B^T B$ of G's *edge-vertex incidence matrix* B. B is constructed as follows: Arbitrarily, direct the edges of G and index them from 1 to m. B is an $m \times n$ matrix, where row i has a -1 in the column corresponding to the vertex at its tail, a 1 in the column corresponding to the vertex at its head, and zeros in all other columns.

Generalized Laplacians can also be expressed as a product involving the edge-vertex incidence matrix B. However, we must introduce an edge weight matrix W (also referred to as the *conductance matrix*). W is a diagonal matrix with entry $[W]_{ii}$ equal to the weight of edge i. It is easy to check that the generalized Laplacian $L = B^T W B$ and that L is independent of the directions of the edges in B.

It is common to think of the set of vectors of length n as the *vertex space*; each entry of the vector $\mathbf{x}$ assigns the value x_i to corresponding vertex i (such an assignment of values is often referred to as a *valuation*). Likewise, the set of all vectors of length m is the *edge space*. Since we will use an electrical analogy below, we sometimes refer to the values assigned to vertices as *potentials*. Note that multiplying a vector of potentials on the left by B produces a vector of *potential differences* in the edge space: each entry in the result is the difference in potentials at the head and tail of some edge in G.

The edge space of the graph can be partitioned into two orthogonal subspaces (see, e.g., [5, 7]). The first is the *cycle space*, the subspace spanned by vectors representing cycles in G. A cycle C is represented by a vector as follows: The edges of C have a natural direction, where each edge is directed toward its successor in the cyclic order. For each edge e in C, the vector $\mathbf{v}_C$ has entry 1 if the direction of e with respect to C and the direction of e used in defining the edge-vertex incidence matrix B are the same, and -1 if they are opposite. The entries for edges not in C are 0.

The cycle space is the orthogonal complement of the *cocycle* or *cut space*, the subspace spanned by vectors representing cuts. Let $(S, \overline{S})$ represent a partition of the vertices of a graph G into two sets. The set of edges with one end in S and the other in $\overline{S}$ is called a *cut*. Cuts are represented as vectors as follows: for each edge in the cut, if the edge has its tail in S (i.e., the edge is directed (according to B) from S to $\overline{S}$), the vector entry for that edge is 1. If the edge has its head in S (i.e., the edge is directed from $\overline{S}$ to S), the vector entry for that edge is -1. Entries for edges not in the cut are 0.

It is well known that, for a connected graph G with m edges on n vertices, the dimension of the cycle space is $m - n + 1$ and the dimension of the cut space is $n - 1$ (see, e.g., [7, Theorem 5, Section II.3]). Some simple but useful results about the

relation of the cycle space and cut spaces to the edge-vertex incidence matrix B are stated in the following lemma (the proofs are easy and left to the reader).

LEMMA 3.2. *Let G be a connected graph with edge-vertex incidence matrix B. Then*

- *the span of the columns of B is the cut space of G, and any $n-1$ columns of B form a basis for the cut space;*
- *the cycle space of G is the null space of B^T.*

3.4. Laplacians and electrical circuits. In this paper we consider the graph of a Laplacian as a resistive electrical circuit, with the edge weights representing conductances between the vertices or nodes (conductance is the reciprocal quantity to resistance). We also calculate quantities such as the current flows on the edges of the graph when unit currents are injected at one point and removed at others. These calculations can be defined in terms of three electrical laws.

The first of these is *Ohm's law*, which connects potential differences and currents. We use it mainly in defining *currents*: As noted above, we can consider any valuation of the vertices as a vector of potentials or *voltages*. For any pair of vertices connected by an edge, we define the *potential difference* across the edge as the value at the head of the edge minus the value at the tail. Thus, given a vector $\mathbf{v}$ of voltages, $B\mathbf{v}$ gives the vector of potential differences, which is in the edge space. Ohm's law says that the current on an edge is the potential difference times the conductance. Thus, for $\mathbf{v}$ we can define the vector $\mathbf{i}$ of currents as $\mathbf{i} = WB\mathbf{v}$, where W is the conductance matrix as discussed in section 3.3.

The second is *Kirchoff's voltage law (KVL)*, which states that the potential drops around any cycle in the graph sum to zero. This requires that for any cycle $\mathbf{c}$, $\mathbf{c}^T B\mathbf{v} = 0$. The second point in Lemma 3.2 shows that this holds for our representation, which is thus consistent with KVL.

The third law is *Kirchoff's current law (KCL)*, which states that the net current flow at any vertex is zero; that is, the sum of the currents over all edges incident to a vertex is the negative of the current injected at that vertex. Note that multiplying a vector of currents (in the edge space) by B^T sums the currents at the vertices. Therefore, if $\mathbf{i}_{\text{ext}}$ is a vector of external currents injected into the circuit, KCL says that the resulting currents in the circuit (represented by the vector $\mathbf{i}$) must obey

$$B^T\mathbf{i} = \mathbf{i}_{\text{ext}}. \tag{3.3}$$

We frequently use KCL as a requirement in calculating flows in the circuit given a unit current injected at one vertex and extracted at another. This is equivalent to solving for $\mathbf{i}$ in (3.3). Note that for a solution to exist, the amount of current injected and extracted in $\mathbf{i}_{\text{ext}}$ must be equal; we must have $\mathbf{i}_{\text{ext}}^T\mathbf{1} = 0$. This is equivalent to the condition that $\mathbf{i}_{\text{ext}}$ be in the column space of B^T, or that it is orthogonal to the null space of B, which is spanned by $\mathbf{1}$ when the graph is connected.

3.5. Graph embeddings. Let G and H be connected graphs such that the vertex set of H is a subset of the vertex set of G. An *embedding* of H into G is a collection Γ of path subgraphs of G such that for each edge $(v_i, v_j) \in E(H)$, the embedding contains a path γ_{ij} from v_i to v_j in G. For full generality, we will allow fractional paths in our embeddings; i.e., an edge $(v_i, v_j) \in E(H)$ can be associated with a finite collection of paths from v_i to v_j in G; each such path has a positive fractional weight associated with it such that the weights add up to 1. If a path γ includes edge e, we say that γ is incident to e.

4. Representing graph embeddings with matrices. Let G be a connected graph, and let Γ be an embedding of H into G. Consider the following observation: Each vertex appears at most once on any path $\gamma \in \Gamma$. Therefore, we can choose one end of the embedded edge $e = (u, v) \in E(H)$ as a source node (say, u) and the other (say, v) as the sink node and order path edges in sequence from u to v. A path (or set of fractional paths) for e in the embedding can thus be viewed as a unit flow from u to v, where the net flow into any intermediate vertex is zero.

These flows can be represented as vectors in the edge space. For the purposes of our representation, we assume that directions have been assigned arbitrarily to the edges of G, and that the vertices and edges of G have been indexed from 1 to n and from 1 to m, respectively. Let $e = (u, v)$ be an edge in H; e is represented in the embedding as a unit flow in G. In the vector for a flow, each edge gets a value with magnitude equal to the amount of the flow it carries. The entry for each edge in the flow also has a sign: positive if the flow direction agrees with the edge direction, and negative if the flow is opposite the edge direction.

The flow vectors can be assembled into a matrix representing an embedding. We will use the symbol Γ to represent the embedding matrix as well as the embedding; the use should be clear from context. The matrix Γ includes one row for each edge e in H; that row is the flow vector for the embedding of e into G. Thus Γ is a $|E(H)| \times |E(G)|$ matrix. Note that Γ depends on both G and H; these graphs will be clear from context, so we will not introduce any notation to express this dependency. This matrix representation is similar to the representation presented in [21], though it differs by including negative entries. We discuss the relationship between these two representations in section 7 below.

The following result holds for arbitrary H embedded into connected graph G.

LEMMA 4.1. *For any embedding matrix Γ representing an embedding of H into G, the product $\Gamma B(G) = B(H)$, where the vertex set of $B(H)$ is $V(G)$.*

Proof. Consider the result of multiplying a row of Γ and a column of $B(G)$. The column of $B(G)$ represents a vertex v of G. It is easy to see from the construction of the rows of Γ that this product gives the net flow out of v. We thus have the following:

- If v is the source of the flow, then the product is -1.
- If v is the sink of the flow, then the product is 1.
- Otherwise, the value is 0.

Thus, the row of $\Gamma B(G)$ corresponding to the flow for edge (u, v) has a -1 in the column for u, a 1 in the column for v, and zeros elsewhere. This is the same as the row for (u, v) in $B(H)$, where the edge directions correspond to the directions of the flows used in the embedding. ∎

It is useful to define the matrix $M = \Gamma^T \Gamma$. M is $|E(G)| \times |E(G)|$.

The results in the rest of the paper are based on a special embedding we will refer to as the *current flow embedding*. This embedding is defined for any generalized Laplacian as follows: The graph G serves as a network of conductances, where the conductance of each edge is its (positive) weight. For each edge (u, v) in H, the flow from u to v is the same as the current flow in G when a unit current is injected at u and extracted at v. Since the currents obey Kirchoff's laws, the set of flows clearly forms an embedding of H into G. We note that this embedding can be extended to cases with a zero Dirichlet boundary condition by adding in a ground node; details of this extension are deferred to section 6 below.

We denote this embedding Γ_{cf}. M_{cf} denotes the matrix M for the embedding matrix Γ_{cf}.

In the remainder of this paper, the following results about Γ_{cf} prove useful.

LEMMA 4.2. *Let G be an unweighted connected graph, and let H be a graph on $V(G)$ with a nonempty edge set. Let Γ_{cf} be as defined above for the embedding of H into G. Then the cycle space of G is in the null space of Γ_{cf}.*

Proof. This lemma is a consequence of each flow in the embedding obeying Kirchoff's laws.

Note that each row of Γ_{cf} is the transpose of a current vector $\mathbf{i}$. Because each such current vector is consistent with KVL and KCL, there exists a potential vector $\mathbf{v}$ (which is not unique) such that $\mathbf{i} = WB\mathbf{v}$ by Ohm's law. To see this, note that the external current flow vector $\mathbf{i}_{\text{ext}}$ used to define Γ_{cf} injects a unit current at one vertex and extracts it at another. Thus $\mathbf{i}_{\text{ext}}$ is orthogonal to the null space of the Laplacian (the vector of all ones), and there is a vector of potentials $\mathbf{v}$ such that $\mathbf{i}_{\text{ext}} = L\mathbf{v} = B^T WB\mathbf{v}$. However, by KCL, $\mathbf{i}_{\text{ext}} = B^T\mathbf{i}$, so either $\mathbf{i} = WB\mathbf{v}$ or $\mathbf{i} = WB\mathbf{v} + \mathbf{c}$, where $\mathbf{c}$ is a nonzero vector in the null space of B^T, which is the cycle space of the graph (Lemma 3.2). Assume the latter. The potential drops for the currents in $\mathbf{i}$ can be written as $W^{-1}\mathbf{i}$; since $\mathbf{c}$ is in the cycle space, KVL implies that

$$0 = \mathbf{c}^T W^{-1}\mathbf{i} = \mathbf{c}^T W^{-1}\left(WB\mathbf{v} + \mathbf{c}\right) = \mathbf{c}^T W^{-1}\mathbf{c}.$$

Since W is a diagonal matrix with positive diagonal entries, the last term cannot be 0, so we must have that $\mathbf{i} = WB\mathbf{v}$, which demonstrates the claim.

For an unweighted graph, $W = I_m$, so this simplifies to $\mathbf{i} = B\mathbf{v}$. As already noted, the cycle space is the null space of B^T and, hence, the left null space of B, so for any vector $\mathbf{c}$ in the cycle space, $\mathbf{c}^T\mathbf{i} = \mathbf{c}^T B\mathbf{v} = 0$. This holds for every row in Γ_{cf}, and hence Lemma 4.2 holds. □

A similar result holds if G is a weighted graph.

LEMMA 4.3. *Let G be a weighted connected graph with conductance matrix W, and let H be a graph on $V(G)$ with a nonempty edge set. Let Γ_{cf} be as defined above for the embedding of H into G. Then for any vector $\mathbf{c}$ in the cycle space of G, $W^{-1}\mathbf{c}$ is in the null space of Γ_{cf}.*

Proof. The proof is the same as that for Lemma 4.2, except that W is no longer the identity matrix. However, the vector $\mathbf{c}$ is now scaled by W^{-1}, and the reader can check that the terms W and W^{-1} cancel each other in the argument. □

This result can be stated in slightly altered form; it is no longer consistent with the electrical analogy, but it is useful in the algebraic arguments below.

COROLLARY 4.4. *Let G be a weighted connected graph with conductance matrix W, and let H be a graph on $V(G)$ with a nonempty edge set. Let Γ_{cf} be as defined above for the embedding of H into G. Then for any vector $\mathbf{c}$ in the cycle space of G, $\sqrt{W^{-1}}\mathbf{c}$ is in the null space of $\Gamma_{cf}\sqrt{W^{-1}}$.*

Proof. Since W^{-1} is a diagonal matrix with positive real entries on the diagonal, we can factor it into $\sqrt{W^{-1}} \cdot \sqrt{W^{-1}}$. The result follows from Lemma 4.3. □

We next consider the special properties of the current flow embedding with respect to embedding particular graphs into G.

5. The clique embedding Γ_{cf}. We now show that there is an exact connection between clique embeddings and Laplacian eigenvalues. Let G be a connected graph on n vertices with positive edge weights and Laplacian $L = B^T WB$, where B is the edge-vertex incidence matrix and W is the conductance matrix as defined in section 3.3. Let Γ_{cf} be the embedding matrix for the current flow embedding of the complete graph into G, with M_{cf} defined as in the previous section. For clarity's sake, we start with the result in the case where G is unweighted (that is, W is the identity matrix).

THEOREM 5.1. *Let G be an unweighted connected graph on n vertices with Laplacian L. Let $M_{cf} = \Gamma_{cf}^T \Gamma_{cf}$, where Γ_{cf} is the current flow embedding of the complete graph on n vertices into G. For any $\lambda > 0$, λ is an eigenvalue of L if and only if $\frac{n}{\lambda}$ is an eigenvalue of M_{cf}. Further, if λ has eigenvector $\mathbf{u}$ for L, then $B\mathbf{u}$ is an eigenvector of M_{cf} with eigenvalue $\frac{n}{\lambda}$.*

Proof. By Lemma 4.1, $B^T \Gamma_{cf}^T \Gamma_{cf} B = K$, where K is the Laplacian of the complete graph on $n = |V(G)|$. It is easy to show that, for any vector $\mathbf{x}$ such that $\mathbf{x}^T \mathbf{1} = 0$, $K\mathbf{x} = n\mathbf{x}$. Thus for any such $\mathbf{x}$, we have the following:

$$B^T M_{cf} B\mathbf{x} = B^T \Gamma_{cf}^T \Gamma_{cf} B\mathbf{x} = n\mathbf{x}. \tag{5.1}$$

Since K has rank $n-1$, the preceding result implies that the ranks of M_{cf}, Γ_{cf}, and Γ_{cf}^T are all at least $n-1$.

Lemma 4.2 shows that the null space of Γ_{cf} contains the cycle space of G. The dimension of the null space is thus at least $m - n + 1$; combining this with the result from the previous paragraph shows that the null space is exactly the cycle space and that the rank of Γ_{cf} (and hence of Γ_{cf}^T) is $n-1$. The null space of M_{cf} contains the null space of Γ_{cf}, and the same results hold for its rank and null space. M_{cf} is clearly symmetric positive semidefinite from its definition.

Let $\mathbf{x}$ be any vector in the cut space of G. Since M_{cf} is symmetric with the cycle space of G as its null space, $M_{cf}\mathbf{x}$ is orthogonal to the cycle space.

Let $\mathbf{u}$ be an eigenvector of L with eigenvalue $\lambda > 0$. Then

$$B^T M_{cf} B\mathbf{u} = n\mathbf{u} = \frac{n}{\lambda} L\mathbf{u} = B^T B \frac{n\mathbf{u}}{\lambda}.$$

Thus

$$B^T \left(M_{cf} B\mathbf{u} - \frac{n}{\lambda} B\mathbf{u} \right) = 0,$$

which implies that $M_{cf} B\mathbf{u} - \frac{n}{\lambda} B\mathbf{u}$ is in the null space of B^T. By Lemma 3.2, this means it is in the cycle space of G. But we showed above that $M_{cf} B\mathbf{u}$ is orthogonal to the cycle space. Further, Lemma 3.2 also implies that $B\mathbf{u}$ lies in the cut space of G and is therefore orthogonal to the cycle space. Thus we must have that $M_{cf} B\mathbf{u} = \frac{n}{\lambda} B\mathbf{u}$. Thus $B\mathbf{u}$ is an eigenvector of M_{cf} with eigenvalue $\frac{n}{\lambda}$.

We still need to show that M_{cf} does not have any other nonzero eigenvalues. Let $\mathbf{u}$ be any eigenvector of M_{cf} with eigenvalue μ. Since the null space of M_{cf} is the cycle space of G, $\mathbf{u}$ is in the cut space. By Lemma 3.2, this is in the column space of B, and we can find a vector $\mathbf{x}$ such that $B\mathbf{x} = \mathbf{u}$; further, we can choose $\mathbf{x}$ such that $\mathbf{x}^T \mathbf{1} = 0$. Then

$$\mu\mathbf{u} = M_{cf}\mathbf{u} = M_{cf} B\mathbf{x} = \mu B\mathbf{x}. \tag{5.2}$$

Since $\mathbf{x}$ is orthogonal to $\mathbf{1}$, we can use (5.1) with (5.2) as follows:

$$n\mathbf{x} = B^T M_{cf} B\mathbf{x} = \mu B^T B\mathbf{x} = \mu L\mathbf{x}.$$

Thus $\mathbf{x}$ is an eigenvector of L with eigenvalue $\lambda = \frac{n}{\mu}$. That is, for any eigenvalue μ of M_{cf}, there is an eigenvalue λ of L such that $\mu = \frac{n}{\lambda}$, and the correspondence between eigenvalues is consistent with the theorem statement. ∎

Note that if $\mathbf{v}$ is an eigenvector corresponding to a nonzero eigenvalue of M_{cf}, $B^T\mathbf{v}$ is an eigenvector of L.

5.1. Extending the results to the weighted case. Extending the results about M_{cf} to the weighted case is straightforward. We now need to take into account the matrix W in the calculations; otherwise, the argument is similar.

THEOREM 5.2. *Let G be a connected graph on n vertices with positive edge weights, conductance matrix W, and generalized Laplacian L. Let $M_{cf} = \Gamma_{cf}^T \Gamma_{cf}$, where Γ_{cf} is the current flow embedding of the complete graph on n vertices into G. For any $\lambda > 0$, λ is an eigenvalue of L if and only if $\frac{n}{\lambda}$ is an eigenvalue of $M_{cf}W^{-1}$. Further, if λ has eigenvector $\mathbf{u}$ for L, then $WB\mathbf{u}$ is an eigenvector of $M_{cf}W^{-1}$ with eigenvalue $\frac{n}{\lambda}$.*

Proof. The theorem is stated so as to emphasize the connection with the electrical interpretation; we will actually prove the result for the similar matrix

$$\sqrt{W^{-1}} M_{cf} \sqrt{W^{-1}} = \sqrt{W^{-1}} \left(M_{cf} W^{-1} \right) \sqrt{W}.$$

This similarity transform corresponds to a change of variable in the theorem statement. We will actually prove that for any eigenvector $\mathbf{u}$ of L with eigenvalue $\lambda > 0$, $\sqrt{W}B\mathbf{u}$ is an eigenvector of $\sqrt{W^{-1}} M_{cf} \sqrt{W^{-1}}$ with eigenvalue $\frac{n}{\lambda}$. We will also show that $\sqrt{W^{-1}} M_{cf} \sqrt{W^{-1}}$ has no other eigenvalues. It is easy to verify that these results imply the theorem.

We again use Lemma 4.1 to show that

$$B^T \sqrt{W} \left(\sqrt{W^{-1}} M_{cf} \sqrt{W^{-1}} \right) \sqrt{W} B = B^T M_{cf} B = B^T \Gamma_{cf}^T \Gamma_{cf} B = K,$$

and hence that for any vector $\mathbf{x}$ such that $\mathbf{x}^T \mathbf{1} = 0$,

$$B^T \sqrt{W} \left(\sqrt{W^{-1}} M_{cf} \sqrt{W^{-1}} \right) \sqrt{W} B \mathbf{x} = n\mathbf{x}.$$

Since K has rank $n-1$, this result implies that the ranks of $\sqrt{W^{-1}} M_{cf} \sqrt{W^{-1}}$, $\Gamma_{cf} \sqrt{W^{-1}}$, and $\sqrt{W^{-1}} \Gamma_{cf}^T$ are all at least $n-1$.

By Corollary 4.4, the null space of $\Gamma_{cf} \sqrt{W^{-1}}$ contains the cycle space of G multiplied by $\sqrt{W^{-1}}$. Because $\sqrt{W^{-1}}$ is an invertible matrix, the dimension of the null space is thus at least $m - n + 1$. Combining this with the result from the previous paragraph shows that this is exactly the null space and that the ranks of $\Gamma_{cf} \sqrt{W^{-1}}$, $\sqrt{W^{-1}} \Gamma_{cf}^T$, and $\sqrt{W^{-1}} M_{cf} \sqrt{W^{-1}}$ are $n-1$.

Through the rest of this proof, we will use the notation C_0 to denote $\sqrt{W^{-1}}$ times the cycle space of G; that is, the null space of $\sqrt{W^{-1}} M_{cf} \sqrt{W^{-1}}$ is C_0. Using Lemma 3.2 and the fact that W is invertible, it is easy to show that C_0 is the null space of $B^T \sqrt{W}$ and is orthogonal to the column space of $\sqrt{W} B$.

Because $\sqrt{W^{-1}} M_{cf} \sqrt{W^{-1}}$ is symmetric with C_0 as its null space, the matrix $\sqrt{W^{-1}} M_{cf} \sqrt{W^{-1}} \mathbf{x}$ is orthogonal to C_0 for any vector $\mathbf{x}$ in the edge space.

Let $\mathbf{u}$ be an eigenvector of L with eigenvalue $\lambda > 0$. Then

$$B^T \sqrt{W} \left(\sqrt{W^{-1}} M_{cf} \sqrt{W^{-1}} \right) \sqrt{W} B \mathbf{u} = B^T M_{cf} B \mathbf{u} = n\mathbf{u} = \frac{n}{\lambda} L \mathbf{u} = B^T W B \frac{n\mathbf{u}}{\lambda}.$$

Thus

$$B^T \sqrt{W} \left(\sqrt{W^{-1}} M_{cf} \sqrt{W^{-1}} (\sqrt{W} B \mathbf{u}) - \frac{n}{\lambda} (\sqrt{W} B \mathbf{u}) \right) = 0,$$

which implies that $\sqrt{W^{-1}} M_{cf} \sqrt{W^{-1}} (\sqrt{W} B \mathbf{u}) - \frac{n}{\lambda} (\sqrt{W} B \mathbf{u})$ is in C_0, the null space of $B^T \sqrt{W}$. But we showed above that the first product is orthogonal to C_0. Further,

the second product is in the column space of $\sqrt{W}B$, which we noted above is orthogonal to C_0. Thus the two products must be equal, and $\sqrt{W}B\mathbf{u}$ is an eigenvector of $\sqrt{W^{-1}}M_{cf}\sqrt{W^{-1}}$ with eigenvalue $\frac{n}{\lambda}$.

We still need to show that $\sqrt{W^{-1}}M_{cf}\sqrt{W^{-1}}$ does not have any other nonzero eigenvalues. Define C_1 to be the subspace that results from transforming the cut space of G by multiplying it by $\sqrt{W}$. Since $\sqrt{W}$ is invertible, C_1 has dimension $n-1$; we noted above that it is orthogonal to C_0. Because (the symmetric matrix) $\sqrt{W^{-1}}M_{cf}\sqrt{W^{-1}}$ has null space C_0, any eigenvector of a nonzero eigenvalue must come from C_1.

Let $\mathbf{u} \in C_1$ be any eigenvector of $\sqrt{W^{-1}}M_{cf}\sqrt{W^{-1}}$ with eigenvalue $\mu > 0$. By the definition of C_1, there is a vector $\mathbf{z}$ in the cut space of G such that $\mathbf{z} = \sqrt{W^{-1}}\mathbf{u}$. We showed in the proof of Theorem 5.1 that for any vector $\mathbf{z}$ in the cut space, there is a vector $\mathbf{x}$ in the vertex space such that $B\mathbf{x} = \mathbf{z}$ and $\mathbf{x}^T\mathbf{1} = 0$. Thus we can write $\mathbf{u} = \sqrt{W}B\mathbf{x}$ for the appropriate such $\mathbf{x}$.

Then

$$\mu\mathbf{u} = \sqrt{W^{-1}}M_{cf}\sqrt{W^{-1}}\mathbf{u} = \sqrt{W^{-1}}M_{cf}\sqrt{W^{-1}}(\sqrt{W}B\mathbf{x}) = \mu\sqrt{W}B\mathbf{x}. \tag{5.3}$$

Since $\mathbf{x}$ is orthogonal to $\mathbf{1}$, we can use (5.1) with (5.3) as follows:

$$n\mathbf{x} = B^T\sqrt{W}\left(\sqrt{W^{-1}}M_{cf}\sqrt{W^{-1}}\right)\sqrt{W}B\mathbf{x} = \mu B^T W B\mathbf{x} = \mu L\mathbf{x}.$$

Thus $\mathbf{x}$ is an eigenvector of L with eigenvalue $\lambda = \frac{n}{\mu}$. That is, for any eigenvalue μ of $\sqrt{W^{-1}}M_{cf}\sqrt{W^{-1}}$, there is an eigenvalue λ of L such that $\mu = \frac{n}{\lambda}$, and the correspondence between eigenvalues is consistent with the theorem statement. ∎

6. The star embedding $\boldsymbol{\Gamma_{cf}}$. We now turn to the case of a graph with a zero Dirichlet boundary condition. Such a graph has a set S_b of boundary vertices; the values of these vertices are constrained to be zero. We wish to solve the problem of determining the set of valuations on the nonboundary (interior) vertices that have the property that the values are scaled by a constant when the Laplacian operator is applied to the valuation.

This problem can be described in matrix terms as follows. Let G be a connected graph with boundary set S_b. Without loss of generality, we can assume that G has no edges between pairs of boundary vertices; such edges can be deleted without changing the solution of the problem. In this section, we will use n to denote the number of interior vertices of G, and m to denote the number of edges of G. Let B be the edge-vertex incidence graph subject to the conditions that all edges between boundary and interior vertices are directed away from the boundary and that the boundary vertices are numbered last. Define $\hat{B}$ as the restricted version of B with all columns corresponding to vertices in S_b deleted. The Laplacian $\hat{L}$ is defined as $\hat{B}^T\hat{B}$. The problem stated in the previous paragraph is equivalent to finding the eigenvalues and eigenvectors of $\hat{L}$.

If the edges of G have positive weights assigned to them, we represent the weights in a conductance matrix W. In the weighted case, $\hat{L} = \hat{B}^T W\hat{B}$.

Note that once the columns of B have been deleted to form $\hat{B}$, the problem does not depend on the number of boundary vertices. For edges incident to the boundary, it is no longer possible to distinguish to which boundary vertex they connect. Because of this, it is useful to think of the problem in terms of a graph G_0 that consists of the interior vertices of G plus a single boundary vertex v_0. The edges of G_0 include

all edges between interior vertices of G, plus an edge (v_i, v_0) for every edge between an interior vertex v_i and some boundary vertex. (If the graph is weighted, the edge in G and the corresponding edge in G_0 have the same weight.) G_0 has m edges and $n+1$ vertices. Thus the size of its cycle space is $m-n$ and the size of its cut space is n. It is easy to check that $\hat{L}(G_0) = \hat{L}(G)$ and $\hat{B}(G_0) = \hat{B}(G)$, so we do not need to distinguish these matrices. Thus working with G_0 does not change the solution we are looking for.

Observe that G_0 might have multiple edges between some vertex v_i and the boundary vertex. There is nothing in our definitions that makes this a problem. Such edges form cycles between v_i and v_0. (We note in passing that a graph similar in structure to G_0 can be defined to eliminate such multiple edges by combining them into a single edge with a weight equal to the sum of the weights of the combined edges. This leads to the same result, but the change in the size of the graph complicates the analysis.)

In this section we will consider embeddings into G based on star embeddings into G_0. In mapping a star to G_0, G_0's single boundary vertex is mapped to the center of the star, and every other vertex of G_0 is mapped to a leaf. The embedding itself consists of a path (or set of fractional paths with weights summing to 1) from each interior vertex to the boundary vertex.

Any star embedding into G_0 has a corresponding embedding into G. The only difference is that the flow on edges between the interior and the boundary terminates at different boundary nodes if G has vertices with multiple edges to the boundary. The flow on each edge is the same, however. We call such an embedding a *star embedding into G*.

The path (or set of paths) for each interior vertex can be represented in a matrix Γ as described in section 4. Since the flow on each edge is the same, Γ is the same for both G and G_0 for any embedding. Note that each row of Γ corresponds to a vertex in the interior; we require that the ordering of the rows in Γ corresponds to the ordering of the vertices in $\hat{B}$.

For the rest of the section, we will assume that G is a connected graph with zero boundary S_b, n interior vertices, and m edges. We also assume that $\hat{B}$ and $\hat{L}$ have been defined consistent with the restrictions specified above and that any embedding matrix Γ has its row order consistent with the vertex order.

The following lemma is analogous to Lemma 4.1.

LEMMA 6.1. *For any star embedding into G,* $\Gamma\hat{B} = I_n$.

Proof. We work with the graph G_0, since that is sufficient to prove the result, and since G_0 maps the center of the star to a single vertex. By Lemma 4.1, $\Gamma B = B_*$ is the edge-vertex incidence matrix for the star. The conditions on ordering ensure that edge i of the star is an edge from the boundary to vertex i. Thus, $[B_*]_{ii} = 1$. The only other entries in row i are in columns associated with boundary vertices. Dropping the boundary columns to produce $\hat{B}$ results in the dropping of the corresponding columns from the product. ∎

The following corollary is useful in the proofs below.

COROLLARY 6.2. *For any star embedding into G,* $\hat{B}^T\Gamma^T\Gamma\hat{B} = I_n$.

Proof. The result is obvious since

$$\hat{B}^T\Gamma^T = \left(\Gamma\hat{B}\right)^T = I_n^T. \qquad \blacksquare$$

The current flow embedding Γ_{cf} of the star is defined as per section 4. In electrical terms, the boundary vertices serve as grounds. Therefore, these vertices are combined

into a single ground node in the electrical circuit. Each current flow is computed with respect to a unit current injected at an interior vertex and extracted at the ground node. Note that the current flows for star embeddings into G_0 and G are the same.

The following theorem gives the connection between Γ_{cf} and the eigenvalues and eigenvectors of $\hat{L}$.

THEOREM 6.3. *Let $M_{cf} = \Gamma_{cf}^T \Gamma_{cf}$. For any $\lambda > 0$, λ is an eigenvalue of $\hat{L}$ if and only if $\frac{1}{\lambda}$ is an eigenvalue of $M_{cf} W^{-1}$. Further, if λ has eigenvector $\mathbf{u}$ for $\hat{L}$, then $W\hat{B}\mathbf{u}$ is an eigenvector of $M_{cf}W^{-1}$ with eigenvalue $\frac{1}{\lambda}$.*

Proof. Because $\hat{L} = (\sqrt{W}B)^T(\sqrt{W}B)$ and W is diagonal with positive entries on the diagonal, all nonzero eigenvalues of $\hat{L}$ are positive, so the theorem applies to all nonzero eigenvalues of $\hat{L}$.

The argument is essentially the same as that for Theorem 5.2. As for that theorem, the theorem statement emphasizes the connection with the electrical interpretation; we will actually prove the result for the similar matrix

$$\sqrt{W^{-1}} M_{cf} \sqrt{W^{-1}} = \sqrt{W^{-1}} \left(M_{cf} W^{-1} \right) \sqrt{W}.$$

This similarity transform corresponds to a change of variable. We will actually prove that for any eigenvector $\mathbf{u}$ of $\hat{L}$ with eigenvalue $\lambda > 0$, $\sqrt{W}\hat{B}\mathbf{u}$ is an eigenvector of $\sqrt{W^{-1}} M_{cf} \sqrt{W^{-1}}$ with eigenvalue $\frac{1}{\lambda}$. We will also show that $\sqrt{W^{-1}} M_{cf} \sqrt{W^{-1}}$ has no other nonzero eigenvalues. It is easy to verify that these results imply the theorem.

By Corollary 6.2,

$$\hat{B}^T \sqrt{W} \left(\sqrt{W^{-1}} M_{cf} \sqrt{W^{-1}} \right) \sqrt{W} \hat{B} = \hat{B}^T M_{cf} \hat{B} = \hat{B}^T \Gamma_{cf}^T \Gamma_{cf} \hat{B} = I_n,$$

and hence that for any vector $\mathbf{x}$ in the vertex space,

$$\hat{B}^T \sqrt{W} \left(\sqrt{W^{-1}} M_{cf} \sqrt{W^{-1}} \right) \sqrt{W} \hat{B} \mathbf{x} = \mathbf{x}. \tag{6.1}$$

Since I_n has rank n, this result implies that the ranks of $\sqrt{W^{-1}} M_{cf} \sqrt{W^{-1}}$, $\Gamma_{cf}\sqrt{W^{-1}}$, and $\sqrt{W^{-1}}\Gamma_{cf}^T$ are all at least n. Since n is one of the dimensions of Γ_{cf}, it is exactly the rank of $\Gamma_{cf}\sqrt{W^{-1}}$ and $\sqrt{W^{-1}}\Gamma_{cf}^T$. $\sqrt{W^{-1}} M_{cf} \sqrt{W^{-1}}$ is the product of these matrices and thus has rank less than or equal to n; this implies that it too has rank exactly n.

A similar argument shows that $\sqrt{W}\hat{B}$ has rank n.

We can apply Lemma 4.3 and Corollary 4.4 to G_0: For any vector $\mathbf{x}$ in the cycle space of G_0, $W^{-1}\mathbf{x}$ is in the null space of Γ_{cf}. Further, for any vector $\mathbf{x}$ in the cycle space of G_0, $\sqrt{W^{-1}}\mathbf{x}$ is in the null space of $\Gamma_{cf}\sqrt{W^{-1}}$.

Because $\sqrt{W^{-1}}$ is an invertible matrix, the dimension of the null space is thus at least $m - n$. Combining this with the result from the previous paragraph shows that this is exactly the null space. Through the rest of this proof, we will use the notation C_0 to denote the the null space of $\Gamma_{cf}\sqrt{W^{-1}}$ and of $\sqrt{W^{-1}} M_{cf} \sqrt{W^{-1}}$.

Using Lemma 3.2 and the fact that W is invertible, it is easy to show that C_0 is the null space of $B^T\sqrt{W}$, and is orthogonal to the column space of $\sqrt{W}B$. Since $\hat{B}$ is derived from B by dropping a column (with respect to G_0), and since $\sqrt{W}\hat{B}$ is an $m \times n$ matrix with rank n, we immediately have that C_0 is the null space of $\hat{B}^T\sqrt{W}$ and is orthogonal to the column space of $\sqrt{W}\hat{B}$.

Because $\sqrt{W^{-1}} M_{cf} \sqrt{W^{-1}}$ is a symmetric matrix with C_0 as its null space, $\sqrt{W^{-1}} M_{cf} \sqrt{W^{-1}}\mathbf{x}$ is orthogonal to C_0 for any vector $\mathbf{x}$ in the edge space.

Let $\mathbf{u}$ be an eigenvector of $\hat{L}$ with eigenvalue $\lambda > 0$. Then

$$\hat{B}^T\sqrt{W}\left(\sqrt{W^{-1}}M_{cf}\sqrt{W^{-1}}\right)\sqrt{W}\hat{B}\mathbf{u} = \hat{B}^T M_{cf}\hat{B}\mathbf{u} = I_n\mathbf{u} = \mathbf{u} = \frac{1}{\lambda}\hat{L}\mathbf{u} = \hat{B}^T W\hat{B}\frac{\mathbf{u}}{\lambda}.$$

Thus

$$\hat{B}^T\sqrt{W}\left(\sqrt{W^{-1}}M_{cf}\sqrt{W^{-1}}(\sqrt{W}\hat{B}\mathbf{u}) - \frac{1}{\lambda}(\sqrt{W}\hat{B}\mathbf{u})\right) = 0,$$

which implies that $\sqrt{W^{-1}}M_{cf}\sqrt{W^{-1}}(\sqrt{W}\hat{B}\mathbf{u}) - \frac{1}{\lambda}(\sqrt{W}\hat{B}\mathbf{u})$ is in C_0, the null space of $\hat{B}^T\sqrt{W}$. But we showed above that the first product is orthogonal to C_0. Further, the second product is in the column space of $\sqrt{W}\hat{B}$, which we noted above is orthogonal to C_0. Thus the two products must be equal, and $\sqrt{W}\hat{B}\mathbf{u}$ is an eigenvector of $\sqrt{W^{-1}}M_{cf}\sqrt{W^{-1}}$ with eigenvalue $\frac{1}{\lambda}$.

We still need to show that $\sqrt{W^{-1}}M_{cf}\sqrt{W^{-1}}$ does not have any other nonzero eigenvalues. Define C_1 to be the subspace that results from transforming the cut space of G_0 by multiplying it by $\sqrt{W}$. By Lemma 3.2, the columns of $\hat{B}$ are a basis for the cut space of G_0, so the columns of $\sqrt{W}\hat{B}$ form a basis for C_1. We noted above that $\sqrt{W}\hat{B}$ is orthogonal to C_0. Since $\sqrt{W}$ is invertible, C_1 has dimension n. Because (the symmetric matrix) $\sqrt{W^{-1}}M_{cf}\sqrt{W^{-1}}$ has null space C_0, any eigenvector of a nonzero eigenvalue must come from C_1.

Let $\mathbf{u} \in C_1$ be any eigenvector of $\sqrt{W^{-1}}M_{cf}\sqrt{W^{-1}}$ with eigenvalue $\mu > 0$. By the definition of C_1, there is a vector $\mathbf{z}$ in the cut space of G_0 such that $\mathbf{z} = \sqrt{W^{-1}}\mathbf{u}$. Since the columns of $\hat{B}$ are a basis for the cut space of G_0, there is a vector $\mathbf{x}$ in the vertex space restricted to the interior vertices such that $\hat{B}\mathbf{x} = \mathbf{z}$. Thus we can write $\mathbf{u} = \sqrt{W}\hat{B}\mathbf{x}$ for the appropriate such $\mathbf{x}$.

Then

$$\mu\mathbf{u} = \sqrt{W^{-1}}M_{cf}\sqrt{W^{-1}}\mathbf{u} = \sqrt{W^{-1}}M_{cf}\sqrt{W^{-1}}(\sqrt{W}\hat{B}\mathbf{x}) = \mu\sqrt{W}\hat{B}\mathbf{x}. \tag{6.2}$$

We can use (6.1) with (6.2) as follows:

$$\mathbf{x} = \hat{B}^T\sqrt{W}\left(\sqrt{W^{-1}}M_{cf}\sqrt{W^{-1}}\right)\sqrt{W}\hat{B}\mathbf{x} = \mu\hat{B}^T W\hat{B}\mathbf{x} = \mu\hat{L}\mathbf{x}.$$

Thus $\mathbf{x}$ is an eigenvector of $\hat{L}$ with eigenvalue $\lambda = \frac{1}{\mu}$. That is, for any eigenvalue μ of $\sqrt{W^{-1}}M_{cf}\sqrt{W^{-1}}$, there is an eigenvalue λ of $\hat{L}$ such that $\mu = \frac{1}{\lambda}$, and the correspondence between eigenvalues is consistent with the theorem statement. ∎

The details of the proof easily provide the following result, which is not new and is easy to show using other techniques.

COROLLARY 6.4. *$\hat{L}$ is positive definite.*

Proof. We can form the matrix $\sqrt{W^{-1}}M_{cf}\sqrt{W^{-1}}$ for $\hat{L}$ as defined above. As shown in the preceding proof, it has rank n. Because $M_{cf} = \Gamma_{cf}^T\Gamma_{cf}$ and W is diagonal with only positive diagonal entries, this matrix is positive semidefinite. It therefore has n positive eigenvalues, the reciprocal of each of which is an eigenvalue of $\hat{L}$ by the preceding theorem. Thus $\hat{L}$ has n positive eigenvalues; it is also an $n \times n$ symmetric matrix, which proves the corollary. ∎

6.1. Current flow embeddings and inverses. We can use Theorem 6.3 to prove the following theorem on the inverse of $\hat{L}$.

THEOREM 6.5. *$\Gamma_{cf}W^{-1}\Gamma_{cf}^T$ is the inverse of $\hat{L}$.*

Proof. $\hat{L}$ is an $n \times n$ real symmetric matrix. Therefore we can write any n-vector $\mathbf{x}$ as a weighted sum of eigenvectors of $\hat{L}$. Showing that $\Gamma_{cf} W^{-1} \Gamma_{cf}^T \hat{L}\mathbf{u} = \mathbf{u}$ for every eigenvector $\mathbf{u}$ of $\hat{L}$ is sufficient to prove the theorem.

$\Gamma_{cf} W^{-1} \Gamma_{cf}^T$ can be rewritten as $(\Gamma_{cf}\sqrt{W^{-1}})(\sqrt{W^{-1}}\Gamma_{cf}^T)$. Applying Lemma 3.1 shows that if $\mathbf{v}$ is an eigenvector of $(\sqrt{W^{-1}}\Gamma_{cf}^T)(\Gamma_{cf}\sqrt{W^{-1}}) = \sqrt{W^{-1}} M_{cf} \sqrt{W^{-1}}$ with eigenvalue $\mu > 0$, then $\Gamma_{cf}\sqrt{W^{-1}}\mathbf{v}$ is an eigenvector of $\Gamma_{cf} W^{-1} \Gamma_{cf}^T$ with eigenvalue μ.

Let $\mathbf{u}$ be any eigenvector of $\hat{L}$ with eigenvalue λ. By Corollary 6.4, $\lambda > 0$. By Theorem 6.3, $\sqrt{W}\hat{B}\mathbf{u}$ is an eigenvector of $\sqrt{W^{-1}} M_{cf} \sqrt{W^{-1}}$ with eigenvalue $\frac{1}{\lambda}$. By the result from the preceding paragraph, $\Gamma_{cf}\sqrt{W^{-1}}\sqrt{W}\hat{B}\mathbf{u}$ is an eigenvector of $\Gamma_{cf} W^{-1} \Gamma_{cf}^T$ with eigenvalue $\frac{1}{\lambda}$. Applying Lemma 6.1 gives the following:

$$\Gamma_{cf}\sqrt{W^{-1}}\sqrt{W}\hat{B}\mathbf{u} = \Gamma_{cf}\hat{B}\mathbf{u} = \mathbf{u}.$$

Hence $\mathbf{u}$ is an eigenvector of $\Gamma_{cf} W^{-1} \Gamma_{cf}^T$ with eigenvalue $\frac{1}{\lambda}$.

Combining these results gives the following:

$$\Gamma_{cf} W^{-1} \Gamma_{cf}^T \hat{L}\mathbf{u} = \lambda \Gamma_{cf} W^{-1} \Gamma_{cf}^T \mathbf{u} = \lambda \frac{\mathbf{u}}{\lambda} = \mathbf{u}.$$

This holds for every eigenvalue of $\hat{L}$, which proves the theorem. ∎

7. Current flow embeddings and lower bound techniques.

7.1. Connections with previous embedding techniques. We now discuss our results in light of previous embedding techniques used to bound Laplacian eigenvalues. For the sake of simplicity, we will focus on unweighted graphs in this section. Details for weighted graphs can be found in the references cited.

We will discuss two general methods. The first, which we call the edge length method, was presented by Kahale in [22]. It works as follows:

- Specify a clique embedding for the graph G and assign each edge in G a positive length.
- Compute the length of each path with respect to the edge lengths.
- For each edge, compute the sum of the lengths of all incident paths divided by the length of that edge. Let $\rho_{\max}$ be the maximum such value taken over all the edges. Then $\frac{n}{\rho_{\max}}$ is a lower bound for λ_2.

The original statement is in terms of an upper bound on the second largest eigenvalue of a reversible Markov chain, but, as Kahale notes, can be reformulated so as to apply to Laplacians.

The second is the path resistance method, presented by Guattery, Leighton, and Miller in [15]. It works as follows:

- Construct a clique embedding into G.
- For each edge e_{ij} in G, compute its congestion c_{ij}. In the unweighted case, the congestion of an edge is the number of paths that are incident to it.
- For each path P and each edge e_{ij} on P, allocate a resistor of size c_{ij} to P.
- For each path P, compute its resistance, i.e., $\sum_{e_{ij} \in P} c_{ij}$. Let r be the maximum resistance over all paths. Then $\frac{n}{r}$ is a lower bound on λ_2.

These methods are duals of each other in the sense that, given an embedding, the best possible bounds produced by the two methods are the same [15]. This is most easily understood in terms of a matrix representation introduced by Kahale [22], who showed that the edge length method could be viewed as an eigenvalue problem. The

representation uses an embedding matrix similar to the one used throughout this paper, except that no edge directions are used, so the embedding matrix is nonnegative. Thus, if Γ is an embedding matrix as defined above, then $|\Gamma|$ is Kahale's form of the embedding matrix for the same embedding. The statement of the bound problem in terms of an eigenvalue problem is as follows: Let Γ be the matrix for a clique embedding into connected graph G. Then

$$\lambda_2(G) \geq \frac{n}{\lambda_{\max}\left(|\Gamma|^T|\Gamma|\right)}. \tag{7.1}$$

The assignment of edge lengths in this method corresponds to multiplying a vector of lengths by the matrix $|\Gamma|^T|\Gamma|$. The eigenvector for $\lambda_{\max}\left(|\Gamma|^T|\Gamma|\right)$ is the best length assignment.

In [15], Guattery, Leighton, and Miller show that the path resistance method can be viewed in terms of the following problem: For Γ as in the paragraph above,

$$\lambda_2(G) \geq \frac{n}{\lambda_{\max}\left(|\Gamma||\Gamma|^T\right)}. \tag{7.2}$$

Note that, by Lemma 3.1, this is the same bound as in (7.1). Note also that these expressions can be applied to weighted graphs (generalized Laplacians) by including the conductance matrix: The matrices in inequalities (7.1) and (7.2) become $|\Gamma|^T|\Gamma|W^{-1}$ and $|\Gamma|W^{-1}|\Gamma|^T$, respectively.

Reference [15] also extends the technique to the zero Dirichlet boundary case through star embeddings: For connected graph G with zero Dirichlet boundary S, let Γ be the matrix for a star embedding. Then

$$\lambda_1(G_0) \geq \frac{1}{\lambda_{\max}\left(|\Gamma||\Gamma|^T\right)}.$$

These results present the problem of bounding an eigenvalue of one matrix in terms of finding an eigenvalue of another; in practice one does not do this, but instead finds a good approximation to the largest eigenvector of $|\Gamma|^T|\Gamma|$. In many cases where the goal is to bound λ_2 asymptotically for a family of graphs with regular structure, a reasonable approximate vector is relatively easy to find. In other cases, the following technique from [21] can be used: Note that the matrices $|\Gamma|^T|\Gamma|$ and $|\Gamma||\Gamma|^T$ are nonnegative. In the clique embedding case, the matrix is irreducible as a result of our assumption that the graph is connected. In the star embedding case, we can consider the connected components that result when the boundary is deleted; the resulting pieces are irreducible. For irreducible nonnegative matrices, the eigenvalue with the largest magnitude is positive; one can start with a positive vector and repeatedly multiply it by the matrix $|\Gamma|^T|\Gamma|$ to get improved estimates.

In the following sections, we present lemmas that connect results from this paper to previous techniques described above. This allows us to show that the earlier techniques can produce exact bounds for certain classes of graphs. Finally, we give an example to demonstrate some of these results. In particular, it shows that including edge directions can give improved bounds when the repeated multiplication method is used.

In the next two sections, we state results for weighted graphs in the interest of generality.

7.2. Technical lemmas. Let G be a connected graph with conductance matrix W and edge-vertex incidence matrix B and Laplacian $L = B^T W B$. In the zero Dirichlet boundary case, we assume that the subgraph induced by the interior vertices is connected, and we use the Laplacian $\hat{L}$ as defined in section 6. As in the previous sections, Γ is an embedding matrix and $M = \Gamma^T\Gamma$. We will use Γ_{cf} and M_{cf} when the embedding is the current flow embedding. In all cases, we assume that all embeddings and B for a particular graph use the same indexing. Whether the embedding is a clique or a star will be clear from context.

The following two lemmas and their corollaries show that we can drop the absolute values from the previous statement of the lower bound problem and state it in terms of $\Gamma^T\Gamma W^{-1}$. Further, for the embedding Γ_{cf}, the result is exact.

LEMMA 7.1. *For any clique embedding* Γ, $\lambda_{\max}(MW^{-1}) \geq \lambda_{\max}(M_{cf}W^{-1})$, *and* $n/\lambda_{\max}(MW^{-1}) \leq \lambda_2(G)$.

Proof. As in Theorem 5.2, we will use the symmetric matrix $\sqrt{W^{-1}}M\sqrt{W^{-1}}$, which is similar to MW^{-1}, to prove the result.

The Courant–Fischer minimax theorem (see, e.g., [12] or [29]) implies the following:

$$\lambda_{\max}(\sqrt{W^{-1}}M\sqrt{W^{-1}}) \geq \max_{\mathbf{x}\neq 0} \frac{\mathbf{x}^T(\sqrt{W^{-1}}M\sqrt{W^{-1}})\mathbf{x}}{\mathbf{x}^T\mathbf{x}}.$$

This inequality holds for $\mathbf{x} = \sqrt{W}B\mathbf{u}_2$, where $\mathbf{u}_2$ is the eigenvector corresponding to $\lambda_2(G)$, and B is the edge-vertex incidence matrix for G. Then we have

$$\begin{aligned}\lambda_{\max}\left(\sqrt{W^{-1}}M\sqrt{W^{-1}}\right) &\geq \frac{\left(\sqrt{W}B\mathbf{u}_2\right)^T\left(\sqrt{W^{-1}}M\sqrt{W^{-1}}\right)\left(\sqrt{W}B\mathbf{u}_2\right)}{\left(\sqrt{W}B\mathbf{u}_2\right)^T\left(\sqrt{W}B\mathbf{u}_2\right)} \\ &= \frac{\mathbf{u}_2^T(B^TMB)\mathbf{u}_2}{\mathbf{u}_2^TL\mathbf{u}_2} = \frac{n\,\mathbf{u}_2^T\mathbf{u}_2}{\lambda_2\,\mathbf{u}_2^T\mathbf{u}_2} = \frac{n}{\lambda_2},\end{aligned}$$

where the next-to-last equality follows from an application of Lemma 4.1 to the numerator as it is applied in the proof of Theorem 5.2. This proves the second claim in the theorem statement. Theorem 5.2 implies

$$\lambda_{\max}(\sqrt{W^{-1}}M_{cf}\sqrt{W^{-1}}) = \frac{n}{\lambda_2},$$

which, when combined with the previous result, proves the first claim. ∎

COROLLARY 7.2. *Let* L *be the generalized Laplacian of connected graph* G *with conductance matrix* W. *For any clique embedding* Γ, $\lambda_2(L) \geq n/\lambda_{\max}(\Gamma^T\Gamma W^{-1})$. *Further, if the current flow embedding* Γ_{cf} *is used, then equality holds in the expression above.*

Proof. This follows directly from Lemma 7.1, with M rewritten as $\Gamma^T\Gamma$. ∎

Similar results hold in the zero Dirichlet boundary case below.

LEMMA 7.3. *In the zero Dirichlet boundary case, for any star embedding* Γ, $\lambda_{\max}(MW^{-1}) \geq \lambda_{\max}(M_{cf}W^{-1})$ *and* $1/\lambda_{\max}(MW^{-1}) \leq \lambda_2(G_0)$.

Proof. The proof is essentially the same as for Lemma 7.1, though it uses Theorem 6.3 rather than Theorem 5.2, and the eigenvector of the smallest eigenvalue of $\hat{L}$ rather than $\mathbf{u}_2$. Details are left to the reader. ∎

COROLLARY 7.4. *Let* L *be the generalized Laplacian of connected graph* G *with conductance matrix* W *and zero boundary* S. *For any star embedding* Γ *into* G_0,

$\lambda_1(\hat{L}) \geq 1/\lambda_{\max}(\Gamma^T \Gamma W^{-1})$. *Further, if the current flow embedding* Γ_{cf} *is used, equality holds in the expression above.*

Proof. This follows directly from Lemma 7.3, with M rewritten as $\Gamma^T\Gamma$. □

The following technical lemma will simplify the proof of Lemma 7.6.

LEMMA 7.5. *Let* X *be an* $m \times n$ *matrix. Then* $\lambda_{\max}\left(\left|X^T X\right|\right) \leq \lambda_{\max}\left(|X|^T |X|\right)$.

Proof. It is well known that, for nonnegative matrices $A \geq B$, $\lambda_{\max}(A) \geq \lambda_{\max}(B)$ (see, e.g., [23, p. 38]). Thus the lemma holds if we show that, for all $1 \leq i \leq n$ and $1 \leq j \leq n$, $\left[\left|X^T X\right|\right]_{ij} \leq \left[|X|^T |X|\right]_{ij}$. Let X_i represent column i of X. For any i and j, it is clear that $\left|X_i^T X_j\right| \leq |X_i|^T |X_j|$. Since the terms in the inequality are the values of the desired matrix entries, the lemma holds. □

We now show that we can use $|\Gamma^T\Gamma|W^{-1}$ to compute lower bounds that are at least as good as those computed using $|\Gamma|^T|\Gamma|W^{-1}$.

LEMMA 7.6. *For any embedding* Γ,

$$\frac{n}{\lambda_{\max}(|\Gamma|^T|\Gamma|W^{-1})} \leq \frac{n}{\lambda_{\max}(|\Gamma^T\Gamma|W^{-1})} \leq \lambda_2.$$

Proof. By Lemma 7.5, $\lambda_{\max}(|\Gamma|^T|\Gamma|W^{-1}) \geq \lambda_{\max}(|\Gamma^T\Gamma|W^{-1})$, which proves the first inequality.

To show the second inequality, note that for real symmetric matrix A, $\lambda_{\max}(A) \leq \lambda_{\max}(|A|)$. (To see this, let $\mathbf{u}$ be a unit-length eigenvector for $\lambda_{\max}(A)$, and compare the products $\mathbf{u}^T A \mathbf{u}$ and $|\mathbf{u}|^T|A||\mathbf{u}|$.) Then Lemmas 7.1 and 7.3 imply the desired result. □

Since $|\Gamma^T\Gamma|W^{-1}$ is a nonnegative matrix, it can be used with the repeated multiplication algorithm.

7.3. Classes for which previous techniques give tight bounds. Note that for trees there is a unique clique embedding. Note also that for trees with a single zero-value boundary point, there is a unique star embedding. In certain cases, it is possible to set the edge orientations in the edge-vertex incidence matrix so that the entries in the embedding matrix Γ are all positive. In that case, $\Gamma^T\Gamma$, $|\Gamma^T\Gamma|$, and $|\Gamma|^T|\Gamma|$ are all the same matrix, and the desired eigenvalue can be computed exactly in terms of $\lambda_{\max}(|\Gamma|^T|\Gamma|)$. Thus the results from the previous section imply that the best lower bounds derived using techniques in [21] and [15] are exact. We will show this for two cases: one for the clique embedding and one for the star.

The first case we cover is the Laplacian form of a result shown in the birth-death chain example from [21]: If L is the generalized Laplacian of any weighted path graph with conductance matrix W, then there is an embedding such that $n/\lambda_{\max}(|\Gamma|^T|\Gamma|W^{-1}) = \lambda_2(L)$. To see this, note that we can index the vertices in order along the path and that we can direct every edge from the lower-index vertex to the higher-index one. Likewise, direct every path in the clique embedding from the lower-index vertex to the higher-index one. There is only one possible embedding, and the choices of path and edge directions ensures that all entries in the embedding matrix Γ are positive. Thus

$$\frac{n}{\lambda_2} = \lambda_{\max}(\Gamma_{cf}^T \Gamma_{cf} W^{-1}) = \lambda_{\max}(|\Gamma|^T|\Gamma|W^{-1}).$$

In the star case, a similar result holds for trees with a single zero-value boundary point. In this case, every edge lies on a path from some vertex to the boundary; the vertex at one end of the edge is closer to the boundary than the vertex at the other. Direct each edge toward its endpoint that is closer to the boundary vertex. Likewise,

direct each path in the embedding toward the center of the star. The embedding is unique, and

$$\frac{1}{\lambda_2} = \lambda_{\max}(\Gamma_{cf}^T \Gamma_{cf} W^{-1}) = \lambda_{\max}(|\Gamma|^T |\Gamma| W^{-1}).$$

7.4. A simple example. We now use a cycle on four vertices to demonstrate the case where previous lower bound techniques are not exact. We number the vertices in clockwise order and also direct the edges in clockwise order. The edge-vertex incidence matrix B and the Laplacian L are shown below:

$$B = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \\ 1 & 0 & 0 & -1 \end{bmatrix}, \qquad L = \begin{bmatrix} 2 & -1 & 0 & -1 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ -1 & 0 & -1 & 2 \end{bmatrix}.$$

The eigenvalues of L are $\lambda = 0$, $\lambda = 2$ with multiplicity 2, and $\lambda = 4$.

The previous embedding techniques cannot produce an exact bound for this problem; as shown in the corollary to the following lemma, the best lower bound they can produce is 4/3.

We use d_{ij} to denote the distance (in terms of number of edges) between vertices v_i and v_j.

LEMMA 7.7. *For L the Laplacian of an unweighted graph, the best lower bound on $\lambda_2(L)$ produced by the edge length or path resistance methods is less than or equal to $mn/\sum_{i,j} d_{ij}^2$.*

Proof. The lower bound for any embedding Γ is $n/\lambda_{\max}(|\Gamma|^T|\Gamma|)$. For any $\mathbf{x}$,

$$\frac{\mathbf{x}^T |\Gamma|^T |\Gamma| \mathbf{x}}{\mathbf{x}^T \mathbf{x}} \leq \lambda_{\max}(|\Gamma|^T |\Gamma|). \tag{7.3}$$

Set $\mathbf{x} = \mathbf{1}$, the vector of length m of all ones. The entries of $|\Gamma|\mathbf{1}$ are the row sums of $|\Gamma|$. If a row represents a single path, there is an entry of 1 for each edge in the path. The sum in this case is the length of the path, which is at least the distance between the endpoints. If a row represents a set of fractional paths, the row sum is the sum over each fractional path of the product of the fractional weight times the path length. Since the fractional weights must add to 1, the row sum is at least the distance between the endpoints of the path. We get one entry in $|\Gamma|\mathbf{x}$ for each edge in K_n.

Note that the numerator of the left-hand side of (7.3) is just the dot product of $|\Gamma|\mathbf{x}$ with itself, which by the previous paragraph is at least $\sum_{i,j} d_{ij}^2$. (Since $d_{ii} = 0$ for all i, we can add these terms into the sum without penalty.)

The denominator $\mathbf{x}^T\mathbf{x}$ is m, the length of the vector. Thus

$$\frac{\sum_{i,j} d_{ij}^2}{m} \leq \lambda_{\max}(|\Gamma|^T |\Gamma|).$$

Applying (7.1) and the duality of the edge length and path resistance methods gives the result in the lemma statement. ∎

COROLLARY 7.8. *For the cycle on four vertices, the best lower bounds from the path resistance or edge length methods are no bigger than 4/3.*

Proof. Apply Lemma 7.7. The cycle has four pairs of vertices at distance 1 and two pairs at distance 2, so the sum of the squares of the distances is 12. Dividing $mn = 16$ by this number gives the result. ∎

Thus, these methods do not produce exact results for our example cycle. The following (undirected) embedding matrix will produce the lower bound of 4/3:

$$|\Gamma| = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{bmatrix}, \qquad |\Gamma|^T|\Gamma| = \frac{1}{2} \cdot \begin{bmatrix} 3 & 1 & 1 & 1 \\ 1 & 3 & 1 & 1 \\ 1 & 1 & 3 & 1 \\ 1 & 1 & 1 & 3 \end{bmatrix}.$$

By adding edge directions, however, we can produce an exact lower bound. Consider the following embedding matrix Γ and the absolute value of the product $\Gamma^T\Gamma$:

$$\Gamma = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{bmatrix}, \qquad |\Gamma^T\Gamma| = \frac{1}{2} \cdot \begin{bmatrix} 3 & 0 & 1 & 0 \\ 0 & 3 & 0 & 1 \\ 1 & 0 & 3 & 0 \\ 0 & 1 & 0 & 3 \end{bmatrix}.$$

The eigenvalues of the $|\Gamma^T\Gamma|$ are $\lambda = 1$ and $\lambda = 2$, each with multiplicity 2. By the results in section 7.2, we get a lower bound of 2 on λ_2 of the cycle, which is exactly equal to the eigenvalue.

As a final example, we show that forming a nonnegative matrix based on the current flow embedding does not necessarily produce the best bounds. This is true in spite of the exact relationship between the current flow embedding eigenvalues and the Laplacian eigenvalues. We show below the current flow embedding matrix and two nonnegative matrices produced from it:

$$\Gamma = \frac{1}{4} \cdot \begin{bmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \\ 2 & 2 & -2 & -2 \\ -2 & 2 & 2 & -2 \end{bmatrix}, \qquad |\Gamma^T\Gamma| = \frac{1}{4} \cdot \begin{bmatrix} 5 & 1 & 3 & 1 \\ 1 & 5 & 1 & 3 \\ 3 & 1 & 5 & 1 \\ 1 & 3 & 1 & 5 \end{bmatrix},$$

and

$$|\Gamma|^T|\Gamma| = \frac{1}{4} \cdot \begin{bmatrix} 5 & 4 & 4 & 4 \\ 4 & 5 & 4 & 4 \\ 4 & 4 & 5 & 4 \\ 4 & 4 & 4 & 5 \end{bmatrix}.$$

The eigenvalues of $|\Gamma^T\Gamma|$ are $\lambda = 1/2$ with multiplicity 2, $\lambda = 3/2$, and $\lambda = 5/2$, yielding a lower bound of 8/5 on λ_2; the previous example gives a better lower bound. The eigenvalues of $|\Gamma^T\Gamma|$ are $\lambda = 1/4$ with multiplicity 3 and $\lambda = 4\,1/4$, yielding a lower bound of 16/17, which is less than 1 and lower than the bounds for our previous examples.

Acknowledgments. The authors wish to thank John Lafferty of Carnegie Mellon University, who provided many valuable comments and suggestions on a previous version of this paper.

REFERENCES

[1] N. Alon, *Eigenvalues and expanders*, Combinatorica, 6 (1986), pp. 83–96.

[2] N. Alon and V. D. Milman, λ_1*, isoperimetric inequalities for graphs, and superconcentrators*, J. Combin. Theory Ser. B, 38 (1985), pp. 73–88.

[3] W. N. Anderson, Jr., and T. D. Morley, *Eigenvalues of the Laplacian of a graph*, Linear and Multilinear Algebra, 18 (1985), pp. 141–145.

[4] O. Axelsson, *Bounds of eigenvalues of preconditioned matrices*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 847–862.

[5] C. Berge, *Graphs and Hypergraphs*, North–Holland, Amsterdam, 1973.

[6] N. Biggs, *Algebraic Graph Theory*, Cambridge University Press, Cambridge, UK, 1974.

[7] B. Bollobás, *Graph Theory: An Introductory Course*, Springer-Verlag, New York, 1979.

[8] A. K. Chandra, P. Raghavan, W. L. Ruzzo, R. Smolensky, and P. Tiwari, *The electrical resistance of a graph captures its commute and cover times*, in Proceedings 21st Annual Symposium on Theory of Computation, Seattle, 1989, pp. 574–586.

[9] P. Diaconis and D. Stroock, *Geometric bounds for eigenvalues of Markov chains*, Ann. Appl. Probab., 1 (1991), pp. 36–61.

[10] M. Fiedler, *Algebraic connectivity of graphs*, Czechoslovak Math. J., 23 (1973), pp. 298–305.

[11] M. Fiedler, *A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory*, Czechoslovak Math. J., 25 (1975), pp. 619–633.

[12] G. H. Golub and C. F. V. Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1989.

[13] K. Gremban, G. Miller, and M. Zagha, *Performance evaluation of a parallel preconditioner*, in Proceedings 9th International Parallel Processing Symposium, Santa Barbara, CA, 1995, pp. 65–69.

[14] K. D. Gremban, *Combinatorial Preconditioners for Sparse, Symmetric, Diagonally Dominant Linear Systems*, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, 1996.

[15] S. Guattery, F. T. Leighton, and G. Miller, *The path resistance method for bounding* λ_2 *of a Laplacian*, in Proceedings 8th ACM-SIAM Symposium on Discrete Algorithms, New Orleans, LA, 1997, pp. 201–210.

[16] S. Guattery and G. Miller, *On the performance of spectral graph partitioning methods*, in Proceedings 6th ACM-SIAM Symposium on Discrete Algorithms, San Francisco, CA, 1995, pp. 233–242.

[17] S. Guattery and G. L. Miller, *On the quality of spectral separators*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 701–719.

[18] L. A. Hageman and D. M. Young, *Applied Iterative Methods*, Computer Science and Applied Mathematics, Academic Press, New York, 1981.

[19] L. Hagen and A. B. Kahng, *New spectral methods for ratio cut partitioning and clustering*, IEEE Trans. Computer-Aided Design, 11 (1992), pp. 1074–1085.

[20] M. Jerrum and A. Sinclair, *Approximating the permanent*, SIAM J. Comput., 18 (1989), pp. 1149–1178.

[21] N. Kahale, *A Semidefinite Bound for Mixing Rates of Markov Chains*, TR 95-41, DIMACS, Piscataway, NJ, 1995.

[22] N. Kahale, *A semidefinite bound for mixing rates of Markov chains*, in Proceedings 5th Integer Programming and Combinatorial Optimization Conference, Lecture Notes in Comput. Sci. 1084, Springer-Verlag, New York, 1996, pp. 190–203.

[23] H. Minc, *Nonnegative Matrices*, John Wiley, New York, 1988.

[24] B. Mohar, *Isoperimetric numbers of graphs*, J. Combin. Theory Ser. B, 47 (1989), pp. 274–291.

[25] A. Pothen, H. D. Simon, and K.-P. Liou, *Partitioning sparse matrices with eigenvectors of graphs*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 430–452.

[26] A. Sinclair, *Improved bounds for mixing rates of Markov chains and multicommodity flow*, Combin. Probab. Comput., 1 (1992), pp. 351–370.

[27] A. Sinclair and M. Jerrum, *Approximate counting, uniform generation and rapidly mixing Markov chains*, Inform. and Comput., 82 (1989), pp. 93–133.

[28] D. A. Spielman and S. H. Teng, *Spectral partitioning works: Planar graphs and finite element meshes*, Techical report UCB CSD-96-898, University of California at Berkeley, Berkeley, CA, 1996.

[29] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 3, pp. 724–742
© 2000 Society for Industrial and Applied Mathematics

REGULARITY OF THE BRUNOVSKY–KRONECKER STRATIFICATION*

J. FERRER†, M. I. GARCÍA†, AND F. PUERTA†

Abstract. We study the partition of the set of pairs of matrices according to the Brunovsky–Kronecker type. We show that it is a constructible stratification, and that it is Whitney regular when the second matrix is a column matrix. We give an application to the obtainment of bifurcation diagrams for few-parameter generic families of linear systems.

Key words. linear system, smooth family, versal deformation, transversality, stratification, bifurcation

AMS subject classifications. 15A21, 93B10, 93B35

PII. S0895479897319052

1. Introduction.

We shall be dealing with the geometric study of the perturbations of linear (time-invariant) systems $\dot{x} = Ax + Bu$—or, more generally, with the study of the local structure of the space of parameters of a generic family of linear systems.

A system $\dot{x} = Ax + Bu$ is identified in a natural way with the pair of matrices (A, B), so that in this paper we consider the space of pairs of matrices of this kind (and the block-similarity equivalence between them), and we study their perturbations. Then we follow the method introduced by Arnold [1] to study perturbations of square matrices, mainly based on the consideration of versal deformations (see also ([13, (8.1.1)]) and ([15, (V)]). A versal deformation for a pair of matrices has been explicitly obtained in [5].

A key role is played by the canonical reduced form of a pair of matrices, usually called Brunovsky form. However, in this paper we will refer to it as the Brunovsky–Kronecker form of a pair (A, B), because it can be easily derived (see (6.5.2) of [10]) from the Kronecker form of the matrix pencil $(\lambda I + A, B)$. In fact, from this point of view of matrix pencils, results equivalent to those in [5] about versal deformations of linear systems have been shown by [2], [3], and [4].

The central goal is to construct a "stratification" of the space of pairs of matrices; that is to say, a partition of it in a locally finite (see Remark 8.3(iii)) family of manifolds, called "strata."As a starting point, the space of pairs of matrices is partitioned into block-similarity classes, that is to say, into the orbits by the action of the feedback group. Thus, the perturbations of a given system are obtained from the local description of this partition, which in turn can be derived from the above mentioned versal deformations. For example, metrical information about these perturbations is obtained in [4].

However, as for square matrices, this partition is not locally finite. Following Arnold's method, we consider "strata" formed by the uncountable union of the orbits which have the same discrete invariants but different eigenvalues. Therefore, the local description of this new partition informs about the perturbation of the controllability

*Received by the editors March 26, 1997; accepted for publication (in revised form) by A. Edelman April 5, 1999; published electronically February 10, 2000.
http://www.siam.org/journals/simax/21-3/31905.html
†Departament de Matemàtica Aplicada I, Universitat Politècnica de Catalunya, Diagonal 647, 08028 Barcelona, Spain (ferrer@ma1.upc.es, igarcia@ma1.upc.es, puerta@ma1.upc.es).

indices, etc., of a system. The key point lies in showing that each stratum is a (regular) manifold, so that the new partition is in fact a stratification, which we call Brunovsky–Kronecker-stratification. Again, the proof follows from the local description given by the versal deformation.

Finally, given a parametrized family of linear systems, the Brunovsky–Kronecker-stratification induces a partition in the space of parameters, known as the bifurcation diagram of the family. In general, if M is a manifold with a stratification $M = \cup_i X_i$, and $f(w) \in M$, $w \in W$, is a W-parametrized differentiable family on it (that is to say, W is another manifold and $f : W \longrightarrow M$ a differentiable map), then the partition $W = \cup_i f^{-1}(X_i)$ is called the "bifurcation diagram"of this family. The bifurcation diagram of a family of linear systems gives precise information about the qualitative properties of the systems arising in the family and about the effects of local perturbations of the parameters. For example, assume that the bifurcation diagram of an 1-parametrized family (say, $W = \mathbb{C}$) consists of a simple point (say, $w = 0$). It means that only two types of discrete invariants appear in this family, having very different "stability properties": the type corresponding to $w \neq 0$ is "stable" under little perturbations of w, whereas the other one disappears under any slight perturbations of the value $w = 0$ of the parameter. In particular, in the numerical computation of the discrete invariants of the element corresponding to $w = 0$ it is probable that one obtains in fact the type corresponding to $w \neq 0$, but no other type can be expected.

Nevertheless, some additional conditions are necessary in order to ensure that the bifurcation diagram is in fact a stratification and that its local structure can be derived from the above description of the Brunovsky–Kronecker-stratification. Transversality provides the required conditions. In fact, because of the generalized implicit function theorem, if the parametrized family is transversal to all the strata, then the bifurcation diagram is also a stratification which besides has the following important property: the codimension of each one of its strata in the parameter space is the same as the codimension of the corresponding stratum in the Brunovsky–Kronecker-stratification. Then, for a transversal family we have a precise limitation about the possibility of changes of the local structure of the family.

The question is then, "Are there *many* families transversal to the Brunovsky–Kronecker-stratification?" At this point is where the Whitney regularity conditions play a key role. In fact, if the stratification is Whitney regular, a theorem of Thom states that the set of families which are transversal to the stratification is open and dense; this gives sense to qualifying as generic such families; this is to say, "almost" all families are transversal to the stratification.

For example, we will see (Example 8.2.i) that any generic 1-parametrized family of single-controlled systems has the bifurcation diagram referred above (a single point); hence, it can never arise three or more different types of discrete invariants in a 1-parametrized family of single-controlled systems. Also, we will see (Example 8.2(ii)) that the only singularity in a generic 2-parametrized bifurcation diagram is the normal crossing. Notice, for example, that in the bifurcation diagram of a generic 2-parametrized family of square matrices (see [1]) cusp singularities also appear.

More in general, one can list all the possible generic bifurcation diagrams for $1, 2, \ldots$ parameters (see Example 8.2). This is clearly interesting from an experimental point of view, because any other bifurcation diagram is "experimentally impossible,"or, following an Arnold's remark, if a parametrized family contains systems of more complicated Brunovsky–Kronecker-form, or the bifurcation diagram has more

complicated singularities, and so on, then they can be removed by an arbitrarily small perturbation of the family.

There are some theoretical tools to verify the Whitney conditions. For example, according to a theorem by Whitney, a (locally finite) algebraic stratification is Whitney regular if the strata are orbits under the action of an algebraic group. Such is the case of the stratification which arises in the classification of linear systems under the action of the full linear group in the space of state variables ([14]; see also [7]). This is also the case of the stratification arising in the classification of reachable linear systems acted on by the feedback group [16], [17].

The situation is more complicated if, as in our case, the partition in orbits is not locally finite, and one must consider strata formed by an infinite union of them. This is the case of the original study of Arnold: the square matrices stratified according to their Jordan type. See [8] for a detailed proof that this stratification is Whitney regular.

In our case, following these techniques, we prove that the Brunovsky–Kronecker stratification of single-controlled linear systems is Whitney regular. Then, as we said, we can obtain precise descriptions of the bifurcation diagrams of generic parametrized families of this kind of systems.

The organization of this paper is as follows. Section 2 is preliminary, and it contains the general definitions and notations needed in the sequel. In section 1 we define the Brunovsky–Kronecker partition of the set $\mathcal{M}_{n,m}$. Previously, we recall the Brunovsky–Kronecker form as well as the way of calculating its invariants. In section 4 we remark that the block-similarity classes can be considered to be the orbits with regard to the action of a Lie group. Some properties can be derived from this representation: constructibility, regularity, homogeneity, and so on. Section 5 is devoted to showing a fundamental property of local triviality along the orbits and to giving explicitly the minitransversal variety to the orbit mentioned above. This is the key point in order to prove, in section 6, that the strata are submanifolds of $\mathcal{M}_{n,m}$ (Theorem 6.2). Moreover, we show that they are constructible and connected. In section 7, we tackle the main theorem (Theorem 7.5): the Brunovsky–Kronecker stratification is Whitney regular when $m = 1$. Also, we remark (see Proposition 7.6) that it verifies the frontier condition for any m. Finally, in section 8 we enumerate the singularities of bifurcation diagrams of few-parameter generic families of linear systems.

2. Preliminaries.

For every integer p, we will denote by M_p the space of p-square complex matrices, and by $Gl(p)$ the *linear group* formed by the invertible matrices of M_p. Also, we will denote $M_{p\times q}$ the space of rectangular complex matrices, having p rows and q columns. We will deal mainly with pairs of matrices (A, B) such that $A \in M_n$ and $B \in M_{n\times m}$. In all the paper, $\mathcal{M}_{n,m}$ will denote the space of such pairs of matrices; $\mathcal{M}_{n,m} = M_n \times M_{n\times m}$.

Throughout this paper the term *manifold* is used as an abbreviation for complex differentiable manifold. Let X, Y be manifolds, $f : X \longrightarrow Y$ a differentiable map, and Z a submanifold of Y. We recall (see, for example, [12]) that the map f is said to be *transversal* to the submanifold Z if at each point x in $f^{-1}(Z)$ the following relation holds:

$$df_x(T_xX) + T_yZ = T_yY,$$

where $y = f(x)$ and T_xX, T_yZ, T_yY are the corresponding tangent spaces.

In particular, if f is the inclusion map the above condition merely states that

$$T_xX + T_xZ = T_xY$$

for every $x \in X \cap Z$. We say then that the submanifolds X and Z are *transversal* (or *minitransversal* if the above sum is direct).

We will use extensively throughout the paper the following result: If the differentiable map $f : X \longrightarrow Y$ is transversal to a submanifold $Z \subset Y$ then $f^{-1}(Z)$ is a submanifold of X. Moreover, the codimension of $f^{-1}(Z)$ in X equals the codimension of Z in Y.

In particular, if X and Z are submanifolds of Y which are transversal, then $X \cap Z$ is a submanifold of Y and $\operatorname{codim}(X \cap Z) = \operatorname{codim} X + \operatorname{codim} Z$.

A *stratification* of a subset V of a manifold M is a partition $\cup_i X_i$ of V into submanifolds X_i of M (called the *strata*) which satisfies the *local finiteness condition*: every point in V has a neighborhood in M which meets only finitely many strata.

Let W, M be manifolds, $f : W \longrightarrow M$ a differentiable map, and $V = \cup_i X_i$ a stratification of M. We say that f is *transversal* to V if it is transversal to each stratum X_i of V.

According to [1] and [8], the space M_p can be partitioned into a finite number of *Segre strata*, each one formed by the matrices having the same *Segre symbol* (or the same *Jordan type*). Thus, each stratum is the uncountable union of similarity classes, differing only in the values of the distinct eigenvalues. If $C \in M_p$, we denote $\sigma_S(C)$, $\mathcal{O}_S(C)$, $E_S(C)$ the Segre symbol, the Segre orbit, and the Segre stratum of C, respectively. If $J \in M_p$ is a Jordan matrix, we denote $\Gamma_S(J)$ the versal deformation of J described in [1]. In fact, it is a linear subvariety of M_p, minitransversal to $\mathcal{O}_S(J)$ at J.

Let M be a finite-dimensional vector space, X, Y submanifolds of M, and $x \in X \cap \overline{Y}$. We say that Y is *Whitney regular* over X at x when the following condition holds: let (x_i), (y_i) be sequences in X, Y, respectively, both converging to x, with $x_i \neq y_i$ for all i. Take L_i to be the line spanned by $x_i - y_i$ and T_i to be the tangent space $T_{y_i}Y$. If (L_i) converges to L (in the Grassmannian of 1-dimensional subspaces of M) and T_i converges to T (in the Grassmannian of q-dimensional subspaces of M, $q = \dim Y$), then $L \subseteq T$.

It is not difficult to see that this condition is invariant under diffeomorphisms. Hence, we can define in an obvious way the Whitney regular condition when M is a manifold.

We say that Y is Whitney regular over X when it is so at every point of $X \cap \overline{Y}$.

Finally, let $\cup_i X_i$ be a stratification of a subset V of a manifold M. We say that this stratification is Whitney regular when every stratum X_i is Whitney regular over X_j $(i \neq j)$.

A closed algebraic subset of $\mathbb{C}^n$ is the set of zeroes of a finite set of polynomials. An open algebraic set of $\mathbb{C}^n$ is the complementary of a closed algebraic subset of $\mathbb{C}^n$.

A subset S of $\mathbb{C}^n$ is *constructible* if S is a disjoint union $T_1 \cup \cdots \cup T_k$, where T_i is locally closed, i.e., it is the intersection of a closed algebraic set with an open algebraic set.

We call *rational mapping* a mapping of a subset of $\mathbb{C}^p$ into $\mathbb{C}^q$ whose components are rational functions, with denominators nowhere zero on the domain. We will need the following theorem of Chevalley: *the image of a constructible set under a rational map is constructible.*

We refer to [8] for some complementary definitions and the results which are needed in what follows.

3. The Brunovsky–Kronecker stratification. We shall partition $\mathcal{M}_{n,m}$ into a finite number of subsets, each one formed by all the pairs of matrices having the same controllability indices and the same Segre symbol (of the Jordan block in the Brunovsky–Kronecker reduced form). Hence, each one of these subsets is an uncountable union of block-similarity classes, differing only in the values of the eigenvalues. In section 6 we shall prove that this partition is in fact a constructible regular stratification. Therefore, we refer to it as the Brunovsky–Kronecker stratification of $\mathcal{M}_{n,m}$. Often, we will abbreviate Brunovsky–Kronecker as BK: BK-stratification, BK-reduced form, and so on.

We recall that two pairs of matrices (A,B), (A',B') of $\mathcal{M}_{n,m}$ are called *block-similar* and we write $(A,B)\sim(A',B')$ if there exist $P\in Gl(n)$, $Q\in Gl(m)$, and $R\in M_{n\times m}$ such that

$$(A',B')=P^{-1}(A,B)\begin{pmatrix}P & 0\\ R & Q\end{pmatrix}.$$

It is an equivalence relation. It is well known that $(A,B)\sim(A',B')$ if and only if $(\lambda I+A,B)$ is strictly equivalent to $(\lambda I+A',B')$. However, notice that the block-similarity equivalence involves only the parameters defining the system. Each pair $(A,B)\in\mathcal{M}_{n,m}$ can be reduced to its BK-*canonical form*

$$\left(\begin{pmatrix}N & 0\\ 0 & J\end{pmatrix},\begin{pmatrix}E & 0\\ 0 & 0\end{pmatrix}\right),$$

where

(a) $N=\operatorname{diag}(N_1,\dots,N_r)$, N_i the nilpotent k_i-matrix $\begin{pmatrix}0 & 1 & 0 & \dots & 0 & 0\\ 0 & 0 & 1 & \dots & 0 & 0\\ \dots & \dots & \dots & \dots & \dots & \dots\\ 0 & 0 & 0 & \dots & 0 & 1\\ 0 & 0 & 0 & \dots & 0 & 0\end{pmatrix}$,

(b) $E=\begin{pmatrix}E_1 & & & \\ & E_2 & & \\ & & \ddots & \\ & & & E_r\end{pmatrix}$, $\quad E_i=(0\dots01)^t\in M_{k_i\times 1}$,

(c) J is a Jordan matrix; we shall write

$$J=\operatorname{diag}(J_1,\dots,J_\nu),\ J_i=\operatorname{diag}(J_{\sigma_1(i)},J_{\sigma_2(i)},\dots),\quad 1\le i\le\nu,$$

$J_{\sigma_j(i)}$ being the $\sigma_j(i)$-matrix

$$J_{\sigma_j(i)}=\begin{pmatrix}\lambda_i & 1 & 0 & \dots & 0\\ 0 & \lambda_i & 1 & \dots & 0\\ \dots & \dots & \dots & \dots & \dots\\ 0 & 0 & 0 & \dots & 1\\ 0 & 0 & 0 & \dots & \lambda_i\end{pmatrix},\quad 1\le i\le\nu,\ j=1,2,\dots.$$

The integers k_i are called the *controllability indices* of (A,B). We assume $k_1\ge\dots\ge k_r$. We shall write $k=(k_1,\dots,k_r)$, $s=k_1+\dots+k_r$. We call $\lambda_1,\dots,\lambda_\nu$ the *eigenvalues of* (A,B). And we write $\sigma(i)=(\sigma_1(i),\sigma_2(i),\dots)$ for the Segre characteristic of λ_i and $\delta_i=\sigma_1(i)+\sigma_2(i)+\dots$. We assume $\sigma_1(i)\ge\sigma_2(i)\ge\dots$ for every $1\le i\le\nu$. We recall (see [6]) the following proposition.

PROPOSITION 3.1.

(i) *The controllability indices of* (A,B), $(k_1,\dots,k_r)$ *are the conjugate partition of* $(r_1,\dots,r_k)$, *where* $r_i=\rho_i-\rho_{i-1}$, $i\geq 1$,

$$\rho_i=\operatorname{rank}\begin{pmatrix}B & AB & \dots & A^{i-1}B\end{pmatrix},\quad i\geq 1,\quad \rho_0=0,$$

and k *is the smallest integer such that* $\rho_{k+1}=\rho_k$.

(ii) *The eigenvalues of* (A,B) *are the complex numbers* λ *for which there exist* $v\in\mathbb{C}^n$, $v\neq 0$ *such that* $A^tv=\lambda v$, $B^tv=0$. *Or equivalently, such that* $\operatorname{rank}(A-\lambda I_n,B)<n$. *For each eigenvalue* λ, *its Segre characteristic is the conjugate partition of* $[n-\nu_1,\nu_1-\nu_2,\dots]$, *where*

$$\nu_j=\operatorname{rank}\begin{pmatrix}B & AB & \dots & A^{k_1-1}B & (A-\lambda I_n)^j\end{pmatrix}$$

and k_1 *is the smallest integer such that*

$$\operatorname{rank}\begin{pmatrix}B & AB & \dots & A^{k_1}B\end{pmatrix}=\operatorname{rank}\begin{pmatrix}B & AB & \dots & A^{k_1-1}B\end{pmatrix}.$$

DEFINITION 3.2. *With the above notation, the BK-symbol* (k,σ) *of the pair* (A,B) *is formed by the controllability indices and the Segre symbol of* J:

$$(k,\sigma)=(k_1,\dots,k_r;\sigma(1),\dots\sigma(\nu)).$$

A BK-stratum, $E(k,\sigma)$, *in* $\mathcal{M}_{n,m}$ *consists of all pairs of matrices having a given BK-symbol* (k,σ). *We denote* $E(A,B)$ *the BK-stratum of the pair* (A,B). *We denote* Σ *the partition* $\cup_{k,\sigma}E(k,\sigma)$ *of* $\mathcal{M}_{n,m}$, *which will be called the BK-stratification.*

When we write (k,σ) we do not exclude the possibility that $(k,\sigma)=(k)$; that is to say, there is no Jordan matrix J or that $(k,\sigma)=(\sigma)$, which means that $B=0$.

Notice that there are only finitely many BK-strata. Each one is an uncountable union of block-similarity equivalence classes, differing only in the values of the eigenvalues $\lambda_1,\dots,\lambda_\nu$.

4. The orbits. If one considers the Lie subgroup of $Gl(n+m)$

$$\mathcal{G}=\left\{\begin{pmatrix}P & 0\\ R & Q\end{pmatrix}\mid P\in Gl(n),Q\in Gl(m),R\in M_{m\times n}\right\},$$

called the *state feedback group*, and its action in $\mathcal{M}_{n,m}$ defined by

$$\alpha\left(\begin{pmatrix}P & 0\\ R & Q\end{pmatrix},(A,B)\right)=P^{-1}(A,B)\begin{pmatrix}P & 0\\ R & Q\end{pmatrix},$$

the equivalence class of a (A,B) with regard to the block-similarity is just its *orbit* under this action:

$$\mathcal{O}(A,B)=\{\alpha(g,(A,B)),g\in\mathcal{G}\}.$$

Following the reasoning in [8], one can prove (see [5]) the following proposition.

PROPOSITION 4.1. *Orbits* $\mathcal{O}(A,B)$ *are constructible complex submanifolds of* $\mathcal{M}_{n,m}$.

A local parametrization of $\mathcal{O}(A,B)$ at (A,B) can be obtained as follows: let $\mathcal{E}st(A,B)$ be the stabilizer of (A,B) under the action of α,

$$\mathcal{E}st(A,B)=\{g\in\mathcal{G}\mid\alpha(g,(A,B))=(A,B)\},$$

and let $V \subset \mathcal{G}$ a submanifold of $\mathcal{G}$ minitransversal to $\mathcal{E}st(A,B)$ at the identity I. Then, it suffices to define $\alpha_1 : V \longrightarrow \mathcal{O}(A,B)$ by $\alpha_1(g) = \alpha(g,(A,B))$.

PROPOSITION 4.2 (see [5, section (II.1.9)]). *With the notations in section* 1*, the codimension of the orbit* $\mathcal{O}(A,B)$ *is*

$$\begin{aligned}\operatorname{codim}\mathcal{O}(A,B) = {} & r(n-s)+(m-r)(s-r)+(m-r)(n-s)\\ & + \sum_{1\le i,j\le r} max\{0,k_j-k_i-1\}\\ & + \sum_{1\le i\le \nu}(\sigma_1(i)+3\sigma_2(i)+5\sigma_3(i)+\cdots).\end{aligned}$$

If $(A',B') \in \mathcal{O}(A,B)$, and we take $g_0 \in \mathcal{G}$ such that $\alpha(g_0,(A,B)) = (A',B')$, then the map $\alpha_2((X,Y)) = \alpha(g_0,(X,Y))$ gives the following homogeneity property which will be used to reduce our study to pairs of matrices in the BK-reduced form.

PROPOSITION 4.3. *Given any two block-similar pairs of matrices* (A,B) *and* (A',B')*, there exists a diffeomorphism* $\alpha_2 : \mathcal{M}_{n,m} \longrightarrow \mathcal{M}_{n,m}$ *which preserves the orbits and such that* $\alpha_2(A,B) = (A',B')$*.*

5. Local triviality along the orbits. A central point in the method of [8] is to reduce the study of the stratification to that of its intersection with a submanifold Γ transversal to the orbits. The key point is the selection of submanifold Γ, in order to have an appropriate description of its intersection with the BK-stratification (see Proposition 5.3). In our case, we select Γ as the versal deformation constructed in [4].

The following lemma provides the desired local trivialization along the orbits. It can be proved by means of the inverse function theorem (see, for example, section (I.2.5) in [5]).

LEMMA 5.1. *Let* $(A,B) \in \mathcal{M}_{n,m}$*,* $\mathcal{O}(A,B)$ *its orbit, and* $\Gamma = (A,B)+T$ *a linear variety minitransversal to* $\mathcal{O}(A,B)$*:*

$$\mathcal{M}_{n,m} = T \oplus T_{(A,B)}\mathcal{O}(A,B).$$

Let $\alpha_1 : V \longrightarrow \mathcal{O}(A,B)$ *be a local parametrization of* $\mathcal{O}(A,B)$ *at* (A,B) *and denote by* $\alpha_1(V)$ *simply as* $\mathcal{O}(A,B)$*. Then the mapping*

$$\begin{aligned}\beta : \Gamma \times \mathcal{O}(A,B) &\longrightarrow \mathcal{M}_{n,m},\\ \beta((A,B)+(X,Y),(A',B')) &= \alpha(\alpha_1^{-1}(A',B'),(A,B)+(X,Y))\end{aligned}$$

is a local diffeomorphism at $((A,B),I)$ *which preserves the orbits.*

In this paper, we will apply this lemma taking Γ to be the linear variety obtained in [5, section (II.2.2)]. With the notations in section 1 we have the following theorem.

THEOREM 5.2. *Let* (A,B) *be a BK-matrix*

$$(A,B) = \left(\begin{pmatrix} N & 0\\ 0 & J\end{pmatrix}, \begin{pmatrix} E & 0\\ 0 & 0\end{pmatrix}\right)$$

and let us consider the set T *of pairs of matrices* (X,Y) *of the form*

$$X = \begin{pmatrix} 0 & 0\\ X_1^2 & X_2^2\end{pmatrix}, \quad Y = \begin{pmatrix} Y_1^1 & Y_2^1\\ 0 & Y_2^2\end{pmatrix},$$

where the block-decomposition corresponds to that of (A,B) *and*

(i) *all the entries in* X_1^2 *are zero except*

$$x_i^{s+1}, \dots, x_i^n, \quad i = 1, k_1 + 1, \dots, k_1 + \cdots + k_{r-1} + 1;$$

(ii) *the matrices* X_2^2 *are such that the variety* $\Gamma_S(J) = J + \{X_2^2\}$ *of* M_{n-s} *is the miniversal deformation of* J *given in* [1, (4.4)]*;*

(iii) *all the entries in* Y_1^1 *are zero except*

$$y_i^j, \quad 2 \le i \le r, \quad k_1 + \cdots + k_{i-2} + k_i + 1 \le j \le k_1 + \cdots + k_{i-2} + k_{i-1} - 1$$

(provided that $k_i \le k_{i-1} - 2$*);*

(iv) Y_2^1 *is such that*

$$y_{r+1}^i = \cdots = y_m^i = 0, \quad i = k_1, k_1 + k_2, \dots, k_1 + \cdots + k_r;$$

(v) *all the entries in* Y_2^2 *are arbitrary.*

Then $\Gamma = (A, B) + T$ *verifies the hypothesis in Lemma* 5.1: $\mathcal{M}_{n,m} = T \oplus T_{(A,B)}\mathcal{O}(A, B)$.

For this particular variety Γ, it is easy to discuss how the BK-symbol of $(A, B) + (X, Y)$ varies according to the values of the entries of (X, Y). For instance (see [4] for more details),

(i) if $X_2^2 \neq 0$, $X_1^2 = Y = 0$, then the controllability indices $k_1, \dots, k_r$ are constant, but the eigenvalues of J or its Segre symbol can vary as in [1];

(ii) if $X_1^2 \neq 0$, $Y = 0$, then r is constant, but s increases;

(iii) if $Y_2^1 \neq 0$ or $Y_2^2 \neq 0$, then r increases;

(iv) if $Y_1^1 \neq 0$, $Y_2^1 = Y_2^2 = 0$, then r and s are constant, but $k_1, \dots, k_r$ vary (in fact, the differences $k_i - k_{i+1}$ tend to decrease).

In particular, examine the following proposition.

PROPOSITION 5.3. *With the notation in Theorem* 5.2,

(a) *if* $(X, Y) \neq (0, 0)$*, then* $(A, B) + (X, Y)$ *is not in* $\mathcal{O}(A, B)$*;*

(b) $(A, B) + (X, Y)$ *is in* $E(A, B)$ *if and only if* $X_1^2 = Y = 0$*, and* $J + X_2^2$ *has the same Segre symbol as* J*.*

6. The strata. Next, we are going to tackle the regularity of the BK-strata (Theorem 6.2) and their dimension (Proposition 6.3). As we have said above, Lemma 5.1 plays an essential role. In fact, thanks to it we can reduce the problem to the intersection with the variety Γ in Theorem 5.2. Then, the result follows from the description of this intersection in Proposition 5.3 and the regularity of the Segre strata. Other properties of the BK-strata are presented in Propositions 6.4 and 6.5. From them, it is possible to give an autonomous proof of the regularity of the BK-strata (without using those of the Segre strata). Moreover, they are of interest in their own right.

LEMMA 6.1. *Let* $(A, B) \in \mathcal{M}_{n,m}$*,* $\mathcal{O}(A, B)$ *be its orbit,* $E(A, B)$ *be its stratum, and* Γ *be as in Lemma* 5.1*. Then, in a neighborhood of* (A, B)*,* $E(A, B)$ *is a submanifold of* $\mathcal{M}_{n,m}$ *if and only if* $E(A, B) \cap \Gamma$ *is.*

Proof. Assume that $E(A, B)$ is regular at (A, B). Being Γ transversal to $\mathcal{O}(A, B)$, it is also transversal to $E(A, B)$. Hence, $E(A, B) \cap \Gamma$ is regular at (A, B).

Conversely, assume that $E(A, B) \cap \Gamma$ is regular at (A, B). According to Lemma 5.1, we have

$$E(A, B) = \beta((E(A, B) \cap \Gamma) \times \mathcal{O}(A, B))$$

locally at (A, B). Therefore, $E(A, B)$ is regular at (A, B). ∎

THEOREM 6.2. *Any BK-stratum is a submanifold of $\mathcal{M}_{n,m}$.*

Proof. Let be $(A,B) \in \mathcal{M}_{n,m}$, $\mathcal{O}(A,B)$ its orbit, and $E(A,B)$ its BK-stratum. We must prove that $E(A,B)$ is regular at (A,B).

Because of Proposition 4.3, we can assume that (A,B) is a BK-matrix:

$$(A,B) = \left(\begin{pmatrix} N & 0 \\ 0 & J \end{pmatrix}, \begin{pmatrix} E & 0 \\ 0 & 0 \end{pmatrix} \right).$$

By Lemma 6.1 it is sufficient to prove that $E(A,B) \cap \Gamma$ is regular at (A,B), where Γ is the particular one in Theorem 5.2. From Proposition 5.3 it follows that $E(A,B) \cap \Gamma$ is formed by pairs of the form

$$\left(\begin{pmatrix} N & 0 \\ 0 & J + X_2^2 \end{pmatrix}, \begin{pmatrix} E & 0 \\ 0 & 0 \end{pmatrix} \right)$$

such that $J + X_2^2$ has the Segre symbol of J, or equivalently, such that $J + X_2^2$ belongs to the Segre stratum $E_S(J)$ of J.

Therefore, the mapping $\phi : M_{n-s} \longrightarrow \mathcal{M}_{n,m}$ defined by

$$\phi(C) = \left(\begin{pmatrix} N & 0 \\ 0 & C \end{pmatrix}, \begin{pmatrix} E & 0 \\ 0 & 0 \end{pmatrix} \right)$$

is a diffeomorphism such that

$$\phi(E_S(J) \cap \Gamma_S(J)) = E(A,B) \cap \Gamma.$$

[8] proves that the Segre strata are regular. Hence, $E_S(J) \cap \Gamma_S(J)$ is regular at J (we recall that $\Gamma_S(J)$ is a linear variety transversal to the Segre orbit of J and hence also transversal to $E_S(J)$ at J), and the proof is completed. □

PROPOSITION 6.3. *Let $(A,B) \in \mathcal{M}_{n,m}$, $\mathcal{O}(A,B)$ be its orbit, and $E(A,B)$ be its stratum. Then*

$$\dim E(A,B) = \nu + \dim \mathcal{O}(A,B),$$

where ν is the number of distinct eigenvalues of (A,B).

Proof. It is sufficient to bear in mind that, in the above proof, $\dim(E_S(J) \cap \Gamma_S(J)) = \nu$ (see [1]). □

The next result is needed for the proof of Lemma 7.4 which is a crucial step in the proof of the main theorem of this paper.

PROPOSITION 6.4. *The BK-strata are constructible connected sets.*

Proof. Let $E(k,\sigma)$ be the BK-stratum corresponding to the BK-symbol $(k,\sigma) = (k,\sigma_1,\ldots,\sigma_\nu)$. Let us consider the set $\mathbb{C}^{(\nu)} = \{(\lambda_1,\ldots,\lambda_\nu) \mid \lambda_i \neq \lambda_j\} \subset \mathbb{C}^\nu$. For each $(\lambda_1,\ldots,\lambda_\nu) \in \mathbb{C}^{(\nu)}$, let be $\mathrm{BK}(k,\sigma;(\lambda_1,\ldots,\lambda_\nu))$ the BK-matrix of $E(k,\sigma)$ having eigenvalues $\lambda_1,\ldots,\lambda_\nu$. Finally, let us consider the mapping $\psi : \mathcal{G} \times \mathbb{C}^{(\nu)} \longrightarrow \mathcal{M}_{n,m}$ defined by

$$\psi(g,(\lambda_1,\ldots,\lambda_\nu)) = \alpha(g, BK(k,\sigma;(\lambda_1,\ldots,\lambda_\nu))).$$

Obviously, $\mathcal{G} \times \mathbb{C}^{(\nu)}$ is a constructible set, ψ is a rational map, and $\psi(\mathcal{G} \times \mathbb{C}^{(\nu)}) = E(k,\sigma)$, so that, according to the Chevalley theorem, $E(k,\sigma)$ is a constructible set. Moreover, it is connected because ψ is continuous and $\mathcal{G} \times \mathbb{C}^{(\nu)}$ is connected. □

PROPOSITION 6.5. *The action of $\mathbf{C}$ in $\mathcal{M}_{n,m}$ defined by $(\lambda,(A,B)) \longrightarrow (A + \lambda I, B)$ preserves the BK-strata. That is to say, if $E(k,\sigma)$ is a BK-stratum, $(A,B) \in E(k,\sigma)$ and $\lambda \in \mathbb{C}$, then*

(1) *μ is an eigenvalue of (A,B) if and only if $\mu+\lambda$ is an eigenvalue of $(A+\lambda I,B)$;*
(2) $(A+\lambda I,B)\in E(k,\sigma)$.

Proof. (1) According to Proposition 3.1, μ is an eigenvalue of (A,B) if and only if rank $(A-\mu I,B)<n$. And obviously it is equivalent to rank $((A+\lambda I)-(\mu+\lambda)I,B)<n$.

(2) To see that (A,B) and $(A+\lambda I,B)$ have the same controllability indices, it is sufficient to prove that

$$\text{rank }(B,AB,A^2B,\dots,A^iB)=\text{rank }(B,(A+\lambda I)B,\dots,(A+\lambda I)^iB)$$

for any $1\le i\le n$. This follows immediately from the fact that $(A+\lambda I)^jB$ is a linear combination of $B,AB,\dots,A^jB$, for any j.

Finally, to see that (A,B) and $(A+\lambda I,B)$ have the same Segre symbol, it is sufficient to prove that

$$\begin{aligned}&\text{rank }(B,AB,\dots,A^{k_1-1}B,(A-\mu I)^j)\\&\quad=\text{rank }(B,(A+\lambda I)B,\dots,(A+\lambda I)^{k_1-1}B,((A+\lambda I)-(\mu+\lambda)I)^j)\end{aligned}$$

for any $1\le j\le n$. We argue as above to conclude the proof. □

7. Regularity of the BK-stratification ($m=1$). Finally, we are going to prove that the BK-stratification $\Sigma=\cup_\sigma E(\sigma)$ is Whitney regular (Theorem 7.5) when $m=1$. Again, as for the regularity of the BK-strata, the key point is the application of Lemma 5.1 to prove Lemma 7.1 which reduces the problem to the stratification induced by the intersection with the variety Γ in Theorem 5.2. But now this induced stratification is not a Segre stratification, so that we cannot conclude simply by using the results in [8]. Previously we studied the particular case of strata called simple (see Definition 7.2); since they have a particular homogeneity property (see Proposition 7.3), the regularity over them follows from the Whitney theorem (see Lemma 7.4). The general case can be viewed as a product of simple strata, because of the explicit descriptions in Proposition 5.3. In addition, we remark (see Proposition 7.6) that for any $m\ge 1$ Σ satisfies the frontier condition.

Given $(A,B)\in\mathcal{M}_{n,m}$ and Γ as in Lemma 5.1, let us consider the induced stratification $\cup_{k,\sigma}(E(k,\sigma)\cap\Gamma)$ of a sufficiently small neighborhood in Γ of (A,B), which we denote $\Sigma\cap\Gamma$. Notice that it is well defined because Γ is not only transversal to $\mathcal{O}(A,B)$ but also to every orbit sufficiently close to (A,B) (see Lemma 5.1) and hence to every BK-stratum sufficiently close to (A,B).

LEMMA 7.1. *With the above notation, Σ is Whitney regular over $E(A,B)$ at (A,B) if and only if $\Sigma\cap\Gamma$ is Whitney regular over $E(A,B)\cap\Gamma$ at (A,B).*

Proof. The necessity follows (see [8, I (1.4)]) from the transversality of Γ to every, stratum sufficiently close to (A,B), which we have remarked above. For the sufficiency, we again use Lemma 5.1:

$$E(k,\sigma)=\beta((E(k,\sigma)\cap\Gamma)\times\mathcal{O}(A,B))$$

locally in (A,B). Obviously, if $\cup_{k,\sigma}(E(k,\sigma)\cap\Gamma)$ is Whitney regular over $E(A,B)\cap\Gamma$ at (A,B), the product stratification $\cup_{k,\sigma}((E(k,\sigma)\cap\Gamma)\times\mathcal{O}(A,B))$ is also Whitney regular over $(E(A,B)\cap\Gamma)\times\mathcal{O}(A,B)$ at (A,B). By means of the above diffeomorphism β, the proof is finished. □

DEFINITION 7.2. *A pair of matrices $(A,B)\in\mathcal{M}_{n,m}$ is called* simple *if it has one eigenvalue, at most. A stratum $E(k,\sigma)$ is called* simple *if its elements are simple.*

The simple strata verify a particular homogeneity property, in some sense the converse to the one in Proposition 6.4.

PROPOSITION 7.3. *Let $E(k,\sigma)$ be a simple stratum. For any $(A,B),(A',B') \in E(k,\sigma)$, there exists a iffeomorphism f of $\mathcal{M}_{n,m}$, preserving strata, and such that $f(A,B)=(A',B')$.*

Proof. If $(k,\sigma)=(k)$, it is trivial. If $(k,\sigma)=(k,\sigma(1))$ or $(k,\sigma)=(\sigma(1))$, let λ,λ' be the eigenvalues of (A,B) and (A',B'), respectively. Then, because of Proposition 6.5, the pair $(A+(\lambda'-\lambda)I,B)$ is block similar to (A',B'). Hence, there exists $g_0 \in \mathcal{G}$ such that $(A',B')=\alpha(g_0,(A+(\lambda'-\lambda)I,B))$. It is straightforward that the mapping

$$f(X,Y)=\alpha(g_0,(X+(\lambda'-\lambda)I,Y))$$

verifies the desired conditions. □

The Whitney theorem states that any stratum of a constructible locally finite stratification has a Whitney regular point. Hence, any stratum $E(k,\sigma)$ has a point (A,B) such that Σ is Whitney regular over $E(k,\sigma)$ at (A,B). In the particular case where $E(k,\sigma)$ is simple, the above homogeneity property implies that all the points of $E(k,\sigma)$ are Whitney regular. Therefore, we have the following lemma, as a first step to tackle the main theorem.

LEMMA 7.4. *Σ is Whitney regular over any simple BK-stratum.*

THEOREM 7.5. *If $m=1$ the BK-stratification is Whitney regular.*

Proof. Let $(A,B) \in \mathcal{M}_{n,1}$, $\mathcal{O}(A,B)$ be its orbit, and $E(A,B)$ be its stratum. We shall prove that Σ is Whitney regular over $E(A,B)$ at (A,B). Let Γ be the linear variety defined in Theorem 5.2, and let us denote with the same symbol the neighborhood of (A,B) where the isomorphism in Lemma 5.1 holds. Then, according to Lemma 7.1, it is sufficient to prove that $\Sigma \cap \Gamma$ is Whitney regular over $E(A,B)\cap\Gamma$ at (A,B).

Because of Proposition 4.3, we can assume that (A,B) is a BK-matrix. First, we suppose that $B \neq 0$. We will discuss later the case $B=0$. Since $B\neq 0$ we have that

$$(A,B)=\left(\begin{pmatrix} N & 0 \\ 0 & J \end{pmatrix},\begin{pmatrix} E \\ 0 \end{pmatrix}\right), \qquad J=\operatorname{diag}(J_1,\ldots,J_\nu),$$

as in Definition 3.2. Remark that N has only one block and $E=\begin{pmatrix} 0 & \ldots & 0 & 1 \end{pmatrix}^t \in M_{s\times 1}$. Let us consider the pairs of matrices:

$$(A_i,B_i)=\left(\begin{pmatrix} N & 0 \\ 0 & J_i \end{pmatrix},\begin{pmatrix} E \\ 0 \end{pmatrix}\right) \in \mathcal{M}_{s+\delta_i,1}=M_{s+\delta_i}\times M_{(s+\delta_i)\times 1}$$

for $1\leq i\leq \nu$. Let $\Sigma_1,\ldots,\Sigma_\nu$ be the BK-stratifications of the respective spaces of matrices $\mathcal{M}_{s+\delta_1,1},\ldots,\mathcal{M}_{s+\delta_\nu,1}$. If we denote $E_i(A_i,B_i)$ the stratum of (A_i,B_i) in Σ_i, $0\leq i\leq\nu$, from Lemma 7.4, it follows that Σ_i is Whitney regular over $E_i(A_i,B_i)\cap\Gamma_i$ at (A_i,B_i), where $\Gamma_1,\ldots,\Gamma_\nu$ are the respective linear varieties defined as in Theorem 5.2:

$$\Gamma_i=(A_i,B_i)+\left\{\left(\begin{pmatrix} 0 & 0 \\ X_1^2(i) & X_2^2(i) \end{pmatrix},\begin{pmatrix} 0 \\ 0 \end{pmatrix}\right)\right\}, \quad 1\leq i\leq\nu.$$

As above, we denote with the same symbol Γ_i the neighborhoods of (A_i,B_i) where the corresponding isomorphisms in Lemma 5.1 hold, so that the induced stratifications $\Sigma_i\cap\Gamma_i$, $0\leq i\leq\nu$, are well defined. In fact, according to Lemma 7.1, $\Sigma_i\cap\Gamma_i$ is Whitney regular over $E_i(A_i,B_i)\cap\Gamma_i$ at (A_i,B_i), $0\leq i\leq\nu$.

Now, let us consider the diffeomorphism

$$\varphi : \Gamma_1 \times \cdots \times \Gamma_\nu \longrightarrow \Gamma$$

$$\varphi\left(\left(\begin{pmatrix} N & 0 \\ X_1 & H_1 \end{pmatrix}, \begin{pmatrix} E \\ 0 \end{pmatrix}\right) \cdots, \left(\begin{pmatrix} N & 0 \\ X_\nu & H_\nu \end{pmatrix}, \begin{pmatrix} E \\ 0 \end{pmatrix}\right)\right)$$
$$= \left(\begin{pmatrix} N & 0 & \dots & 0 \\ X_1 & H_1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ X_\nu & 0 & \dots & H_\nu \end{pmatrix}, \begin{pmatrix} E \\ 0 \\ \vdots \\ 0 \end{pmatrix}\right),$$

where $X_i = X_1^2(i)$ and $H_i = J_i + X_2^2(i)$, $1 \le i \le \nu$.

We know that $\Pi = (\Sigma_1 \cap \Gamma_1) \times \cdots \times (\Sigma_\nu \cap \Gamma_\nu)$ is a stratification which is Whitney regular over the stratum $(E_1(A_1, B_1) \cap \Gamma_1) \times \cdots \times (E_\nu(A_\nu, B_\nu) \cap \Gamma_\nu)$ at $((A_1, B_1), \dots, (A_\nu, B_\nu))$ [8, (1.2)]. Hence, because φ is a diffeomorphism, to conclude the proof it is sufficient to show that φ preserves strata locally at (A, B). That is to say, given two points

$$z = \left(\begin{pmatrix} N & 0 \\ X_i & H_i \end{pmatrix}, \begin{pmatrix} E \\ 0 \end{pmatrix}\right)_{1 \le i \le v}$$

and

$$z' = \left(\begin{pmatrix} N & 0 \\ X'_i & H'_i \end{pmatrix}, \begin{pmatrix} E \\ 0 \end{pmatrix}\right)_{1 \le i \le v}$$

belonging to the same stratum of Π, then the images $\varphi(z)$, $\varphi(z')$ belong to the same stratum of $\Sigma \cap \Gamma$, provided that they are sufficiently close to (A, B).

To prove this, we are going to see that the controllability indices and the Segre symbol of $\varphi(z)$ and $\varphi(z')$ are the same.

(i) In order to prove the first assertion it is sufficient to show that if we write $\varphi(z) = (C, D)$,

$$C = \begin{pmatrix} N & 0 & \dots & 0 \\ X_1 & H_1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ X_\nu & 0 & \dots & H_v \end{pmatrix}, \quad D = \begin{pmatrix} E \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

and $\varphi(z') = (C', D)$, then

$$\begin{aligned} r_j &= \text{rank } (D, CD, \dots, C^{j-1}D) \\ &= \text{rank } (D, C'D, \dots C'^{j-1}D) = r'_j, \quad 1 \le j \le n. \end{aligned}$$

A simple computation shows that $r_j = r'_j = j$ for $1 \le j \le s$. Thus, we can assume $j > s$.

If we denote Y_i the first column of the matrix X_i, $1 \le i \le \nu$, it is easy to check that

$$r_j = \text{rank} \begin{pmatrix} E & NE & \dots & N^{s-1}E & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & Y_1 & H_1Y_1 & \dots & H_1^\ell Y_1 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 & Y_\nu & H_\nu Y_\nu & \dots & H_\nu^\ell Y_\nu \end{pmatrix},$$

where $\ell = j - s - 1$.

We remark that, because z and z' belong to the same stratum, one has $Y_i = 0$ if and only if ${Y'}_i = 0$. Hence, if $Y_i = 0$ for $1 \leq i \leq \nu$ it is obvious that $r_j = {r'}_j$ for $s+1 \leq j \leq n$. Consequently, we assume now that $Y_i \neq 0$ for some i. Renumbering if necessary, we can suppose that $Y_i \neq 0$ (hence ${Y'}_i \neq 0$) for $1 \leq i \leq t$, $t \leq \nu$.

Since $r_j \leq r_{j+1} \leq r_j + 1$, in order to see that $r_j = {r'}_j$ for $s < j \leq n$ it is enough to prove that if j is the first index such that $r_j = r_{j-1}$, index j has also the same property with regard to ${r'}_j$.

Therefore, let $j = s + \ell + 1$ be the first index such that

$$r_j = r_{j-1}.$$

Then there exist a_k, $0 \leq k \leq \ell - 1$, satisfying

$$\begin{pmatrix} H_1^{\ell} Y_1 \\ \vdots \\ H_t^{\ell} Y_t \end{pmatrix} + a_{\ell-1} \begin{pmatrix} H_1^{\ell-1} Y_1 \\ \vdots \\ H_t^{\ell-1} Y_t \end{pmatrix} + \cdots + a_0 \begin{pmatrix} Y_1 \\ \vdots \\ Y_t \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

and

$$\psi(x) = a_0 + \cdots + a_{\ell-1} x^{\ell-1} + x^{\ell}$$

is the minimal polynomial of the vector

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_t \end{pmatrix}$$

with regard to the matrix H=diag$(H_1, \ldots, H_t)$.

Now, let $\varphi_i(x)$ be the minimal polynomial of H_i and $\psi_i(x)$ the minimal polynomial of Y_i, with regard to H_i, for $1 \leq i \leq t$. Making the transversal Γ_i still a bit smaller, if necessary, we can assume that g.c.d. $(\varphi_i(x), \varphi_k(x)) = 1$ for $i \neq k$. We are going to see that $\psi(x) = \psi_1(x) \ldots \psi_t(x)$. In fact, since $\varphi_i(H_i)(Y_i) = 0$ one has that $\psi_i(x) \mid \varphi_i(x)$ for $1 \leq i \leq t$, so that g.c.d$(\psi_i(x), \psi_k(x)) = 1$ if $i \neq k$. On the other hand, as $\psi(Y) = 0$, we have $\psi(Y_i) = 0$; hence, $\psi_i(x) \mid \psi(x)$ for $1 \leq i \leq t$. Consequently, $\psi_1(x) \ldots \psi_t(x) \mid \psi(x)$. Conversely, it is easy to check that $\psi_1(H) \ldots \psi_t(H)(Y) = 0$, so that $\psi(x) \mid \psi_1(x) \ldots \psi_t(x)$, and the claimed equality holds.

Next, if we denote $deg \psi_i(x) = m_i$ it is obvious that

$$\ell = m_1 + \cdots + m_t.$$

But m_i is the minimal integer such that

$$\begin{aligned} &\operatorname{rank} \begin{pmatrix} E & NE & \ldots & N^{s-1}E & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & Y_i & H_i Y_i & \ldots & H_i^{m_i} Y_i \end{pmatrix} \\ &\quad = \operatorname{rank} \begin{pmatrix} E & NE & \ldots & N^{s-1}E & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & Y_i & H_i Y_i & \ldots & H_i^{m_i-1} Y_i \end{pmatrix} = s + m_i. \end{aligned}$$

Hence, taking into account that $\left(\left(\begin{smallmatrix} N & 0 \\ X_i & H_i \end{smallmatrix}\right), \binom{E}{0}\right)$ and $\left(\left(\begin{smallmatrix} N & 0 \\ {X'}_i & {H'}_i \end{smallmatrix}\right), \binom{E}{0}\right)$ belong to the same stratum, it follows that the minimal polynomial of ${Y'}_i$ (first column of ${X'}_i$) has also

degree m_i, $i = 1 \dots t$. Therefore, the degree of the minimal polynomial of

$$Y' = \begin{pmatrix} {Y'}_1 \\ \vdots \\ {Y'}_t \end{pmatrix}$$

is $m_1 + \cdots + m_t = \ell$ and the first assertion above is proved.

(ii) We must now prove that the Segre symbols of $\varphi(z)$ and $\varphi(z')$ are the same.

First, we remark that μ is an eigenvalue of $\varphi(z) = (C, D)$ if and only if it is an eigenvalue of $\left(\left(\begin{smallmatrix} N & 0 \\ X_i & H_i \end{smallmatrix}\right), \left(\begin{smallmatrix} E \\ 0 \end{smallmatrix}\right)\right)$ for some $1 \le i \le \nu$. To see this, it is sufficient to bear in mind the characterization in Proposition 3.1: μ is an eigenvalue of $\varphi(z) = (C, D)$ if and only if rank $(C - \mu I, D) < n$; since rank $\left(\begin{pmatrix} N & 0 \end{pmatrix})(E)\right) = s$, it means that rank $\left(\begin{pmatrix} X_i & H_i - \mu I \end{pmatrix}(0)\right) < \delta_i$, for some $1 \le i \le \nu$; hence rank $\left(\left(\begin{smallmatrix} N-\mu I & 0 \\ X_i & H_i - \mu I \end{smallmatrix}\right)\left(\begin{smallmatrix} E \\ 0 \end{smallmatrix}\right)\right) < s + \delta_i$, so that μ is an eigenvalue of $\left(\left(\begin{smallmatrix} N & 0 \\ X_i & H_i \end{smallmatrix}\right)\left(\begin{smallmatrix} E \\ 0 \end{smallmatrix}\right)\right)$. The opposite is true also.

On the other hand, since $\left(\left(\begin{smallmatrix} N & 0 \\ X_i & H_i \end{smallmatrix}\right), \left(\begin{smallmatrix} E \\ 0 \end{smallmatrix}\right)\right)$, and $\left(\left(\begin{smallmatrix} N & 0 \\ X'_i & H'_i \end{smallmatrix}\right), \left(\begin{smallmatrix} E \\ 0 \end{smallmatrix}\right)\right)$ are in the same stratum, there is an eigenvalue μ' of the last pair with the same Segre characteristic as that of μ. As we have just seen, μ' is also an eigenvalue of $\varphi(z')$. Then, taking into account Proposition 3.1, it can be checked that μ and μ' have the same Segre characteristic, and the theorem is completely proved if $B \neq 0$.

Now we suppose that $B = 0$. In this case we have to make the following changes in the first part of the proof:

- $(A, B) = (J, 0)$,
- $(A_i, B_i) = (J_i, 0)$,
- $\Gamma_i = (J_i, 0) + \{(X_2^2(i), Y_2^2(i))\}$,
-

$$\varphi\left((H_1, Z_1), \dots (H_\nu, Z_\nu)\right) = \left(\begin{pmatrix} H_1 & & \\ & \ddots & \\ & & H_\nu \end{pmatrix}\begin{pmatrix} Z_1 \\ \vdots \\ Z_\nu \end{pmatrix}\right),$$

  where $Z_i = Y_2^2(i)$.
- $Z = ((H_1, Z_i))_{1 \le i \le \nu}$, $Z' = ((H'_1, Z'_i))_{1 \le i \le \nu}$.

Then, the proof of (i) is quite similar to the former case with

$$C = \begin{pmatrix} H_1 & & \\ & \ddots & \\ & & H_\nu \end{pmatrix}, \quad D = \begin{pmatrix} Z_1 \\ \vdots \\ Z_\nu \end{pmatrix}.$$

As for (ii), the reasoning is also the same: μ is an eigenvalue of (C, D) if and only if it is an eigenvalue of (H_i, Z_i) for some i, (as in the former case we have to bear in mind that (H_i, Z_i) belongs to a neighborhood of $(J_i, 0)$ and that $J_1, \dots, J_\nu$ have different eigenvalues). The theorem now is completely proved. □

We recall that a stratification Σ satisfies the *frontier condition* if for any strata $E, E' \in \Sigma$ such that $\overline{E} \cap E' = \emptyset$ then $\overline{E} \supset E'$. That is to say, the frontier of a stratum is a union of strata. It follows from (5.5), (4.4), and [9] that the BK-stratification satisfies the frontier condition, if $m = 1$. In fact, the Theorems 4.7 and 5.6 of [11] show that this condition holds for all values of m. Therefore, we have the following proposition.

PROPOSITION 7.6. *The BK-stratification verifies the frontier condition.*

REMARK 7.7. *Notice that the limitation $m = 1$ in the above theorem is only used in part* (i) *of the proof of Theorem* (7.5). *We conjecture that the theorem is also true for $m > 1$, but in this case our approach should be modified because Y is not a column matrix. Unfortunately, until now we have not succeeded in finding an alternative approach which works in the general case.*

8. Bifurcation diagrams. The space $\mathcal{M}_{n,1}$ is equipped with a Whitney stratification, so that we can make use of the Thom transversality theorem: let W be a d-dimensional manifold, and $C^\infty(W, \mathcal{M}_{n,1})$ the space of the W-parametrized differentiable families of pairs of matrices of $\mathcal{M}_{n,1}$; that is to say, the differentiable maps $\varphi : W \longrightarrow \mathcal{M}_{n,1}$; then the subset formed by the ones transversal to the BK-stratification is open and dense. These families will be called "generic."

Let $\varphi : W \longrightarrow \mathcal{M}_{n,1}$ be a generic family. For all $E(k,\sigma) \subset \mathcal{M}_{n,1}$, $\varphi^{-1}(E(k,\sigma))$ is a submanifold of W and

$$\operatorname{codim} \varphi^{-1}(E(k,\sigma)) = \operatorname{codim} E(k,\sigma).$$

Moreover, $\cup\varphi^{-1}(E(k,\sigma))$ is a Whitney regular stratification for W, called the *bifurcation diagram* of the (generic) family.

The local description at $w \in W$ of the bifurcation diagram can be derived from the one of a versal deformation of $(A,B) = \varphi(w)$. Because of Proposition 4.3, we can assume (A,B) to be a BK-matrix and then consider its versal deformation Γ in Theorem 5.2.

REMARK 8.1. *Notice that the linear subvariety of Γ defined by trace $X_i = 0$, for all diagonal blocks X_i of X_2^2, is transversal to $E(A,B)$ at (A,B) (but not to $\mathcal{O}(A,B)$).*

Example 8.2. As an example, we enumerate the singularities of bifurcation diagrams of few-parameter generic families with $B \neq 0$.

For the commodity of the reader we write down the codimension of $E(A,B)$ for $m = 1$ (see Propositions 4.2 and 6.3):

$$\operatorname{codim} E(A,B) = n - k_1 + \sum_{1 \le i \le \nu} (\sigma_1(i) + 3\sigma_2(i) + 5\sigma_3(i) + \cdots) - \nu.$$

Notice that $\operatorname{codim} E(A,B) = 0$ if and only if $n = k_1$, which corresponds to the open and dense subset of $\mathcal{M}_{n-1}$ formed by the pairs (A,B) equivalents to the controllable pair

$$\left(\begin{pmatrix} 0 & I_{n-1} \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right). \tag{8.1}$$

Usually, it is also called bifurcation diagram of φ the variety $\cup\varphi^{-1}(E(k,\sigma))$ with $\operatorname{codim} E(A,B) > 0$.

(i) *One-parameter families.* The only stratum of codimension one in the space $\mathcal{M}_{n,1}$ is the one formed by the pairs of matrices equivalent to

$$(A,B) = \left(\begin{pmatrix} N & 0 \\ 0 & \alpha \end{pmatrix}, \begin{pmatrix} E \\ 0 \end{pmatrix} \right), \tag{8.2}$$

where N is the nilpotent $(n-1)$-matrix, and $\alpha \in \mathbb{C}$. A transversal family is (see Remark 8.1)

$$\left(\begin{pmatrix} N & 0 \\ X & \alpha \end{pmatrix}, \begin{pmatrix} E \\ 0 \end{pmatrix} \right) \tag{8.3}$$

with $X = \begin{pmatrix} x & 0 & \dots & 0 \end{pmatrix} \in M_{1\times(n-1)}(\mathbb{C})$. One checks easily that for $x \neq 0$ the above pair (8.3) is equivalent to the pair (8.1). Therefore, the bifurcation diagram for a generic 1-parameter family is $x = \{0\}$.

Notice that x is a complex number so that $\mathbb{C} - \{0\}$ is connected.

Therefore, for a generic 1-parameter family there are only pairs equivalents to (8.1) and for an isolated value of the parameter a pair equivalent to (8.2).

(ii) *Two-parameter families.* The only stratum of codimension two is the one formed by the pairs of matrices equivalent to

$$\left(\begin{pmatrix} N & 0 & 0 \\ 0 & \alpha & 0 \\ 0 & 0 & \beta \end{pmatrix}, \begin{pmatrix} E \\ 0 \\ 0 \end{pmatrix} \right), \tag{8.4}$$

where N is the nilpotent $(n-2)$-matrix, and $\alpha, \beta \in \mathbb{C}$. A transversal family is

$$\left(\begin{pmatrix} N & 0 & 0 \\ X & \alpha & 0 \\ Y & 0 & \beta \end{pmatrix}, \begin{pmatrix} E \\ 0 \\ 0 \end{pmatrix} \right) \tag{8.5}$$

with $X = \begin{pmatrix} x & 0 & \dots & 0 \end{pmatrix} \in M_{1\times(n-2)}(\mathbb{C})$, $Y = \begin{pmatrix} y & 0 & \dots & 0 \end{pmatrix} \in M_{1\times(n-2)}(\mathbb{C})$. Then, it is easily seen that for a pair of (8.5) there are only the following possibilities: if $x \neq 0$ and $y \neq 0$, (8.5) is equivalent to (8.1); if $xy = 0$ but $x \neq 0$ or $y \neq 0$, (8.5) is equivalent to (8.2). Therefore, the bifurcation diagram for a generic 2-parameter family is the usual stratification associated with $xy = 0$; that is to say,

$$\{x = y = 0\} \cup \{xy = 0\,,\ x \neq 0 \text{ or } y \neq 0\} \cup \{x \neq 0\,,\ y \neq 0\}.$$

Therefore, for a generic 2-parameter family there are only three possibilities:

- The pairs corresponding to points outside the axes are equivalent to (8.1).
- To the points of the axes, except 0, correspond pairs equivalent to (8.2).
- To the origin 0 corresponds a pair equivalent to (8.4).

(iii) *Three-parameter families.* There are two strata of codimension three:

(a) the one corresponding to

$$\left(\begin{pmatrix} N & 0 & 0 & 0 \\ 0 & \alpha & 0 & 0 \\ 0 & 0 & \beta & 0 \\ 0 & 0 & 0 & \gamma \end{pmatrix}, \begin{pmatrix} E \\ 0 \\ 0 \\ 0 \end{pmatrix} \right), \tag{8.6}$$

where N is the nilpotent $(n-3)$-matrix, and $\alpha, \beta, \gamma \in \mathbb{C}$;

(b) the one corresponding to

$$\left(\begin{pmatrix} N & 0 & 0 \\ 0 & \alpha & 1 \\ 0 & 0 & \alpha \end{pmatrix}, \begin{pmatrix} E \\ 0 \\ 0 \end{pmatrix} \right), \tag{8.7}$$

where N is the nilpotent $(n-2)$-matrix, and $\alpha, \beta \in \mathbb{C}$.

A transversal family is, in the first case,

$$\left(\begin{pmatrix} N & 0 & 0 & 0 \\ X & \alpha & 0 & 0 \\ Y & 0 & \beta & 0 \\ Z & 0 & 0 & \gamma \end{pmatrix}, \begin{pmatrix} E \\ 0 \\ 0 \\ 0 \end{pmatrix} \right) \tag{8.8}$$

with $X = (x\ \ 0\ \ \dots\ \ 0) \in M_{1\times(n-3)}(\mathbb{C})$, $Y = (y\ \ 0\ \ \dots\ \ 0) \in M_{1\times(n-3)}(\mathbb{C})$, $Z = (z\ \ 0\ \ \dots\ \ 0) \in M_{1\times(n-3)}(\mathbb{C})$.

For the second case,

$$\left(\begin{pmatrix} N & 0 & 0 \\ X & \alpha & 1 \\ Y & z & \alpha \end{pmatrix}, \begin{pmatrix} E \\ 0 \\ 0 \end{pmatrix}\right) \tag{8.9}$$

with $X = \begin{pmatrix} x & 0 & \dots & 0 \end{pmatrix} \in M_{1\times(n-2)}(\mathbb{C})$, $Y = \begin{pmatrix} y & 0 & \dots & 0 \end{pmatrix} \in M_{1\times(n-2)}(\mathbb{C})$.

We have to discuss each one of these cases when a pair is equivalent to (8.1), (8.2), (8.4), (8.6), or (8.7). As in the former cases we compute the controllability matrix, which is, in the first case

$$\begin{pmatrix} 0 & & & 0 & & \\ \vdots & 0 & \dots & 1 & 0 & 0 \\ & & & \vdots & & \\ 0 & 1 & & & & \\ 1 & 0 & & & & \\ 0 & 0 & & 0 & x & \alpha x \\ 0 & 0 & & 0 & y & \beta y \\ 0 & 0 & & 0 & z & \gamma \end{pmatrix}$$

and in the second case

$$\begin{pmatrix} 0 & & & 0 & & \\ \vdots & 0 & \dots & 1 & 0 & 0 \\ & & & \vdots & & \\ 0 & 1 & & & & \\ 1 & 0 & & & & 0 \\ 0 & 0 & & 0 & x & \alpha x + y \\ 0 & 0 & & 0 & y & zx + \alpha y \end{pmatrix}.$$

Then two bifurcation diagrams are possible for a generic three-parameter family, corresponding to

(a) $xyz = 0$,

(b) $x^2 z - y^2 = 0$.

Notice that the last one is the well-known umbrella singularity introduced by Whitney.

These diagrams are stratified in the following way:

$$\begin{aligned} \text{(a)}\ & \underbrace{\{x = y = z = 0\}}_{(\alpha)} \cup \underbrace{\{x = y = 0, z = 0\} \cup \{x = z = 0, y = 0\} \cup \{y = z = 0, x = 0\}}_{(\beta)} \\ & \cup \underbrace{\{x = 0, y = 0, z = 0\} \cup \{y = 0, x = 0, x = 0\} \cup \{z = 0, x = 0, y = 0\}}_{(\gamma)} \\ & \cup \underbrace{\{x = 0, y = 0, z = 0\}}_{(\delta)}. \end{aligned}$$

$$(b)\ \underbrace{\{x=y=z=0\}}_{(\alpha)} \cup \underbrace{\{zx^2-y^2=0, x=0 \text{ or } y=0\}}_{(\beta)} \cup \underbrace{\{x=y=0, z=0\}}_{(\gamma)} \cup \underbrace{\{zx^2-y^2=0\}}_{(\delta)}.$$

To each point of the stratum α, β, γ, δ corresponds, equivalently, a pair equivalent to (8.1), (8.2), (8.4), or (8.6) for (a) or (8.1), (8.2), (8.4), or (8.7) for (b).

Therefore, a generic three-parameter family only contains pairs equivalent to the ones mentioned above. Notice, moreover, that the bifurcation diagram tells us how these pairs are distributed.

REMARK 8.3.

(i) *We know that the Whitney regular condition implies that a small perturbation of a generic parametrized family will not affect the structure of the family. We think that it would be very interesting to deduce from the stratification a numerical knowledge of how far the family is from a change of structure.*

(ii) *It is also worth mentioning that a generic d-parameter family cannot contain any structures of codimension $d+1$ or higher, but it may not contain any structures of codimension d either if the family is overparametrized. For example, the two-parameter family (clearly overparametrized)*

$$\left(\begin{pmatrix} 0 & 0 \\ x+y & \alpha \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix}\right)$$

only contains structures of codimension 0 or 1.

(iii) *There are several reasons to include the local finiteness condition in the definition of a stratification. First of all, if we remove this condition, we do not know if the fundamental theorems of Thom about stratification, such as transversality, isotopy, and so on, hold. Besides, notice that without that condition the bifurcation diagram would have infinitely many components so that when we perturb a family we can get infinitely many possibilities. This is poor information. Also, if the union of strata of codimension d is a manifold of codimension $d-1$, the generic families would not be transversal to the strata. Consider, for example, a disc in $\mathbb{C}$ which is the union of all its interior points.*

Acknowledgments. We are very grateful to the referees for their valuable suggestions and comments.

REFERENCES

[1] V.I. ARNOLD, *On matrices depending on parameters*, Uspekhi Mat. Nauk, 26 (1971), pp. 101–114.

[2] J.M. BERG AND H.G. KWANTNY, *A canonical parameterization of the Kronecker form of a matrix pencil*, Automatica J. IFAC, 31 (1995), pp. 669–680.

[3] J.M. BERG AND H.G. KWANTNY, *Unfolding the zero structure of a linear control system*, Linear Algebra Appl., 258 (1997), pp. 19–39.

[4] A. EDELMAN, E. ELMROTH, AND B. KÅGSTRÖM, *A geometric approach to perturbation theory of matrices and matrix pencils. Part* I: *Versal deformations*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 653–692.

[5] J. FERRER, M.I. GARCÍA, AND F. PUERTA, *Brunovsky local form of a holomorphic family of pairs of matrices*, Linear Algebra Appl., 253 (1997), pp. 175–198.

[6] J. FERRER AND F. PUERTA, *Similarity of non-everywhere defined linear maps*, Linear Algebra Appl., 168 (1992), pp. 27–55.

[7] S. Friedland, *Canonical forms*, in Frequency Domain and State Space Methods for Linear Systems, C.I. Byrnes and A. Linquist, eds., North–Holland, Amsterdam, New York, 1986, pp. 115–121.
[8] C.G. Gibson, *Regularity of the Segre stratification*, Math. Proc. Cambridge Philos. Soc., 80 (1976), pp. 91–97.
[9] C.G. Gibson, K. Wirthmüller, A.A. du Plessis, and E.J.N. Looijenga, *Topological Stability of Smooth Mappings*, Springer-Verlag, Berlin, New York, 1976.
[10] I. Goberg, P. Lancaster, and L. Rodman, *Invariant Subspaces of Matrices with Applications*, John Wiley, New York, 1986.
[11] J.M. Gracia, I. de Hoyos, and I. Zaballa, *Perturbation of linear control systems*, Linear Algebra Appl. 121 (1989), pp. 353–383.
[12] V. Guillemin and A. Pollack, *Differential Topology*, Prentice–Hall, Englewood Cliffs, NJ, 1974.
[13] M. Hazewinkel, *(Fine) moduli (spaces) for linear systems: What are they and what are they good for?*, in Geometrical Methods for the Theory of Linear Systems, D. Reidel, Dordrecht, Boston, 1980, pp. 125–193.
[14] U. Helmke, *The topology of a moduli space for linear dynamical systems*, Comment. Math. Helv., 60 (1985), pp. 630–655.
[15] A. Tannenbaum, *Invariance and System Theory: Algebraic and Geometric Aspects*, Lecture Notes in Math. 845, Springer-Verlag, Berlin, New York, 1981.
[16] K. Tchon, *Bifurcation diagrams for feedback families of linear system*, Systems Control Lett., 5 (1985), pp. 397–401.
[17] W.W. Udilov, *The method of invariant manifolds in control of dynamical systems*, Kibernet. i Vychisl. Tekhn., 47 (1980), pp. 10–24.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 3, pp. 743–767
© 2000 Society for Industrial and Applied Mathematics

A STATE-SPACE APPROACH TO INDEFINITE SPECTRAL FACTORIZATION*

D. J. CLEMENTS†

Abstract. We use state-space methods to show that any proper, rational indefinite spectral density invertible at infinity has a proper, rational, stable, and minimum phase spectral factor of some degree, and we write down a formula for such factors.

Key words. spectral factorization, rational matrices

AMS subject classifications. 15A22, 15A23, 47A68

PII. S0895479897324469

1. Introduction. For an $m \times n$ rational matrix R, we define the Hermitian conjugate R^* to be the $n \times m$ rational matrix formed by taking the transpose of R and replacing all coefficients of the rational function entries with their complex conjugates. The para-Hermitian conjugate R_* is then defined by

$$R_*(s) = R^*(-s).$$

A square, rational matrix Φ is said to be Hermitian if $\Phi^* = \Phi$ and para-Hermitian if $\Phi_* = \Phi$.

It is well known that an $n \times n$ para-Hermitian matrix Φ with rank r has a factorization of the form

$$\Phi(s) = G_*(s)G(s)$$

for some $r \times n$ rational matrix G if and only if Φ is positive semidefinite at almost every point on the imaginary axis. This basic result is known as spectral factorization in the engineering literature, the name originally deriving from the fact that the spectral density of a stationary, random process is a proper, rational, para-Hermitian matrix Φ, positive semidefinite on the imaginary axis. A standard reference for spectral factorization for rational matrices is [26].

The *inertia* $i\{H\}$ of an Hermitian matrix H is the triple $\{p, n, z\}$ of the number of positive, negative, and zero eigenvalues, respectively. A para-Hermitian Φ is Hermitian at every point on the imaginary axis that is not a pole of Φ. We say that a para-Hermitian matrix Φ has *constant inertia* if it has the same inertia at almost every point on the imaginary axis. In particular, it is clear that a para-Hermitian matrix Φ has constant inertia if it is positive semidefinite at almost every point on the imaginary axis.

We shall call the rational matrix Φ a *spectral density* if it is para-Hermitian and has constant inertia $\{p, n, z\}$, even if it is indefinite on the imaginary axis, and we shall say that Φ has a (right) *spectral factor* G if

$$\Phi(s) = G_*(s)JG(s), \tag{1.1}$$

*Received by the editors July 14, 1997; accepted for publication (in revised form) by A. C. M. Ran April 29, 1999; published electronically February 10, 2000. This work was completed while the author was visiting the Mathematisches Institut, Universität Würzburg, Würzburg, Germany.
http://www.siam.org/journals/simax/21-3/32446.html
†School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney 2052, Australia (D.Clements@unsw.edu.au)

where J is an invertible, constant, Hermitian matrix with inertia $\{p, n, 0\}$. Of course, we have the standard spectral density and spectral factorization problem if J is positive definite, that is, $n = 0$. There are various constraints that one may impose on the possible factors G to reduce the nonuniqueness. For example, one may require G to be both *stable* (i.e., analytic in the open right half-plane) and *minimum phase* (i.e., full row rank everywhere in the open right half-plane).

The basic result for indefinite spectral factorization is of fairly recent origin [10, 21]. In its simplest form, the result can be stated as the following theorem.

THEOREM 1.1. *A rational matrix Φ is a spectral density if and only if it can be spectrally factored as in* (1.1).

Preliminary work on factorization of indefinite spectral densities appears to be due to [6, 15, 16], but the result as stated in the Theorem 1.1 was first proved for invertible polynomial matrices in [10]. The extension to arbitrary rational matrices is straightforward and is written down in the introduction to [21]. A different and simpler proof of the theorem is given in [7].

The work reported in this paper is restricted to the class of spectral densities which are proper and invertible at infinity. It is a combination of previous work of the author [5], in which the most general definite spectral factorization is studied using Hermitian pencils; a previous attempt by the author contained in the unpublished report [4]; and the paper [21], in which a class of indefinite spectral densities, which have spectral factors which are minimal at all points except one, are studied.

In [21], formulas for part of the almost minimal factors are presented, but the formulas for the factors associated with the exceptional point are missing. Moreover, the results in [21] are restricted to spectral densities for which the minimal realization is completely separable. In this paper, we give a formula for spectral factors which includes the exceptional point and is valid for any indefinite spectral density, proper and invertible at infinity. For the class of spectral densities studied in [21], our formula has the same minimality properties.

In the recent engineering literature on H_∞-control theory, indefinite spectral factorization is known as J-spectral factorization. For example, in the invertible H^∞-control problem [8], the design essentially reduces to the solution of two invertible J-spectral factorization problems, one for the state-feedback gain and one for the state-estimator gain. A direct J-spectral approach to this problem is investigated in [12, 13], whereas an alternative polynomial approach [18] reduces to a pair of polynomial J-spectral factorizations.

In these applications, the spectral density to be factored is usually proper with no zeros on the imaginary axis and is invertible at infinity. Moreover, these problems are usually canonical, i.e., the spectral factor G in (1.1) has full row rank and is analytic in the closed right half-plane. The more difficult noncanonical case (having to relax the full rank and analyticity requirements on G to the open right half-plane) is usually ruled out by the problem assumptions. The engineering literature on the general J-spectral factorization is limited. Other than [15] the main sources appear to be [19, 22, 23], which all focus on the general polynomial J-spectral factorization problem but again, for applications, the authors are only interested in algorithms for canonical factorization. One paper that looks at the H_∞ problem in a general context is [25], where it is shown that a polynomial J-spectral factorization is an essential element.

Indefinite spectral factorization is simply a special case of the more general rational matrix factorization problems. There is a substantial literature on the state-space

approach to factorization problems in general and spectral factorization problems in particular. Some relevant papers of interest here include [2, 5, 20], while [1, 9, 11] contain a wealth of material.

The organization of this paper is as follows. In section 2, we discuss a simple example to illustrate some of the differences between indefinite spectral factorization and the special case of indefinite spectral factorization. In section 3, we revise some material on realizations of para-Hermitian matrices and introduce Hermitian matrices. In section 4, the canonical form for Hermitian pencils is discussed together with some results on the inertia of Hermitian matrices on the real axis. Section 5 presents a basic result for spectral densities in terms of the pole and zero pencils while, in section 6, some structural results relating minimal, structured realizations are developed. The main result of the paper is presented in section 7. Section 8 contains some discussion of the results. Finally, a number of proofs are contained in the appendix.

2. A simple example. Suppose that Φ is a definite, proper, rational spectral density with degree $2n$. Then we know that every spectral factor is proper and that there exists a stable, minimum phase spectral factor with degree n. Neither of these statements is true for an indefinite spectral density. The simple example that follows highlights the essential differences between indefinite spectral factorization and the more familiar definite spectral factorization.

Example 2.1. Consider the degree 2 spectral density Φ defined by

$$\Phi(s) = \begin{bmatrix} 0 & \frac{-s-b}{-s} \\ \frac{s-b}{s} & \frac{2}{s^2} \end{bmatrix},$$

where b is a real parameter. It is clear that $\Phi(\jmath\omega)$ has one positive and one negative eigenvalue for each nonzero value ω, so Φ is an indefinite spectral density. For b nonzero, one spectral factorization of Φ is

$$\Phi(s) = G_*(s)JG(s) = \begin{bmatrix} 1 & 0 \\ \frac{1}{bs} & \frac{s-b}{s} \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & -\frac{1}{bs} \\ 0 & \frac{-s-b}{-s} \end{bmatrix},$$

where G has degree 1. For $b > 0$, G is stable and minimum phase, but for $b < 0$, G is no longer minimum phase, though it remains stable. Indeed, it can be shown that there does not exist a stable, minimum phase spectral factor of degree 1. An alternative spectral factorization of Φ is

$$\Phi(s) = \tilde{G}_*(s)J\tilde{G}(s) = \begin{bmatrix} \frac{-s-b}{-s} & 0 \\ \frac{1}{s^2} & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{s-b}{s} & \frac{1}{s^2} \\ 0 & 1 \end{bmatrix},$$

where $\tilde{G}$ has degree 2, but now, for $b \leq 0$, $\tilde{G}$ is both stable and minimum phase.

In summary, for $b \neq 0$, Φ has a spectral factor of degree 1, but this spectral factor is both stable and minimum phase only for $b < 0$. For $b = 0$, there does not exist any spectral factor with degree 1. For $b \leq 0$, the smallest stable, minimum phase spectral factor has degree 2.

One further observation is that there even exists an improper, stable, and minimum phase spectral factor $\hat{G}$ with degree 2 for this example when $b < 0$:

$$\Phi(s) = \hat{G}_*(s)J\hat{G}(s) = \begin{bmatrix} -s-b & 0 \\ -\frac{1}{s} & \frac{1}{s} \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} s-b & \frac{1}{s} \\ 0 & -\frac{1}{s} \end{bmatrix}.$$

This contrasts with the case of definite factorization where proper, definite spectral densities can have only proper spectral factors.

3. Realizations. In this section, we discuss some aspects of realizations of rational matrices. It is well known that any rational matrix $R(s)$ can be represented (realized) as

$$R(s) = D + C(sE - A)^{-1}B$$

for certain appropriately sized matrices A, B, C, D, and E. It is now fairly standard to indicate the underlying system for which $R(s)$ is the transfer function with the notation

$$R(s) = \left[\begin{array}{c|c} A - sE & B \\ \hline C & D \end{array}\right].$$

We shall use this notation throughout this paper.

The indefinite spectral factorization problem as stated in the introduction is equivalent to a symmetric factorization problem [5]. We note that Φ is para-Hermitian if and only if Ψ, defined by $\Psi(s) = \Phi(\imath s)$, is Hermitian. This transformation rotates the s-plane through $-\pi/2$. Consequently, the imaginary axis, the open left half-plane, and the open right half-plane map to the real axis, the open upper half-plane, and the open lower half-plane, respectively. We shall also transfer the use of the words stable, minimum phase, inertia, spectral density, and spectral factor in an obvious manner.

We shall call the rational matrix Ψ a *spectral density* if it is Hermitian and has constant inertia $\{p, n, z\}$ on the real axis, even if it is indefinite, and we shall say that Ψ has a (right) *spectral factor* G if

$$\Psi(s) = G^*(s)JG(s), \tag{3.1}$$

where J is an invertible, constant, Hermitian matrix with inertia $\{p, n, 0\}$.

For rational matrices which are either para-Hermitian $\Phi_* = \Phi$ or Hermitian $\Psi^* = \Psi$, it is always possible to choose a state-space realization which reflects the symmetry of the rational matrix. Although in this paper we strictly need only the proper case, which is discussed in detail in [5], we state the general result for the Hermitian case here. There is an obvious corresponding result for the para-Hermitian case.

LEMMA 3.1. *Let Ψ be a (not necessarily proper) rational, Hermitian matrix. Then there exists a realization*

$$\Psi(s) = \left[\begin{array}{c|c} \mathcal{A} - s\mathcal{E} & \mathcal{B} \\ \hline \mathcal{B}^* & J \end{array}\right] \tag{3.2}$$

with $J^* = J$, $\mathcal{E}^* = \mathcal{E}$, and $\mathcal{A}^* = \mathcal{A}$. The realization is irreducible if

(i) *there are no uncontrollable finite or infinite modes, i.e., $\left[\begin{array}{cc} t\mathcal{A} - s\mathcal{E} & \mathcal{B} \end{array}\right]$ has full row rank for all s and t not both zero, and*

(ii) *there are no nondynamic modes, i.e., $\mathcal{A}\ker(\mathcal{E}) \subseteq \operatorname{im}(\mathcal{E})$.*

The realization is minimal if and only if it is irreducible.

Proof. See Appendix A.1. ∎

The realization (3.2) of Ψ has pole pencil

$$P(s) = \mathcal{A} - s\mathcal{E}$$

and zero pencil

$$Z(s) = \left[\begin{array}{cc} \mathcal{A} - s\mathcal{E} & \mathcal{B} \\ \mathcal{B}^* & J \end{array}\right].$$

It is particularly important to note that these pencils are also both Hermitian since our factorization method is based on a detailed understanding of the structure of these two Hermitian pencils.

We also note that if Ψ is proper, then $\mathcal{E}$ can be chosen to be invertible and is, in fact, guaranteed to be so for a minimal realization.

4. Hermitian pencils. Essential to our development are the inertia (equivalently, signature) properties of Hermitian pencils along the real axis. Some of these results are more general than strictly needed for this paper, but we include them for completeness. First, however, we state a standard result on the (Jordan) canonical form for Hermitian pencils [24]. Let O_m denote the $m \times m$ zero matrix, let S_m denote the $m \times m$ matrix with entries $\delta_{i,m+1-j}$, and let T_m denote the $m \times m$ matrix with entries $\delta_{i,m+2-j}$, where $\delta_{i,j}$ is the Kronecker delta function. Let $\mathcal{E}$ and $\mathcal{A}$ be a pair of $n \times n$ Hermitian matrices. Then there exists an $n \times n$ invertible matrix D such that

$$D^*(\mathcal{A} - s\mathcal{E})D = \bigoplus_{i=1}^{p} \mathcal{B}_i(s),$$

where each $\mathcal{B}_i$ has one of the four following canonical forms:

Type 1:

$$\mathcal{B}_i(s) = \epsilon_i^f \left[(a_i - s)S_{m_i} + T_{m_i}\right]$$

is an $m_i \times m_i$ block, with sign $\epsilon_i^f = \pm 1$, a_i is real, and m_i is a positive integer.

Type 2:

$$\mathcal{B}_i(s) = \begin{bmatrix} O_{m_i} & (a_i - s)S_{m_i} + T_{m_i} \\ (\bar{a}_i - s)S_{m_i} + T_{m_i} & O_{m_i} \end{bmatrix}$$

is a $2m_i \times 2m_i$ block, with $\bar{a}_i \neq a_i$, and m_i is a positive integer.

Type 3:

$$\mathcal{B}_i(s) = \epsilon_i^\infty \left[S_{m_i} - sT_{m_i}\right]$$

is an $m_i \times m_i$ block, with sign $\epsilon_i^\infty = \pm 1$, and m_i is a positive integer.

Type 4:

$$\mathcal{B}_i(s) = \left[\hat{S}_{2m_i+1} - sT_{2m_i+1}\right],$$

where $\hat{S}_{2m_i+1}$ is simply S_{2m_i+1} with the entry in the (m_i+1, m_i+1) position changed from 1 to 0, and m_i is a nonnegative integer.

The inertia of an Hermitian matrix H has two simple but useful properties. First, the inertia is invariant under congruence, and second, the inertia of a direct sum of Hermitian matrices is the sum of the inertias, regarded as three-dimensional vectors.

Table 4.1 summarizes the properties of inertias of the fundamental blocks in the canonical form as a function of real, finite values s distinct from the real eigenvalues. The cases for m even and m odd are treated separately. (In the table, N.A. means "not applicable" and $i\{\epsilon^f(a-s)\}$ refers to the inertia of the scalar matrix (i.e., the real number) $\epsilon^f(a-s)$, and likewise for $i\{\epsilon^\infty\}$.)

We need some additional notation. For the Hermitian pencil $\mathcal{A} - s\mathcal{E}$, we denote the set of real eigenvalues by Λ. For each $a \in \Lambda$, $\mathrm{Odd}(a)$ is the set of indices of the corresponding (Type 1) odd-sized blocks in the canonical form. The sets of indices of the even- and odd-sized Type 3 blocks in the canonical form are denoted by Even_∞ and Odd_∞, respectively. Further, we say that

TABLE 4.1
Generic inertias on the finite real axis.

$\mathcal{A}-s\mathcal{E}$	Type 1	Type 2	Type 3	Type 4
$m=2n$	$\left\{\begin{array}{c} n \\ n \\ 0 \end{array}\right\}$	$\left\{\begin{array}{c} n \\ n \\ 0 \end{array}\right\}$	$\left\{\begin{array}{c} n \\ n \\ 0 \end{array}\right\}$	N.A.
$m=2n+1$	$i\{\epsilon^f(a-s)\}+\left\{\begin{array}{c} n \\ n \\ 0 \end{array}\right\}$	N.A.	$i\{\epsilon^\infty\}+\left\{\begin{array}{c} n \\ n \\ 0 \end{array}\right\}$	$\left\{\begin{array}{c} n \\ n \\ 1 \end{array}\right\}$

(i) a Hermitian pencil $\mathcal{A}-s\mathcal{E}$ has *constant inertia* if it has the same inertia at all but a finite number of points on the real axis,

(ii) a Hermitian matrix has *symmetric inertia* if the numbers of positive and negative eigenvalues are equal,

(iii) a Hermitian pencil $\mathcal{A}-s\mathcal{E}$ has *symmetric inertia* if it has symmetric inertia at all but a finite number of points on the real axis, and

(iv) a Hermitian pencil $\mathcal{A}-s\mathcal{E}$ has *symmetric inertia at infinity* if the Hermitian matrix $\mathcal{E}$ has symmetric inertia.

The next three lemmas are central to a number of the subsequent results. The first lemma, Lemma 4.1, says that a pencil has constant inertia if and only if, for each real eigenvalue, the number of Type 1 blocks with odd size and positive sign equals the number of Type 1 blocks with odd size and negative sign. The second lemma, Lemma 4.2, says that a pencil has symmetric inertia if the pencil has constant inertia and the number of Type 3 blocks with odd size and positive sign equals the number of Type 3 blocks with odd size and negative sign, while the third lemma, Lemma 4.3, gives a condition for a pencil to have symmetric inertia at infinity.

LEMMA 4.1. *A Hermitian pencil $\mathcal{A}-s\mathcal{E}$ has constant inertia if and only if*

$$\sum_{i\in\mathrm{Odd}(a)} \epsilon_i^f = 0 \quad \text{for each } a\in\Lambda.$$

Proof. From Table 4.1, the inertia of each block of Type 2, 3, or 4 is independent of real s. Thus, the inertia of the pencil $\mathcal{A}-s\mathcal{E}$ is constant if and only if the contribution to the inertia from the Type 1 blocks is constant. In the neighborhood of a real eigenvalue a, the Type 1 blocks not corresponding to a have constant inertia. Thus, we only need look at the change in inertia in the neighborhood of a real eigenvalue a.

It is clear from Table 4.1 that the change in the inertia as s varies from just less than a to just greater than a only depends on the blocks with odd size and equals

$$\sum_{i\in\mathrm{Odd}(a)} \left[i\{\epsilon_i^f\}-i\{-\epsilon_i^f\}\right],$$

which is zero if and only if

$$\sum_{i\in\mathrm{Odd}(a)} \epsilon_i^f = 0. \qquad \square$$

TABLE 4.2
Inertias of blocks at infinity.

$\mathcal{E}$	Type 1	Type 2	Type 3	Type 4
$m = 2n$	$\left\{ \begin{array}{c} n \\ n \\ 0 \end{array} \right\}$	$\left\{ \begin{array}{c} n \\ n \\ 0 \end{array} \right\}$	$i\{\epsilon^\infty\} + \left\{ \begin{array}{c} n-1 \\ n-1 \\ 1 \end{array} \right\}$	N.A.
$m = 2n+1$	$i\{\epsilon^f\} + \left\{ \begin{array}{c} n \\ n \\ 0 \end{array} \right\}$	N.A.	$\left\{ \begin{array}{c} n \\ n \\ 1 \end{array} \right\}$	$\left\{ \begin{array}{c} n \\ n \\ 1 \end{array} \right\}$

LEMMA 4.2. *A Hermitian pencil $\mathcal{A} - s\mathcal{E}$ has symmetric inertia if and only if*

$$\sum_{i \in \mathrm{Odd}(a)} \epsilon_i^f = 0 \quad \text{for each } a \in \Lambda$$

and

$$\sum_{i \in \mathrm{Odd}_\infty} \epsilon_i^\infty = 0.$$

Proof. Again, the proof proceeds by examination of Table 4.1. First, we observe that symmetric inertia implies constant inertia, since a pencil is a rational matrix and hence has the same rank at all but a finite number of points. It is therefore sufficient to evaluate the inertia at any point s on the real axis, $s \notin \Lambda$. Choose s less than all $a \in \Lambda$. Then the inertia of the pencil is symmetric if and only if

$$\sum_{a \in \Lambda} \sum_{i \in \mathrm{Odd}(a)} \epsilon_i^f + \sum_{i \in \mathrm{Odd}_\infty} \epsilon_i^\infty = 0.$$

Now the result follows from the Lemma 4.1, since the pencil has constant inertia. □

LEMMA 4.3. *A Hermitian pencil $\mathcal{A} - s\mathcal{E}$ has symmetric inertia at infinity if and only if*

$$\sum_{i \in \mathrm{Even}_\infty} \epsilon_i^\infty + \sum_{a \in \Lambda} \sum_{i \in \mathrm{Odd}(a)} \epsilon_i^f = 0.$$

Proof. Examination of the inertia of $\mathcal{E}$ in terms of the canonical blocks leads to Table 4.2.

Since $\mathcal{E}$ has symmetric inertia if and only if $i\{\mathcal{E}\} = i\{-\mathcal{E}\}$, examination of Table 4.2 shows that $\mathcal{E}$ has symmetric inertia precisely when

$$\sum_{i \in \mathrm{Even}_\infty} [i\{\epsilon_i^\infty\} - i\{-\epsilon_i^\infty\}] + \sum_{a \in \Lambda} \sum_{i \in \mathrm{Odd}(a)} \left[i\{\epsilon_i^f\} - i\{-\epsilon_i^f\} \right] = 0,$$

which is equivalent to the required condition. □

COROLLARY 4.4. *A Hermitian pencil $\mathcal{A} - s\mathcal{E}$ has both constant inertia and an $\mathcal{E}$ with symmetric inertia if and only if*

$$\sum_{i \in \mathrm{Odd}(a)} \epsilon_i^f = 0 \quad \text{for each } a \in \Lambda$$

and

$$\sum_{i\in \mathrm{Even}_\infty} \epsilon_i^\infty = 0.$$

COROLLARY 4.5. *A Hermitian pencil $\mathcal{A} - s\mathcal{E}$ has both symmetric inertia and an $\mathcal{E}$ with symmetric inertia if and only if*

$$\sum_{i\in \mathrm{Odd}(a)} \epsilon_i^f = 0 \quad \text{for each } a \in \Lambda$$

and

$$\sum_{i\in \mathrm{Even}_\infty} \epsilon_i^\infty = 0$$

and

$$\sum_{i\in \mathrm{Odd}_\infty} \epsilon_i^\infty = 0.$$

We now introduce two definitions that appear to be fundamental to factorization problems of Hermitian matrices. We say that a Hermitian pencil $\mathcal{A} - s\mathcal{E}$ is *separable* if there exists an invertible D such that

$$D^*(\mathcal{A} - s\mathcal{E})D = \begin{bmatrix} 0 & 0 & A - sE \\ 0 & J_1 & A_1 - sE_1 \\ A^* - sE^* & A_1^* - sE_1^* & Q \end{bmatrix}$$

for some matrices A, E, A_1, E_1, $J_1^* = J_1$, and $Q^* = Q$ such that $A - sE$ has full column rank and J_1 is invertible. We say that a Hermitian pencil $\mathcal{A} - s\mathcal{E}$ is *completely separable* if there exists an invertible D such that

$$D^*(\mathcal{A} - s\mathcal{E})D = \begin{bmatrix} 0 & A - sE \\ A^* - sE^* & Q \end{bmatrix}$$

for some matrices A, E, and $Q^* = Q$ such that $A - sE$ has full column rank. Before we relate separability and complete separability of a Hermitian pencil to properties of its inertia, we need the following lemma.

LEMMA 4.6. *The following statements hold.*

(i) *The pencil formed from the direct sum of two Type* 1 *blocks corresponding to the same real eigenvalue, both with odd size but with opposite signs, is completely separable.*

(ii) *The pencil formed from the direct sum of two Type* 3 *blocks, both with odd size but with opposite signs, is completely separable.*

(iii) *The pencil formed from the direct sum of two Type* 3 *blocks, both with even size but with opposite signs, is completely separable.*

Proof. See Appendix A.2. ∎

We now have the two main results of this section.

LEMMA 4.7. *A Hermitian pencil $\mathcal{A} - s\mathcal{E}$ is separable if and only if it has constant inertia and $\mathcal{E}$ has symmetric inertia.*

Proof. The necessity is clear. The sufficiency follows by showing that the result holds for various combinations of canonical blocks. The only nontrivial cases to consider are a pair of Type 1 odd-sized blocks with opposite signs corresponding to a real eigenvalue and a pair of even-sized blocks of Type 3 with opposite signs. These two cases are handled by Lemma 4.6(i) and(iii), respectively. ∎

LEMMA 4.8. *A Hermitian pencil $\mathcal{A} - s\mathcal{E}$ is completely separable if and only if it has symmetric inertia and $\mathcal{E}$ has symmetric inertia.*

Proof. The proof is along the lines of Lemma 4.7. We now also need to consider the case of a pair of odd-sized blocks of Type 3 with opposite signs. This case is handled by Lemma 4.6(ii). ∎

For $\mathcal{E}$ invertible, the notions of separability and complete separability coincide. We then have the following corollary.

COROLLARY 4.9. *For a Hermitian pencil $\mathcal{A} - s\mathcal{E}$ with $\mathcal{E}$ invertible, constant inertia, symmetric inertia, separability, and complete separability are all equivalent.*

Proof. Since $\mathcal{E}$ is invertible, the canonical form for $\mathcal{A} - s\mathcal{E}$ contains blocks of Types 1 and 2 only. The result then follows from the above lemmas. ∎

Remark. We have chosen to work with the inertia of rational, Hermitian matrices. We could have worked equally well with the signature, the difference being between the numbers of positive and negative eigenvalues and of rational, Hermitian matrices. In this case, constant inertia corresponds to constant signature and symmetric inertia corresponds to zero signature.

5. Inertia result for spectral densities. It is of interest to know whether the inertia properties of a Hermitian, rational matrix Ψ can be connected with those of the pole and zero pencils of a symmetric realization of Ψ. First, we consider the proper case. So let Ψ be a proper, rational, Hermitian matrix with minimal, symmetric realization

$$\Psi(s) = \left[\begin{array}{c|c} \mathcal{A} - s\mathcal{E} & \mathcal{B} \\ \hline \mathcal{B}^* & J \end{array}\right]$$

with $J^* = J$, $\mathcal{E}^* = \mathcal{E}$ and is invertible, and $\mathcal{A}^* = \mathcal{A}$. Suppose that Ψ has constant inertia, i.e., Ψ is a spectral density. In the case that Ψ is a semidefinite spectral density, then it is known [5] that in the canonical structure of the pole pencil $\mathcal{A} - s\mathcal{E}$ all Type 1 blocks must have even size and indeed the same sign. By Lemma 4.1, this means that the pole pencil itself is a spectral density, so it is tempting to think that this property may carry over to an arbitrary spectral density. Unfortunately, it is not true. As the following example illustrates, we can find an indefinite spectral density with no restriction on the sizes and signs of the Type 1 blocks.

Example 5.1. Consider the degree n indefinite spectral density Ψ defined by

$$\Psi(s) = \left[\begin{array}{cc} 0 & 1 \\ 1 & \frac{2\epsilon}{(s-a)^n} \end{array}\right],$$

where a is real and n is a positive integer. The pole pencil of any minimal, symmetric realization of Ψ consists of exactly one block of Type 1 corresponding to the pole at a and has size n and sign ϵ.

We conclude that it is not generally possible to impose the spectral density condition on the pole pencil of a minimal, symmetric realization. This immediately means that minimal spectral factorization is generally not possible for the indefinite case. This is an important observation.

So we ask whether there is any symmetric realization for which the pole pencil is a spectral density. The answer is clearly yes. For the pole pencil of any symmetric realization, the addition of extra uncontrollable and unobservable Type 1 blocks of odd size and of appropriate sign will ensure that the pole pencil is a spectral density. Moreover, to ensure the pole pencil is separable, we further add just enough uncontrollable and unobservable Type 3 blocks of even size. Finally, to ensure the pole pencil is completely separable, we still further add just enough uncontrollable and unobservable Type 3 blocks of odd size.

If we start with a minimal, symmetric realization, we can construct a symmetric realization of minimal size for which the pole pencil is a spectral density by adding just enough Type 1 blocks of size 1. Similarly, we can construct a symmetric realization of minimal size for which the pole pencil is completely separable by further adding just enough Type 3 blocks of sizes 1 and 2.

We shall say that a symmetric realization of Ψ is a *(completely) separable realization* if the pole pencil $\mathcal{A} - s\mathcal{E}$ is (completely) separable. We now obtain the following simple, but useful, result.

LEMMA 5.2. *A rational, Hermitian matrix Ψ is a spectral density if and only if there exists a symmetric realization*

$$\Psi(s) = \left[\begin{array}{c|c} \mathcal{A} - s\mathcal{E} & \mathcal{B} \\ \hline \mathcal{B}^* & J \end{array}\right]$$

for which the pole pencil P is completely separable and the zero pencil Z is separable.

Proof. From the discussion above, we can always assume that Ψ has a symmetric realization

$$\Psi(s) = \left[\begin{array}{c|c} \mathcal{A} - s\mathcal{E} & \mathcal{B} \\ \hline \mathcal{B}^* & J \end{array}\right]$$

for which the pole pencil $P(s) = \mathcal{A} - s\mathcal{E}$ is completely separable. This means that the pole pencil has symmetric inertia $i\{P\}$ and symmetric inertia at infinity. Since

$$Z(s) = \left[\begin{array}{cc} \mathcal{A} - s\mathcal{E} & \mathcal{B} \\ \mathcal{B}^* & J \end{array}\right],$$

we have that Z also has symmetric inertia at infinity, and since

$$i\{Z\} = i\{P\} + i\{\Psi\}$$

except at the real poles, we conclude that $i\{Z\}$ is separable if and only if Ψ is a spectral density. This completes the proof. □

6. Completely separable realizations. In the remainder of this paper, we restrict attention to a proper, Hermitian, rational matrix invertible at infinity. In particular, this means that we shall always have $\mathcal{E}$ invertible. In this section, we concentrate on symmetric realizations of both Ψ and Ψ^{-1} and the connections between them.

LEMMA 6.1. *Let Ψ be a proper, spectral density, invertible at infinity. Let Ψ have a minimal, symmetric realization*

$$\Psi(s) = \left[\begin{array}{c|c} \mathcal{A} - s\mathcal{E} & \mathcal{B} \\ \hline \mathcal{B}^* & J \end{array}\right]$$

and let Ψ^{-1} *have a minimal, symmetric realization*

$$\Psi^{-1}(s) = \left[\begin{array}{c|c} \tilde{\mathcal{A}} - s\tilde{\mathcal{E}} & \tilde{\mathcal{B}} \\ \hline \tilde{\mathcal{B}}^* & J^{-1} \end{array}\right].$$

Then there exist unique, invertible matrices $\mathcal{S}$ *and* $\mathcal{T}$ *such that*

$$\mathcal{S}\mathcal{A} - \tilde{\mathcal{A}}\mathcal{T} = \tilde{\mathcal{B}}\mathcal{B}^*, \tag{6.1}$$

$$\mathcal{S}\mathcal{E} - \tilde{\mathcal{E}}\mathcal{T} = 0, \tag{6.2}$$

$$\mathcal{S}\mathcal{B} = \tilde{\mathcal{B}}J, \tag{6.3}$$

$$-\tilde{\mathcal{B}}^*\mathcal{T} = J^{-1}\mathcal{B}^*, \tag{6.4}$$

$$\mathcal{T} = -\mathcal{S}^{-*}. \tag{6.5}$$

Proof. The proof uses completely standard arguments. First, two minimal, state-space realizations for Ψ^{-1} are given by

$$\Psi^{-1}(s) = \left[\begin{array}{c|c} \tilde{\mathcal{A}}\tilde{\mathcal{E}}^{-1} - sI & \tilde{\mathcal{B}} \\ \hline \tilde{\mathcal{B}}^*\tilde{\mathcal{E}}^{-1} & J^{-1} \end{array}\right]$$

and

$$\Psi^{-1}(s) = \left[\begin{array}{c|c} (\mathcal{A} - \mathcal{B}J^{-1}\mathcal{B}^*)\mathcal{E}^{-1} - sI & \mathcal{B}J^{-1} \\ \hline -J^{-1}\mathcal{B}^*\mathcal{E}^{-1} & J^{-1} \end{array}\right],$$

and so there exists a unique, invertible matrix $\mathcal{S}$ relating these two minimal realizations giving

$$\begin{aligned} \mathcal{S}(\mathcal{A} - \mathcal{B}J^{-1}\mathcal{B}^*)\mathcal{E}^{-1}\mathcal{S}^{-1} &= \tilde{\mathcal{A}}\tilde{\mathcal{E}}^{-1}, \\ \mathcal{S}\mathcal{B}J^{-1} &= \tilde{\mathcal{B}}, \\ -J^{-1}\mathcal{B}^*\mathcal{E}^{-1}\mathcal{S}^{-1} &= \tilde{\mathcal{B}}^*\tilde{\mathcal{E}}^{-1}. \end{aligned}$$

Taking Hermitian transposes of these equations and rearranging, it is easy to show that they are also satisfied by $-\tilde{\mathcal{E}}\mathcal{S}^{-*}\mathcal{E}^{-1}$, which, by uniqueness, must equal $\mathcal{S}$. Defining $\mathcal{T} = \tilde{\mathcal{E}}^{-1}\mathcal{S}\mathcal{E}$, we obtain the result. ∎

As we showed in section 5, it is not necessarily the case that the pole pencil of a minimal, symmetric realization of a proper, indefinite spectral density is also a spectral density. But, even for nonminimal realizations, we would like to retain the symmetry properties of Lemma 6.1. To accomplish this, we can do the following. Let Ψ and Ψ^{-1} have minimal, symmetric realizations as in Lemma 6.1, but suppose that the pole pencil $\mathcal{A} - s\mathcal{E}$ is not separable. Let $\mathcal{A}_1 - s\mathcal{E}_1$ be any pencil such that the pencil $(\mathcal{A} - s\mathcal{E}) \oplus (\mathcal{A}_1 - s\mathcal{E}_1)$ is separable. Then a (nonminimal) realization for Ψ is

$$\Psi(s) = \left[\begin{array}{cc|c} \mathcal{A} - s\mathcal{E} & 0 & \mathcal{B} \\ 0 & \mathcal{A}_1 - s\mathcal{E}_1 & 0 \\ \hline \mathcal{B}^* & 0 & J \end{array}\right],$$

and if we define a (nonminimal) realization for Ψ^{-1} by

$$\Psi^{-1}(s) = \left[\begin{array}{cc|c} \tilde{\mathcal{A}} - s\tilde{\mathcal{E}} & 0 & \tilde{\mathcal{B}} \\ 0 & -\mathcal{A}_1 + s\mathcal{E}_1 & 0 \\ \hline \tilde{\mathcal{B}}^* & 0 & J^{-1} \end{array}\right],$$

then these nonminimal, symmetric realizations are connected as in Lemma 6.1 via $\mathcal{S}$ and $\mathcal{T}$ given by $\mathcal{S} \oplus I$ and $\mathcal{T} \oplus (-I)$, respectively.

LEMMA 6.2. *Let Ψ be a proper, Hermitian, rational matrix invertible at infinity. Suppose that Ψ has a symmetric realization*

$$\Psi(s) = \left[\begin{array}{c|c} \mathcal{A} - s\mathcal{E} & \mathcal{B} \\ \hline \mathcal{B}^* & J \end{array}\right],$$

that Ψ^{-1} has a symmetric realization

$$\Psi^{-1}(s) = \left[\begin{array}{c|c} \tilde{\mathcal{A}} - s\tilde{\mathcal{E}} & \tilde{\mathcal{B}} \\ \hline \tilde{\mathcal{B}}^* & J^{-1} \end{array}\right],$$

and that there exist square, invertible matrices $\mathcal{S}$ and $\mathcal{T}$ such that (6.1)–(6.5) *are satisfied. If Ψ is a spectral density, then the pole pencil $\mathcal{A} - s\mathcal{E}$ of Ψ is completely separable if and only if the pole pencil $\tilde{\mathcal{A}} - s\tilde{\mathcal{E}}$ of Ψ^{-1} is completely separable.*

Proof. Consider the zero pencil

$$Z(s) = \left[\begin{array}{cc} \mathcal{A} - s\mathcal{E} & \mathcal{B} \\ \mathcal{B}^* & J \end{array}\right].$$

Using congruence arguments, we have

$$i\{Z(s)\} = i\{\mathcal{A} - s\mathcal{E}\} + i\{\Psi(s)\} = i\{\mathcal{A} - \mathcal{B}J^{-1}\mathcal{B}^* - s\mathcal{E}\} + i\{J\},$$

but $i\{\Psi(s)\} = i\{J\}$ since Ψ is a spectral density, and so

$$i\{\mathcal{A} - s\mathcal{E}\} = i\{\mathcal{A} - \mathcal{B}J^{-1}\mathcal{B}^* - s\mathcal{E}\}.$$

Since $\mathcal{A} - s\mathcal{E}$ is completely separable by assumption, it follows from Corollary 4.9 that $\mathcal{A} - \mathcal{B}J^{-1}\mathcal{B}^* - s\mathcal{E}$ is completely separable also. Moreover, we have

$$\begin{aligned} \mathcal{S}(\mathcal{A} - \mathcal{B}J^{-1}\mathcal{B}^* - s\mathcal{E})\mathcal{S}^* &= (\mathcal{S}\mathcal{A} - \tilde{\mathcal{B}}\mathcal{B}^* - s\mathcal{S}\mathcal{E})\mathcal{S}^* \\ &= (\tilde{\mathcal{A}} - s\tilde{\mathcal{E}})\mathcal{T}\mathcal{S}^* \\ &= -(\tilde{\mathcal{A}} - s\tilde{\mathcal{E}}), \end{aligned}$$

which shows that the pole pencil $\tilde{\mathcal{A}} - s\tilde{\mathcal{E}}$ is completely separable if and only if the pencil $\mathcal{A} - \mathcal{B}J^{-1}\mathcal{B}^* - s\mathcal{E}$ is completely separable. This completes the proof. ∎

To complete this section, we describe basis changes under which these patterns remain invariant. The following statements can be easily shown. If, for an invertible matrix $\mathcal{X}$, we change the symmetric realization for Ψ according to

$$(\mathcal{A}, \mathcal{E}, \mathcal{B}, J) \longrightarrow (\mathcal{X}^*\mathcal{A}\mathcal{X}, \mathcal{X}^*\mathcal{E}\mathcal{X}, \mathcal{X}^*\mathcal{B}, J),$$

then

$$(\mathcal{S}, \mathcal{T}) \longrightarrow \left(\mathcal{S}\mathcal{X}^{-*}, \mathcal{T}\mathcal{X}\right).$$

Similarly, if, for an invertible matrix $\mathcal{Y}$, we change the symmetric realization for Ψ^{-1} according to

$$\left(\tilde{\mathcal{A}}, \tilde{\mathcal{E}}, \tilde{\mathcal{B}}, J^{-1}\right) \longrightarrow \left(\mathcal{Y}^*\tilde{\mathcal{A}}\mathcal{Y}, \mathcal{Y}^*\tilde{\mathcal{E}}\mathcal{Y}, \mathcal{Y}^*\tilde{\mathcal{B}}, J^{-1}\right),$$

then

$$(\mathcal{S}, \mathcal{T}) \longrightarrow \left(\mathcal{Y}^*\mathcal{S}, \mathcal{Y}^{-1}\mathcal{T}\right).$$

7. Extraction of spectral factors. Theorem 7.1, our main result, is a restricted version of the basic factorization result for spectral densities. For this restricted version, we use state-space arguments to construct the desired spectral factor explicitly. In general, the factorization constructed in this section is not minimal. We shall briefly comment on this lack of minimality in the next section.

THEOREM 7.1. *Let Ψ be a proper, rational, Hermitian matrix, invertible at infinity. The following conditions are equivalent.*

(i) *Ψ is a spectral density.*

(ii) *There exists a realization of Ψ for which both the pole pencil and the zero pencil are spectral densities.*

(iii) *There exists a proper, stable, minimum phase, rational matrix W and a constant, invertible, symmetric J such that $\Psi = W^*JW$.*

The proof of the equivalence of (i) and (ii) is provided in the previous section. Moreover, it is obvious that (i) and (ii) each follow from (iii). Here, we show how (iii) follows from (ii). So, let us assume that we have realizations

$$\Psi(s) = \left[\begin{array}{c|c} \mathcal{A} - s\mathcal{E} & \mathcal{B} \\ \hline \mathcal{B}^* & J \end{array}\right] \quad \text{and} \quad \Psi^{-1}(s) = \left[\begin{array}{c|c} \tilde{\mathcal{A}} - s\tilde{\mathcal{E}} & \tilde{\mathcal{B}} \\ \hline \tilde{\mathcal{B}}^* & J^{-1} \end{array}\right],$$

where the pole pencils $\mathcal{A} - s\mathcal{E}$ and $\tilde{\mathcal{A}} - s\tilde{\mathcal{E}}$ are completely separable. As discussed in the previous section, we can, in fact, also assume that the separable realization of Ψ is the smallest such realization and, even further, we can assume that there exist invertible matrices $\mathcal{S}$ and $\mathcal{T}$ such that (6.1)–(6.5) hold. With change of basis matrices $\mathcal{X}$ and $\mathcal{Y}$, we can therefore assume separable realizations of the form

$$\Psi(s) = \left[\begin{array}{cc|c} 0 & A - sI & B \\ A^* - sI & Q & C \\ \hline B^* & C^* & J \end{array}\right] \quad \text{and} \quad \Psi^{-1}(s) = \left[\begin{array}{cc|c} 0 & F - sI & G \\ F^* - sI & \tilde{Q} & H \\ \hline G^* & H^* & J^{-1} \end{array}\right],$$

where the eigenvalues of A are contained in the closed upper half-plane and the eigenvalues of F are contained in the closed lower half-plane. We further claim that (A, B) and (F, G) are controllable. See Appendix A.3 for the details.

Since

$$\mathcal{E} = \tilde{\mathcal{E}} = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix},$$

then (6.2), with appropriate partitioning, becomes

$$\mathcal{S} = \begin{bmatrix} S_1 & S_2 \\ S_3 & S_4 \end{bmatrix} \quad \text{and} \quad \mathcal{T} = \begin{bmatrix} S_4 & S_3 \\ S_2 & S_1 \end{bmatrix}.$$

From (6.1) in Lemma 6.1, we have

$$\begin{bmatrix} S_1 & S_2 \\ S_3 & S_4 \end{bmatrix} \begin{bmatrix} 0 & A \\ A^* & Q \end{bmatrix} - \begin{bmatrix} 0 & F \\ F^* & \tilde{Q} \end{bmatrix} \begin{bmatrix} S_4 & S_3 \\ S_2 & S_1 \end{bmatrix} = \begin{bmatrix} G \\ H \end{bmatrix} \begin{bmatrix} B^* & C^* \end{bmatrix} \tag{7.1}$$

and from (6.3) in Lemma 6.1, we have

$$\begin{bmatrix} S_1 & S_2 \\ S_3 & S_4 \end{bmatrix} \begin{bmatrix} B \\ C \end{bmatrix} = \begin{bmatrix} G \\ H \end{bmatrix} J, \tag{7.2}$$

from which we extract, in particular, the coupling equation

$$S_2A^* - FS_2 = GB^*. \tag{7.3}$$

In Appendix A.4, we show that we can make further basis changes $\mathcal{X}$ and $\mathcal{Y}$ to display $\mathcal{S}$ in the form

$$\mathcal{S} = \begin{bmatrix} S_1 & S_2 \\ S_3 & S_4 \end{bmatrix} = \left[\begin{array}{cc|cc} 0 & 0 & 0 & I \\ -I & 0 & 0 & 0 \\ \hline 0 & -I & 0 & 0 \\ 0 & 0 & I & 0 \end{array}\right].$$

Then, with corresponding partitioning of the various other matrices, we have

$$A = \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix}, \quad Q = \begin{bmatrix} Q_1 & Q_2 \\ Q_2^* & Q_3 \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad C = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix}$$

and

$$F = \begin{bmatrix} F_1 & F_2 \\ F_3 & F_4 \end{bmatrix}, \quad \tilde{Q} = \begin{bmatrix} \tilde{Q}_1 & \tilde{Q}_2 \\ \tilde{Q}_2^* & \tilde{Q}_3 \end{bmatrix}, \quad G = \begin{bmatrix} G_1 \\ G_2 \end{bmatrix}, \quad H = \begin{bmatrix} H_1 \\ H_2 \end{bmatrix}.$$

Equation (7.1) expands to

$$\begin{aligned} S_2A^* - FS_2 &= GB^*, \\ S_1A + S_2Q - FS_1 &= GC^*, \\ S_4A^* - F^*S_4 - \tilde{Q}S_2 &= HB^*, \\ S_3A + S_4Q - F^*S_3 - \tilde{Q}S_1 &= HC^*, \end{aligned}$$

which in turn expand to

$$\begin{aligned} A_2^* &= G_1B_1^*, & A_4^* - F_1 &= G_1B_2^*, \\ 0 &= G_2B_1^*, & -F_3 &= G_2B_2^*, \\ \\ Q_2^* + F_2 &= G_1C_1^*, & Q_3 &= G_1C_2^*, \\ -A_1 + F_4 &= G_2C_1^*, & -A_2 &= G_2C_2^*, \\ \\ -F_3^* &= H_1B_1^*, & -\tilde{Q}_1 &= H_1B_2^*, \\ A_1^* - F_4^* &= H_2B_1^*, & A_3^* - \tilde{Q}_2^* &= H_2B_2^*, \\ \\ -A_3 + \tilde{Q}_2 &= H_1C_1^*, & -A_4 + F_1^* &= H_1C_2^*, \\ Q_1 + \tilde{Q}_3 &= H_2C_1^*, & Q_2 + F_2^* &= H_2C_2^*, \end{aligned}$$

and writing out the equations (7.2) in full, we obtain

$$C_2 = G_1J, \qquad -B_1 = G_2J, \qquad -B_2 = H_1J, \qquad C_1 = H_2J.$$

We next extend the coupling equation (7.3), i.e., we choose matrices K, L, X_1, X_2, Y_1, Y_2, S, and T to satisfy the extended coupling equation

$$\begin{aligned} &\left[\begin{array}{cc|c} 0 & I & 0 \\ 0 & 0 & I \\ \hline I & 0 & 0 \end{array}\right] \left[\begin{array}{cc|c} A_1^* & A_3^* & X_1 \\ A_2^* & A_4^* & X_2 \\ \hline 0 & 0 & S \end{array}\right] - \left[\begin{array}{cc|c} F_1 & F_2 & 0 \\ F_3 & F_4 & 0 \\ \hline Y_1 & Y_2 & T \end{array}\right] \left[\begin{array}{cc|c} 0 & I & 0 \\ 0 & 0 & I \\ \hline I & 0 & 0 \end{array}\right] \\ &\quad = \left[\begin{array}{c} G_1 \\ G_2 \\ \hline -L \end{array}\right] \left[\begin{array}{cc|c} B_1^* & B_2^* & K \end{array}\right], \end{aligned} \tag{7.4}$$

which is equivalent to

$$\begin{aligned} A_2^* &= G_1 B_1^*, & A_4^* - F_1 &= G_1 B_2^*, & X_2 - F_2 &= G_1 K, \\ 0 &= G_2 B_1^*, & -F_3 &= G_2 B_2^*, & S - F_4 &= G_2 K, \\ A_1^* - T &= -L B_1^*, & A_3^* - Y_1 &= -L B_2^*, & X_1 - Y_2 &= -LK. \end{aligned}$$

It remains to choose the matrices K, L, X_1, X_2, Y_1, Y_2, S, and T, which we do as follows.

1. Choose K to place all the eigenvalues of $S = F_4 + G_2 K$ on the real axis. (One possibility is to place all the eigenvalues at the origin.)
2. Set $L = -(C_1 + K^*)J^{-1}$.
3. This choice of L specifies $T = A_1^* + LB_1^*$.
4. The choice of K specifies $X_2 = F_2 + G_1 K$.
5. The choice of L specifies $Y_1 = A_3^* + LB_2^*$.
6. Choose X_1 so that $X_1 + X_1^* = LJL^* - Q_1$.
7. The choice of X_1 specifies $Y_2 = X_1 + LK$.

Note that the choice of K in step 1 is possible since (F_4, G_2) is controllable. This follows directly from the controllability of (F, G) and the fact that $-F_3 = G_2 B_2^*$.

This explicit construction allows us to write down an expression for a spectral factor of Ψ. Set

$$W(s) = \left[\begin{array}{cc|c} A - sI & 0 & B \\ X^* & S^* - sI & K^* \\ \hline \hat{G}^* & G_2^* & I \end{array}\right], \tag{7.5}$$

where $\hat{G}^* = \begin{bmatrix} -L^* & G_1^* \end{bmatrix}$ and $X^* = \begin{bmatrix} X_1^* & X_2^* \end{bmatrix}$.

LEMMA 7.2. *We have*

$$\Psi(s) = W^*(s) J W(s), \tag{7.6}$$

where W is defined in (7.5). *The poles of W are given by the union of the eigenvalues of A and of the eigenvalues of S^*, which are all stable, while the zeros of W are given by the union of the eigenvalues of F^* and of the eigenvalues of S, which are also all stable.*

Proof. For a proof of (7.6) by direct evaluation, see Appendix A.5. With regard to the remaining statements, we first define

$$W^{-1}(s) = \left[\begin{array}{cc|c} F^* - sI & Y^* & \hat{B} \\ 0 & T^* - sI & -B_1 \\ \hline G^* & -L^* & I \end{array}\right], \tag{7.7}$$

where $\hat{B}^* = \begin{bmatrix} -B_2^* & -K^* \end{bmatrix}$ and $Y = \begin{bmatrix} Y_1 & Y_2 \end{bmatrix}$. From the extended coupling equation (7.4), it is easy to show that $W(s)W^{-1}(s) = I$. Thus, the zeros of W, which are identical to the poles of W^{-1}, are given by the union of the eigenvalues of F^* and of the eigenvalues of T^*. Further, we have

$$\begin{aligned} T &= A_1^* + LB_1^* \\ &= A_1^* - (C_1 + K^*)J^{-1}B_1^* \\ &= A_1^* + (C_1 + K^*)G_2^* \\ &= F_4^* + K^* G_2^* \\ &= S^*, \end{aligned}$$

and so the eigenvalues of T^* and S coincide. ∎

8. Discussion. It is believed that state-space based methods offer the best chance of constructing computational algorithms for spectral factorization. This paper gives a state-space-based proof of necessary and sufficient conditions for the existence of spectral factors for indefinite spectral densities invertible at infinity, together with an explicit formula for a spectral factor. Unfortunately, this paper does not really provide an appropriate computational procedure. Indeed, one of the conclusions of this work is that indefinite spectral factorization, especially the noncanonical problem, is much more difficult numerically than definite spectral factorization. One reason for this is that the numerical construction of a separable realization for spectral factorization can be achieved by deflation methods, choosing at each stage of the deflation any eigenvector corresponding to a given eigenvalue [5], whereas for the indefinite spectral factorization problem we know only that there exists an eigenvector corresponding to a given eigenvalue for which the construction can be achieved. This means that Jordan structure may need to be computed. It is well known that this should be avoided whenever possible. A second reason is that one may need to solve a poorly conditioned pole placement problem.

Factorization of a polynomial spectral density Ψ with $\Psi(0)$ invertible is straightforward, given the results of this paper. Simply define a proper, rational spectral density $\tilde{\Psi}$ by $\tilde{\Psi}(s) = \Psi(\frac{1}{s})$ and in the case of a noncanonical factorization, add poles and zeros at zero. It follows that $\tilde{W}(\frac{1}{s})$ is polynomial where $\tilde{W}$ is the spectral factor of $\tilde{\Psi}$. In particular, we have an algorithm for the indefinite factorization of unimodular, Hermitian, polynomial matrices.

Indeed, any invertible, rational spectral density can be transformed by a bilinear transformation to a proper, invertible, rational spectral density, invertible at infinity, and hence can be factored using the above technique. However, exactly how the method of this paper extends to noninvertible, rational matrices is not clear.

An important problem is to compute a spectral factor of minimal degree. We have provided a formula for a spectral factor which may be of minimal degree, and often is, but is not guaranteed to be. It is simply not clear what the exact relationship is between the degree of the spectral density and the degree of a spectral factor for an indefinite spectral density. The lack of minimality is introduced in two ways: (i) when increasing the degree of a minimal realization to construct a minimal, separable realization, and (ii) when creating an extended coupling equation. In particular, for a spectral density with degree $2n$ (for a minimal, separable realization), what is an upper bound on the minimal degree of a spectral factor? It is clear that it is simply the size of the minimal extension computed in (7.4). In turn, an upper bound for this is obviously $2n$. As to the relationship between the sizes of a minimal realization and a minimal, separable realization, the answer is less clear. It is certainly possible to have a minimal realization of size m and a minimal, separable realization of size $2m$.

The inertia results developed in sections 4 and 5 are of some interest in their own right but also simplify parts of the proof in [5]. These results may also prove useful for the general spectral factorization problem with Ψ not necessarily proper.

Appendix A.

A.1. Existence of minimal Hermitian realizations. We are given a rational, Hermitian matrix Ψ and we wish to show the existence of a minimal, Hermitian realization.

First, we show the existence of a Hermitian realization. Let K be any invertible, rational matrix such that $\tilde{\Psi} = K^*\Psi K$ is proper. For example, $K(s) = I/s^n$ for

sufficiently large n would suffice. Now let

$$K(s) = \left[\begin{array}{c|c} F - sM & G \\ \hline H & L \end{array}\right]$$

be any realization of K with $F - sM$ nonsingular. It follows easily that

$$K^{-1}(s) = \left[\begin{array}{cc|c} F - sM & G & 0 \\ H & L & -I \\ \hline 0 & I & 0 \end{array}\right]$$

is a realization of K^{-1}. Now suppose that

$$\tilde{\Psi}(s) = \left[\begin{array}{c|c} A - sE & B \\ \hline B^* & D \end{array}\right]$$

is a Hermitian realization for the proper matrix $\tilde{\Psi}$ with $A - sE$ nonsingular. See, for example, the material in [5] which treats this problem. Then

$$\begin{aligned}
\Psi(s) &= K^{-*}\tilde{\Psi}K^{-1} \\
&= \left[\begin{array}{cc|c} F^* - sM^* & H^* & 0 \\ G^* & L^* & I \\ \hline 0 & -I & 0 \end{array}\right] \left[\begin{array}{c|c} A - sE & B \\ \hline B^* & D \end{array}\right] \left[\begin{array}{cc|c} F - sM & G & 0 \\ H & L & -I \\ \hline 0 & I & 0 \end{array}\right] \\
&= \left[\begin{array}{ccccc|c} 0 & 0 & 0 & F - sM & G & 0 \\ 0 & 0 & 0 & H & L & -I \\ 0 & 0 & A - sE & 0 & B & 0 \\ F^* - sM^* & H^* & 0 & 0 & 0 & 0 \\ G^* & L^* & B^* & 0 & 0 & 0 \\ \hline 0 & -I & 0 & 0 & 0 & D \end{array}\right]
\end{aligned}$$

is a Hermitian realization for Ψ, noting that the pole pencil for this realization is nonsingular.

Thus we may assume, without loss of generality, that

$$\Psi(s) = \left[\begin{array}{c|c} A - sE & B \\ \hline B^* & D \end{array}\right]$$

is a Hermitian, but not necessarily minimal, realization with nonsingular pole pencil $A - sE$. It remains to show that a minimal realization can be constructed. Suppose that there is a finite, uncontrollable mode, i.e., there exist a and $x \neq 0$ such that $x^* \left[\begin{array}{cc} A - aE & B \end{array}\right] = 0$. First, suppose that $x^*Ex = 0$. Then $x^*Ax = 0$ and therefore there exists a change of state-space basis such that

$$\begin{aligned}
\Psi(s) &= \left[\begin{array}{ccc|c} 0 & 0 & (a - s)e & 0 \\ 0 & A_1 - sE_1 & * & B_1 \\ (\bar{a} - s)\bar{e} & * & * & * \\ \hline 0 & B_1^* & * & D \end{array}\right] \\
&= \left[\begin{array}{c|c} A_1 - sE_1 & B_1 \\ \hline B_1^* & D \end{array}\right],
\end{aligned}$$

where e is a nonzero (because the pole pencil $A - sE$ is nonsingular) constant and each $*$ denotes a "don't care" entry. Second, if $x^*Ex \neq 0$, then there exists a change

of state-space basis such that

$$\Psi(s) = \left[\begin{array}{cc|c} (a-s)e & 0 & 0 \\ 0 & A_1 - sE_1 & B_1 \\ \hline 0 & B_1^* & D \end{array}\right] = \left[\begin{array}{c|c} A_1 - sE_1 & B_1 \\ \hline B_1^* & D \end{array}\right],$$

again with e a nonzero constant. In both cases, we have reduced the state-space dimension.

Now, suppose that there is an infinite uncontrollable mode, i.e., there exists $x \neq 0$ such that $x^* \left[\begin{array}{cc} E & B \end{array}\right] = 0$. First, suppose that $x^*Ax = 0$. Then, there exists a change of state-space basis such that

$$\Psi(s) = \left[\begin{array}{ccc|c} 0 & 0 & a & 0 \\ 0 & A_1 - sE_1 & * & B_1 \\ \bar{a} & * & * & * \\ \hline 0 & B_1^* & * & D \end{array}\right] = \left[\begin{array}{c|c} A_1 - sE_1 & B_1 \\ \hline B_1^* & D \end{array}\right],$$

where a is a nonzero (because the pole pencil $A - sE$ is nonsingular) constant. Second, if $x^*Ax \neq 0$, then there exists a change of state-space basis such that

$$\Psi(s) = \left[\begin{array}{cc|c} a & 0 & 0 \\ 0 & A_1 - sE_1 & B_1 \\ \hline 0 & B_1^* & D \end{array}\right] = \left[\begin{array}{c|c} A_1 - sE_1 & B_1 \\ \hline B_1^* & D \end{array}\right],$$

again with a a nonzero constant. In both cases, we have reduced the order of the state-space and, importantly, if $\left[\begin{array}{cc} A - sE & B \end{array}\right]$ has full row rank for all finite s, then so does $\left[\begin{array}{cc} A_1 - sE_1 & B_1 \end{array}\right]$. This means that the removal of infinite uncontrollable modes cannot introduce finite uncontrollable modes.

The final step is the removal of any nondynamic modes. If the realization for Ψ has nondynamic modes, then with state-space basis changes, we have

$$\Psi(s) = \left[\begin{array}{cc|c} A_1 - sE_1 & A_3 & B_1 \\ A_3^* & A_2 & B_2 \\ \hline B_1^* & B_2^* & D \end{array}\right],$$

where A_2 is nonsingular and so

$$\Psi = \left[\begin{array}{c|c} A_1 - A_3A_2^{-1}A_3^* - sE_1 & B_1 - A_3A_2^{-1}B_2 \\ \hline B_1^* - B_2^*A_2^{-1}A_3^* & D - B_2^*A_2^{-1}B_2 \end{array}\right].$$

Again, it is easily checked that the removal of the nondynamic nodes does not introduce any uncontrollable finite or infinite modes.

Finally, from the theory of minimal realizations for descriptor systems [3, 14, 17], we have a minimal (indeed, Hermitian) realization of Ψ, as our realization has no unobservable or uncontrollable finite or infinite modes and has no nondynamic modes.

A.2. Proof of Lemma 4.6. We use the notation from section 4. In addition, let P_{ij} denote an $i \times j$ matrix which is zero except for the (1,1)-entry, which is 1.

We first prove part (i) of Lemma 4.6. For the real eigenvalue a, suppose we have two Type 1 blocks of sizes $2m+n$ and n and signs $+1$ and -1, respectively. It is sufficient to prove the result for this case. We have

$$\begin{aligned}\mathcal{A}-s\mathcal{E} &= [(a-s)S_{2m+n}+T_{2m+n}] \oplus [-(a-s)S_n - T_n] \\ &= \begin{bmatrix} 0 & 0 & (a-s)S_m+T_m \\ 0 & (a-s)S_n+T_n & P_{nm} \\ (a-s)S_m+T_m & P_{mn} & 0 \end{bmatrix} \\ &\quad \oplus [-(a-s)S_n - T_n]\,,\end{aligned}$$

which is congruent to

$$\begin{bmatrix} 0 & 0 & 0 & (a-s)S_m+T_m \\ 0 & (a-s)S_n+T_n & 0 & P_{nm} \\ 0 & 0 & -(a-s)S_n-T_n & 0 \\ (a-s)S_m+T_m & P_{mn} & 0 & 0 \end{bmatrix}.$$

With

$$D_n = \begin{bmatrix} I_n & \frac{1}{2}I_n \\ -I_n & \frac{1}{2}I_n \end{bmatrix}$$

we have the congruence

$$\begin{aligned}&D_n^* \begin{bmatrix} (a-s)S_n+T_n & 0 \\ 0 & -(a-s)S_n-T_n \end{bmatrix} D_n \\ &\qquad = \begin{bmatrix} 0 & (a-s)S_n+T_n \\ (a-s)S_n+T_n & 0 \end{bmatrix}\end{aligned}$$

and we also have

$$D_n^* \begin{bmatrix} P_{nm} \\ 0 \end{bmatrix} = \begin{bmatrix} P_{nm} \\ \frac{1}{2}P_{nm} \end{bmatrix}.$$

We then see that $\mathcal{A}-s\mathcal{E}$ is congruent to

$$\begin{bmatrix} 0 & 0 & 0 & (a-s)S_m+T_m \\ 0 & 0 & (a-s)S_n+T_n & P_{nm} \\ 0 & (a-s)S_n+T_n & 0 & \frac{1}{2}P_{nm} \\ (a-s)S_m+T_m & P_{mn} & \frac{1}{2}P_{mn} & 0 \end{bmatrix}.$$

This proves the result, since this pencil is completely separable.

We prove parts (ii) and (iii) of Lemma 4.6 simultaneously. Even though the proof is essentially the same as the proof of part (i), except for an additional congruence, we shall write it out completely. Suppose we have two Type 3 blocks of sizes $2m+n$ and n and signs $+1$ and -1, respectively. It is sufficient to prove the result for this case. We have

$$\begin{aligned}\mathcal{A}-s\mathcal{E} &= [S_{2m+n}-sT_{2m+n}] \oplus [S_n - sT_n] \\ &= \begin{bmatrix} 0 & 0 & S_m-sT_m \\ 0 & S_n-sT_n & sP_{nm} \\ S_m-sT_m & sP_{mn} & 0 \end{bmatrix} \oplus [-S_n+sT_n]\,,\end{aligned}$$

which is congruent to

$$\begin{bmatrix} 0 & 0 & 0 & S_m - sT_m \\ 0 & S_n - sT_n & 0 & sP_{nm} \\ 0 & 0 & -S_n + sT_n & 0 \\ S_m - sT_m & sP_{mn} & 0 & 0 \end{bmatrix}.$$

With D_n as defined above, we have the congruence

$$D_n^* \begin{bmatrix} S_n - sT_n & 0 \\ 0 & -S_n + sT_n \end{bmatrix} D_n = \begin{bmatrix} 0 & S_n - sT_n \\ S_n - sT_n & 0 \end{bmatrix}$$

and also

$$D_n^* \begin{bmatrix} sP_{nm} \\ 0 \end{bmatrix} = \begin{bmatrix} sP_{nm} \\ \frac{1}{2} sP_{nm} \end{bmatrix}.$$

We then see that $\mathcal{A} - s\mathcal{E}$ is congruent to

$$\begin{bmatrix} 0 & 0 & 0 & S_m - sT_m \\ 0 & 0 & S_n - sT_n & sP_{nm} \\ 0 & S_n - sT_n & 0 & \frac{1}{2} sP_{nm} \\ S_m - sT_m & sP_{mn} & \frac{1}{2} sP_{mn} & 0 \end{bmatrix},$$

which is, in turn, congruent to

$$\begin{bmatrix} 0 & 0 & 0 & S_m - sT_m \\ 0 & 0 & S_n - sT_n & sP_{nm} \\ 0 & S_n - sT_n & -\frac{1}{2}(P_{nn}S_n + S_n P_{nn}) & 0 \\ S_m - sT_m & sP_{mn} & 0 & 0 \end{bmatrix},$$

since $P_{nn}P_{nm} = P_{nm}$ and $P_{nn}T_n = 0$. This proves the result, since this pencil is completely separable.

A.3. Controllability of (A, B) and (F, G). Suppose that (F, G) is not controllable. Then there exists an $x \neq 0$ and a such that

$$x^* \begin{bmatrix} F - aI & G \end{bmatrix} = 0,$$

from which, with $y^* = \begin{bmatrix} x^* & 0 \end{bmatrix}$, it follows easily that

$$y^* \begin{bmatrix} \tilde{\mathcal{A}} - a\tilde{\mathcal{E}} & \tilde{\mathcal{B}} \end{bmatrix} = 0 \quad \text{and} \quad y^* \tilde{\mathcal{E}} y = 0.$$

Then, with $z = \mathcal{S}^* y$, and using (6.1), (6.2), and (6.3), we conclude that

$$z^* \begin{bmatrix} \mathcal{A} - a\mathcal{E} & \mathcal{B} \end{bmatrix} = 0 \quad \text{and} \quad z^* \mathcal{E} z = 0. \tag{A.1}$$

Alternatively, suppose that (A, B) is not controllable. Then there exists an $x \neq 0$ and a such that

$$x^* \begin{bmatrix} A - aI & B \end{bmatrix} = 0,$$

from which, with $z^* = \begin{bmatrix} x^* & 0 \end{bmatrix}$, it again follows that (A.1) holds.

Hence, assume that (A.1) holds. Proceeding in a manner identical to the construction used in Appendix A.1 to reduce a nonminimal realization to a minimal one, we see that

$$\Psi = \left[\begin{array}{c|c} \mathcal{A}-s\mathcal{E} & \mathcal{B} \\ \hline \mathcal{B}^* & J \end{array}\right] = \left[\begin{array}{ccc|c} 0 & 0 & (a-s)e & 0 \\ 0 & A_1-sE_1 & * & B_1 \\ (\bar{a}-s)\bar{e} & * & * & * \\ \hline 0 & B_1^* & * & J \end{array}\right],$$

which means that

$$\left[\begin{array}{c|c} A_1-sE_1 & B_1 \\ \hline B_1^* & J \end{array}\right]$$

is also a separable realization of Ψ but of lower degree. This is a contradiction.

A.4. Structured form of the $\mathcal{S}$ matrix. Suppose that the Hermitian pencil $\mathcal{A}-s\mathcal{E}$ is completely separable and has the form

$$\mathcal{A}-s\mathcal{E} = \left[\begin{array}{cc} 0 & A-sI \\ A^*-sI & Q \end{array}\right].$$

We seek a useful class of nonsingular matrices $\mathcal{X}$ such that the $\mathcal{X}^*(\mathcal{A}-s\mathcal{E})\mathcal{X}$ has the same shape but with different A and Q. In particular, this means that we require $\mathcal{X}^*\mathcal{E}\mathcal{X}=\mathcal{E}$. Matrices of the form

$$\mathcal{X} = \left[\begin{array}{cc} U & 0 \\ 0 & U^{-*} \end{array}\right] \quad \text{or} \quad \mathcal{X} = \left[\begin{array}{cc} I & V \\ 0 & I \end{array}\right],$$

where U is nonsingular and $V+V^*=0$, satisfy these requirements. In the following discussion, we refer to these matrices as having Form I and Form II, respectively.

We now show that a sequence of choices of the matrices $\mathcal{X}$ and $\mathcal{Y}$ allow us to transform $\mathcal{S}$ to a particular form, namely,

$$\mathcal{S} = \left[\begin{array}{cc} S_1 & S_2 \\ S_3 & S_4 \end{array}\right] = \left[\begin{array}{cc|cc} 0 & 0 & 0 & I \\ -I & 0 & 0 & 0 \\ \hline 0 & -I & 0 & 0 \\ 0 & 0 & I & 0 \end{array}\right],$$

while maintaining the complete separability structure and the constraint $\mathcal{S}\mathcal{E}\mathcal{S}^* = -\mathcal{E}$. In the displayed form of $\mathcal{S}$ at various stages of the construction, each "don't care" entry is indicated by $*$.

Step 1. Find nonsingular matrices U and V such that

$$V^*S_2U = \left[\begin{array}{cc} 0 & I \\ 0 & 0 \end{array}\right].$$

Use U to construct an $\mathcal{X}$ matrix of Form I and V to construct a $\mathcal{Y}$ matrix of Form II. The transformed matrix $\mathcal{S}$ then has the shape

$$\mathcal{S} = \left[\begin{array}{cc} S_1 & S_2 \\ S_3 & S_4 \end{array}\right] = \left[\begin{array}{cc|cc} * & * & 0 & I \\ * & * & 0 & 0 \\ \hline * & * & * & * \\ * & * & * & * \end{array}\right].$$

Step 2. Since $\mathcal{S}\mathcal{E}\mathcal{S}^* = -\mathcal{E}$, we have $S_1 S_2^* + S_2 S_1^* = 0$. If we partition S_1 as

$$S_1 = \begin{bmatrix} V_1 & V_2 \\ V_3 & V_4 \end{bmatrix},$$

we have $V_2 + V_2^* = 0$ and $V_4 = 0$. With

$$V = \begin{bmatrix} 0 & V_1^* \\ -V_1 & -V_2 \end{bmatrix}$$

construct an $\mathcal{X}$ matrix of Form II. The transformed matrix $\mathcal{S}$ then has the shape

$$\mathcal{S} = \begin{bmatrix} S_1 & S_2 \\ S_3 & S_4 \end{bmatrix} = \left[\begin{array}{cc|cc} 0 & 0 & 0 & I \\ V_3 & 0 & 0 & 0 \\ \hline * & * & * & * \\ * & * & * & * \end{array}\right].$$

Step 3. This step is similar to Step 2. Since $\mathcal{S}\mathcal{E}\mathcal{S}^* = -\mathcal{E}$ and since $\mathcal{E}$ is invertible with $\mathcal{E}^{-1} = \mathcal{E}$, we also have $\mathcal{S}^*\mathcal{E}\mathcal{S} = -\mathcal{E}$, from which we conclude $S_2^* S_4 + S_4^* S_2 = 0$. If we partition S_4 as

$$S_4 = \begin{bmatrix} W_1 & W_2 \\ W_3 & W_4 \end{bmatrix},$$

we have $W_2 + W_2^* = 0$ and $W_1 = 0$. With

$$V = \begin{bmatrix} W_2 & -W_4^* \\ W_4 & 0 \end{bmatrix}$$

construct a $\mathcal{Y}$ matrix of Form II. The transformed matrix $\mathcal{S}$ then has the shape

$$\mathcal{S} = \begin{bmatrix} S_1 & S_2 \\ S_3 & S_4 \end{bmatrix} = \left[\begin{array}{cc|cc} 0 & 0 & 0 & I \\ V_3 & 0 & 0 & 0 \\ \hline Z_1 & Z_2 & 0 & 0 \\ Z_3 & Z_4 & W_3 & 0 \end{array}\right].$$

Step 4. Since $\mathcal{S}$ is nonsingular, we immediately obtain that V_3 and W_3 are nonsingular. Now evaluating $\mathcal{S}^*\mathcal{E}\mathcal{S} = -\mathcal{E}$, we obtain $Z_1 = 0$, $Z_4 = 0$, and $Z_2 = -I$. Also, we have $V_3 = -W_3^{-*}$. With

$$U = \begin{bmatrix} W_3^{-1} & 0 \\ 0 & I \end{bmatrix}$$

construct an $\mathcal{X}$ matrix of Form I. The transformed matrix $\mathcal{S}$ then has the shape

$$\mathcal{S} = \begin{bmatrix} S_1 & S_2 \\ S_3 & S_4 \end{bmatrix} = \left[\begin{array}{cc|cc} 0 & 0 & 0 & I \\ -I & 0 & 0 & 0 \\ \hline 0 & -I & 0 & 0 \\ Z_3 & 0 & I & 0 \end{array}\right].$$

Step 5. Finally, reevaluating $\mathcal{S}^*\mathcal{E}\mathcal{S} = -\mathcal{E}$, we obtain $Z_3 + Z_3^* = 0$. With

$$V = \begin{bmatrix} -Z_3 & 0 \\ 0 & 0 \end{bmatrix}$$

construct an $\mathcal{X}$ matrix of Form II. The transformed matrix $\mathcal{S}$ finally has the shape

$$\mathcal{S} = \begin{bmatrix} S_1 & S_2 \\ S_3 & S_4 \end{bmatrix} = \left[\begin{array}{cc|cc} 0 & 0 & 0 & I \\ -I & 0 & 0 & 0 \\ \hline 0 & -I & 0 & 0 \\ 0 & 0 & I & 0 \end{array}\right],$$

as required.

A.5. Proof of Lemma 7.2. The proof is by direct evaluation. First,

$$W^*(s)JW(s) = \left[\begin{array}{cc|c} A^* - sI & X & \hat{G} \\ 0 & S - sI & G_2 \\ \hline B^* & K & I \end{array}\right] J \left[\begin{array}{cc|c} A - sI & 0 & B \\ X^* & S^* - sI & K^* \\ \hline \hat{G}^* & G_2^* & I \end{array}\right],$$

which equals

$$\left[\begin{array}{cccc|c} 0 & 0 & A - sI & 0 & B \\ 0 & 0 & X^* & S^* - sI & K^* \\ A^* - sI & X & \hat{G}J\hat{G}^* & \hat{G}JG_2^* & \hat{G}J \\ 0 & S - sI & G_2J\hat{G}^* & G_2JG_2^* & G_2J \\ \hline B^* & K & \hat{G}^* & G_2^* & I \end{array}\right],$$

which expands out to

$$\left[\begin{array}{cccccc|c} 0 & 0 & 0 & A_1 - sI & A_2 & 0 & B_1 \\ 0 & 0 & 0 & A_3 & A_4 - sI & 0 & B_2 \\ 0 & 0 & 0 & X_1^* & X_2^* & S^* - sI & K^* \\ A_1^* - sI & A_3^* & X_1 & LJL^* & -LJG_1^* & -LJG_2^* & -LJ \\ A_2^* & A_4^* - sI & X_2 & -G_1JL^* & G_1JG_1^* & G_1JG_2^* & G_1J \\ 0 & 0 & S - sI & -G_2JL^* & G_2JG_1^* & G_2JG_2^* & G_2J \\ \hline B_1^* & B_2^* & K & -JL^* & JG_1^* & JG_2^* & J \end{array}\right].$$

We add the first block row to the sixth block row, and subtract the third block column from the fourth block column. Then, using the facts that $A_1 - S - G_2JL^* = A_1 - F_4 - G_2K - G_2JL^* = -G_2(C_1 + K^* + LJ)^* = 0$, $G_2JG_1^* + A_2 = 0$, $G_2JG_2^* = 0$, and $G_2J + B_1 = 0$, we obtain

$$\left[\begin{array}{cccccc|c} 0 & 0 & 0 & A_1 - sI & A_2 & 0 & B_1 \\ 0 & 0 & 0 & A_3 & A_4 - sI & 0 & B_2 \\ 0 & 0 & 0 & X_1^* & X_2^* & S^* - sI & K^* \\ A_1^* - sI & A_3^* & X_1 & LJL^* - X_1 & -LJG_1^* & -LJG_2^* & -LJ \\ A_2^* & A_4^* - sI & X_2 & -G_1JL^* - X_2 & G_1JG_1^* & G_1JG_2^* & G_1J \\ 0 & 0 & S - sI & 0 & 0 & 0 & 0 \\ \hline B_1^* & B_2^* & K & -JL^* - K & JG_1^* & JG_2^* & J \end{array}\right],$$

which reduces to

$$\left[\begin{array}{ccccc|c} 0 & 0 & A_1 - sI & A_2 & 0 & B_1 \\ 0 & 0 & A_3 & A_4 - sI & 0 & B_2 \\ 0 & 0 & X_1^* & X_2^* & S^* - sI & K^* \\ A_1^* - sI & A_3^* & LJL^* - X_1 & -LJG_1^* & -LJG_2^* & -LJ \\ A_2^* & A_4^* - sI & -G_1JL^* - X_2 & G_1JG_1^* & G_1JG_2^* & G_1J \\ \hline B_1^* & B_2^* & -JL^* - K & JG_1^* & JG_2^* & J \end{array}\right].$$

We subtract the third block row from the fourth block row and add the first block column to the fifth block column. Again, noting that $-LJG_2^* - S^* + A_1^* = (A_1 - S - G_2JL^*)^* = 0$, $G_1JG_2^* + A_2^* = (G_2JG_1^* + A_2)^* = 0$, and $G_2^* + B_1^* = (G_2J + B_1)^* = 0$, we obtain

$$\left[\begin{array}{ccccc|c} 0 & 0 & A_1 - sI & A_2 & 0 & B_1 \\ 0 & 0 & A_3 & A_4 - sI & 0 & B_2 \\ 0 & 0 & X_1^* & X_2^* & S^* - sI & K^* \\ A_1^* - sI & A_3^* & LJL^* - X_1 - X_1^* & -LJG_1^* - X_2^* & 0 & -LJ - K^* \\ A_2^* & A_4^* - sI & -G_1JL^* - X_2 & G_1JG_1^* & 0 & G_1J \\ \hline B_1^* & B_2^* & -JL^* - K & JG_1^* & 0 & J \end{array}\right],$$

which reduces to

$$\left[\begin{array}{cccc|c} 0 & 0 & A_1 - sI & A_2 & B_1 \\ 0 & 0 & A_3 & A_4 - sI & B_2 \\ A_1^* - sI & A_3^* & LJL^* - X_1 - X_1^* & -LJG_1^* - X_2^* & -LJ - K^* \\ A_2^* & A_4^* - sI & -G_1JL^* - X_2 & G_1JG_1^* & G_1J \\ \hline B_1^* & B_2^* & -JL^* - K & JG_1^* & J \end{array}\right],$$

which equals

$$\left[\begin{array}{cccc|c} 0 & 0 & A_1 - sI & A_2 & B_1 \\ 0 & 0 & A_3 & A_4 - sI & B_2 \\ A_1^* - sI & A_3^* & Q_1 & Q_2 & C_1 \\ A_2^* & A_4^* - sI & Q_2^* & Q_3 & C_2 \\ \hline B_1^* & B_2^* & C_1^* & C_2^* & J \end{array}\right].$$

This completes the proof.

Acknowledgment. The author would like to thank both reviewers for their detailed reading of the manuscript and the helpful comments provided.

REFERENCES

[1] H. Bart, I. Gohberg, and M. A. Kaashoek, *Minimal Factorization of Matrix and Operator Functions*, Oper. Theory Adv. Appl. 1, Birkhäuser, Basel, 1979.

[2] H. Bart, I. Gohberg, M. A. Kaashoek, and P. Van Dooren, *Factorization of transfer functions*, SIAM J. Control Optim., 18 (1980), pp. 675–696.

[3] P. Bernhard, *On singular implicit linear dynamical systems*, SIAM J. Control Optim., 20 (1982), pp. 612–633.

[4] D. J. Clements, *A State-Space Approach to J-Spectral Factorization*, Internal Report: UNSW, University of New South Wales, Sydney, Australia, 1993.

[5] D. J. Clements and K. Glover, *Spectral factorization via Hermitian pencils*, Linear Algebra Appl., 122–124 (1989), pp. 797–846.

[6] W. A. Coppel, *Linear Systems*, Notes in Pure Mathematics 6, Australian National University, Canberra, 1972.

[7] D. Z. Dokovic, *Factorization of Hermitian matrix polynomials with constant signature*, Linear Algebra Appl., 194 (1993), pp. 85–90.

[8] J. C. Doyle, K. Glover, P. Khargonekar, and B. Francis, *State-space solutions to standard H_2 and H_∞ control problems*, IEEE Trans. Automat. Control, 32 (1987), pp. 831–847.

[9] I. Gohberg and M. A. Kaashoek, eds., *Constructive Methods of Wiener-Hopf Factorization*, Oper. Theory Adv. Appl. 21, Birkhäuser, Basel, 1986.

[10] I. Gohberg, P. Lancaster, and L. Rodman, *Factorization of selfadjoint matrix polynomials with constant signature*, Linear and Multilinear Algebra, 11 (1982), pp. 209–214.

[11] I. Gohberg, P. Lancaster, and L. Rodman, *Matrix Polynomials*, Computer Science and Applied Mathematics, Academic Press, New York, 1982.

[12] M. Green, *H_∞ controller synthesis by J-lossless coprime factorization*, SIAM J. Control Optim., 30 (1992), pp. 522–547.
[13] M. Green, K. Glover, D. J. N. Limebeer, and J. C. Doyle, *A J-spectral factorization approach to H_∞ control*, SIAM J. Control Optim., 28 (1990), pp. 1350–1371.
[14] J. Grimm, *Realization and canonicity for implicit systems*, SIAM J. Control Optim., 26 (1988), pp. 1331–1347.
[15] V. K. Jacubovic, *Factorization of symmetric matrix polynomials*, Soviet Mathematics Doklady, 11 (1970), pp. 1261–1264.
[16] B. D. Jjubacevskii, *A representation of a symmetric polynomial matrix*, Vestnik Leningrad Univ. Math., 5 (1978), pp. 245–249.
[17] M. Kuijper and J. Schumacher, *Minimality of descriptor representations under external equivalence*, Automatica, 27 (1991), pp. 985–995.
[18] H. Kwakernaak, *Robust control and H_∞-optimization—tutorial paper*, Automatica, 29 (1993), pp. 255–273.
[19] H. Kwakernaak and M. Sebek, *Polynomial J-spectral factorization*, IEEE Trans. Automat. Control, 39 (1994), pp. 315–328.
[20] P. Lancaster and L. Rodman, *Existence and uniqueness theorems for the algebraic Riccati equation*, Internat. J. Control, 32 (1980), pp. 285–309.
[21] A. C. M. Ran and L. Rodman, *On symmetric factorizations of rational matrix functions*, Linear and Multilinear Algebra, 29 (1991), pp. 243–261.
[22] M. Sebek, *An Algorithm for Spectral Factorization of Polynomial Matrices with Any Signature*, Technical Report 912, University of Twente, The Netherlands, 1990.
[23] M. Sebek and H. Kwakernaak, *J-spectral factorization*, in Proceedings of the 30th IEEE Decision and Control Conference, Brighton, UK, 1991, pp. 1278–1283.
[24] R. C. Thompson, *The characteristic polynomial of a principal subpencil of an Hermitian matrix pencil*, Linear Algebra Appl., 14 (1976), pp. 135–177.
[25] H. Trentelman and J. Willems, *H_∞ control in a behavioral context*, in Proceedings of the 33rd IEEE Decision and Control Conference, Lake Buena Vista, FL, 1994, pp. 3695–3700.
[26] D. C. Youla, *On the factorization of rational matrices*, IRE Transactions on Information Theory, IT-7 (1961), pp. 172–189.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 3, pp. 768–796
© 2000 Society for Industrial and Applied Mathematics

MULTIPLE RECURSION CONJUGATE GRADIENT ALGORITHMS PART I: SUFFICIENT CONDITIONS*

TERI BARTH† AND THOMAS MANTEUFFEL‡

Abstract. In this paper we show that a conjugate gradient method can be implemented using a short multiple recursion for a wider class of matrices than that demonstrated by Faber and Manteuffel [*SIAM J. Numer. Anal.*, 21 (1984), pp. 352–362]. In part I of this paper, sufficient conditions are determined in order that a conjugate gradient method can be implemented using this form of multiple recursion. Part II of this paper addresses necessary conditions. This is the first known short recursion for a conjugate gradient method that can be applied to a nonnormal matrix of nontrivial degree.

Key words. linear system of equations, conjugate gradient method, short recurrence, nonnormal matrices

AMS subject classifications. 15A06, 65F10, 65F25

PII. S089547989833245X

1. Introduction.

A conjugate gradient method for the solution of an $n \times n$ linear system of equations $A\underline{x} = \underline{b}$ is defined with respect to an inner product matrix B and the matrix A. It is a Krylov subspace method in which the B-norm of the error is minimized at each step (see section 2.1). A conjugate gradient method is implemented via the construction of an ascending B-orthogonal basis $\{\underline{p}_i\}_{i=0}^{d-1}$ for $\mathcal{K}_d(\underline{r}_0, A)$, where

$$\mathcal{K}_d(\underline{r}_0, A) = \mathrm{sp}\{\underline{r}_0,\ A\underline{r}_0, \ldots,\ A^{d-1}\underline{r}_0\}$$

is the maximum dimension Krylov subspace generated by the initial residual $\underline{r}_0$ and the matrix A. The $\underline{p}_i$'s are called direction vectors. There are many algorithms for implementing the method, for example: Orthodir, Orthomin, Orthores, and GMRES [13].

If a B-orthogonal basis can be constructed with some form of a short recursion, then the work and storage requirements to implement the method are minimal, and we say an economical conjugate gradient algorithm exists. In 1984, Faber and Manteuffel [5] proved that the class of matrices for which a conjugate gradient method can be implemented via a single short $(s+2)$-term recursion of the form

$$\underline{p}_{j+1} = A\underline{p}_j - \sum_{i=j-s}^{j} \sigma_{i,j}\underline{p}_i$$

for the direction vectors is limited to matrices that are B-normal(s) or to matrices with only a small number of distinct eigenvalues (see section 2.3).

Gragg [9] demonstrated that if A is unitary with respect to some inner product, then a B-orthonormal basis can be constructed using a pair of short recurrence

*Received by the editors January 12, 1998; accepted for publication (in revised form) by A. Greenbaum March 2, 1999; published electronically February 10, 2000. This work was sponsored by the Department of Energy, Applied Math Program grant DE-FG03-94ER25217, and National Science Foundation grant DMS-8704169.

http://www.siam.org/journals/simax/21-3/33245.html

†TRW, Aurora, CO (tbarth@auc.trw.com).

‡Department of Mathematics, University of Colorado at Boulder, Boulder, CO 50309 (tmanteuf@boulder.colorado.edu).

formulas. The unitary matrix is B-normal ($B = I$), but it is not B-normal(s) for small degree s; thus a short $(s+2)$-term recurrence cannot be used to construct an orthonormal basis. Using Gragg's algorithm for constructing an orthonormal basis, Jagels and Reichel [11], [12] derived an efficient minimal-residual algorithm for unitary and shifted unitary matrices (see section 2.3). This motivates the exploration of alternate forms of short recurrences. By considering a more general form of recursion, we will show that the class of matrices for which a practical conjugate gradient algorithm is known to exist can be extended.

We notice that if A is unitary, its adjoint is

$$A^* = \bar{A}^T = A^{-1} = \frac{p_0(A)}{q_1(A)},$$

where $p_0(A) = I$ and $q_1(A) = A$. (Throughout this paper, we will be using the notation $\frac{p_\ell(A)}{q_m(A)} = q_m(A)^{-1}p_\ell(A)$ when $q_m(A)^{-1}$ exists.) In section 3.1 we define a class of normal matrices, called B-normal(ℓ, m) matrices. The B-adjoint of these matrices, $A^\dagger$, can be expressed as the ratio of two polynomials,

$$A^\dagger = \frac{p_\ell(A)}{q_m(A)},$$

of degrees ℓ and m, respectively. We refer to an extension of this class as generalized B-normal(ℓ, m) matrices. These matrices satisfy the relationship

$$A^\dagger q_m(A) - p_\ell(A) = Q_B(A)$$

for some polynomials, p_ℓ and q_m, and $Q_B(A)$, a low rank matrix (see section 3.2). We show that low rank perturbations of B-normal(ℓ, m) matrices satisfy the above relationship.

The double recursion [9] for unitary and shifted unitary matrices can be rewritten as a single (s, t)-recursion of the form

$$\underline{p}_{j+1} = \sum_{i=j-t}^{j} \beta_{i,j} A\underline{p}_i - \sum_{i=j-s}^{j} \sigma_{i,j}\underline{p}_i.$$

If A is unitary, $t = 1$ and $s = 0$; if A is shifted unitary, $t = 1$ and $s = 1$. In section 4, we show that in the absence of a condition we call breakdown, a short (s, t)-recursion can be used to construct a B-orthogonal basis for either B-normal(ℓ, m) or generalized B-normal(ℓ, m) matrices. Since breakdown is possible using the single (s, t)-recursion, we show how the recursion can be reformulated as a multiple recursion that avoids the possibility of breakdown. Sufficient conditions on the matrix A are then given in order that a B-orthogonal basis can be constructed using this form of multiple recursion. Necessary conditions for $\ell = m \leq 1$ are given in [4].

Details about the implementation are included in section 5.1. Since low rank perturbations of B-normal(ℓ, m) matrices can be nonnormal, this work demonstrates that a conjugate gradient method can be implemented with a short recurrence for certain nonnormal matrices of nontrivial degree. A numerical example of this type is given in section 5.2.

1.1. Notation. The notation used throughout this paper is summarized below.

$\mathcal{R}^n$, $\mathcal{C}^n$	–	Vector spaces of real and complex n-tuples.
$\mathcal{R}^{n\times m}$, $\mathcal{C}^{n\times m}$	–	Vector spaces of real and complex $n \times m$ matrices.
$\mathcal{K},\ \mathcal{V}, \ldots$	–	Other calligraphic letters denote subspaces of $\mathcal{R}^n$ or $\mathcal{C}^n$.
$A, B, \ldots$	–	Uppercase Roman (Greek) letters denote matrices.
$\underline{x}, \underline{p}, \ldots$	–	Underlined lowercase Roman (Greek) letters denote column vectors.
$\alpha, \beta, \ldots$	–	Lowercase Roman (Greek) letters denote scalars.
$\langle \cdot, \cdot \rangle,\ \| \cdot \|$	–	Euclidean inner product on $\mathcal{C}^n$ and induced norm.
$\langle B\cdot, \cdot \rangle,\ \| \cdot \|_B$	–	B-inner product on $\mathcal{C}^n$ and induced B-norm.
$\perp_B$	–	Represents orthogonality in the B-inner product.
A^*	–	Euclidean adjoint of A, $A^* = \bar{A}^T$.
$A^\dagger$	–	B-adjoint of A, $A^\dagger = B^{-1}A^*B$.
$\mathrm{sp}\{\underline{x}_j\}$	–	The linear span of the vectors $\underline{x}_j$.
$\Sigma(A)$	–	Spectrum of A.
$\mathcal{K}_\ell(\underline{r}_0, A)$	–	$\mathrm{sp}\{\underline{r}_0, A\underline{r}_0, \ldots, A^{\ell-1}\underline{r}_0\}$. Krylov subspace of dimension ℓ.
$d = d(\underline{r}_0, A)$	–	Degree of the minimal annihilating polynomial of $\underline{r}_0$.
$d(A)$	–	Degree of the minimal polynomial of A.

2. The conjugate gradient method.

2.1. Definition. Consider the linear system of equations

$$A\underline{x} = \underline{b}, \tag{2.1}$$

where A is a general $n \times n$ nonsingular matrix and $\underline{x}$ and $\underline{b} \in \mathcal{C}^n$. A conjugate gradient method for the solution of (2.1) is a Krylov subspace method in which the error is minimized at every step in an inner product norm. We define this method with respect to a Hermitian positive definite inner product matrix B and the system matrix A and denote it as $\mathcal{CG}(B, A)$.

DEFINITION 2.1. *A conjugate gradient method, $\mathcal{CG}(B, A)$, is a Krylov subspace method whose iterates are defined uniquely as follows:*

$$\begin{array}{l} \underline{x}_{j+1} = \underline{x}_j + \underline{z}_j, \qquad \underline{z}_j \in \mathcal{K}_{j+1}(\underline{r}_0, A), \\ \|\underline{e}_{j+1}\|_B \quad \textit{is minimized over } \mathcal{K}_{j+1}(\underline{r}_0, A), \end{array} \tag{2.2}$$

where $\underline{r}_0$ denotes the initial residual and $\underline{e}_{j+1}$ is the error at step $j+1$. The second condition is equivalent to requiring

$$\underline{e}_{j+1} \perp_B \underline{z} \qquad \forall\, \underline{z} \in \mathcal{K}_{j+1}(\underline{r}_0, A).$$

For example, if A is symmetric positive definite, $\langle A\cdot, \cdot \rangle$ defines an inner product, and we could take $B = A$. In this case, $\mathcal{CG}(B, A)$ minimizes the A-norm of the error, which results in the original conjugate gradient method of Hestenes and Stiefel [10]. Another example of a conjugate gradient method is given by taking $B = A^*A$. Notice that since we are assuming that A is nonsingular, A^*A is Hermitian positive definite. In this case, $\mathcal{CG}(B, A)$ minimizes

$$\|\underline{e}_{j+1}\|_{A^*A} = \langle A^*A\underline{e}_{j+1}, \underline{e}_{j+1} \rangle^{\frac{1}{2}} = \langle \underline{r}_{j+1}, \underline{r}_{j+1} \rangle^{\frac{1}{2}},$$

yielding a minimal residual method. A detailed taxonomy of conjugate gradient methods can be found in [1].

2.2. Implementation. A conjugate gradient method is implemented via the construction of an ascending B-orthogonal basis $\{\underline{p}_i\}_{i=0}^{d-1}$ for $\mathcal{K}_d(\underline{r}_0, A)$, where $d = d(\underline{r}_0, A)$ is the degree of the minimal annihilating polynomial of $\underline{r}_0$ with respect to A. The $\underline{p}_i$'s are called direction vectors. Given $\underline{p}_0 = \gamma\ \underline{r}_0$, at each step compute

$$\underline{p}_j \in \mathcal{K}_{j+1}(\underline{p}_0, A) \quad \text{such that } \underline{p}_j \perp_B \mathcal{K}_j(\underline{p}_0, A).$$

The iterate is then updated from the previous step by adding a scalar (step length) times the new direction vector,

$$\underline{x}_{j+1} = \underline{x}_j + \alpha_j \underline{p}_j, \qquad \alpha_j = \frac{\langle B\underline{e}_j, \underline{p}_j\rangle}{\langle B\underline{p}_j, \underline{p}_j\rangle} = \frac{\langle \underline{e}_j, \underline{p}_j\rangle_B}{\langle \underline{p}_j, \underline{p}_j\rangle_B}. \tag{2.3}$$

Since the computation of α_j involves the unknown quantity of the error, in order to have a computable conjugate gradient method, the inner product matrix B must be chosen so that α_j is computable. For example, if A is Hermitian positive definite, we could choose $B = A$ and

$$\alpha_j = \frac{\langle \underline{r}_j, \underline{p}_j\rangle}{\langle A\underline{p}_j, \underline{p}_j\rangle}.$$

By choosing $B = A^*A$, we obtain

$$\alpha_j = \frac{\langle r_j, A\underline{p}_j\rangle}{\langle A\underline{p}_j, A\underline{p}_j\rangle}.$$

For a complete derivation of the implementation details, see [5].

From the above, we see that the work involved in implementing a conjugate gradient method is in the construction of a B-orthogonal basis $\{\underline{p}_i\}_{i=0}^{d-1}$ for $\mathcal{K}_d(\underline{r}_0, A)$. There may be many different ways for constructing such a basis. However, since this basis is unique up to scale, the basis can always be derived by a rescaling of the Gram–Schmidt process:

$$\begin{aligned} \underline{p}_0 &= \underline{r}_0, \\ &\vdots \\ \underline{p}_{j+1} &= A\underline{p}_j - \sum_{i=0}^{j} \sigma_{i,j}\underline{p}_i, \qquad \sigma_{i,j} = \frac{\langle BA\underline{p}_j, \underline{p}_i\rangle}{\langle B\underline{p}_i, \underline{p}_i\rangle} = \frac{\langle \underline{p}_j, A^\dagger \underline{p}_i\rangle_B}{\langle \underline{p}_i, \underline{p}_i\rangle_B}. \end{aligned} \tag{2.4}$$

Suppose that $d = n$, where n is the dimension of A. The matrix form of this recursion at step n is given by

$$A\begin{bmatrix} \underline{p}_0 & \underline{p}_1 & \cdots & \underline{p}_{n-1} \end{bmatrix} = \begin{bmatrix} \underline{p}_0 & \underline{p}_1 & \cdots & \underline{p}_{n-1} \end{bmatrix} \begin{bmatrix} \sigma_{0,0} & \sigma_{0,1} & \cdots & \cdots & \sigma_{0,n-1} \\ 1 & \sigma_{1,1} & & & \vdots \\ & 1 & \ddots & & \vdots \\ & & \ddots & & \vdots \\ & & & 1 & \sigma_{n-1,n-1} \end{bmatrix}, \tag{2.5}$$

which we denote as

$$AP_n = P_n H_n. \tag{2.6}$$

In general, H_n is a full upper Hessenberg matrix. If $d < n$, then $\underline{p}_j = \underline{0}$ for $j \geq d$ and $\sigma_{i,j} = 0$ for $j \geq d$. Equation (2.5) remains valid; in fact, $\underline{p}_j$ and $\sigma_{i,j}$ are bounded rational functions of $\underline{p}_0$ [4], [5].

The above recursion (2.4) utilizes all the previous direction vectors to construct $\underline{p}_{j+1}$. In the next section, we will see that for certain kinds of matrices, a B-orthogonal basis can be constructed with a short recursion that uses only a few previous direction vectors.

2.3. Economical conjugate gradient algorithms. We review some terminology on normal matrices that will be used throughout this paper. First, the concept of an adjoint of a matrix is generalized to the B-inner product. The B-adjoint of A is the unique matrix $A^\dagger$ satisfying

$$\langle A\underline{x},\, \underline{y}\rangle_B = \langle BA\underline{x},\, \underline{y}\rangle = \langle B\underline{x},\, A^\dagger\underline{y}\rangle = \langle \underline{x},\, A^\dagger\underline{y}\rangle_B \qquad \forall\ \underline{x},\ \underline{y} \in C^n.$$

This yields $A^\dagger = B^{-1}A^*B$. The following conditions are generalizations of those for normal matrices.

A matrix A is B-normal if it satisfies the following equivalent conditions:

(1) $AA^\dagger = A^\dagger A$.

(2) A and $A^\dagger$ have the same complete set of B-orthogonal eigenvectors.

(3) $A^\dagger$ can be written as a polynomial of some degree in the matrix A.

We say that A is B-normal(s) if it is B-normal and s is the degree of the polynomial $p_s(A)$ of smallest degree for which $A^\dagger = p_s(A)$. For example, if A is B-normal(1), then $A^\dagger$ can be written as a first-degree polynomial in A. A special case of B-normal(1) matrices are B-self-adjoint matrices. They satisfy

$$A^\dagger = A \quad \text{or} \quad BA = (BA)^* = A^*B.$$

Consider an $(s+2)$-term recursion of the form

$$\underline{p}_{j+1} = A\underline{p}_j - \sum_{i=j-s}^{j} \sigma_{i,j}\underline{p}_i, \qquad \sigma_{i,j} = \frac{\langle A\underline{p}_j,\, \underline{p}_i\rangle_B}{\langle \underline{p}_i,\, \underline{p}_i\rangle_B}. \tag{2.7}$$

DEFINITION 2.2. *The matrix A is in the class $\mathcal{CG}(s)$ if for every $\underline{r}_0$ a single $(s+2)$-term recursion of the form given in* (2.7) *can be used to construct a B-orthogonal basis for $\mathcal{K}_d(\underline{r}_0, A)$.*

THEOREM 2.1. *$A \in \mathcal{CG}(s)$ if and only if*

(1) *A is B normal(s), or*

(2) *the degree of the minimal polynomial of A, $d(A)$, is less than or equal to $s+2$.*

Proof. The proof is a result from Faber and Manteuffel [5]. □

In their 1984 paper, Faber and Manteuffel characterized B-normal(s) matrices. These results are given in the following lemma.

LEMMA 2.1. *If A is B-normal(s), then*

(1) *if $s > 1$, then $d(A) \leq s^2$;*

(2) *if $s = 1$, then either $d(A) = 1$, $A = A^\dagger$, or*

$$A = e^{i\theta}\left(i\frac{r}{2}I + G\right),$$

where $i = \sqrt{-1}$, $r \geq 0$ is a real number, $0 \leq \theta \leq 2\pi$, and $G = G^\dagger$.

Proof. The proof is given in [5]. □

To summarize, this lemma says that if A is B-normal(s) and $s > 1$, then A has at most s^2 distinct eigenvalues. If $s = 1$ and A has more than one distinct eigenvalue, then all the eigenvalues of A lie on some straight line in the complex plane. Note that in the case that A is B-self-adjoint, all the eigenvalues of A are real. Otherwise, the eigenvalues of A can be obtained by shifting and then rotating the eigenvalues of a B-self-adjoint matrix G, yielding colinear eigenvalues.

It follows that an economical recursion for the direction vectors of the form given by (2.7) can be applied only to a very small class of matrices. Of this class, the only matrices with more than a few distinct eigenvalues are B-normal(1) matrices. In this case, a B-orthogonal basis can be computed with a three-term recursion of the form in (2.7).

There are other forms of short recurrences that can't be written as a short $(s+2)$-term recursion of the form given in (2.7). By using a short double recursion, Jagels and Reichel [11], [12] demonstrated an efficient minimal-residual algorithm for unitary and shifted unitary matrices (see below). It follows that the class of matrices for which a short recursion of some form admits a conjugate gradient method is wider than the class described by Faber and Manteuffel [5].

A matrix U is B-unitary or isometric with respect to the inner product $\langle B\cdot, \cdot\rangle$ if

$$\langle U\underline{x}, U\underline{y}\rangle_B = \langle BU\underline{x}, U\underline{y}\rangle = \langle B\underline{x}, \underline{y}\rangle = \langle \underline{x}, \underline{y}\rangle_B \quad \forall\ \underline{x},\ \underline{y} \in \mathcal{C}^n.$$

Gragg [9] showed that if U is B-unitary, then a B-orthonormal basis for the Krylov subspace can be constructed using two short recursions at each step:

$$\begin{aligned} \sigma_{j+1}\underline{p}_{j+1} &= U\underline{p}_j + \gamma_{j+1}\tilde{\underline{p}}_j, \\ \tilde{\underline{p}}_{j+1} &= \sigma_{j+1}\tilde{\underline{p}}_j + \bar{\gamma}_{j+1}\underline{p}_{j+1}, \\ \gamma_{j+1} &= -\langle BU\underline{p}_j, \tilde{\underline{p}}_j\rangle, \\ \sigma_{j+1} &= (1 - |\gamma_{j+1}|^2)^{\frac{1}{2}}. \end{aligned} \tag{2.8}$$

Another derivation of this double recursion for unitary matrices is given by Watkins [14].

A shifted unitary matrix has the form

$$\rho U + \xi I, \tag{2.9}$$

where U is I-unitary matrix (I is the identity matrix) and ρ and ξ are complex scalars. Jagels and Reichel observed that when A is shifted unitary, $\mathcal{K}_d(\underline{r}_0, A) = \mathcal{K}_d(\underline{r}_0, U)$. Thus, the double recursion (2.8) can be used to construct an orthonormal basis for $\mathcal{K}_d(\underline{r}_0, A)$. They continued, showing how to obtain an efficient minimal-residual algorithm for shifted unitary matrices. See [11], [12] for a complete presentation.

Although unitary and shifted unitary matrices are normal matrices, they are not normal(s) for small degree s. That is, the adjoint of A cannot be written as

$$A^* = p_s(A) \quad \text{for } s \text{ small.}$$

It follows from Theorem 2.1 that a short $(s + 2)$-term recursion of the form (2.7) cannot be used to construct an orthonormal basis. However, notice that if A is unitary, $A^* = A^{-1}$. If λ_j is an eigenvalue of A, then $\frac{1}{\lambda_j}$ is an eigenvalue of A^*, and we obtain

$$A^* = A^{-1} = \frac{p_0(A)}{q_1(A)} = \frac{I}{A}, \qquad \bar{\lambda}_j = \frac{p_0(\lambda_j)}{q_1(\lambda_j)} = \frac{1}{\lambda_j}, \tag{2.10}$$

where $p_0(\mu) = 1$ and $q_1(\mu) = \mu$.

If A is shifted unitary and $\hat{\lambda}_j$ is an eigenvalue of U, then $\lambda_j = \rho\hat{\lambda}_j + \xi$ is an eigenvalue of A. Thus, the eigenvalues of shifted unitary matrices lie on a circle in the complex plane. Further, note that

$$A^* = \frac{p_1(A)}{q_1(A)}, \qquad \bar{\lambda}_j = \frac{p_1(\lambda_j)}{q_1(\lambda_j)}, \tag{2.11}$$

where $p_1(\mu) = \bar{\xi}\mu + (\bar{\rho}\rho - \bar{\xi}\xi)$ and $q_1(\mu) = \mu - \xi$.

The double recursion (2.8) for unitary and shifted unitary matrices can be rewritten as a single $(s,\, t)$-recursion of the form

$$\underline{p}_{j+1} = \sum_{i=j-t}^{j} \beta_{i,j} A\underline{p}_i - \sum_{i=j-s}^{j} \sigma_{i,j}\underline{p}_i,$$

where $(s,\, t) = (0,\, 1)$ if A is unitary and $(s,\, t) = (1,\, 1)$ if A is shifted unitary. Notice that the terms s and t correspond to the degrees of the polynomials in (2.10) if A is unitary and (2.11) if A is shifted unitary. To see how this is done, at step $j+1$ of the iteration (2.8), form

$$\sigma_{j+1}\underline{p}_{j+1} = U\underline{p}_j + \gamma_{j+1}\tilde{\underline{p}}_j.$$

For $\tilde{\underline{p}}_j$, substitute the second equation at step j in (2.8) to obtain

$$\sigma_{j+1}\underline{p}_{j+1} = U\underline{p}_j + \gamma_{j+1}(\sigma_j\tilde{\underline{p}}_{j-1} + \bar{\gamma}_j\underline{p}_j).$$

Solving the first equation at step j in (2.8) for $\tilde{\underline{p}}_{j-1}$, and making this substitution into the above, yields the recursion

$$\underline{p}_{j+1} = \frac{1}{\sigma_{j+1}}U\underline{p}_j - \frac{\gamma_{j+1}\sigma_j}{\gamma_j\sigma_{j+1}}U\underline{p}_{j-1} + \left(\frac{\gamma_{j+1}\sigma_j^2}{\gamma_j\sigma_{j+1}} + \frac{\gamma_{j+1}\bar{\gamma}_j}{\sigma_{j+1}}\right)\underline{p}_j. \tag{2.12}$$

For unitary and shifted unitary matrices, the double recursion (2.8) can be rewritten as the above single recursion as long as the γ's are nonzero. When A is shifted unitary, another way of writing the single recursion follows from substituting $U = \frac{1}{\rho}(A - \xi I)$ into (2.12), yielding

$$\begin{aligned}\underline{p}_{j+1} = {} & \frac{1}{\rho\sigma_{j+1}}A\underline{p}_j - \frac{\gamma_{j+1}\sigma_j}{\rho\gamma_j\sigma_{j+1}}A\underline{p}_{j-1} \\ & + \left(\frac{\gamma_{j+1}\sigma_j^2}{\gamma_j\sigma_{j+1}} - \frac{\xi}{\rho\sigma_{j+1}} + \frac{\gamma_{j+1}\bar{\gamma}_j}{\sigma_{j+1}}\right)\underline{p}_j + \frac{\gamma_{j+1}\sigma_j\xi}{\rho\gamma_j\sigma_{j+1}}\,\underline{p}_{j-1}.\end{aligned} \tag{2.13}$$

This motivates the development in sections 3 and 4.

3. B-normal$(\ell,\, m)$ and generalized B-normal$(\ell,\, m)$ matrices.

3.1. B-normal$(\ell,\, m)$ matrices. In section 2.3, we saw that a short double recursion could be used to construct an orthonormal basis for the Krylov subspace for unitary and shifted unitary matrices. It was noted that a short $(s+2)$-term recursion of the form given in (2.7) could not be used to construct this basis, since

these matrices are not B-normal(s) for some small degree s. However, the adjoint of these matrices can be expressed as given in (2.10) if A is unitary and (2.11) if A is shifted unitary. This motivates the exploration of a more general class of normal matrices whose B-adjoint can be written as a ratio of two polynomials. We call this class B-normal(ℓ, m) matrices.

DEFINITION 3.1. *A is B-normal(ℓ, m) if A is B-normal and there exists polynomials $p_\ell(\lambda)$ and $q_m(\lambda)$ of degree ℓ and m, respectively, such that*

$$A^\dagger q_m(A) = p_\ell(A). \tag{3.1}$$

Since $A^\dagger = \frac{p_\ell(A)}{q_m(A)}$, we say A is B-normal with rational degree ℓ/m.

Notice that if $m = 0$, (3.1) becomes

$$A^\dagger = \hat{p}_\ell(A)$$

and A is B-normal(ℓ).

B-normal(ℓ, m) matrices satisfy

$$A^\dagger = \frac{p_\ell(A)}{q_m(A)} = \frac{a_\ell A^\ell + \cdots + a_1 A + a_0 I}{b_m A^m + \cdots + b_1 A + b_0 I} \tag{3.2}$$

for some polynomials

$$p_\ell(\mu) = a_\ell \mu^\ell + \cdots + a_1 \mu + a_0 \quad \text{and} \quad q_m(\mu) = b_m \mu^m + \cdots + b_1 \mu + b_0$$

of degree ℓ and m, respectively. In the following theorem, which characterizes B-normal(ℓ, m) matrices, we will assume that ℓ and m are the smallest degrees for which (3.2) holds. This means that a_ℓ, $b_m \neq 0$ and that $p_\ell(\mu)$ and $q_m(\mu)$ have no common roots, since, otherwise, (3.2) would hold for some smaller degrees ℓ' and m'.

THEOREM 3.1. *Let A be B-normal(ℓ, m), where ℓ and m are the smallest degrees for which*

$$A^\dagger = \frac{p_\ell(A)}{q_m(A)}.$$

Let $d(A)$ be the degree of the minimal polynomial of A. Then

1. *if $\ell > m + 1$, then $d(A) \leq \ell^2$;*
2. *if $\ell = m + 1$, then $d(A) \leq \ell^2$ or A is B-normal$(1, 0)$;*
3. *if $\ell < m$ and $b_0 \neq 0$, then $d(A) \leq m^2 + 1$;*
4. *if $\ell < m - 1$ and $b_0 = 0$, then $d(A) \leq m^2$;*
5. *if $\ell = m - 1$ and $b_0 = 0$, then $d(A) \leq m^2$ or A is B-normal$(0, 1)$;*
6. *if $\ell = m$, then $d(A) \leq m^2 + 1$ or A is B-normal$(1, 1)$.*

Proof. The proof can be found in [3] and [4]. ∎

The above theorem says that if A is B-normal(ℓ, m) and either ℓ or m are greater than 1, then A has a relatively small number of distinct eigenvalues. Furthermore, if ℓ, $m \leq 1$, the only matrices that have more than two distinct eigenvalues are B-normal(1), B-normal$(0, 1)$ (e.g., B-unitary), and B-normal$(1, 1)$ (e.g., shifted B-unitary) matrices.

Consider the B-orthogonal basis vectors $\{\underline{p}_i\}_{i=0}^{d-1}$ for $\mathcal{K}_d(\underline{p}_0, A)$. Since for every i, $\underline{p}_i \in \mathcal{K}_{i+1}(\underline{p}_0, A)$, $\underline{p}_i$ can be written as $\psi_i(A)\underline{p}_0$ for some polynomial ψ_i of exact degree i. For $i > m$, there exist polynomials, $\tilde{\psi}_{i-m}$ and $\zeta_{m-1}^{(i)}$, such that

$$\underline{p}_i = \psi_i(A)\underline{p}_0 = q_m(A)\tilde{\psi}_{i-m}(A)\underline{p}_0 + \zeta_{m-1}^{(i)}(A)\underline{p}_0,$$

where the second equality follows upon division of $\psi_i(A)$ by $q_m(A)$, resulting in a quotient term, $\tilde{\psi}_{i-m}(A)$, and a remainder term, $\zeta_{m-1}^{(i)}(A)$. If A is B-normal(ℓ, m), $A^\dagger q_m(A) = p_\ell(A)$, and it follows that

$$\begin{aligned} A^\dagger \underline{p}_i &= A^\dagger q_m(A)\tilde{\psi}_{i-m}(A)\underline{p}_0 + A^\dagger \zeta_{m-1}^{(i)}(A)\underline{p}_0 \\ &= p_\ell(A)\tilde{\psi}_{i-m}(A)\underline{p}_0 + A^\dagger \zeta_{m-1}^{(i)}(A)\underline{p}_0. \end{aligned} \tag{3.3}$$

This relationship will be utilized later in the proofs on sufficient conditions in section 4.

3.2. Generalized B-normal(ℓ, m) matrices. A more general class of matrices that includes the B-normal(ℓ, m) matrices as a special case is given in the following definition.

DEFINITION 3.2. *A matrix A is a generalized B-normal(ℓ, m) matrix if there exist polynomials $p_\ell(\lambda)$ and $q_m(\lambda)$ of degree ℓ and m, respectively, such that*

$$A^\dagger q_m(A) - p_\ell(A) = Q_B(A), \qquad \textit{where} \quad \operatorname{rank}(Q_B(A)) = \kappa. \tag{3.4}$$

If $\operatorname{rank}(Q_B(A)) = 0$, *then A is B-normal(ℓ, m).*

For example, matrices that are B-normal(ℓ, m) plus a low rank perturbation are of this form.

LEMMA 3.1. *Suppose the matrix A has the form*

$$A = \widehat{A} + R_z,$$

where $\widehat{A}$ is B-normal(ℓ, m) and R_z is a rank z matrix. Then A satisfies (3.4) *with* $\kappa \le (\ell + m + 1)z$.

Proof. It follows from Definition 3.1 that

$$\widehat{A}^\dagger q_m(\widehat{A}) - p_\ell(\widehat{A}) = 0.$$

Notice that

$$q_m(A) = q_m(\widehat{A} + R_z) = a_m(\widehat{A} + R_z)^m + \cdots + a_1(\widehat{A} + R_z) + a_0 I.$$

By expanding each of these terms, we see that

$$q_m(A) = q_m(\widehat{A}) + R_{zm},$$

where R_{zm} is an accumulation of all terms involving R_z. It is easy to show that

$$\operatorname{Range}(R_{zm}) \subseteq \bigcup \sum_{j=0}^{m-1} \operatorname{Range}(\widehat{A}^j \ R_z).$$

Each of the terms involving R_z has rank at most z. Since there are m of these terms, the rank of $R_{zm} \le zm$.

Similarly,

$$p_\ell(A) = p_\ell(\widehat{A} + R_z) = p_\ell(\widehat{A}) + R_{z\ell},$$

where the rank of $R_{z\ell} \le z\ell$. Using this information, we obtain

$$\begin{aligned} Q_B(A) &= A^\dagger q_m(A) - p_\ell(A) \\ &= (\widehat{A}^\dagger + R_z^\dagger)[q_m(\widehat{A}) + R_{zm}] - [p_\ell(\widehat{A}) + R_{z\ell}] \\ &= [\widehat{A}^\dagger q_m(\widehat{A}) - p_\ell(\widehat{A})] + \widehat{A}^\dagger R_{zm} + R_z^\dagger [q_m(\widehat{A}) + R_{zm}] - R_{z\ell} \\ &= \widehat{A}^\dagger R_{zm} + R_z^\dagger [q_m(\widehat{A}) + R_{zm}] - R_{z\ell}, \end{aligned}$$

which yields $\text{rank}(Q_B(A)) \leq (\ell + m + 1)z$. ∎

An important example of this type is a matrix that results from the discretization of a partial differential equation that is formally self-adjoint, positive definite, where nonstandard boundary conditions are applied to a few points on the boundary. A numerical example of this type is given in section 5.2.

From the previous section, we know that only certain B-normal(ℓ, m) matrices have more than a few distinct eigenvalues. Suppose $A = \widehat{A} + R_z$, where $\widehat{A}$ is B-normal(ℓ, m) with $d(\widehat{A}) = k$ and R_z has low rank z. There exists a polynomial p_k of degree k such that $p_k(\widehat{A}) = 0$. Notice that

$$p_k(\widehat{A} + R_z) = p_k(\widehat{A}) + R_{kz},$$

where the $\text{rank}(R_{kz}) \leq kz$. It follows that there exists a polynomial q_{kz} of degree at most kz such that $q_{kz}(R_{kz}) = 0$. By combining this information, we obtain

$$q_{kz}(p_k(\widehat{A} + R_z)) = 0,$$

which implies that $d(\widehat{A} + R_z) \leq k^2 z$. This means that any low rank perturbation of a matrix with only a few distinct eigenvalues still has only a few distinct eigenvalues. For this reason, we will be primarily interested in matrices of the form

$$A = \widehat{A} + R_z,$$

where $\widehat{A}$ is either B-normal(1), B-unitary, or shifted B-unitary, and R_z has low rank.

Consider the B-orthogonal basis vectors $\{\underline{p}_i\}_{i=0}^{d-1}$ for $\mathcal{K}_d(\underline{p}_0, A)$. Since for every i, $\underline{p}_i \in \mathcal{K}_{i+1}(\underline{p}_0, A)$, $\underline{p}_i$ can be written as $\psi_i(A)\underline{p}_0$ for some polynomial ψ_i of exact degree i. Analogous to the B-normal(ℓ, m) case, for $i > m$ there exist polynomials $\tilde{\psi}_{i-m}$ and $\zeta_{m-1}^{(i)}$ such that

$$\underline{p}_i = \psi_i(A)\underline{p}_0 = q_m(A)\tilde{\psi}_{i-m}(A)\underline{p}_0 + \zeta_{m-1}^{(i)}(A)\underline{p}_0,$$

where the second equality follows from dividing the polynomial $\psi_i(A)$ by $q_m(A)$. Multiplying through by $A^\dagger$ and then adding and subtracting $p_\ell(A)\tilde{\psi}_{i-m}(A)\underline{p}_0$ yields

$$\begin{aligned} A^\dagger \underline{p}_i &= [A^\dagger q_m(A) - p_\ell(A)]\tilde{\psi}_{i-m}(A)\underline{p}_0 + p_\ell(A)\tilde{\psi}_{i-m}(A)\underline{p}_0 + A^\dagger \zeta_{m-1}^{(i)}(A)\underline{p}_0 \\ &= Q_B(A)\tilde{\psi}_{i-m}(A)\underline{p}_0 + p_\ell(A)\tilde{\psi}_{i-m}(A)\underline{p}_0 + A^\dagger \zeta_{m-1}^{(i)}(A)\underline{p}_0. \end{aligned} \tag{3.5}$$

This relationship will be useful in the proofs on sufficient conditions in the next section.

4. Sufficient conditions.

4.1. Introduction. In this section, we begin by considering general recursions of the form

$$\underline{p}_{j+1} = \sum_{i=j-t}^{j} \beta_{i,j} A\underline{p}_i - \sum_{i=j-s}^{j} \sigma_{i,j}\underline{p}_i,$$

where s and t can be any integers ≥ 0. We will refer to this type of a recursion as a single (s, t)-recursion. In section 2.3, we showed how the double recursion for unitary and shifted unitary matrices could be written in this form. By taking $t = 0$, we notice that the $(s + 2)$-term recursion in (2.7) for B-normal(s) matrices is also of this form.

Thus, we can view the single $(s,\, t)$-recursion as a more general form of short recursion, since it includes the short recurrences for B-normal(s) matrices as well as those for unitary and shifted unitary matrices.

At each step in the $(s,\, t)$-iteration, the term $\beta_{j,j} A\underline{p}_j$ brings the recursion up to the next higher dimension Krylov subspace, yielding $\underline{p}_{j+1} \in \mathcal{K}_{j+2}(\underline{r}_0, A)$. In order for the above recursion to yield a B-orthogonal basis, $\beta_{j,j} \neq 0$ for every j. Since $\beta_{j,j}$ can be used as a normalization constant, for simplicity of this presentation we will assume $\beta_{j,j} = 1$ for every j and denote the single $(s,\, t)$-recursion as

$$\underline{p}_{j+1} = A\underline{p}_j + \sum_{k=j-t}^{j-1} \beta_{k,j} A\underline{p}_k - \sum_{i=j-s}^{j} \sigma_{i,j} \underline{p}_i, \tag{4.1}$$

where s and t can be any integers ≥ 0.

Notice that $\underline{p}_{j+1}$ can always be computed by a recursion using all previous $\underline{p}_i$'s; that is,

$$\underline{p}_{j+1} = A\underline{p}_j + \sum_{k=j-t}^{j-1} \beta_{k,j} A\underline{p}_k - \sum_{i=0}^{j} \sigma_{i,j} \underline{p}_i$$

because, given any $\{\beta_{k,j}\}_{k=j-t}^{j-1}$, $\underline{p}_{j+1}$ can be constructed so it is B-orthogonal to $\mathrm{sp}\{\underline{p}_0, \ldots,\, \underline{p}_j\}$ by choosing

$$\sigma_{i,j} = \frac{\left\langle A\underline{p}_j + \displaystyle\sum_{k=j-t}^{j-1} \beta_{k,j} A\underline{p}_k, \underline{p}_i \right\rangle_B}{\langle \underline{p}_i,\, \underline{p}_i \rangle_B}, \qquad i = 0, \ldots, j.$$

However, we are interested in determining when a B-orthogonal basis for $\mathcal{K}_d(\underline{r}_0, A)$ can be constructed using a short recursion of the form (4.1), that is, with both s and t small. Notice that in order for this to be the case, at each step $j+1$, $\sigma_{i,j} = 0$ for $i = 0, \ldots, j-s-1$. If a B-orthogonal basis for $\mathcal{K}_d(\underline{r}_0, A)$ can be constructed using the single $(s,\, t)$-recursion, we say $A \in \mathcal{CG}_{\mathrm{SR}}(s,\, t)$. To be more precise, we make the following definition.

DEFINITION 4.1. *$A \in \mathcal{CG}_{SR}(s,\, t)$ if, for almost every $\underline{p}_0$ and every $0 \leq j \leq d(\underline{p}_0, A) - 2$, there exist coefficients $\{\beta_{k,j}\}_{k=j-t}^{j-1}$ such that*

$$\sigma_{i,j} = \frac{\left\langle A\underline{p}_j + \displaystyle\sum_{k=j-t}^{j-1} \beta_{k,j} A\underline{p}_k, \underline{p}_i \right\rangle_B}{\langle \underline{p}_i,\, \underline{p}_i \rangle_B} = 0 \quad \textit{for} \quad i = 0, \ldots, j-s-1$$

or, equivalently,

$$\left\langle \underline{p}_j + \sum_{k=j-t}^{j-1} \beta_{k,j} \underline{p}_k, A^\dagger \underline{p}_i \right\rangle_B = 0 \quad \textit{for} \quad i = 0, \ldots, j-s-1. \tag{4.2}$$

Notice that the condition in the definition for $A \in \mathcal{CG}_{\mathrm{SR}}(s,\, t)$ must hold for almost every $\underline{p}_0$. This is a technical point that will become important in part II of this paper

in the proof of necessary conditions. In the absence of a condition we call breakdown, section 4.2 demonstrates that a single (s, t)-recursion can be used to construct a B-orthogonal basis for $\mathcal{K}_d(\underline{r}_0, A)$ when A is either B-normal(ℓ, m) or generalized B-normal(ℓ, m). Since breakdown is possible in the single (s, t)-recursion, which is illustrated by an example given in section 4.3, we show how the computations can be reorganized as a set of multiple recursions that avoids the problem of breakdown. Sufficient conditions on the system matrix A are then given in order for a B-orthogonal basis to be constructed using this form of multiple recursion.

4.2. A single (s, t)-recursion for generalized B-normal(ℓ, m) matrices. In order for a B-orthogonal basis to be constructed using a single (s, t)-recursion, there must exist at each step coefficients $\{\beta_{k,j}\}_{k=j-t}^{j-1}$ satisfying (4.2). If A is a generalized B-normal(ℓ, m) matrix, the expression for $A^\dagger \underline{p}_i$ given in (3.5) can be substituted in (4.2). It follows that for $i = 0, \ldots, j-s-1$, the β's must satisfy

$$\left\langle \left(\underline{p}_j + \sum_{k=j-t}^{j-1} \beta_{k,j} \underline{p}_k \right), \left(Q_B(A)\tilde{\psi}_{i-m}(A)\underline{p}_0 + p_\ell(A)\tilde{\psi}_{i-m}(A)\underline{p}_0 + A^\dagger \zeta_{m-1}^{(i)}(A)\underline{p}_0 \right) \right\rangle_B = 0. \tag{4.3}$$

Since $\zeta_{m-1}^{(i)}(A)$ is a polynomial of degree $\leq m-1$,

$$A^\dagger \zeta_{m-1}^{(i)}(A)\underline{p}_0 \in \mathrm{sp}\{A^\dagger \underline{p}_0, \ldots, A^\dagger \underline{p}_{m-1}\}.$$

Next, by letting $\{\underline{\phi}_0, \underline{\phi}_1, \ldots, \underline{\phi}_{\kappa-1}\}$ be a basis for the range of $Q_B(A)$, we can write

$$Q_B(A)\tilde{\psi}_{i-m}(A)\underline{p}_0 = (\underline{\tau}_i)_{\kappa-1}\ \underline{\phi}_{\kappa-1} + \cdots + (\underline{\tau}_i)_0\ \underline{\phi}_0$$

for some vector of coefficients $\underline{\tau}_i = [(\underline{\tau}_i)_0, \ldots, (\underline{\tau}_i)_{\kappa-1}]^T \in \mathcal{C}^\kappa$.

By choosing s and t to satisfy

$$s - \ell \geq t - m \geq \kappa, \tag{4.4}$$

we have for $i = 0, \ldots, j-s-1$ that

$$\begin{aligned} i - m + \ell &\leq (j-s-1) - m + \ell = j - 1 - m - (s-\ell) \\ &\leq j - 1 - m - (t-m) = j - 1 - t, \end{aligned}$$

and thus,

$$p_\ell(A)\tilde{\psi}_{i-m}(A)\underline{p}_0 \in \mathrm{sp}\{\underline{p}_0, \ldots, \underline{p}_{i-m+\ell}\} \subseteq \mathrm{sp}\{\underline{p}_0, \underline{p}_1, \ldots, \underline{p}_{j-t-1}\}.$$

Therefore,

$$p_\ell(A)\tilde{\psi}_{i-m}(A)\underline{p}_0 \perp_B \left(\underline{p}_j + \sum_{k=j-t}^{j-1} \beta_{k,j}\underline{p}_k \right), \tag{4.5}$$

and it follows that we need to choose only the $\beta_{k,j}$'s so that for $i = 0, \ldots, j-s-1$,

$$\left\langle \left(\underline{p}_j + \sum_{k=j-t}^{j-1} \beta_{k,j}\underline{p}_k \right), \left(A^\dagger \zeta_{m-1}^{(i)}(A)\underline{p}_0 + Q_B(A)\tilde{\psi}_{i-m}(A)\underline{p}_0 \right) \right\rangle_B = 0.$$

This can be accomplished by choosing $\{\beta_{k,j}\}_{k=j-t}^{j-1}$ to satisfy

$$\begin{bmatrix} \langle \underline{p}_{j-t}, A^\dagger \underline{p}_0 \rangle_B & \cdots & \langle \underline{p}_{j-1}, A^\dagger \underline{p}_0 \rangle_B \\ \vdots & & \vdots \\ \langle \underline{p}_{j-t}, A^\dagger \underline{p}_{m-1} \rangle_B & \cdots & \langle \underline{p}_{j-1}, A^\dagger \underline{p}_{m-1} \rangle_B \\ \langle \underline{p}_{j-t}, \underline{\phi}_0 \rangle_B & \cdots & \langle \underline{p}_{j-1}, \underline{\phi}_0 \rangle_B \\ \vdots & & \vdots \\ \langle \underline{p}_{j-t}, \underline{\phi}_{\kappa-1} \rangle_B & \cdots & \langle \underline{p}_{j-1}, \underline{\phi}_{\kappa-1} \rangle_B \end{bmatrix} \begin{pmatrix} \beta_{j-t,j} \\ \vdots \\ \beta_{j-1,j} \end{pmatrix} \tag{4.6}$$

$$= - \begin{pmatrix} \langle \underline{p}_j, A^\dagger \underline{p}_0 \rangle_B \\ \vdots \\ \langle \underline{p}_j, A^\dagger \underline{p}_{m-1} \rangle_B \\ \langle \underline{p}_j, \underline{\phi}_0 \rangle_B \\ \vdots \\ \langle \underline{p}_j, \underline{\phi}_{\kappa-1} \rangle_B \end{pmatrix}.$$

This system has $m + \kappa$ equations in t unknowns. From (4.4) we see that t was chosen so that $\kappa \leq t - m$, which means the system has $t - (m + \kappa)$ degrees of freedom. Up to this point, we have only specified that s and t must satisfy (4.4). There is some flexibility in choosing s and t. To make s and t as small as possible, yielding the most economical (s, t)-recursion, we would choose

$$t = \kappa + m \quad \text{and} \quad s = \kappa + \ell$$

so the system in (4.6) is $(m+\kappa) \times (m+\kappa)$. If this system is nonsingular for every step, then an (s, t)-recursion will yield a B-orthogonal basis. Notice that the number of terms s and t depends on the rank of $Q_B(A)$ as well as the degrees of the polynomials p_ℓ and q_m. Breakdown occurs in the single (s, t)-recursion if the above system becomes inconsistent at some step in the iteration.

This process of demonstrating a single (s, t)-recursion simplifies for the special case when A is B-normal(ℓ, m). First, the expression given in (3.3) is substituted into (4.2), yielding (4.3) without the $Q_B(A)\tilde{\psi}_{i-m}(A)\underline{p}_0$ term. By choosing $(s, t) = (\ell, m)$ we see that the orthogonality condition in (4.5) still holds. It remains to choose the β's so that

$$\left\langle \left(\underline{p}_j + \sum_{k=j-m}^{j-1} \beta_{k,j} \underline{p}_k \right), \left(A^\dagger \zeta_{m-1}^{(i)}(A) \underline{p}_0 \right) \right\rangle_B = 0 \quad \text{for} \quad i = 0, \ldots, j - \ell - 1.$$

This yields an $m \times m$ system of equations to be solved at each step for the β_{kj}'s (the first m rows of the system in (4.6)). Breakdown occurs if this system is inconsistent at some step. For more details, see [3] or [4].

4.3. Breakdown in the single (s, t)-recursion. The next example illustrates that breakdown can occur during the (s, t)-iteration for B-normal(ℓ, m) matrices. Recall that breakdown occurs in the (s, t)-recursion for some $\underline{p}_0$ if the system (4.6) is

inconsistent at some step in the iteration. Consider the unitary matrix

$$U = \begin{bmatrix} 0 & 0 & \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Since U is unitary with respect to the standard Euclidean inner product, $B = I$, we have

$$U^\dagger = U^* = \frac{I}{U} = \frac{p_0(U)}{q_1(U)}.$$

By Definition 3.1, U is I-normal(0, 1). In the absence of breakdown, section 4.2 shows an I-orthogonal basis $\{\underline{p}_j\}_{j=0}^4$ for $\mathcal{K}_5(\underline{p}_0, U)$ can be computed with the single (0, 1)-recursion:

$$\underline{p}_{j+1} = U\underline{p}_j + \beta_{j-1,j} U\underline{p}_{j-1} - \sigma_{j,j}\underline{p}_j \quad \text{for} \quad j = 0, \ldots, 3. \tag{4.7}$$

We set $\beta_{-1,0} = 0$. If $j \geq 1$, $\beta_{j-1,j}$ is chosen to satisfy

$$\langle U\underline{p}_{j-1}, \underline{p}_0\rangle \beta_{j-1,j} = -\langle U\underline{p}_j, \underline{p}_0\rangle,$$

then $\sigma_{j,j}$ is computed by

$$\sigma_{j,j} = \frac{\langle U\underline{p}_j + \beta_{j-1,j} U\underline{p}_{j-1}, \underline{p}_j\rangle}{\langle \underline{p}_j, \underline{p}_j\rangle}.$$

Since an I-orthogonal basis is unique up to scale, the same basis can be obtained using any other recursion that constructs an I-orthogonal basis, for example, the full Gram–Schmidt process given by (2.4).

Let $\underline{p}_0 = [1,\ 0,\ 0,\ 0,\ 0]^T$. By computing $\{\underline{p}_j\}_{j=0}^4$ using (2.4), we obtain

$$\{\underline{p}_0,\ \underline{p}_1,\ \underline{p}_2,\ \underline{p}_3,\ \underline{p}_4\} = \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1/\sqrt{2} \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1/\sqrt{2} \end{pmatrix} \right\}. \tag{4.8}$$

The computation of $\underline{p}_1$ is identical using the (0, 1)-recursion (4.7). To compute $\underline{p}_2$, we must choose $\beta_{0,1}$ to satisfy

$$\langle U\underline{p}_0, \underline{p}_0\rangle \beta_{0,1} = -\langle U\underline{p}_1, \underline{p}_0\rangle.$$

Since $\langle U\underline{p}_0, \underline{p}_0\rangle = \langle U\underline{p}_1, \underline{p}_0\rangle = 0$, $\beta_{0,1}$ can be chosen arbitrarily. Any choice of $\beta_{0,1}$ will yield $\underline{p}_2 = [0, 0, 1, 0, 0]^T$. Similarly, to compute $\underline{p}_3$, we choose $\beta_{1,2}$ to satisfy

$$\langle U\underline{p}_1, \underline{p}_0\rangle \beta_{1,2} = -\langle U\underline{p}_2, \underline{p}_0\rangle.$$

Since $\langle U\underline{p}_1, \underline{p}_0\rangle = 0$ and $\langle U\underline{p}_2, \underline{p}_0\rangle = \frac{1}{\sqrt{2}}$, this is impossible and breakdown occurs at this step.

4.4. A multiple recursion for generalized B-normal(ℓ, m)matrices. This section shows how to reformulate the single $(s,\ t)$ iteration for generalized B-normal$(\ell,\ m)$ matrices as a multiple recursion that avoids the problem of breakdown.

Suppose $d(\underline{p}_0, A) = n$. Recall that a B-orthogonal basis can be computed using a Gram–Schmidt process (2.4) which yields at step n the corresponding matrix equation (2.5). Using (3.5) in the formula for the $\sigma_{i,j}$'s in (2.4) yields

$$\sigma_{i,j} = \frac{\left\langle \underline{p}_j,\ \left(Q_B(A)\tilde{\psi}_{i-m}(A)\underline{p}_0 + p_\ell(A)\tilde{\psi}_{i-m}(A)\underline{p}_0 + A^\dagger \zeta^{(i)}_{m-1}(A)\underline{p}_0 \right) \right\rangle_B}{\langle \underline{p}_i,\ \underline{p}_i \rangle_B}. \tag{4.9}$$

Since $p_\ell(A)\tilde{\psi}_{i-m}(A)\underline{p}_0 \in \mathrm{sp}\{\underline{p}_0,\ \underline{p}_1, \ldots,\ \underline{p}_{i-m+\ell}\}$ for $j > i - m + \ell$,

$$\underline{p}_j \perp_B p_\ell(A)\tilde{\psi}_{i-m}(A)\underline{p}_0 \tag{4.10}$$

and

$$\sigma_{i,j} = \frac{\left\langle \underline{p}_j,\ \left(Q_B(A)\tilde{\psi}_{i-m}(A)\underline{p}_0 + A^\dagger \zeta^{(i)}_{m-1}(A)\underline{p}_0 \right) \right\rangle_B}{\langle \underline{p}_i,\ \underline{p}_i \rangle_B}. \tag{4.11}$$

Since $\zeta^{(i)}_{m-1}(A) \in \mathrm{sp}\{\underline{p}_0,\ \underline{p}_1, \ldots,\ \underline{p}_{m-1}\}$, we obtain that

$$\begin{aligned} \langle \underline{p}_j,\ A^\dagger \zeta^{(i)}_{m-1}(A)\underline{p}_0 \rangle_B &= \langle A\underline{p}_j,\ \zeta^{(i)}_{m-1}(A)\underline{p}_0 \rangle_B \\ &= (\bar{\underline{\rho}}_i)_{m-1}\langle A\underline{p}_j,\ \underline{p}_{m-1}\rangle_B + \cdots + (\bar{\underline{\rho}}_i)_0 \langle A\underline{p}_j,\ \underline{p}_0 \rangle_B \\ &= \langle \underline{\eta}_j,\ \underline{\rho}_i \rangle, \end{aligned}$$

where

$$\underline{\rho}_i = [(\underline{\rho}_i)_0, \ldots,\ (\underline{\rho}_i)_{m-1}]^T \in \mathcal{C}^m \tag{4.12}$$

and

$$\underline{\eta}_j = [\langle A\underline{p}_j,\ \underline{p}_0\rangle_B,\ \langle A\underline{p}_j,\ \underline{p}_1\rangle_B, \ldots,\ \langle A\underline{p}_j,\ \underline{p}_{m-1}\rangle_B]^T \in \mathcal{C}^m. \tag{4.13}$$

By letting $\{\underline{\phi}_0,\ \underline{\phi}_1, \ldots,\ \underline{\phi}_{\kappa-1}\}$ denote a basis for the range of $Q_B(A)$, we can write

$$Q_B(A)\tilde{\psi}_{i-m}(A)\underline{p}_0 = (\underline{\tau}_i)_0\ \underline{\phi}_0 + \cdots + (\underline{\tau}_i)_{\kappa-1}\ \underline{\phi}_{\kappa-1}$$

for some vector

$$\underline{\tau}_i = [(\underline{\tau}_i)_0, \ldots,\ (\underline{\tau}_i)_{\kappa-1}]^T \in \mathcal{C}^\kappa. \tag{4.14}$$

Next, denote

$$\underline{\mu}_j = [\langle \underline{p}_j,\ \underline{\phi}_0\rangle_B,\ \langle \underline{p}_j,\ \underline{\phi}_1\rangle_B, \ldots,\ \langle \underline{p}_j,\ \underline{\phi}_{\kappa-1}\rangle_B]^T \in \mathcal{C}^\kappa \tag{4.15}$$

and notice that

$$\langle \underline{p}_j,\ Q_B(A)\tilde{\psi}_{i-m}(A)\underline{p}_0 \rangle_B = \langle \underline{\mu}_j,\ \underline{\tau}_i \rangle.$$

Therefore, the entries of H_n simplify as

$$h_{i,j} = \begin{cases} \dfrac{\langle \underline{\eta}_j, \underline{\rho}_i \rangle + \langle \underline{\mu}_j, \underline{\tau}_i \rangle}{\langle \underline{p}_i, \underline{p}_i \rangle_B}, & j > \max\{i-1, i-m+\ell\}, \\ \dfrac{\langle A\underline{p}_j, \underline{p}_i \rangle_B}{\langle \underline{p}_i, \underline{p}_i \rangle_B}, & j = i, \ldots, i-m+\ell \ \ (\text{if } \ell \geq m), \\ 1, & j = i-1, \\ 0, & j < i-1. \end{cases} \tag{4.16}$$

Consider the following decomposition of the Hessenberg matrix H_n:

$$H_n = T + U,$$

where U is upper triangular with entries

$$u_{i,j} = \begin{cases} \dfrac{\langle \underline{\eta}_j, \underline{\rho}_i \rangle + \langle \underline{\mu}_j, \underline{\tau}_i \rangle}{\langle \underline{p}_i, \underline{p}_i \rangle_B}, & j \geq i, \\ 0, & j < i \end{cases} \tag{4.17}$$

and T is upper Hessenberg with entries

$$t_{i,j} = \begin{cases} 0, & j > \max\{i-1, i-m+\ell\}, \\ \dfrac{\langle A\underline{p}_j, \underline{p}_i \rangle_B - \langle \underline{\eta}_j, \underline{\rho}_i \rangle - \langle \underline{\mu}_j, \underline{\tau}_i \rangle}{\langle \underline{p}_i, \underline{p}_i \rangle_B}, & j = i, \ldots, i-m+\ell \ \ (\text{if } \ell \geq m), \\ 1, & j = i-1, \\ 0, & j < i-1. \end{cases} \tag{4.18}$$

Notice that T has an upper bandwidth of $\max\{0, \ell - m + 1\}$, so that if $m > \ell$, T has only a nonzero subdiagonal consisting of all ones. The matrix U can be further decomposed as

$$U = U_1 + U_2,$$

where

$$U_1 = \begin{bmatrix} \dfrac{\langle \underline{\eta}_0, \underline{\rho}_0 \rangle}{\langle \underline{p}_0, \underline{p}_0 \rangle_B} & \cdots & \dfrac{\langle \underline{\eta}_{n-1}, \underline{\rho}_0 \rangle}{\langle \underline{p}_0, \underline{p}_0 \rangle_B} \\ & \ddots & \vdots \\ & & \dfrac{\langle \underline{\eta}_{n-1}, \underline{\rho}_{n-1} \rangle}{\langle \underline{p}_{n-1}, \underline{p}_{n-1} \rangle_B} \end{bmatrix},$$

$$U_2 = \begin{bmatrix} \dfrac{\langle \underline{\mu}_0, \underline{\tau}_0 \rangle}{\langle \underline{p}_0, \underline{p}_0 \rangle_B} & \cdots & \dfrac{\langle \underline{\mu}_{n-1}, \underline{\tau}_0 \rangle}{\langle \underline{p}_0, \underline{p}_0 \rangle_B} \\ & \ddots & \vdots \\ & & \dfrac{\langle \underline{\mu}_{n-1}, \underline{\tau}_{n-1} \rangle}{\langle \underline{p}_{n-1}, \underline{p}_{n-1} \rangle_B} \end{bmatrix}. \tag{4.19}$$

The matrices U_1 and U_2 can be factored as follows:

$$\begin{aligned} U_1 &= \begin{bmatrix} \dfrac{\underline{\bar{\rho}}_0^T}{\langle \underline{p}_0, \underline{p}_0 \rangle_B} & & \\ & \ddots & \\ & & \dfrac{\underline{\bar{\rho}}_{n-1}^T}{\langle \underline{p}_{n-1}, \underline{p}_{n-1} \rangle_B} \end{bmatrix} \begin{bmatrix} I_m & \cdots & I_m \\ & \ddots & \vdots \\ & & I_m \end{bmatrix} \begin{bmatrix} \underline{\eta}_0 & & \\ & \ddots & \\ & & \underline{\eta}_{n-1} \end{bmatrix} \\ &= M\ E\ D, \end{aligned} \tag{4.20}$$

$$\begin{aligned} U_2 &= \begin{bmatrix} \dfrac{\underline{\bar{\tau}}_0^T}{\langle \underline{p}_0, \underline{p}_0 \rangle_B} & & \\ & \ddots & \\ & & \dfrac{\underline{\bar{\tau}}_{n-1}^T}{\langle \underline{p}_{n-1}, \underline{p}_{n-1} \rangle_B} \end{bmatrix} \begin{bmatrix} I_\kappa & \cdots & I_\kappa \\ & \ddots & \vdots \\ & & I_\kappa \end{bmatrix} \begin{bmatrix} \underline{\mu}_0 & & \\ & \ddots & \\ & & \underline{\mu}_{n-1} \end{bmatrix} \\ &= \widehat{M}\ \widehat{E}\ \widehat{D}. \end{aligned}$$

The dimensions of the matrices in the factorizations are as follows:

$$\begin{array}{llll} M & \text{is} & n \times (n \times m), & \widehat{M} \quad \text{is} \quad n \times (n \times \kappa), \\ E & \text{is} & (n \times m) \times (n \times m), & \widehat{E} \quad \text{is} \quad (n \times \kappa) \times (n \times \kappa), \\ D & \text{is} & (n \times m) \times n, & \widehat{D} \quad \text{is} \quad (n \times \kappa) \times n. \end{array}$$

The matrices E ($\widehat{E}$) consist of the blocks I_m (I_κ) on and above the diagonal. Each block I_m (I_κ) is an $m \times m$ ($\kappa \times \kappa$) identity matrix. Notice that the matrix E ($\widehat{E}$), being upper triangular with all 1s on the diagonal, is nonsingular. It is easily verified that the inverses of E and $\widehat{E}$ are given by the upper triangular, banded matrices,

$$E^{-1} = \begin{bmatrix} I_m & -I_m & & \\ & \ddots & \ddots & \\ & & \ddots & -I_m \\ & & & I_m \end{bmatrix}, \qquad \hat{E}^{-1} = \begin{bmatrix} I_\kappa & -I_\kappa & & \\ & \ddots & \ddots & \\ & & \ddots & -I_\kappa \\ & & & I_\kappa \end{bmatrix}. \tag{4.21}$$

Using the above factorizations for U_1 and U_2, we obtain the matrix equation

$$\begin{aligned} AP_n &= P_n H_n \\ &= P_n T + P_n U_1 + P_n U_2 \\ &= P_n T + P_n MED + P_n \widehat{M}\widehat{E}\widehat{D}. \end{aligned}$$

Next, we define two matrices of auxiliary vectors:

$$\begin{aligned} Q &= P_n ME, \\ \widehat{Q} &= P_n \widehat{M}\widehat{E}. \end{aligned}$$

Making these substitutions into the above matrix equation yields

$$AP_n = P_n T + QD + \widehat{Q}\widehat{D}.$$

Notice that the dimensions of Q and $\widehat{Q}$ are $n \times (n \times m)$ and $n \times (n \times \kappa)$, respectively. In particular, we will denote

$$\begin{aligned} Q &= \left\{ \underline{q}_{0_0}, \ldots, \underline{q}_{0_{\mathrm{m}-1}},\ \underline{q}_{1_0}, \ldots, \underline{q}_{1_{\mathrm{m}-1}},\ \cdots\ \cdots,\ \underline{q}_{n-1_0}, \ldots, \underline{q}_{n-1_{\mathrm{m}-1}} \right\}, \\ \widehat{Q} &= \left\{ \underline{\hat{q}}_{0_0}, \ldots, \underline{\hat{q}}_{0_{\kappa-1}},\ \underline{\hat{q}}_{1_0}, \ldots, \underline{\hat{q}}_{1_{\kappa-1}},\ \cdots\ \cdots,\ \underline{\hat{q}}_{n-1_0}, \ldots, \underline{\hat{q}}_{n-1_{\kappa-1}} \right\}. \end{aligned} \tag{4.22}$$

Multiplying the matrix equations for Q and $\widehat{Q}$ through by E^{-1} and $\widehat{E}^{-1}$, respectively, we obtain

$$\begin{aligned} QE^{-1} &= P_n M, \\ \widehat{Q}\widehat{E}^{-1} &= P_n \widehat{M}. \end{aligned}$$

It follows that the equations

$$\begin{aligned} AP_n &= P_n T + QD + \widehat{Q}\widehat{D}, \\ QE^{-1} &= P_n M, \\ \widehat{Q}\widehat{E}^{-1} &= P_n \widehat{M} \end{aligned} \tag{4.23}$$

define the matrix form of a set of multiple recursions that can be used to construct a B-orthogonal basis for $\mathcal{K}_n(\underline{p}_0, A)$. Equating the columns in the three matrix equations yields the formulas for the recursions at each step of the iteration. These are given by

$$\underline{p}_{j+1} = A\underline{p}_j - \sum_{i=j-(\ell-m)}^{j} t_{i,j}\ \underline{p}_i - \left[\underline{q}_{j_0}\ \cdots\ \underline{q}_{j_{\mathrm{m}-1}} \right]\ \underline{\eta}_j - \left[\underline{\hat{q}}_{j_0}\ \cdots\ \underline{\hat{q}}_{j_{\kappa-1}} \right]\ \underline{\mu}_j,$$

$$\underline{q}_{j+1_{\mathrm{i}}} = \frac{(\underline{\bar{\rho}}_{j+1})_i}{\langle \underline{p}_{j+1}, \underline{p}_{j+1} \rangle_B}\ \underline{p}_{j+1} + \underline{q}_{j_{\mathrm{i}}} \quad \text{for} \quad i = 0, \ldots, m-1, \tag{4.24}$$

$$\underline{\hat{q}}_{j+1_{\mathrm{i}}} = \frac{(\underline{\bar{\tau}}_{j+1})_i}{\langle \underline{p}_{j+1}, \underline{p}_{j+1} \rangle_B}\ \underline{p}_{j+1} + \underline{\hat{q}}_{j_{\mathrm{i}}} \quad \text{for} \quad i = 0, \ldots, \kappa-1,$$

where the $t_{i,j}$'s are the elements of the matrix T (4.18) and $\underline{\eta}_j$, $\underline{\rho}_j$, $\underline{\mu}_j$, and $\underline{\tau}_j$ are specified in (4.12)–(4.15). Note that these quantities stay bounded; thus, the problem of breakdown in the corresponding single (s, t)-recursion is avoided.

If A is a generalized B-normal(ℓ, m) matrix, with $m = 0$, the multiple recursion (4.24) simplifies; the $\underline{q}$'s are all $\underline{0}$. Another simplification occurs in the special case when A is B-normal$(\bar{\ell}, m)$; that is, $\mathrm{rank}(Q_B(A)) = 0$. In this case, the $\underline{\hat{q}}$'s are all $\underline{0}$. Details for deriving the multiple recursion for the B-normal(ℓ, m) case can be found in [4].

These results are summarized in the following theorem.

THEOREM 4.1. *If there exists polynomials $p_\ell(\lambda)$ and $q_m(\lambda)$ of degree ℓ and m, respectively, such that*

$$Q_B(A) = A^\dagger q_m(A) - p_\ell(A) \quad \text{and} \quad \mathrm{rank}(Q_B(A)) = \kappa, \tag{4.25}$$

then, for every $\underline{r}_0$, a B-orthogonal basis for $\mathcal{K}_d(\underline{r}_0, A)$ can be constructed using the multiple recursion given in (4.24).

Proof. See above discussion. ∎

COROLLARY 4.1. *If $A = \widehat{A} + R_z$, where $\widehat{A}$ is B-normal(ℓ, m) and R_z has rank z, then a B-orthogonal basis can be constructed using the multiple recursions given in* (4.24)*, where $\kappa \leq (\ell + m + 1)z$.*

Proof. The proof follows from Theorem 4.1 and Lemma 3.1. ∎

The recursions given by (4.24) could be applied to any rank z perturbation of a B-normal(ℓ, m) matrix. Since

$$\text{rank}(Q_B(A)) = \kappa \leq (\ell + m + 1)z,$$

if either z, ℓ, or m is large, κ can be large, and the number of recursions and the number of terms needed to compute the basis increases. Thus, the work involved with this implementation is dependent on the degree of the polynomials p_ℓ and q_m, as well as the rank of the perturbation z. This means the practical application of this form of recursion is limited to low rank perturbations of B-normal(ℓ, m) matrices.

5. Implementation and numerical results. This section focuses on the implementation details for the multiple recursion given in (4.24) for generalized B-normal(ℓ, m) matrices. In particular, we describe how to compute a basis for the range of $Q_B(A)$, as well as the computations for the $t_{i,j}$'s and the vectors $\underline{\eta}_j$, $\underline{\mu}_j$, $\underline{\rho}_j$, and $\underline{\tau}_j$ used in the iteration. The multiple recursion given in (4.24) must be modified because all of the information specified at a given step in this recursion cannot be computed until a later step in the iteration. The algorithm is given in Figure 5.1 to clarify the order of the computations. A numerical example is presented to validate the theory developed and to compare the algorithm to other iterative solvers.

5.1. Multiple recursion implementation details.

Computation of the vectors $\underline{\eta}_j$ and $\underline{\mu}_j$. For every j, the computation of the vector $\underline{\eta}_j$ can be done directly using (4.13).

The computation for each $\underline{\mu}_j$ involves the basis vectors for the range of $Q_B(A)$, $\{\underline{\phi}_i\}_{i=0}^{\kappa-1}$. Recall that $A^\dagger = B^{-1}A^*B$, so that

$$Q_B(A) = (B^{-1}A^*B)\ q_m(A) - p_\ell(A).$$

In the general case, a basis for the range of $Q_B(A)$ can be obtained by first computing $Q_B(A)$; then, since $Q_B(A)$ is rank deficient, a basis for its range can be obtained using a QR decomposition routine with column pivoting (see [6, pp. 233–236]). Then each $\underline{\mu}_j$ can be computing directly using (4.15).

The calculations for the vectors $\underline{\mu}_j$ can be simplified in some cases. For example, if

$$A = \widehat{A} + R_z,$$

where $\widehat{A}$ is I-self-adjoint and R_z is a rank z matrix that can be written as an outer product,

$$R_z = W_z V_z^*$$

for some $n \times z$ matrices V_z and W_z of rank z. Since $\widehat{A}$ is I-self-adjoint,

$$\widehat{A}^* = \frac{p_\ell(\widehat{A})}{q_m(\widehat{A})} = \widehat{A}.$$

This yields $q_m(\widehat{A}) = I$ and $p_\ell(\widehat{A}) = \widehat{A}$. Suppose $B = A^*A$. Notice that

$$A^\dagger = B^{-1}A^*B = (A^*A)^{-1}A^*A^*A = A^{-1}A^*A$$

and

$$\begin{aligned} Q_{A^*A}(A) &= A^\dagger q_m(A) - p_\ell(A) = A^\dagger - A \\ &= A^{-1}A^*A - A = A^{-1}(A^* - A)A \\ &= A^{-1}(V_z W_z^* - W_z V_z^*)A. \end{aligned}$$

Denote $\{\underline{\hat{v}}_s\}_{s=0}^{t-1}$ as the linearly independent columns from the matrices V_z and W_z. Notice that

$$\text{Range}\,(Q_{A^*A}(A)) = \text{sp}\{A^{-1}\underline{\hat{v}}_s\}_{s=0}^{t-1} = \text{sp}\{\underline{\phi}_0, \ldots, \underline{\phi}_{\kappa-1}\}, \qquad \text{where} \quad t = \kappa.$$

Using this information, the components of the vectors $\underline{\mu}_j$ (4.15) can be computed as

$$\langle \underline{p}_j, \underline{\phi}_s\rangle_B = \langle A^*A\underline{p}_j, \underline{\phi}_s\rangle = \langle A\underline{p}_j, A(A^{-1}\underline{\hat{v}}_s)\rangle = \langle A\underline{p}_j, \underline{\hat{v}}_s\rangle, \qquad s = 0, \ldots, \kappa - 1.$$

Computation of the vectors $\underline{\tau}_j$. At any step $j+1$, the recursion for the direction vector in (4.24) involves both sets of auxiliary vectors $\underline{q}_j$'s and $\underline{\hat{q}}_j$'s. The vector $\underline{\tau}_j$ is used in the computation of the $\underline{\hat{q}}_j$'s. The method we will describe for computing this vector is not possible until a later step in the iteration. This means the multiple recursion (4.24) must be modified in a way that allows the computation of the $\underline{\hat{q}}_j$'s to be delayed.

Recall from the development of the multiple recursion for generalized B-normal(ℓ, m) matrices that the entries in the Hessenberg matrix H_n simplify, yielding

$$h_{i,j} = \frac{\langle A\underline{p}_j, \underline{p}_i\rangle_B}{\langle \underline{p}_i, \underline{p}_i\rangle_B} = \frac{\langle \underline{\eta}_j, \underline{\rho}_i\rangle + \langle \underline{\mu}_j, \underline{\tau}_i\rangle}{\langle \underline{p}_i, \underline{p}_i\rangle_B} \quad \text{if} \quad j > \max\{i-1,\, i-m+\ell\}. \tag{5.1}$$

Denote

$$\vartheta = \begin{cases} 1 & \text{if } m > \ell, \\ m - \ell & \text{if } m \le \ell. \end{cases}$$

Notice that if $j > i - \vartheta$, we can compute

$$\langle \underline{\mu}_j, \underline{\tau}_i\rangle = \langle A\underline{p}_j, \underline{p}_i\rangle_B - \langle \underline{\eta}_j, \underline{\rho}_i\rangle.$$

All the entries to the right of column $(i - \vartheta)$ in the ith row of H_n involve the same vector $\underline{\tau}_i$. By moving κ places to the right, we can obtain κ equations involving $\underline{\tau}_i$. At step $j + 1 = i - \vartheta + \kappa + 1$, we obtain the system

$$\left[\underline{\mu}_j \cdots \underline{\mu}_{j-\kappa+1}\right]^\tau = \begin{pmatrix} \langle A\underline{p}_j, \underline{p}_i\rangle_B - \langle \underline{\eta}_j, \underline{\rho}_i\rangle \\ \vdots \\ \langle A\underline{p}_{j-\kappa+1}, \underline{p}_i\rangle_B - \langle \underline{\eta}_{j-\kappa+1}, \underline{\rho}_i\rangle \end{pmatrix} \tag{5.2}$$

for $\bar{\underline{\tau}}_i$. For generalized B-normal(ℓ, m) matrices, theory guarantees that (5.1) holds. Thus, if we have chosen $\{\underline{\phi}_0, \ldots, \underline{\phi}_{\kappa-1}\}$ properly—that is, if they form a basis for the range of $Q_B(A)$—it follows that the above system is consistent. We note here that

the computation for the $\underline{\tau}_i$'s requires knowledge of the $\underline{\rho}_i$'s. These calculations will be described below.

From the above, we see that when $j = \kappa - \vartheta$, we can compute $\bar{\underline{\tau}}_0$ using (5.2), and when $j = \kappa - \vartheta + 1$, we can compute $\bar{\underline{\tau}}_1$, and so forth. Denote

$$\theta = \kappa - \vartheta.$$

Whenever $j \geq \theta$, $\bar{\underline{\tau}}_{j-\theta}$ is computable using (5.2), which allows the computation of the $\hat{\underline{q}}_{j-\theta}$'s. However, at step $j+1$ ($j \geq \theta$), the multiple recursion in (4.24) requires us to compute the $\hat{\underline{q}}_j$'s. This is not possible unless $\theta = 0$. The recursion for $\underline{p}_{j+1}$ in (4.24) can be rewritten by successively substituting in the expressions given in (4.24) for the $\hat{\underline{q}}_t$'s for $t = j,\ j-1, \ldots, j-\theta+1$. The same process is repeated using the $\underline{q}_t$'s. This yields

$$\begin{aligned} \underline{p}_{j+1} = A\underline{p}_j - \sum_{i=j-(\ell-m)}^{j} t_{i,j}\ \underline{p}_i - \delta_j\ \underline{p}_j - \cdots - \delta_{j-\theta+1}\ \underline{p}_{j-\theta+1} \\ - \left[\underline{q}_{j-\theta_0} \cdots \underline{q}_{j-\theta_{m-1}}\right] \underline{\eta}_j - \left[\hat{\underline{q}}_{j-\theta_0} \cdots \hat{\underline{q}}_{j-\theta_{\kappa-1}}\right] \underline{\mu}_j, \end{aligned} \tag{5.3}$$

where $\bar{\underline{\tau}}_{j-\theta}$ can be computed using (5.2) and

$$\begin{aligned} \underline{q}_{j-\theta_i} &= \frac{(\bar{\underline{\rho}}_{j-\theta})_i}{\langle \underline{p}_{j-\theta},\ \underline{p}_{j-\theta}\rangle_B}\ \underline{p}_{j-\theta} + \underline{q}_{j-\theta-1_i} \quad \text{for} \quad i = 0, \ldots, m-1, \\ \hat{\underline{q}}_{j-\theta_i} &= \frac{(\bar{\underline{\tau}}_{j-\theta})_i}{\langle \underline{p}_{j-\theta},\ \underline{p}_{j-\theta}\rangle_B}\ \underline{p}_{j-\theta} + \hat{\underline{q}}_{j-\theta-1_i} \quad \text{for} \quad i = 0, \ldots, \kappa-1. \end{aligned} \tag{5.4}$$

The $t_{i,j}$'s are given by (4.18) and

$$\delta_t = \left(\sum_{i=0}^{m-1} \frac{(\bar{\underline{\rho}}_t)_i}{\langle \underline{p}_t,\ \underline{p}_t\rangle_B} (\underline{\eta}_j)_i + \sum_{i=0}^{\kappa-1} \frac{(\bar{\underline{\tau}}_t)_i}{\langle \underline{p}_t,\ \underline{p}_t\rangle_B} (\underline{\mu}_j)_i \right) \quad \text{for} \quad t = j-\theta+1, \ldots, j.$$

At first glance, it looks like the information needed to compute the δ_t's and the $t_{i,j}$'s is not available. Let

$$\varphi = \begin{cases} j - \theta + 1 & \text{if } m > \ell, \\ \min\{j - (\ell - m),\ j - \theta + 1\} & \text{if } m \leq \ell. \end{cases} \tag{5.5}$$

Instead of calculating the δ_t's and $t_{i,j}$'s as specified, we rewrite (5.3) as

$$\begin{aligned} \underline{y}_{j+1} &= A\underline{p}_j - \left[\underline{q}_{j-\theta_0} \cdots \underline{q}_{j-\theta_{m-1}}\right] \underline{\eta}_j - \left[\hat{\underline{q}}_{j-\theta_0} \cdots \hat{\underline{q}}_{j-\theta_{\kappa-1}}\right] \underline{\mu}_j, \\ \underline{p}_{j+1} &= \underline{y}_{j+1} - \sum_{i=\varphi}^{j} \hat{t}_{i,j}\ \underline{p}_i, \end{aligned} \tag{5.6}$$

where the $\hat{t}_{i,j}$'s must enforce B-orthogonality of $\underline{p}_{j+1}$ to $\mathrm{sp}\{\underline{p}_\varphi, \ldots, \underline{p}_j\}$, which yields

$$\hat{t}_{i,j} = \frac{\langle \underline{y}_{j+1}, \underline{p}_i\rangle_B}{\langle \underline{p}_i, \underline{p}_i\rangle_B}. \tag{5.7}$$

Computation of the vectors $\underline{\rho}_j$. Unlike the computation for the $\underline{\tau}_j$'s, which must be delayed, the $\underline{\rho}_j$'s can be computed at each step. This is done recursively once the information is available to start the process.

Suppose $\underline{v} = \psi(A)\underline{p}_0$, where $\psi(A)$ is some polynomial of A. Denote

$$\text{rem}(\underline{v}) = \zeta(A)\underline{p}_0,$$

where $\zeta(A)$ is the remainder that results from the division of $\psi(A)$ by $q_m(A)$. Now, $\text{rem}(\underline{v}) \in \text{sp}\{\underline{p}_0, \ldots, \underline{p}_{m-1}\}$, that is,

$$\text{rem}(\underline{v}) = \nu_{m-1}\underline{p}_{m-1} + \cdots + \nu_0\underline{p}_0$$

for some coefficients $\nu_0, \ldots, \nu_{m-1}$. Define

$$\underline{rv} = [\nu_0, \ldots, \nu_{m-1}]^T \in C^m$$

to be the vector composed of these coefficents. In this notation, the vectors $\underline{\rho}_j$ defined in section 4.4 can be written

$$\underline{rp}_j = \underline{\rho}_j.$$

Before we can begin the recursive computation for the vectors $\underline{\rho}_j$, some startup information is required. This information can be obtained by computing the first $m+1$ direction vectors $\{\underline{p}_0, \underline{p}_1, \ldots, \underline{p}_m\}$ using a full Gram–Schmidt process (2.4) or some modified version of it. Recall that for every k, $\underline{p}_k$ can be written as

$$\underline{p}_k = \psi_k(A)\underline{p}_0 = [\alpha_k A^k + \cdots + \alpha_1 A + \alpha_0 I]\underline{p}_0$$

for some polynomial of exact degree k.

No computation is necessary to obtain the vectors $\underline{\rho}_0, \ldots, \underline{\rho}_{m-1}$. For $k = 0, \ldots, m-1$, the remainder of $\underline{p}_k$ upon division by $q_m(A)$ is itself, $\text{rem}(\underline{p}_k) = 1\ \underline{p}_k$, which yields

$$\underline{rp}_k = \underline{\rho}_k = [0 \cdots 0\ 1\ 0 \cdots 0]^T \in \mathcal{C}^m,$$

which is a vector of all zeros, except for a 1 in the kth position.

The computation of $\underline{\rho}_m$ is more complicated. To facilitate this, we write the polynomial $q_m(A)$ given in (3.2) as a linear combination of the polynomials $\{\psi_k\}_{k=0}^m$

$$\begin{aligned} q_m(A) &= \gamma_m\psi_m(A) + \cdots + \gamma_1\psi_1(A) + \gamma_0\psi_0(A) \\ &= b_m A^m + \cdots + b_0 I. \end{aligned} \tag{5.8}$$

This requires that we accumulate the polynomials $\{\psi_0, \ldots, \psi_m\}$ during the iteration (2.4). The coefficients of each $\psi_k(A)$ are stored in a vector of length $m+1$, which we denote

$$\underline{cp}_k = [\alpha_0,\ \alpha_1, \ldots,\ \alpha_k,\ 0, \ldots,\ 0]_{m+1}^T.$$

Starting with $\underline{p}_0 = \psi_0(A)\underline{p}_0 = 1\ \underline{p}_0$, we obtain

$$\underline{cp}_0 = [1\ 0\ \cdots\ 0]_{m+1}^T.$$

Using (2.4), we see that the vector of coefficients for the polynomials ψ_{j+1} for $j = 0, \ldots, m-1$ can be updated using

$$\underline{cp}_{j+1} = \underline{cAp}_j - \sum_{i=0}^{j} \sigma_{i,j}\ \underline{cp}_i.$$

Since

$$A\underline{p}_j = A\psi_j(A)\underline{p}_0 = [\alpha_j A^{j+1} + \cdots + \alpha_1 A^2 + \alpha_0 A]\underline{p}_0,$$

the corresponding vector of coefficients is obtained by shifting the vector $\underline{cp}_j$ one place to the right, i.e.,

$$\underline{cAp}_j = [0,\ \alpha_0,\ \alpha_1, \ldots,\ \alpha_j,\ 0, \ldots,\ 0]^T_{m+1}.$$

The vectors $\underline{cp}_0, \ldots, \underline{cp}_m$ are stored in the matrix K, which has the following form:

$$K = \begin{bmatrix} | & | & & | \\ \underline{cp}_0 & \underline{cp}_1 & \cdots & \underline{cp}_m \\ | & | & & | \end{bmatrix} = \begin{bmatrix} 1 & * & \cdots & \cdots & * \\ 0 & 1 & \ddots & & \vdots \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & * \\ 0 & \cdots & \cdots & 0 & 1 \end{bmatrix},$$

where the $*$'s denote possible nonzero entries. Finally, the γ's in (5.8) can be obtained by solving the system

$$K[\gamma_0, \ldots, \gamma_m]^T = [b_0, \ldots, b_m]^T \tag{5.9}$$

using backsubstitution.

Using the form given in (5.8) for $q_m(A)$, it is easy to see that the remainder of $\underline{p}_m$ upon division by $q_m(A)$ is given by

$$\begin{aligned} \operatorname{rem}(\underline{p}_m) &= -\tfrac{\gamma_{m-1}}{\gamma_m}\ \underline{p}_{m-1} - \cdots - \tfrac{\gamma_1}{\gamma_m}\ \underline{p}_1 - \tfrac{\gamma_0}{\gamma_m}\ \underline{p}_0, \\ \underline{\rho}_m &= \left[-\tfrac{\gamma_0}{\gamma_m},\ -\tfrac{\gamma_1}{\gamma_m}, \ldots, -\tfrac{\gamma_{m-1}}{\gamma_m}\right]^T. \end{aligned} \tag{5.10}$$

When $j \geq m$, we begin updating $\underline{\rho}_{j+1}$ recursively. Since $\underline{p}_0$ and each subsequent direction vector $\underline{p}_{j+1}$ can be written as a polynomial in A times $\underline{p}_0$ for some polynomial ψ_{j+1} of exact degree $j+1$, each auxiliary vector, being the sum of the current direction vector and a previous auxiliary vector, can also be expressed as a polynomial in A times $\underline{p}_0$.

If $m \leq j < \theta$, $\underline{p}_{j+1}$ cannot be computed using (5.6). Instead, $\underline{p}_{j+1}$ is computed using (2.4). Each of the terms in (2.4) can be written as a polynomial in A times $\underline{p}_0$. The remainder of $\underline{p}_{j+1}$ upon division by $q_m(A)$ can be obtained by dividing the individual polynomials by $q_m(A)$, then adding their corresponding remainder terms. This yields

$$\begin{aligned} \operatorname{rem}(\underline{p}_{j+1}) &= \operatorname{rem}(A\underline{p}_j) - \textstyle\sum_{i=0}^{j} \sigma_{ij}\ \operatorname{rem}(\underline{p}_j), \\ \underline{rp}_{j+1} &= \underline{rAp}_j - \textstyle\sum_{i=0}^{j} \sigma_{i,j}\ \underline{rp}_i. \end{aligned} \tag{5.11}$$

If $j \geq \theta$, $\bar{\tau}_{j-\theta}$ is computable using (5.2), the $\underline{q}_{j-\theta}$'s and the $\hat{\underline{q}}_{j-\theta}$'s can be computed using (5.4), and $\underline{p}_{j+1}$ can be computed using (5.6). The remainder terms

are updated using

$$
\begin{aligned}
\underline{r\hat{q}}_{j-\theta_i} &= \frac{(\bar{\tau}_{j-\theta})_i}{\langle \underline{p}_{j-\theta},\ \underline{p}_{j-\theta}\rangle_B}\ \underline{rp}_{j-\theta} + \underline{r\hat{q}}_{j-\theta-1_i} \quad \text{for} \quad i = 0, \ldots, \kappa-1,\\
\underline{rq}_{j-\theta_i} &= \frac{(\bar{\rho}_{j-\theta})_i}{\langle \underline{p}_{j-\theta}, \underline{p}_{j-\theta}\rangle_B}\ \underline{rp}_{j-\theta} + \underline{rq}_{j-\theta-1_i} \quad \text{for} \quad i = 0, \ldots, m-1,\\
\underline{ry}_{j+1} &= \underline{rAp}_j - \sum_{i=0}^{m-1} (\underline{\eta}_j)_i\ \underline{rq}_{j-\theta_i} - \sum_{i=0}^{\kappa-1} (\underline{\mu}_j)_i\ \underline{r\hat{q}}_{j-\theta_i},\\
\underline{rp}_{j+1} &= \underline{ry}_{j+1} - \sum_{i=\varphi}^{j} \hat{t}_{i,j}\ \underline{rp}_i.
\end{aligned}
\tag{5.12}
$$

The recursive update of $\underline{\rho}_{j+1}$ using either (5.12) if $j \geq \theta$ or (5.11) if $j < \theta$ involves the term $\underline{rAp}_j$. To compute this, notice that

$$
\begin{aligned}
\text{rem}(A\underline{p}_j) &= \text{rem}(A\psi_j(A)\underline{p}_0)\\
&= \text{rem}\left(A\left[\tilde{\psi}_{j-m}(A)q_m(A)\underline{p}_0 + \zeta^{(j)}_{m-1}(A)\underline{p}_0\right]\right)\\
&= \text{rem}\ (A\ \text{rem}(\underline{p}_j)).
\end{aligned}
$$

Since $\text{rem}(\underline{p}_j) = (\underline{\rho}_j)_0\underline{p}_0 + \cdots + (\underline{\rho}_j)_{m-1}\underline{p}_{m-1}$,

$$
\begin{aligned}
A\ \text{rem}(\underline{p}_j) &= A[\underline{p}_0, \ldots, \underline{p}_{m-1}]\ \underline{\rho}_j = [\underline{p}_0 \cdots \underline{p}_m]H_{m+1,m}\ \underline{\rho}_j\\
&= c_0\underline{p}_0 + c_1\underline{p}_1 + \cdots + c_m\underline{p}_m,
\end{aligned}
$$

where the second equality follows by using (2.5), and

$$
\underline{c}_j = [c_0,\ c_1, \ldots,\ c_m]^T = H_{m+1,m}\ \underline{\rho}_j. \tag{5.13}
$$

Therefore,

$$
\begin{aligned}
\text{rem}(A\underline{p}_j) &= \text{rem}(A\ \text{rem}(\underline{p}_j))\\
&= c_0\ \text{rem}(\underline{p}_0) + c_1\ \text{rem}(\underline{p}_1) + \cdots + c_m\ \text{rem}(\underline{p}_m)\\
&= c_0\underline{p}_0 + c_1\underline{p}_1 + \cdots + c_m\left[-\tfrac{\gamma_{m-1}}{\gamma_m}\underline{p}_{m-1} - \cdots - \tfrac{\gamma_0}{\gamma_m}\underline{p}_0\right],
\end{aligned}
$$

which yields

$$
\underline{rAp}_j = c_0\underline{\rho}_0 + c_1\underline{\rho}_1 + \cdots + c_m\underline{\rho}_m. \tag{5.14}
$$

The multiple recursion conjugate (MRCG) algorithm for generalized B-normal$(\ell,\ m)$ matrices is given in Figure 5.1.

5.2. Numerical results. One useful application of the multiple recursion given in (4.24) is on matrices that result from a discretization of a partial differential equation that is formally self-adjoint, where nonstandard boundary conditions are applied to a portion of the boundary. For example, suppose

$$
A = \widehat{A} + R_z,
$$

Input: A, initial guess $\underline{x}_0$, $\underline{p}_0 = \underline{r}_0$, coefficients of $q_m(A)$, $\{\underline{\phi}_0, \ldots, \underline{\phi}_{\kappa-1}\}$,
$\theta,\ \kappa,\ \ell,\ m,\ \underline{cp}_0 = [1\ 0\ \cdots\ 0]^T_{m+1}, \quad \underline{\rho}_0 = [1\ 0\ \cdots\ 0]^T_m,$
for $j = 0, \ldots, \theta - 1,$

$\underline{x}_{j+1} = \underline{x}_j + \alpha_j \underline{p}_j, \qquad \underline{r}_{j+1} = \underline{r}_j - \alpha_j A\underline{p}_j, \qquad \alpha_j = \frac{\langle B\underline{e}_j, \underline{p}_j\rangle}{\langle B\underline{p}_j, \underline{p}_j\rangle},$

$\underline{p}_{j+1} = A\underline{p}_j - \sum_{i=0}^{j} \sigma_{i,j}\ \underline{p}_i, \quad \sigma_{i,j} = \frac{\langle BA\underline{p}_j, \underline{p}_i\rangle}{\langle B\underline{p}_i, \underline{p}_i\rangle},$

if $j < m$,

$\underline{cp}_{j+1} = \underline{cAp}_j - \sum_{i=0}^{j} \sigma_{i,j}\ \underline{cp}_j, \qquad \underline{cAp}_j = [0\ \underline{cp}_j]^T_{m+1},$

if $j < m-1$, $\underline{\rho}_{j+1} = [0 \cdots 0\ 1\ 0 \cdots 0]^T_m,$

if $j = m-1$, solve (5.9) for $\underline{\gamma}$, $\underline{\rho}_{j+1} = \left[-\frac{\gamma_0}{\gamma_m}, \ldots, -\frac{\gamma_{m-1}}{\gamma_m}\right]^T_m,$

if $j \geq m$,

$\underline{c}_j = H_{m+1,m}\ \underline{\rho}_j, \quad \underline{rAp}_j = c_0\underline{\rho}_0 + \cdots + c_m\underline{\rho}_m,$

$\underline{\rho}_{j+1} = \underline{rAp}_j - \sum_{i=0}^{j} \sigma_{i,j}\ \underline{\rho}_i,$

for $j = \theta, \ldots,$

$\underline{x}_{j+1} = \underline{x}_j + \alpha_j \underline{p}_j, \qquad \underline{r}_{j+1} = \underline{r}_j - \alpha_j A\underline{p}_j, \qquad \alpha_j = \frac{\langle B\underline{e}_j, \underline{p}_j\rangle}{\langle B\underline{p}_j, \underline{p}_j\rangle},$

solve (5.2) for $\underline{\bar{\tau}}_{j-\theta}$,

$\underline{q}_{j-\theta_i} = \frac{(\underline{\bar{\rho}}_{j-\theta})_i}{\langle \underline{p}_{j-\theta}, \underline{p}_{j-\theta}\rangle_B}\ \underline{p}_{j-\theta} + \underline{q}_{j-\theta-1_i}, \qquad i = 0, \ldots, m-1,$

$\hat{\underline{q}}_{j-\theta_i} = \frac{(\underline{\bar{\tau}}_{j-\theta})_i}{\langle \underline{p}_{j-\theta}, \underline{p}_{j-\theta}\rangle_B}\ \underline{p}_{j-\theta} + \hat{\underline{q}}_{j-\theta-1_i}, \qquad i = 0, \ldots, \kappa-1,$

$\underline{rq}_{j-\theta_i} = \frac{(\underline{\bar{\rho}}_{j-\theta})_i}{\langle \underline{p}_{j-\theta}, \underline{p}_{j-\theta}\rangle_B}\ \underline{\rho}_{j-\theta} + \underline{rq}_{j-\theta-1_i}, \quad i = 0, \ldots, m-1,$

$\underline{r\hat{q}}_{j-\theta_i} = \frac{(\underline{\bar{\tau}}_{j-\theta})_i}{\langle \underline{p}_{j-\theta}, \underline{p}_{j-\theta}\rangle_B}\ \underline{\rho}_{j-\theta} + \underline{r\hat{q}}_{j-\theta-1_i}, \quad i = 0, \ldots, \kappa-1,$

if $j < m$,

$\underline{p}_{j+1} = A\underline{p}_j - \sum_{i=0}^{j} \sigma_{i,j}\ \underline{p}_j, \qquad \sigma_{i,j} = \frac{\langle BA\underline{p}_j, \underline{p}_i\rangle}{\langle B\underline{p}_i, \underline{p}_i\rangle},$

$\underline{cp}_{j+1} = \underline{cAp}_j - \sum_{i=0}^{j} \sigma_{i,j}\ \underline{cp}_j, \qquad \underline{cAp}_j = [0\ \underline{cp}_j]^T_{m+1},$

if $j < m-1$, $\underline{\rho}_{j+1} = [0 \cdots 0\ 1\ 0 \cdots 0]^T_m,$

if $j = m-1$, solve (5.9) for $\underline{\gamma}$, $\underline{\rho}_{j+1} = \left[-\frac{\gamma_0}{\gamma_m}, \ldots, -\frac{\gamma_{m-1}}{\gamma_m}\right]^T_m,$

if $j \geq m$,

$\underline{\eta}_j = [\langle A\underline{p}_j, \underline{p}_0\rangle_B, \ldots, \langle A\underline{p}_j, \underline{p}_{m-1}\rangle_B]^T_m,$

$\underline{\mu}_j = [\langle \underline{p}_j, \underline{\phi}_0\rangle_B, \ldots, \langle \underline{p}_j, \underline{\phi}_{\kappa-1}\rangle_B]^T_\kappa,$

$\underline{y}_{j+1} = A\underline{p}_j - \left[\underline{q}_{j-\theta_0} \cdots \underline{q}_{j-\theta_{m-1}}\right] \underline{\eta}_j - \left[\hat{\underline{q}}_{j-\theta_0} \cdots \hat{\underline{q}}_{j-\theta_{\kappa-1}}\right] \underline{\mu}_j,$

$\underline{p}_{j+1} = \underline{y}_{j+1} - \sum_{i=\varphi}^{j} \hat{t}_{i,j}\ \underline{p}_i, \qquad \hat{t}_{i,j} = \frac{\langle B\underline{y}_{j+1}, \underline{p}_i\rangle}{\langle B\underline{p}_i, \underline{p}_i\rangle},$

$\underline{c}_j = H_{m+1,m}\ \underline{\rho}_j, \qquad \underline{rAp}_j = c_0\underline{\rho}_0 + \cdots + c_m\underline{\rho}_m,$

$\underline{ry}_{j+1} = \underline{rAp}_j - \sum_{i=0}^{m-1} (\underline{\eta}_j)_i\ \underline{rq}_{j-\theta_i} - \sum_{i=0}^{\kappa-1} (\underline{\mu}_j)_i\ \underline{r\hat{q}}_{j-\theta_i},$

$\underline{\rho}_{j+1} = \underline{ry}_{j+1} - \sum_{i=\varphi}^{j} \hat{t}_{i,j}\ \underline{\rho}_i.$

FIG. 5.1. *MRCG algorithm.*

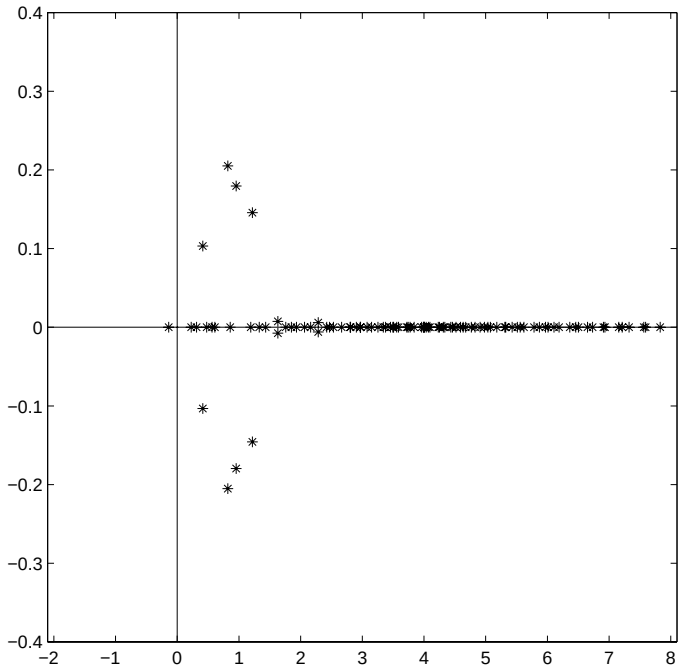


FIG. 5.2. *Spectrum of matrix* $A = \widehat{A} + R_z$.

where $\widehat{A}$ is the symmetric pentadiagonal matrix that results from discretizing Poisson's equation on the unit square with homogeneous Dirichlet boundary conditions,

$$\begin{aligned} -\Delta \underline{u} &= \underline{f} \quad \text{on} \quad (0,1) \times (0,1), \\ \underline{u} &= \underline{0} \quad \text{on} \quad \partial\Omega, \end{aligned}$$

using a five-point difference scheme with a uniform mesh size, $h_x = h_y = \frac{1}{11}$, yielding $n = 100$. The matrix R_z is a nonsymmetric, rank 8 matrix that was chosen to correspond to perturbing the boundary conditions at a few points near the four corners of the domain Ω. Specifically, the matrix A was obtained from the matrix $\hat{A}$, which has the five-point stencil $[-1, \ldots, -1, 4, -1, \ldots, -1]$, by setting the first upper diagonal to 0 and the diagonal to 1 in the first five rows and in the 91st row, and also setting the first subdiagonal to 0 and the diagonal to 1 in the 10th and 100th rows. Note that the matrix A is I-normal(1) (self-adjoint) plus rank 8. The spectrum is plotted in Figure 5.2. Notice that the matrix is slightly indefinite, with a few eigenvalues off the real line.

From section 3.2 we know that there exist polynomials $p_1(A) = A$ and $q_0(A) = I$ such that

$$A^* q_0(A) - p_1(A) = A^* - A = R_z^* - R_z = Q_I(A), \quad \text{where} \quad \operatorname{rank}(Q_I(A)) \leq 2z = 16.$$

Recall that a conjugate gradient method ($\mathcal{CG}(B, A)$ with $B = I$) does not yield a computable algorithm, since at each step we must compute the step length α_j (2.3) which involves $\underline{e}_j$, the error at step j (see section 2.2). To obtain a computable

algorithm, let $B = A^*A$ and notice that

$$\begin{aligned} A^\dagger &= B^{-1}A^*B = A^{-1}A^*A, \\ Q_{A^*A}(A) &= A^\dagger q_m(A) - p_\ell(A) \\ &= A^{-1}\left[A^* q_m(A) - p_\ell(A)\right] A \\ &= A^{-1} Q_I(A) A, \end{aligned}$$

and thus $\mathrm{rank}(Q_{A^*A}(A)) = \mathrm{rank}(Q_I(A)) \leq 16$. It follows from Theorem 4.1 that a minimal-residual method, $\mathcal{CG}(A^*A, A)$, can be implemented using the multiple recursion (4.24), with $\ell = 1, m = 0$, and $\kappa = 16$.

Figure 5.3 compares the convergence measured in the relative residual norm of $\mathcal{CG}(A^*A, A)$, implemented using the multiple recursion conjugate gradient (MRCG) algorithm given in Figure 5.1, to that obtained using a full conjugate gradient iteration, where the direction vectors are computed using a full Gram–Schmidt process (2.4) (Orthodir algorithm). As theory predicts, these two curves are identical. In addition, we plot the convergence using a three-term recursion, (2.7) with $s = 1$ (Odir algorithm). Recall that the Odir algorithm will yield the same iterates in exact arithmetic as Orthodir if the matrix A is B-normal(1). This would be the case if we did not add the low rank matrix R_z and $A = \widehat{A}$. Note that Odir makes no progress after the first few iterations.

For this problem, we might consider how the MRCG algorithm compares with other iterative solvers for nonsymmetric matrices. Figure 5.3 also plots the convergence of the QMR (quasi-minimal residual) iteration of Freund and Nachtigal [7]. In contrast to the conjugate gradient method $\mathcal{CG}(B, A)$ with $B = A^*A$, which minimizes the 2-norm of the residual, QMR is based upon a quasi minimization of the residual norm. QMR can be viewed as a "variable metric" conjugate gradient method [2]. This is a conjugate gradient method where the inner product matrix B is dependent upon the initial residual. Regardless of how we view this method for nonsymmetric matrices, QMR is not a true minimal-residual method. Thus, when convergence is measured in the residual norm, QMR cannot do any better than $\mathcal{CG}(A^*A, A)$. For this problem, the number of QMR iterations is comparable to the number of iterations used by the MRCG algorithm. If we consider the main work involved in these two implementations to be the number of matrix vector multiplications, then the MRCG algorithm produces quite a savings over the QMR iteration. QMR is implemented via the nonsymmetric Lanczos process using a "look-ahead" strategy [8]. This process generates two sequences of vectors, $\{\underline{v}_i\}_{i=1}^n$ and $\{\underline{w}_i\}_{i=1}^n$, such that

$$\begin{aligned} \mathrm{sp}\{\underline{v}_1, \ldots, \underline{v}_n\} &= \mathcal{K}_n(\underline{v}_1, A), \\ \mathrm{sp}\{\underline{w}_1, \ldots, \underline{w}_n\} &= \mathcal{K}_n(\underline{w}_1, A^*). \end{aligned}$$

This involves two matrix vector multiplications at each step. The MRCG algorithm requires only one matrix vector multiplication at each step.

6. Conclusion. The conjugate gradient method is implemented via the construction of a B-orthogonal basis for the underlying Krylov subspace. When this construction can be accomplished using some form of a short recurrence, an economical conjugate gradient algorithm exists.

The theory from Faber and Manteuffel [5] shows that a practical implementation of a conjugate gradient method using a single $(s+2)$-term recursion (2.7) is limited to a very small class of matrices called B-normal(s) matrices. The work done by Gragg, Jagels, and Reichel [9], [11], [12] on unitary and shifted unitary matrices proves that

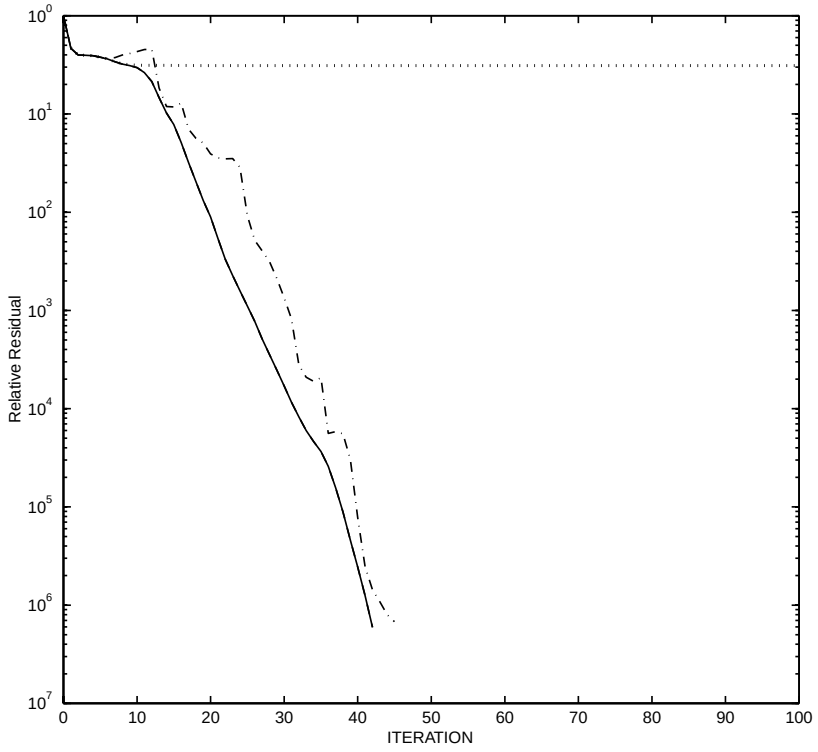


FIG. 5.3. *Convergence of Orthodir algorithm (solid line), MRCG algorithm (dashed line, plotted on top of solid line), the Odir algorithm (dotted line), and the QMR algorithm (dashed-dotted line).*

the single $(s+2)$-term recursion is not general enough to include all possible forms of short recurrences. Unitary and shifted unitary matrices are examples of matrices that are not B-normal(s) for small degree s; thus, no short $(s+2)$-term recursion exists. However, an orthonormal basis can be constructed using a short double recursion. This double recursion can be rewritten as a single (s, t)-recursion (4.1), where, if A is unitary, $(s, t) = (0, 1)$, and if A is shifted unitary, $(s, t) = (1, 1)$. The single (s, t)-recursion is a more general form of short recursion; by letting $t = 0$, we obtain the $(s+2)$-term recursion for B-normal(s) matrices.

A more general class of normal matrices, called B-normal (ℓ, m) matrices, was defined. These are normal matrices whose B-adjoint can be written as the ratio of two polynomials in A,

$$A^\dagger = \frac{p_\ell(A)}{q_m(A)}.$$

Generalized B-normal(ℓ, m) matrices satisfy

$$A^\dagger q_m(A) - p_\ell(A) = Q_B(A), \qquad \text{rank}(Q_B(A)) = \kappa$$

for some polynomials q_m and p_ℓ of degrees m and ℓ, respectively. These include matrices of the form $A = \hat{A} + R_z$, where $\hat{A}$ is B-normal(ℓ, m) and R_z has rank z. These matrices also include B-normal(ℓ, m) matrices as a special case, where $\text{rank}(Q_B(A)) = 0$.

We showed that in the absence of a condition called breakdown, if A is a generalized B-normal(ℓ, m) matrix, an (s, t)-recursion can be used to construct a B-orthogonal basis. The number of terms s and t depends on ℓ and m, as well as the rank of $Q_B(A)$. For implementation purposes, we showed how to reformulate the problem as a multiple recursion that won't break down. Since matrices of this form can be nonnormal, this work demonstrates a short recurrence for certain nonnormal matrices of nontrivial degree.

This paper has focused on sufficient conditions for a conjugate gradient method to be implemented with either a single (s, t)-recursion or a corresponding multiple recursion. More will be said about the relationship between these two forms of recursions in part II, which focuses on necessary conditions for a conjugate gradient method to be implemented with an (s, t)-recursion or a multiple recursion of a specific form. In addition to extending the class of matrices for which a practical conjugate gradient algorithm exists, this research opens the door to the possibility that short recurrences exist for even a wider class of matrices.

Acknowledgments. The authors thank the referees for the many useful suggestions that helped improve the presentation of this material.

REFERENCES

[1] S. F. Ashby, T. A. Manteuffel, and P. E. Saylor, *A taxonomy for conjugate gradient methods*, SIAM J. Numer. Anal., 27 (1990), pp. 1542–1568.

[2] T. Barth and T. A. Manteuffel, *Variable metric conjugate gradient methods*, in PCG '94: Matrix Analysis and Parallel Computing, M. Natori and T. Nodera, eds., Yokohama, 1994.

[3] T. Barth and T. A. Manteuffel, *Conjugate gradient algorithms using multiple recursions*, in Linear and Nonlinear Conjugate Gradient-Related Methods, L. Adams and J. L. Nazareth, eds., SIAM, Philadelphia, PA, 1996, pp. 107–123.

[4] T. Barth, *Implementation of the Conjugate Gradient Method Using Short Multiple Recursions*, Ph.D. thesis, Center for Computational Mathematics, University of Colorado at Denver, Denver, CO, 1996.

[5] V. Faber and T. A. Manteuffel, *Necessary and sufficient conditions for the existence of a conjugate gradient method*, SIAM J. Numer. Anal., 21 (1984), pp. 352–362.

[6] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1989.

[7] R. W. Freund and N. M. Nachtigal, *QMR: A quasi-minimal residual method for non-Hermitian linear systems*, Numer. Math., 60 (1991), pp. 315–339.

[8] R. W. Freund, M. H. Gutknecht, and N. M. Nachtigal, *An implementation of the look-ahead Lanczos algorithm for non-Hermitian matrices*, SIAM J. Sci. Comput., 14 (1993), pp. 137–158.

[9] W. B. Gragg, *Positive definite Toeplitz matrices, the Arnoldi process for isometric operators, and Gaussian quadrature on the unit circle*, J. Comput. Appl. Math, 46 (1993), pp. 183–198.

[10] M. R. Hestenes and E. L. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research. Nat. Bur. Standards, 49 (1952), pp. 409–436.

[11] C. F. Jagels and L. Reichel, *The isometric Arnoldi process and an application to iterative solution of large linear systems*, in Iterative Methods in Linear Algebra, R. Beauwens and P. de Groen, eds., Elsevier, Amsterdam, 1992, pp. 361–369.

[12] C. F. Jagels and L. Reichel, *A fast minimal residual algorithm for shifted unitary matrices*, Numer. Linear Algebra Appl., 1 (1994), pp. 555–570.

[13] Y. Saad and M. H. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[14] D. S. Watkins, *Some perspectives on the eigenvalue problem*, SIAM Rev., 35 (1993), pp. 430–471.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 3, pp. 797–808
© 2000 Society for Industrial and Applied Mathematics

ON THE OPTIMALITY OF THE BACKWARD GREEDY ALGORITHM FOR THE SUBSET SELECTION PROBLEM*

CHRISTOPHE COUVREUR† AND YORAM BRESLER‡

Abstract. The following linear inverse problem is considered: Given a full column rank $m \times n$ data matrix A and a length m observation vector b, find the best least-squares solution to $Ax = b$ with at most $r < n$ nonzero components. The backward greedy algorithm computes a sparse solution to $Ax = b$ by removing greedily columns from A until r columns are left. A simple implementation based on a QR downdating scheme using Givens rotations is described. The backward greedy algorithm is shown to be optimal for the subset selection problem in the sense that it selects the "correct" subset of columns from A if the perturbation of the data vector b is small enough. The results generalize to any other norm of the residual.

Key words. subset selection, sparse least-squares solutions, backward greedy algorithm, NP-hard

AMS subject classifications. 65F30, 05A05, 65U05, 65F05

PII. S0895479898332928

1. Introduction.

The problem of computing a sparse approximate solution to a set of linear equations is important in matrix computations, where it is known as the subset selection problem [4]. Such sparse approximate solutions have applications in statistical regression [6], function interpolation [10], and signal processing [2, 5, 7].

In this paper, the following formulation of the subset selection problem is considered. Given a data matrix $A \in \mathbb{C}^{m \times n}$, $m \geq n$, and an observation vector $b \in \mathbb{C}^m$, find the best least-squares solution to $Ax = b$ with at most r nonzero components. That is, find the subset of r columns from A that gives the best approximation of b in the sense that $\|Ax - b\|_2$ is minimized over all vectors x with at most r nonzero components. Alternately, the problem can be viewed as estimating $x_0 \in \mathbb{C}^n$ from b and the a priori knowledge that x_0 is sparse when

$$\begin{aligned} b &= Ax_0 + \eta \\ &= b_0 + \eta \end{aligned} \tag{1}$$

and η is an unknown noise vector. For instance, in [2], b is a noisy observation of a linear mixture of signal vectors. A dictionary of possible signals is available, but the subset of signals from the dictionary present in the mixture is unknown. Their mixing proportions are also unknown. It is desired to find the signals that are effectively present in b and their mixing proportions. This estimation and detection problem can be naturally formulated as a subset selection problem of the form (1) by letting the columns of A and the elements of x_0 represent the possible signals in the dictionary

*Received by the editors January 22, 1998; accepted for publication (in revised form) by D. P. O'Leary March 2, 1999; published electronically February 16, 2000.

http://www.siam.org/journals/simax/21-3/33292.html

†Physics Department, Faculté Polytechnique de Mons, Mons, Belgium (christophe.couvreur@lhs.be). Current address: Lernout & Hauspie Speech Products, Konig Albert I laan 64, B-1780 Wemmel, Belgium. This author was supported in part by the National Fund for Scientific Research of Belgium (F.N.R.S.) and by a Francqui Fellowship of the Belgian-American Educational Foundation.

‡Coordinated Science Laboratory and Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, 1308 W. Main St., Urbana, IL 61801 (ybresler@uiuc.edu). This author was supported in part by National Science Foundation grant MIP 97-07633.

and the associated mixing proportions (which are equal to zero for absent signals), respectively.

The best subset of columns from A can be found by exhaustive evaluation of the least-squares residual for all possible subsets. It is possible to take advantage of the structure of the problem to reduce the size of the search space by means of branch-and-bound-type algorithms [8, 9]. However, when n increases, algorithms based on exhaustive searches (even with branch-and-bound restrictions of the search space) become rapidly impractical. Therefore, several heuristics which do not require exhaustive searches have been proposed [6]. One such heuristic is the well-known "greedy" algorithm of Golub and Van Loan [4]. The greedy algorithm is a sequential forward selection scheme. The idea of the greedy algorithm is to start by finding the column of A whose span is closest to b, and then add columns one by one until r columns have been selected, each time adding the column that gives the largest decrement of the least-squares residual. Some theoretical arguments have recently been given in favor of the greedy algorithm [10]. However, there are situations in which the greedy algorithm will fail even if an exact sparse solution to $Ax = b$ exists. Consider the following simple example:

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \qquad x = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \end{bmatrix}, \qquad r = 2.$$

The greedy algorithm will start by erroneously selecting the first column of A and then will add the second column. In this case, the correct set of columns (second and third) cannot be selected by the greedy algorithm. In general, it is always possible to find situations in which one or all of the heuristics that have been proposed so far will fail [6].

An alternative to the sequential forward selection used in the greedy algorithm is sequential backward selection [6]. Rather than adding columns to the solution one by one, it is possible to start with all columns present (i.e., with the complete matrix A) and remove one column at a time until r columns are left. The column that is removed should be chosen to minimize the increment in the least-squares residual. We call this alternative approach the backward greedy algorithm. To avoid confusion, the standard greedy algorithm will be referred to as the forward greedy algorithm in the following.

From the simple example above, it is obvious that the backward greedy algorithm will correctly remove the first column from A. In this case, the backward greedy algorithm correctly yields the optimal solution. More generally, as will be shown in section 3, the backward greedy algorithm is guaranteed to yield the correct subset of nonzero components in x if the perturbation η in (1) is small enough. This can be observed in the example of the previous paragraph. Let the perturbation vector η be equal to $[0, 0, \epsilon, 0]^H$. It is easily verified that the backward greedy algorithm yields the correct solution as long as $\|\eta\| = |\epsilon| < 1/2$.

In the following section, the backward greedy algorithm for subset selection is formally defined and a simple QR-based implementation is described. Section 3 presents our main result about the backward greedy algorithm in the form of a theorem proving that it always finds the correct subset of nonzero components of x in the small

perturbation case. In this sense, the backward greedy algorithm is an optimal[1] polynomial-time algorithm, in contrast to other optimal algorithms such as exhaustive search or branch-and-bound, which are exponential-time. Some implications of this theorem are then discussed. Numerical results illustrating the properties of the backward greedy algorithm are given in section 4. The paper concludes with some remarks on the choice of the number of columns r and the NP-hardness of the subset selection problem and a discussion of some possible extensions of our result.

2. The backward greedy algorithm. Formally, the backward greedy algorithm can be defined as follows. Let $\Xi = \{1, \ldots, n\}$ denote the ordered set of column indices of A, and $\Gamma = \{\gamma_1, \ldots, \gamma_{c(\Gamma)}\}$, $\Gamma \subseteq \Xi$, an ordered subset of Ξ, of cardinality $c(\Gamma)$. The colon notation $A(:, \Gamma)$ is used to designate the matrix formed from the columns of A whose indices are in Γ. Similarly, $x(\Gamma)$ designates the Γ-indexed elements of the vector x, and $x(i:j)$ designates the subvector with elements indexed by i through j. Denote by

$$\rho(\Gamma) = \min_{z \in \mathbb{C}^{c(\Gamma)}} \|A(:, \Gamma)z - b\|_2 \tag{2}$$

the least-squares residual associated with the sparse LS solution of $Ax = b$ based on the Γ-indexed subset of columns of A. The backward greedy algorithm for subset selection is initialized by taking $\Gamma = \Xi$. Elements are then removed from Γ one by one by repeating the iteration

$$\Gamma \leftarrow \Gamma \setminus \{k^*\} : k^* \in \arg\min_{k \in \Gamma} \rho(\Gamma \setminus \{k\}) \tag{3}$$

until $c(\Gamma) = r$. The column k^* that is removed at each iteration is chosen to minimize the increment in the least-squares residual ρ. Once the last iteration has been performed, the sparse least-squares solution associated with the column indices left in Γ is computed. The subset of indices obtained at the last iteration of the backward greedy algorithm and the associated sparse least-squares solution will be denoted by $\hat{\Gamma}_{\mathrm{s}}$ and $\hat{x}_{\mathrm{s}}$, respectively. The backward greedy algorithm can be summarized as follows.

BACKWARD GREEDY ALGORITHM.
Input: matrix A, column vector b, integer $r < n$.
Output: index set $\hat{\Gamma}_{\mathrm{s}}$, column vector $\hat{x}_{\mathrm{s}}$.
Subset selection step:
    i = n;
    while $i > r$ do
        $i \leftarrow i - 1$;
        $k^* \in \arg\min_{k \in \Gamma} \rho(\Gamma \setminus \{k\})$;
        $\Gamma \leftarrow \Gamma \setminus \{k^*\}$;
    end.
Solution step:
    $\hat{\Gamma}_{\mathrm{s}} \leftarrow \Gamma$;
    $z^* = \arg\min_{z \in \mathbb{C}^r} \|A(:, \Gamma)z - b\|_2$;
    $\hat{x}_{\mathrm{s}}(\Gamma) \leftarrow z^*$; $\hat{x}_{\mathrm{s}}(\Gamma^c) \leftarrow 0$.

[1] Our focus in this paper is on finding the optimal (best) solution to the subset selection problem. Thus we will say that an algorithm is optimal when it provides the optimal solution; not necessarily when its run time or numerical properties are optimal.

The QR downdating-based implementation of the algorithm presented in this section is reasonably efficient but it is not the only possibility, nor is it necessarily the best one. For instance, fast implementations of the backward greedy algorithm based on a different approach that does not require a QR decomposition have been proposed in [11] and [6]. The choice of a particular implementation of the backward greedy algorithm should be based on constraints such as number of operations, memory requirements, possibility of parallelization, or numerical stability. Since this paper focuses on the ability of the backward greedy algorithm to find the correct solution, we do not discuss the issues which depend on the specifics of an implementation. The QR-based implementation that is described here should be viewed as a simple and straightforward one, without any attempt at optimization for a particular set of constraints.

The forward greedy algorithm is usually implemented by means of the QR algorithm for least-squares solution of linear systems. In essence, the forward greedy algorithm is a QR factorization of A by the Gram–Schmidt procedure in which the column pivot is chosen greedily with respect to the right-hand side b of the matrix equation $Ax = b$ and the Gram–Schmidt procedure is interrupted once r columns have been selected (see [10, 4]). Similarly, a possible implementation of the backward greedy algorithm can be straightforwardly obtained by combining the QR algorithm for least-squares problems with a column-deletion QR downdating step based on Givens rotations [1, 4].

For a practical implementation of the backward greedy algorithm, it is necessary to evaluate the least-squares residuals ρ associated with matrices obtained by gradual removal of columns in A. Consider first the initial step of the backward greedy algorithm. Let us assume that the QR factorization of A is available: $Q^H A = R$, where Q^H denotes the Hermitian transpose of Q. The associated least-squares residual is given by $\rho(\Xi) = \|e(n+1:m)\|_2$, $e = Q^H b$. Since $A(:,\Xi\setminus\{k\})$ is the matrix obtained by deleting the kth column from $A(:,\Xi)$, its QR factorization $(Q^{(k)})^H A(:,\Xi\setminus\{k\}) = R^{(k)}$ can be computed by downdating Q and R for the column deletion by a sequence of Givens rotations. The downdating operation is [4, p. 595]

$$G_{n-1}^H \cdots G_k^H R(:,\Xi\setminus\{k\}) = R^{(k)},$$
$$G_{n-1}^H \cdots G_k^H e = e^{(k)},$$

where G_i is a rotation in planes i, $i+1$ for $i = k:n-1$, and $R(:,\Xi\setminus\{k\})$ is the upper Hessenberg matrix obtained by deleting the kth column of R. The downdated least-squares residual is given by

$$\rho(\Xi\setminus\{k\}) = \|e^{(k)}(n:m)\|_2.$$

Note that, unless numerical stability is a critical issue [1], it is not necessary to explicitly compute the downdated QR factorization $Q^{(k)}R^{(k)}$. It is simply necessary to apply the sequence of Givens rotations to e to find the downdated least-squares residual.

Once the minimum downdated $\rho(\Xi\setminus\{k^*\})$ has been found, the k^*th column is removed from A and the process repeated recursively until only r columns are left. The sparse least-squares solution $\hat{x}_s$ can then be computed by the usual QR approach. The QR-based implementation of the backward greedy algorithm can thus be written in pseudocode as shown below.

BACKWARD GREEDY ALGORITHM—IMPLEMENTATION.
Input: matrix A, column vector b, integer $r < n$.
Output: index set $\hat{\Gamma}_{\mathrm{s}}$, column vector $\hat{x}_{\mathrm{s}}$.
Subset selection step:
$i \leftarrow n$; $\Gamma \leftarrow \{1, \ldots, n\}$;
compute $Q^H A = R$;
$e \leftarrow Q^H b$;
while $i > r$ do
$i \leftarrow i - 1$;
for $j = 1 : i$ do
compute the Givens rotations downdating $G_{i-1}^H \cdots G_j^H H^{(j)} = R^{(j)}$
where $H^{(j)} = R$ with column j deleted
and $R^{(j)}$ upper triangular;
$e^{(j)} \leftarrow G_{i-1}^H \cdots G_j^H e$;
$\rho(\Gamma \setminus \{\gamma_j\}) \leftarrow \|e^{(j)}(i : m)\|$ where γ_j is the jth element of Γ;
end
choose k, $1 \le k \le i$, such that the residual $\rho(\Gamma \setminus \{\gamma_k\})$ is minimum;
$\Gamma \leftarrow \Gamma \setminus \{\gamma_k\}$;
$R \leftarrow R^{(k)}$;
$e \leftarrow e^{(k)}$;
end.
Solution step:
$\hat{\Gamma}_{\mathrm{s}} \leftarrow \Gamma$;
compute the solution of $Rz = e(1 : r)$ by back-substitution;
$\hat{x}_{\mathrm{s}}(\Gamma) \leftarrow z$; $\hat{x}_{\mathrm{s}}(\Gamma^c) \leftarrow 0$.

In general, the global cost of the QR-based implementation of the backward greedy algorithm will be dominated by that of the original QR factorization and QR downdating operations during the subset selection step. The original QR factorization of A has a cost of $O(mn^2)$. Downdating a QR factorization (i.e., computing the Givens rotations and applying them to e at step i) requires $O(i^2)$ operations (recall that $i = n : -1 : r$) [4]. The complete ith selection iteration will require $O(i^3)$ operations for the i QR downdating operations, plus $O(i(m - i))$ operations for the evaluation of the i least-squares residuals. Once the last selection iteration has been performed, computing $\hat{x}_{\mathrm{s}}$ still requires $O(r^2)$ operations.

We conclude this section with a short discussion of the comparative computational costs of the QR-based implementations of the backward greedy algorithm and of the forward greedy algorithm. The backward greedy algorithm starts with n columns and removes them one by one until only r columns are left. The first step in the algorithm is the solution of the nonsparse least-squares problem $Ax = b$ by QR factorization. Recall the equivalence between the forward greedy algorithm and interrupted QR factorization of A with column pivoting. Obviously, if r is small with respect to n, the computational cost of the backward greedy algorithm will far exceed that of the forward greedy algorithm. Because of the overhead of the forward greedy algorithm compared to a "regular" QR factorization, the cost of the backward algorithm might be similar to or even slightly lower than that of the forward algorithm if r is close to n. Even if r is not too close to n and the cost of the backward greedy algorithm is higher than that of the forward greedy algorithm, it may still be advantageous to use the backward algorithm. The faster forward algorithm is prone to producing the *wrong solution* in many situations as demonstrated in section 1. However, as will be

seen in the next section, the slower (but still polynomial-time) backward algorithm is guaranteed (under appropriate conditions) to produce the right solution.

3. Main result. Consider the alternative interpretation of the subset selection problem (1). In this interpretation, there exists a sparse vector x_0 such that $Ax_0 = b_0$, or $A(:,\Gamma_0)x_0(\Gamma_0) = b_0$ and $x_0(\Gamma_0^c) = 0$, where Γ_0 is the subset of indices of nonzero elements in x_0. The subset selection problem can be viewed as finding the r components that are "active" in x_0 (i.e., finding Γ_0) and estimating their values from a noisy observation of $b = b_0 + \eta$. That is, the subset selection is a detection and estimation problem with a priori knowledge about the solution in the form of a sparsity constraint.

The main result of this paper, presented in Theorem 1, is that, for any full rank matrix A and any sparse vector x_0, there exists a bound on the perturbation η that guarantees that the backward greedy algorithm will select the correct subset of components, i.e., $\hat{\Gamma}_s = \Gamma_0$. This means that the backward greedy algorithm is optimal for the subset selection problem, at least for small noise levels. In order to prove this theorem, we will need the following definition and two lemmas.

As usual we define the distance $d(x,\mathcal{S})$ between a point x and a set $\mathcal{S}$ as the distance between x and the closest point to it in $\mathcal{S}$. That is,

$$d(x,\mathcal{S}) = \min_{y\in\mathcal{S}} d(x,y),$$

where $d(x,y) = \|x-y\|$. When $\mathcal{S}$ is a subspace (and the Euclidean norm is used), $d(x,\mathcal{S})$ is the distance between x and its orthogonal projection on $\mathcal{S}$. Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be two subspaces of $\mathbb{C}^n$. The *bisector* of $\mathcal{S}_1$ and $\mathcal{S}_2$, denoted by $\mathcal{H}$, is defined as the set of points located at equal distance from $\mathcal{S}_1$ and $\mathcal{S}_2$,

$$\mathcal{H} = \{x : x \in \mathbb{C}^n,\ d(x,\mathcal{S}_1) = d(x,\mathcal{S}_2)\}.$$

In general, $\mathcal{H}$ is a conic.

The first lemma will be instrumental in proving Lemma 2. It can be interpreted as a statement of the fact that the bisector $\mathcal{H}$ of two subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$ partitions $\mathbb{C}^n$ into points that are closer to $\mathcal{S}_1$ and points that are closer to $\mathcal{S}_2$. As usual, we define an open ball (of radius ξ) in $\mathbb{C}^n$ centered at b as $V_\xi(b) = \{x \in \mathbb{C}^n : \|x-b\| < \xi\}$.

LEMMA 1. *Let $\mathcal{S}_1$, $\mathcal{S}_2$, and $\mathcal{H}$ denote two subspaces of $\mathbb{C}^n$ and their bisector. Let b be a point strictly closer to $\mathcal{S}_1$ than to $\mathcal{S}_2$. If $V_\xi(b)$ is a ball in $\mathbb{C}^n$ centered at b that does not intersect with $\mathcal{H}$, then all points in $V_\xi(b)$ are closer to $\mathcal{S}_1$ than they are to $\mathcal{S}_2$.*

Proof. Let $x \in V_\xi(b)$ and suppose that x is closer to $\mathcal{S}_2$ than to $\mathcal{S}_1$. Then, $d(x,\mathcal{S}_2) - d(x,\mathcal{S}_1) < 0$, while $d(b,\mathcal{S}_2) - d(b,\mathcal{S}_1) > 0$. It follows by continuity that there must exist $z \in V_\xi(b)$ defined by $z = (1-\alpha)b + \alpha x$, for some $\alpha \in [0,1]$, such that $d(z,\mathcal{S}_2) - d(z,\mathcal{S}_1) = 0$, and therefore $z \in \mathcal{H}$. But since $z \in V_\xi(b)$, this contradicts the assumption that $V_\xi(b)$ does not intersect with $\mathcal{H}$. ∎

The second lemma considers one iteration of the backward greedy algorithm, i.e., the removal of one column from the current subset of columns Γ. It states that there exists a nontrivial bound on the norm of the perturbation $\|b-b_0\|$ such that, for any perturbation smaller than the bound, one iteration of the backward greedy algorithm is guaranteed to remove a column that is not part of the correct subset of columns Γ_0, provided that the current subset of columns Γ contains the correct subset Γ_0.

LEMMA 2. *Let*

$$\hat{\Gamma} \in \arg \min_{\substack{\lambda\subset\Gamma \\ c(\lambda)=c(\Gamma)-1}} \rho(\lambda), \tag{4}$$

where $\rho(\lambda)$ is defined in (2)*, the matrix A in* (2) *has full column rank, and $\Gamma \subseteq \Xi$. Denote by Γ_0 the subset of indices associated with the sparse solution x_0 in* (1)*. If $\Gamma \supset \Gamma_0$, then there exists $\delta > 0$ such that $\|b - b_0\| < \delta$, $\|b_0\| > 0$, implies $\hat{\Gamma} \supseteq \Gamma_0$. The value of δ is given by*

$$\delta = \min_{\substack{\tau \subset \Gamma, \tau \not\supseteq \Gamma_0, c(\tau) = c(\Gamma) - 1 \\ \lambda \subset \Gamma, \lambda \supseteq \Gamma_0, c(\lambda) = c(\Gamma) - 1}} d(b_0, \mathcal{H}(\tau, \lambda)), \tag{5}$$

where $\mathcal{H}(\tau, \lambda)$ denotes the bisector of $\mathcal{R}(A(:, \tau))$ and $\mathcal{R}(A(:, \lambda))$.

Proof. Let $\mathcal{H}(\tau, \lambda)$, $\tau \subset \{1, \ldots, n\}$, $\lambda \subset \{1, \ldots, n\}$, $c(\tau) = c(\lambda)$, denote the bisector of $\mathcal{R}(A(:, \tau))$ and $\mathcal{R}(A(:, \lambda))$. Let δ be the radius of the largest ball $V_\delta(b_0)$ centered at $b_0 \in \mathcal{R}(A(:, \Gamma_0))$ that does not intersect any of the bisectors $\mathcal{H}(\tau, \lambda)$, $\tau \not\supseteq \Gamma_0$, $\tau \subset \Gamma$, $\lambda \supseteq \Gamma_0$, $\lambda \subset \Gamma$. That is, δ is given by (5). By Lemma 1, all the points in $V_\delta(b_0)$ are closer to at least one of the $\mathcal{R}(A(:, \lambda))$ for some $\lambda \supseteq \Gamma_0$ than they are to any of the $\mathcal{R}(A(:, \tau))$ for all $\tau \not\supseteq \Gamma_0$. Thus, if $|b - b_0| < \delta$, we are guaranteed to select $\hat{\Gamma}$ among the λ's, which means $\hat{\Gamma} \supseteq \Gamma_0$ as desired.

All that remains to be shown is that $\delta > 0$. This comes as a consequence of the linear independence of the columns of A. Suppose $\delta = 0$. Then there exists $\tau \not\supseteq \Gamma_0$ and $\lambda \supseteq \Gamma_0$ such that $d(b_0, \mathcal{H}(\tau, \lambda)) = 0$, i.e., $b_0 \in \mathcal{H}(\tau, \lambda))$. This implies that $b_0 \in \mathcal{R}(A(:, \tau))$ because, by the definition of $\mathcal{H}(\tau, \lambda)$, $d[b_0, \mathcal{R}(A(:, \lambda))] = d[b_0, \mathcal{R}(A(:, \tau))]$ and, by assumption, $b_0 \in \mathcal{R}(A(:, \lambda))$. Given that $\|b_0\| > 0$, this would contradict the hypothesis that A is full rank because $\tau \not\supseteq \Gamma_0$ while $\lambda \supseteq \Gamma_0$. ∎

Equipped with Lemma 2, we can now state our main theorem.

THEOREM 1. *For any full column rank matrix A and for any sparse vector x_0 with r nonzero components (or, alternately, for any $b_0 = Ax_0$), there exists $\delta > 0$ such that $\|b - b_0\| < \delta$ guarantees that the backward greedy algorithm for solving $Ax = b$ will select the correct subset of components Γ_0. The value of δ is given by*

$$\delta = \min_{r \le k < n} \min_{\substack{\Gamma \subseteq \Xi \\ c(\Gamma) = k+1}} \min_{\substack{\tau \subset \Gamma, \tau \not\supseteq \Gamma_0, c(\tau) = k \\ \lambda \subset \Gamma, \lambda \supseteq \Gamma_0, c(\lambda) = k}} d(b_0, \mathcal{H}(\tau, \lambda)). \tag{6}$$

Furthermore, the corresponding sparse least-squares solution $\hat{x}_{\mathrm{s}}$ satisfies $\|\hat{x}_{\mathrm{s}} - x_0\| \le \delta / \sigma_{\min}[A(:, \Gamma_0)]$ where $\sigma_{\min}[A]$ denotes the smallest singular value of A.

Proof. The first part of the theorem can be proven by an inductive argument based on Lemma 2. Let $\Gamma^{(k)}$ denote the subset of columns of A selected at step k of the backward greedy algorithm, $k = n, n-1, \ldots, r$. The base for the induction is the observation that $\Gamma^{(n)} \supset \Gamma_0$ since $\Gamma^{(n)} = \Xi$ by definition of the backward greedy algorithm. The induction step consists in showing that if $\Gamma^{(k+1)} \supset \Gamma_0$, then $\Gamma^{(k)} \supseteq \Gamma_0$ provided that $\|b - b_0\| < \delta^{(k)}$ for some $\delta^{(k)} > 0$, for $i = n-1, n-2, \ldots, r$. The induction step follows directly from Lemma 2. The bound $\delta^{(k)}$ is given by (5) with $\Gamma = \Gamma^{(k+1)}$. It follows that the backward greedy algorithm will select the correct subset of columns of A, i.e., $\hat{\Gamma}_{\mathrm{s}} = \Gamma_0$, if $\|b - b_0\| < \delta$ with $\delta = \min_{r \le k < n} \delta^{(k)}$, the last relation yielding the bound (6). The last part of the theorem follows directly from

$$\|\hat{x}_{\mathrm{s}} - x_0\| = \|A^\dagger(:, \Gamma_0)(\hat{b} - b_0)\| \le \|A^\dagger(:, \Gamma_0)\| \delta,$$

where $\hat{b} = A(:, \Gamma_0)\hat{x}_{\mathrm{s}}$ and $A^\dagger$ denotes the pseudoinverse of A. ∎

Our specific motivation for studying the backward greedy algorithm for subset selection arises from an estimation and detection problem in statistical signal processing [3, 2]. In this application, b^ℓ is a sequence of random vectors converging with probability one to b_0 as $\ell \to \infty$. We are interested in the properties of the

sequence of estimates of Γ_0 and x_0 obtained by solving $Ax = b^\ell$ by the backward greedy algorithm. Note that for the application described in [3, 2], finding Γ_0 is at least as important as finding x_0. Using Theorem 1, it is easy to establish that the estimates $\hat{\Gamma}_{\mathrm{s}}^\ell$ and $\hat{x}_{\mathrm{s}}^\ell$ obtained by the backward greedy algorithm are strongly consistent. That is, they converge to the true values Γ_0 and x_0 with probability one as $\ell \to \infty$. To see this, note that the backward greedy algorithm is guaranteed to pick the correct subset of components if the perturbation η is small enough, and recall that $\|\hat{x}_{\mathrm{s}} - x_0\| \leq \|b - b_0\|/\sigma_{\min}[A(:,\Gamma_0)]$. We thus get directly the following optimal detection corollary.

COROLLARY 1. *Let $Ax_0 = b_0$ with x_0 having q nonzero components. Let b^ℓ be a sequence of strongly consistent estimates of b_0, i.e., b^ℓ converges with probability one to b_0 as $k \to \infty$. Let $\hat{x}_{\mathrm{s}}^\ell$ be the corresponding sequences of estimates of x_0 with q nonzero components produced by the backward greedy algorithm applied to the sequence of equations $Ax = b^\ell$. Then, $\hat{x}_{\mathrm{s}}^\ell$ is a sequence of strongly consistent estimates of x_0. That is, the correct subset of nonzero components is selected with probability one for ℓ sufficiently large, and the estimated values of the nonzero components converge with probability one to the true values.*

So far it has been assumed that a true sparse solution x_0 exists and that the subset selection problem consisted of finding this sparse solution. This detection and estimation view led to Theorem 1 and its first corollary. There are, however, situations in which this formulation is not adequate: There is no underlying true solution and the subset selection problem is simply that of finding the subset of r columns from A such that the least-squares residual is minimized. In this case, let Γ_{s} and x_{s} be defined as the optimal sparse solution. That is, let

$$\Gamma_{\mathrm{s}} = \arg \min_{\substack{\Gamma \subseteq \Xi \\ c(\Gamma) = r}} \rho(\Gamma), \tag{7}$$

$$x_{\mathrm{s}} = \arg \min_{\substack{x \in \mathbb{C}^n \\ x(\Gamma_{\mathrm{s}}^c) = 0}} \|Ax - b\|. \tag{8}$$

In practice, Γ_{s} and x_{s} can always be found with an exhaustive search. Recall that $\hat{\Gamma}_{\mathrm{s}}$ and $\hat{x}_{\mathrm{s}}$ are the solutions provided by the backward greedy algorithm. The following corollary of Theorem 1 implies that, if the residual $\rho(\hat{\Gamma}_{\mathrm{s}})$ is small enough, the backward greedy algorithm will give the same solution as an exhaustive search.

COROLLARY 2. *The backward greedy algorithm computes the optimal solution, i.e., $\hat{\Gamma}_{\mathrm{s}} = \Gamma_{\mathrm{s}}$ and $\hat{x}_{\mathrm{s}} = x_{\mathrm{s}}$, if $\rho(\hat{\Gamma}_{\mathrm{s}})$ is small enough, that is, if $\rho(\hat{\Gamma}_{\mathrm{s}}) < \delta$ where δ is defined by* (6) *in which $b_0 = A\hat{x}_{\mathrm{s}}$ and $\Gamma_0 = \hat{\Gamma}_{\mathrm{s}}$.*

Proof. By hypothesis, $\|b - A\hat{x}_{\mathrm{s}}\| < \delta$. From the definition of δ, this implies that b is closer to $\mathcal{R}(A(:,\hat{\Gamma}_{\mathrm{s}}))$ than to $\mathcal{R}(A(:,\Gamma)$ for any subset Γ with r or fewer components since b is on the $\hat{\Gamma}_{\mathrm{s}}$-side of all the bisectors $\mathcal{H}(\hat{\Gamma}_{\mathrm{s}}, \Gamma)$. Hence, $d(b, \mathcal{R}(A(:,\hat{\Gamma}_{\mathrm{s}}))) < d(b, \mathcal{R}(A(:,\Gamma)))$ or $\rho(\hat{\Gamma}_{\mathrm{s}}) < \rho(\Gamma)$, which implies $\hat{\Gamma}_{\mathrm{s}} = \Gamma_{\mathrm{s}}$ and $\hat{x}_{\mathrm{s}} = x_{\mathrm{s}}$. ∎

Using Corollary 2, a procedure for checking if the output of the backward greedy algorithm $\hat{\Gamma}_{\mathrm{s}}$ is the optimal sparse solution Γ_{s} can be suggested: simply compare $\rho(\hat{\Gamma}_{\mathrm{s}})$ to δ given by (6). If $\rho(\hat{\Gamma}_{\mathrm{s}}) < \delta$, then $\hat{\Gamma}_{\mathrm{s}} = \Gamma_{\mathrm{s}}$. Unfortunately, computing δ from (6) with $\hat{\Gamma}_{\mathrm{s}}$, A, and b is not practical, so this procedure is of limited usefulness in general and the implication of Corollary 2 is more of a qualitative nature.

The implication of Corollary 2 is that the backward greedy algorithm is optimal for the subset selection problem when its solution corresponds to a small enough residual. That is, if $\rho(\Gamma_{\mathrm{s}})$ is small enough then Γ_{s} is guaranteed to be found by the

algorithm. For larger residuals, the backward greedy algorithm is suboptimal; it may or may not find the correct solution, depending on the values of A and b.

4. Numerical results. Figure 1 illustrates Theorem 1 and Corollary 2. A 10×7 matrix A with real random i.i.d. Gaussian coefficients was generated:

$$A = \begin{bmatrix} -0.54 & -0.10 & 0.44 & -0.39 & -1.71 & -1.16 & -0.06 \\ -1.77 & 0.80 & -1.38 & 0.73 & 0.29 & 0.59 & -0.25 \\ 0.08 & -0.60 & 0.78 & 0.78 & -1.12 & 0.64 & 1.20 \\ 0.28 & 2.25 & 1.64 & 0.65 & 1.33 & 1.73 & -1.91 \\ -1.20 & -0.15 & -0.70 & 1.41 & -0.82 & 0.59 & 0.43 \\ 1.25 & 0.12 & 0.82 & 0.79 & 0.57 & -0.79 & 0.90 \\ 1.22 & 1.19 & 1.24 & 1.21 & 0.18 & 0.36 & 0.85 \\ -0.03 & 0.23 & -1.61 & -1.33 & -0.91 & -0.98 & 0.71 \\ 0.48 & -0.09 & -1.51 & 0.04 & -2.26 & 0.04 & -0.06 \\ 0.14 & -0.88 & 0.62 & -0.23 & -0.05 & 0.03 & 0.74 \end{bmatrix}.$$

The first three columns of the matrix were linearly combined to yield the vector b_0, i.e., $b_0 = Ax_0$ with $x_0 = [1, 1, 1, 0, 0, 0, 0]^H$. One hundred random perturbation vectors η were generated and added to b_0 to yield $b = b_0 + \eta$. The perturbation vectors η were uniformly distributed on a hypersphere of given radius, i.e., the norm $\|\eta\|$ of the perturbation vectors was fixed. For each perturbation vector, the sparse least-squares solution to $Ax = b$ with $r = 3$ nonzero elements was computed by the backward greedy algorithm and by exhaustive search. The experiment was repeated for several values of $\|\eta\|$. Figure 1 gives the number of times the backward greedy algorithm found the true subset of nonzero components $\Gamma_0 = \{1, 2, 3\}$ as a function of the ratio $\|\eta\|/\|b_0\|$ (i.e., of the inverse of the signal-to-noise ratio). Figure 1 also gives the number of times the sparse solution obtained by exhaustive search (Γ_s) was equal to Γ_0 or to $\hat{\Gamma}_s$ (recall that for large perturbations, Γ_s does not necessarily need to be equal to Γ_0). According to Theorem 1 and Corollary 2, we should expect $\hat{\Gamma}_s = \Gamma_0 = \Gamma_s$ when $\|\eta\| < \delta$. This can indeed be verified on Figure 1 where the value of δ corresponding to A and b_0 is marked by a circle. The value of δ was computed from (6).

The value of δ depends on the matrix A and the exact vector b_0. Two general observations can be made about δ. First, it is easily verified that, for given A, δ should be proportional to $\|b_0\|$. Second, recall that δ can be viewed as the radius of the largest hypersphere centered on the sparse vector b_0 that can be fitted in the n-dimensional cone including b_0 whose facets are the bisectors of pairs of subspaces spanned by subsets of columns of A in (6). Intuitively, when the principal angles between these subspaces are small, the principal angles between the bisectors will also be small, the n-dimensional cone will be narrower, and δ will be smaller. The principal angles between subspaces spanned by subsets of columns of A will be small when these columns are nearly linearly dependent, i.e., when the condition number of the matrix A is large. Thus one would expect to see the ratio $\delta/\|b_0\|$ vary inversely with the condition number of the matrix A. In order to verify this hypothesis, one thousand 10×7 matrices with random real coefficients were generated. By manipulating the singular values of the matrices, the condition number $\kappa(A)$ was set to 5 for the first 500 matrices and to 50 for the next 500 matrices. For each matrix, the vector b_0 was again computed from the first three columns of A as $b_0 = A[1, 1, 1, 0, 0, 0, 0]^H$, and the bound δ was evaluated from (6). The histograms of Figure 2 give the distributions of δ obtained for $\kappa(A) = 5$ and $\kappa(A) = 50$. It can be verified that $\delta/\|b_0\|$ is generally higher when the condition number of the matrix is small.

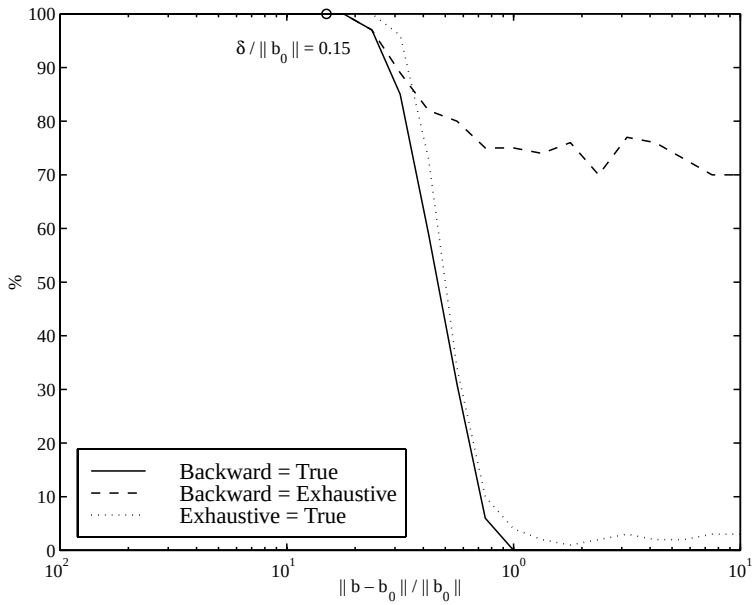


FIG. 1. *Performance of the backward greedy algorithm as a function of the ratio* $\|b - b_0\|/\|b_0\|$. *The value of the bound* δ *is represented by a circle.*

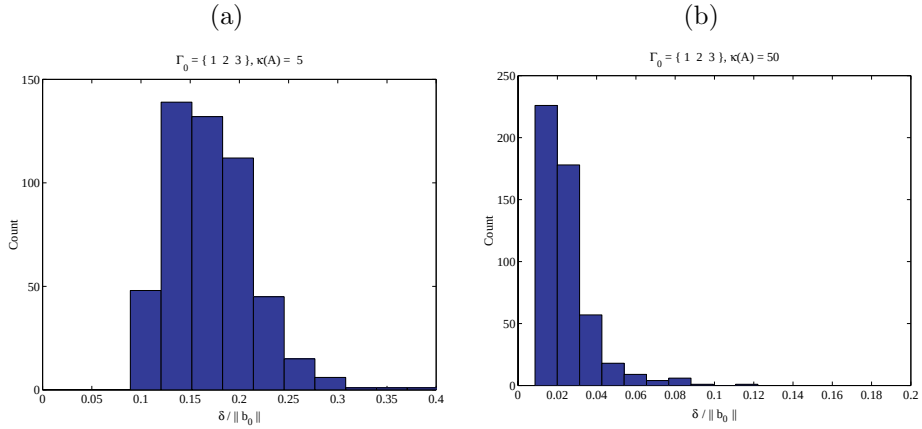


FIG. 2. *Distribution of the bound* δ *for* 10×7 *random matrices with given condition number: (a)* $\kappa(A) = 5$ *and (b)* $\kappa(A) = 50$.

Given A and b_0, a method for computing an estimate of the bound δ that does not require exhaustive evaluation of the distances between b_0, and bisectors in (6) can be suggested. This method can be used when the matrix A has a large number of columns. Using the same approach as in the first experiment, random perturbation vectors $\|\eta\|$ can be added to b_0 and the backward greedy algorithm can be used to compute $\hat{\Gamma}_s$ for increasing values of $\|\eta\|$. When a perturbation vector η is found such that $\hat{\Gamma}_s \neq \Gamma_0$, Theorem 1 implies that $\|\eta\| \geq \delta$. If a large number of random vectors η are generated for each value of $\|\eta\|$, this Monte Carlo–type method can yield tight approximations of δ. Since it is based on the backward greedy algorithm, the method

can yield an estimate of δ to any desired confidence level in polynomial time.

This same Monte Carlo method can be applied for given A and b to $b_0 = A\hat{x}_s$, where $\hat{x}_s$ is the estimate produced by the backward greedy algorithm. In view of Corollary 2, comparing the estimate of δ so obtained to $\rho = \|b - b_0\|$ will determine, to any desired confidence level, whether $\hat{x}_s$ is the optimal solution, in polynomial time.

5. Remarks.

5.1. Determination of the number of components r. In the variant of the subset selection problem considered here, the number of desired nonzero components r was assumed to be known a priori. It is possible to consider other formulations of the problem. For example, in [10], Natarajan states the following version of the subset selection problem ("Natarajan's problem" in the following): *Given A, b, and $\epsilon > 0$, find the vector x satisfying $\|Ax - b\|_2 \leq \epsilon$ if such exists, such that x has the fewest nonzero entries of all such vectors.*

In the version considered in [2], neither ϵ nor r are known a priori. Instead, the subset selection problem is defined as finding the sparse vector x that realizes the best trade-off between the match to the observation vector b and the number of nonzero components. This trade-off is measured by a figure of merit of the form

$$\min_x \{\|Ax - b\|_2 + f(x)\}, \tag{9}$$

where $f(x)$ is a complexity penalty monotonically increasing with the number of nonzero components in x. Note that because $\kappa(A(:, \Gamma_0)) < \kappa(A)$, a sparsity constraint reduces the sensitivity of the solution of the linear inverse problem $Ax = b$ to perturbations of the observation vector b_0, at least for η small enough ($\eta < \delta$). For this reason, solving a linear inverse problem with a sparsity constraint like in (9) is sometimes known as regularization by sparsity [5].

The backward greedy algorithm presented in section 2 can be readily applied to the above problems by solving for successively decreasing r and looking for the sparsest solution satisfying $\|Ax - b\| < \epsilon$ in the case of Natarajan's problem, or the minimizer of the figure of merit (9) in the case of regularization by sparsity.

5.2. NP-hardness of subset selection. In [10], Natarajan considered the first variant of the subset selection problem described in section 5.1. He showed that finding the sparsest solution to $\|Ax - b\| \leq \epsilon$ for a given ϵ, if such a solution exists, is NP-hard. This may seem to be in contradiction with our result that the backward greedy algorithm can correctly solve the subset selection problem in polynomial time in the small residual case. Indeed, for Natarajan's formulation of the subset selection problem, we have the following corollary to our main result.

COROLLARY 3. *Let ϵ_1 ($\epsilon_1 \leq \epsilon$), $\Gamma_0(\epsilon)$, and $r(\epsilon)$ be the residual, component indices, and their number, respectively, in the solution to the subset selection problem (Natarajan's problem) for given A of full column rank, b, and ϵ, if the solution exists. Suppose $\epsilon_1 < \delta$, where δ is given by (6) with $r = r(\epsilon)$ and $\Gamma_0 = \Gamma_0(\epsilon)$. Then the backward greedy algorithm will provide the optimal solution $\hat{\Gamma}_s = \Gamma_0(\epsilon)$ if stopped at the smallest r satisfying $\rho(\hat{\Gamma}_s) \leq \epsilon$.*

Proof. The proof follows directly from the formulation of the subset selection problem by Natarajan and from Corollary 2. ∎

Corollary 3 states that if the threshold ϵ is small enough, the solution to Natarajan's problem will be found in polynomial time by the backward greedy algorithm.

The contradiction between Natarajan's result on NP-hardness of the subset selection problem and the last statement is only apparent. We do not have a polynomial-time algorithm to compute δ. Furthermore, in view of Natarajan's result, we suspect that such an algorithm does not exist; that is, computing δ is an NP-hard problem. Thus, even if the backward greedy algorithm yields a solution in polynomial time which can be expected to be the optimal solution if the residual is small enough, the verification that the residual is indeed small enough cannot be performed in polynomial time. Nonetheless, the Monte Carlo algorithm discussed in section 4 suggests a polynomial-time method to perform this verification to any desired level of statistical significance.

5.3. Extensions of the results. The results that have been presented in the previous sections can be extended to variants of the subset selection problem based on any other norm for the residual than the two-norm (and induced Euclidean distance). Indeed, it is easily verified that Lemma 1, Lemma 2, and the first part of Theorem 1 remain valid mutatis mutandis.

Acknowledgments. The authors would like to thank the anonymous referees for their suggestions and for pointing out an error in an earlier version of this manuscript.

REFERENCES

[1] A. Björck, H. Park, and L. Eldén, *Accurate downdating of least squares solutions*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 549–568.
[2] C. Couvreur and Y. Bresler, *Dictionary-based decomposition of linear mixtures of Gaussian processes*, in Proceedings ICASSP, IEEE, Piscataway, NJ, 1996, pp. 2519–2522.
[3] C. Couvreur and Y. Bresler, *Optimal decomposition and classification of linear mixtures of ARMA processes*, 1998, submitted.
[4] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd. ed., The Johns Hopkins University Press, Baltimore, MD, 1989.
[5] G. Harikumar and Y. Bresler, *A new algorithm for computing sparse solutions to linear inverse problems*, in Proceedings ICASSP, IEEE, Piscataway, NJ, 1996, pp. 1331–1334.
[6] A. J. Miller, *Subset Selection in Regression*, Chapman and Hall, London, UK, 1990.
[7] M. Nafie, M. Ali, and A. H. Tewfik, *Optimal subset selection for adaptive signal representation*, in Proceedings ICASSP, IEEE, Piscataway, NJ, 1996.
[8] G. M. Furnival and R. W. Wilson, Jr., *Regression by leaps and bounds*, Technometrics, 16 (1974), pp. 499–511.
[9] A. H. Feiveson, *Finding the best regression subset by Reduction in nonfull-rank cases*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 194–204.
[10] B. K. Natarajan, *Sparse approximate solutions to linear systems*, SIAM J. Comput., 24 (1995), pp. 227–234.
[11] S. J. Reeves, *An efficient implementation of the backward greedy algorithm for sparse signal reconstruction*, IEEE Signal Process. Lett., 6 (1999), pp. 266–268.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 3, pp. 809–824
© 2000 Society for Industrial and Applied Mathematics

ON THE PERTURBATION THEORY FOR UNITARY EIGENVALUE PROBLEMS*

B. BOHNHORST†, A. BUNSE-GERSTNER‡, AND H. FAßBENDER‡

Abstract. Some aspects of the perturbation theory for eigenvalues of unitary matrices are considered. Making use of the close relation between unitary and Hermitian eigenvalue problems a Courant–Fischer-type theorem for unitary matrices is derived and an inclusion theorem analogous to the Kahan theorem for Hermitian matrices is presented. Implications for the special case of unitary Hessenberg matrices are discussed.

Key words. unitary eigenvalue problem, perturbation theory

AMS subject classifications. 15A18, 65F99

PII. S0895479898333235

1. Introduction.

New numerical methods to compute eigenvalues of unitary matrices have been developed during the last 10 years. Unitary QR-type methods [19, 9], a divide-and-conquer method [20, 21], a bisection method [10], and some special methods for the real orthogonal eigenvalue problem [1, 2] have been presented. Interest in this task arose from problems in signal processing [11, 29, 33], in Gaussian quadrature on the unit circle [18], and in trigonometric approximations [31, 16] which can be stated as eigenvalue problems for unitary matrices, often in Hessenberg form. As those numerical methods exploit the rich mathematical structure of unitary matrices, which is closely analogous to the structure of Hermitian matrices, the methods are efficient and deliver very good approximations to the desired eigenvalues.

There exist, however, only a few perturbation results for the unitary eigenvalue problem, which can be used to derive error bounds for the computed eigenvalue approximations. A thorough and complete treatment of the perturbation aspects associated with the numerical methods for unitary eigenvalue problems is still missing.

The following perturbation results have been obtained so far. If U and $\widetilde{U}$ are unitary matrices with spectra $\sigma(U) = \{\lambda_j\}$ and $\sigma(\widetilde{U}) = \{\widetilde{\lambda}_j\}$, respectively, we can arrange the eigenvalues in diagonal matrices Λ and $\widetilde{\Lambda}$, respectively, and consider as a measure for the distance of the spectra

$$d_\mu(\sigma(U), \sigma(\widetilde{U})) := \min_P ||\Lambda - P^T \widetilde{\Lambda} P||_\mu, \qquad \mu = 2, F, \tag{1.1}$$

where the minimum is taken over all permutation matrices P and the norm is either the spectral or the Frobenius norm. By the Hoffman–Wielandt theorem (see, e.g., [34]) we get

$$d_F(\sigma(U), \sigma(\widetilde{U})) \leq ||U - \widetilde{U}||_F.$$

Bhatia and Davis [5] proved the corresponding result for the spectral norm

$$d_2(\sigma(U), \sigma(\widetilde{U})) \leq ||U - \widetilde{U}||_2.$$

*Received by the editors January 26, 1998; accepted for publication (in revised form) by R. Bhatia January 25, 1999; published electronically February 16, 2000.
http://www.siam.org/journals/simax/21-3/33323.html

†Schmidt, Vogel & Partner Consult, Gesellschaft für Organisation und Managementberatung mbH, Gadderbaumerstraße 19, 33602 Bielefeld, Germany.
‡Universität Bremen, Fachbereich 3—Mathematik und Informatik, 28334 Bremen, Germany (angelika@math.uni-bremen.de, heike@math.uni-bremen.de).

Elsner and He consider a relative error in [15]. They use the measure

$$\widetilde{d}_\mu(\sigma(U),\sigma(\widetilde{U})) := \min_P ||(\Lambda + P^T\widetilde{\Lambda}P)^{-1}(\Lambda - P^T\widetilde{\Lambda}P)||_\mu, \tag{1.2}$$

where again $\mu = 2$ or $\mu = F$. They prove that

$$\widetilde{d}_\mu(\sigma(U),\sigma(\widetilde{U})) \le ||C(U^H\widetilde{U})||_\mu,$$

where $C(U) = i(I+U)^{-1}(I-U)$ is the Cayley transformation of U (assuming here that $-1 \notin \sigma(U)$).

To each eigenvalue λ of U, where $-1 \notin \sigma(U)$, we can associate an angle θ_λ by defining

$$\theta_\lambda = \arctan[\sqrt{-1}(1+\lambda)^{-1}(1-\lambda)]$$

with $-\pi/2 \le \theta_\lambda < \pi/2$. It is the angle formed by the line from -1 through λ and the real axis (see also section 2). With respect to their angles the eigenvalues of U and $\widetilde{U}$ have a natural ordering on the unit circle. Elsner and He give sine and tangent interpretations of the above inequality in terms of these angles. Furthermore, they show that with respect to a certain cutting point ζ on the unit circle the eigenvalues of U and $\widetilde{U}$ have a natural ordering $\{\lambda_j(\zeta)\}$ and $\{\widetilde{\lambda}_j(\zeta)\}$ on the unit circle such that

$$\max_j |\lambda_j(\zeta) - \widetilde{\lambda}_j(\zeta)| \le ||U - \widetilde{U}||_2.$$

An interlacing theorem for unitary matrices is also presented in [15], showing that the eigenvalues of suitably modified principal submatrices of a unitary matrix interlace those of the complete matrix on the unit circle (see section 2).

In this paper we consider further aspects of the perturbation problem for the eigenvalues of a unitary matrix U. In section 2 we show how the angles ζ are related to the eigenvalues of the Cayley transform of U. With the aid of this relation we can give a min-max characterization for the angles of U's eigenvalues in analogy to the Courant–Fischer theorem for Hermitian matrices. We also show that tangents of these angles can be characterized by usual Rayleigh quotients corresponding to the generalized eigenvalue problem

$$\sqrt{-1}(I+U^H)(I-U)x = \mu(I+U^H)(I+U)x.$$

Furthermore, we prove a Kahan-like inclusion theorem showing that the eigenvalues of a certain modified leading principal submatrix of U determine arcs on the unit circle such that each arc contains an eigenvalue of U. In applications unitary matrices are often of Hessenberg form. In section 3 we recall that a unitary unreduced Hessenberg matrix H has a unique parameterization $H = H(\gamma_1,\dots,\gamma_n)$, where the reflection parameters $\gamma_1,\dots,\gamma_n \in \mathbb{C}$, with $|\gamma_i| < 1$ for $i = 1,\dots,n-1$ and $|\gamma_n| = 1$, determine H completely. We show the implications of the results in section 2 for the special case of unitary Hessenberg matrices. In particular it will be seen that the modified kth leading principal submatrix in this special case is just $H(\gamma_1,\dots,\gamma_{k-1},\zeta)$, where $|\zeta| = 1$. We discuss the dependence of the eigenvalues on this last reflection parameter ζ. Finally section 4 will give numerical examples which elucidate the statements proved in section 3.

2. Perturbation results for unitary matrices. Unitary matrices have a rich mathematical structure that is closely analogous to that of Hermitian matrices. In this section we first discuss the intimate relationship between unitary and Hermitian matrices which indicates that one can hope to find unitary analogues for the good numerical methods and for the theoretical results that exist for the symmetric/Hermitian eigenvalue problem. We will adapt some eigenvalue bounds for Hermitian matrices to the unitary case.

Let ρ be a complex unimodular number. The Cayley transformation with respect to ρ maps the unitary matrices whose spectrum does not include ρ onto the Hermitian matrices. The *Cayley transformation with respect to* ρ for a unitary matrix $U \in \mathbb{C}^{n\times n}$ is defined as

$$C(U) = \imath(\rho I_n - U)^{-1}(\rho I_n + U),$$

where ρ is not an eigenvalue of U and $\imath = \sqrt{-1}$. I_n denotes the $n \times n$ identity matrix. A simple calculation shows that $C(U)$ is Hermitian. The mapping is one to one and the *inverse Cayley transformation with respect to* ρ for a Hermitian matrix X is given by

$$C^{-1}(X) = \rho(X + \imath I_n)^{-1}(X - \imath I_n).$$

The symmetric/Hermitian eigenproblem has been extensively studied; see, e.g., [28, 17, 24, 26]. Due to this relation between Hermitian and unitary matrices, one can hope to get similar results for unitary matrices.

With the aid of the Cayley transformation we can order the eigenvalues conveniently. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of U numbered starting at ρ moving counterclockwise along the unit circle. Let $\mu_1 \le \mu_2 \le \cdots \le \mu_n$ be the eigenvalues of $X = C(U)$. Then for $k = 1, \ldots, n$

$$\mu_k = \imath\frac{\rho + \lambda_k}{\rho - \lambda_k}. \tag{2.1}$$

For simplicity assume that $\rho = -1$. Then

$$\begin{aligned}\mu_k &= \imath\frac{1 - \lambda_k}{1 + \lambda_k}\\ &= \frac{\mathrm{Im}(\lambda_k)}{1 + \mathrm{Re}(\lambda_k)}\\ &= \tan(\theta_k),\end{aligned}$$

where $\theta_k = \arg(1+\lambda_k)$. For $z \in \mathbb{C}\setminus\{0\}$ the argument of z, $\arg(z) \in (-\pi, \pi]$ is defined by

$$\arg(z) = \begin{cases} \arctan\left(\dfrac{\mathrm{Im}(z)}{\mathrm{Re}(z)}\right) + \pi, & \mathrm{Re}(z) < 0,\\ \arctan\left(\dfrac{\mathrm{Im}(z)}{\mathrm{Re}(z)}\right), & \mathrm{Re}(z) > 0,\\ \frac{\pi}{2}, & \mathrm{Re}(z) = 0, \mathrm{Im}(z) > 0,\\ -\frac{\pi}{2}, & \mathrm{Re}(z) = 0, \mathrm{Im}(z) < 0. \end{cases}$$

The Cayley transformation of λ_k is the tangent of the angle $\theta_k = \arg(1 + \lambda_k)$, which is formed by the real axis and the straight line through λ_k and -1, as shown in Figure 2.1.

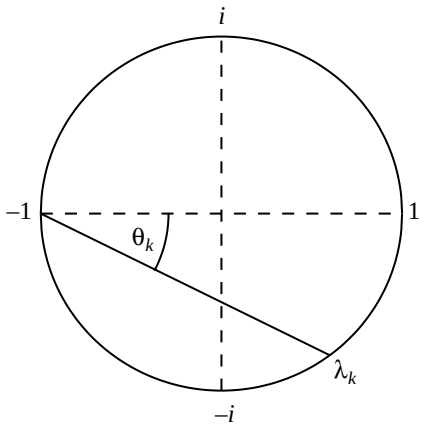


FIG. 2.1.

Hence it is reasonable to define

$$\lambda_i \leq \lambda_j \quad \text{if} \quad \imath\frac{1-\lambda_i}{1+\lambda_i} \leq \imath\frac{1-\lambda_j}{1+\lambda_j}.$$

This also gives a complete ordering of the points on the unit circle with respect to the cutting point -1. Note that the complete ordering excludes the cutting point -1. For a different cutting point the orders of the eigenvalues are changed only cyclically.

If ζ_1, ζ_2 are complex unimodular numbers such that $\zeta_1 < \zeta_2$, then (ζ_1, ζ_2) will denote the *open arc* from the point ζ_1 to the point ζ_2 on the unit circle (moving counterclockwise).

The Courant–Fischer theorem (see, e.g., [17, Theorem 8.1.2]) characterizes the eigenvalues of Hermitian matrices by Rayleigh quotients. A similar characterization can be given for the eigenvalues of unitary matrices. Let $U \in \mathbb{C}^{n\times n}$ be a unitary matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$. Assume that -1 is not an eigenvalue of U and number the eigenvalues starting at -1 moving counterclockwise along the unit circle. Let $v_1, \ldots, v_n$ be an orthonormal basis in $\mathbb{C}^{n\times n}$ of eigenvectors of U. Let $z \in \mathbb{C}^n$ with $||z||_2 = 1$. Then we can expand z as

$$z = \sum_{i=1}^{n} \alpha_i v_i, \qquad \alpha_i \in \mathbb{C},\ i = 1, \ldots, n.$$

From $||z||_2 = 1$ we obtain

$$\sum_{i=1}^{n} |\alpha_i|^2 = 1.$$

Because of

$$z^H U z = \sum_{i=1}^{n} |\alpha_i|^2 \lambda_i$$

we see that the Rayleigh quotient $z^H U z$ lies in the convex polygon which is spanned by the eigenvalues of U, as depicted in Figure 2.2 for a 5 x 5 example.

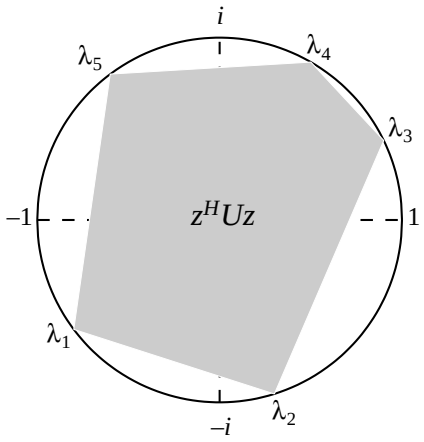


FIG. 2.2.

THEOREM 2.1. *With the notation given above we obtain for* $\theta_k = \arg(1+\lambda_k), k = 1, \ldots, n$

$$\begin{aligned}\theta_k &= \max_{V \in V_{n-k+1}} \min_{z \in V, ||z||_2 = 1} \arg(1 + z^H U z) \\ &= \min_{V \in V_k} \max_{z \in V, ||z||_2 = 1} \arg(1 + z^H U z),\end{aligned}$$

where V_k *denotes the set of all* k*-dimensional subspaces of* $\mathbb{C}^n$.

In particular,

$$\begin{aligned}\theta_1 &= \min_{z \in \mathbb{C}^n, ||z||_2 = 1} \arg(1 + z^H U z), \\ \theta_n &= \max_{z \in \mathbb{C}^n, ||z||_2 = 1} \arg(1 + z^H U z).\end{aligned}$$

Proof. Let $v_1, \ldots, v_n$ be an orthonormal basis of eigenvectors of U. Let $V \in V_{n-k+1}$. Then $V \cap \{v_1, \ldots, v_k\} \neq \emptyset$. Let z be a vector in this intersection, $||z||_2 = 1$. Then

$$z = \sum_{i=1}^{k} \alpha_i v_i, \qquad \sum_{i=1}^{k} |\alpha_i|^2 = 1, \qquad \alpha_i \in \mathbb{C},\ i = 1, \ldots, k.$$

Hence, the Rayleigh quotient $z^H U z$ lies in the convex polygon spanned by the eigenvalues $\lambda_1, \ldots, \lambda_k$. Therefore

$$\arg(1 + z^H U z) \leq \arg(1 + \lambda_k)$$

and

$$\max_{V \in V_{n-k+1}} \min_{z \in V, ||z||_2 = 1} \arg(1 + z^H U z) \leq \theta_k.$$

Now consider the subspace of dimension $n - k + 1$ which is spanned by $v_k, \ldots, v_n$. A vector $z, ||z||_2 = 1$ in this subspace can be written as

$$z = \sum_{i=k}^{n} \gamma_i v_i, \qquad \sum_{i=k}^{n} |\gamma_i|^2 = 1, \qquad \gamma_i \in \mathbb{C},\ i = k, \ldots, n.$$

Hence, the Rayleigh quotient $z^H U z$ lies in the convex polygon spanned by the eigenvalues $\lambda_k, \ldots, \lambda_n$. Therefore

$$\arg(1 + z^H U z) \geq \arg(1 + \lambda_k)$$

and

$$\max_{V \in V_{n-k+1}} \min_{z \in V, ||z||_2 = 1} \arg(1 + z^H U z) = \theta_k.$$

The second equation can be shown analogously. ∎

COROLLARY 2.2. *With the notation given above we define for* $z \in \mathbb{C}^n$ *with* $||z||_2 = 1$:

$$R(z) = \frac{z^H[\imath(I_n + U^H)(I_n - U)]z}{z^H[(I_n + U^H)(I_n + U)]z}.$$

Then

$$\tan(\arg(1 + z^H U z)) = R(z) \tag{2.2}$$

and for $k = 1, \ldots, n$

$$\begin{aligned} \tan(\theta_k) &= \max_{V \in V_{n-k+1}} \min_{z \in V, ||z||_2 = 1} R(z) \\ &= \min_{V \in V_k} \max_{z \in V, ||z||_2 = 1} R(z). \end{aligned}$$

Proof. A simple calculation yields (2.2). The rest of the corollary follows from Theorem 2.1 and the monotonicity of the function tan in $(-\frac{\pi}{2}, \frac{\pi}{2})$. ∎

The corollary shows that the angles θ_k can be characterized by usual Rayleigh quotients. $R(z)$ can be interpreted as the Rayleigh quotient corresponding to the generalized eigenvalue problem

$$\{\imath(I_n + U^H)(I_n - U) - \mu(I_n + U^H)(I_n + U)\}x = 0. \tag{2.3}$$

Since U is unitary, $\imath(I_n + U^H)(I_n - U)$ is Hermitian and $(I_n + U^H)(I_n + U)$ is Hermitian and positive definite. Equation (2.3) is equivalent to the eigenvalue problem

$$\imath(I_n + U)^{-1}(I_n - U)x = \mu x$$

for the Cayley transformation of U.

Remark 2.3. For ease of notation the above theorem and corollary are formulated for the case that $\rho = -1$ is not an eigenvalue of U. This restriction is not necessary—one can prove the corresponding statements for any cutting point $\rho \in \mathbb{C}, |\rho| = 1, \rho$ not an eigenvalue of U.

In [15], the Cauchy interlacing theorem for Hermitian matrices is generalized to the unitary case. The Cauchy interlace theorem shows that the eigenvalues of the $k \times k$ leading principal submatrix of a Hermitian matrix X interlace the eigenvalues of X. Adapting this theorem to the unitary case, one has to deal with the problem that leading principal submatrices of unitary matrices are in general not unitary and that their eigenvalues lie inside the unit circle. In [15] it is shown that certain modified leading principal submatrices of a unitary matrix U have the property that their eigenvalues interlace with those of U.

THEOREM 2.4 (see Elsner and He [15, Theorems 5.2 and 5.3]). *Let*

$$U = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix} \tag{2.4}$$

be an $n \times n$ unitary matrix, U_{11} the $k \times k$ leading principal submatrix of U, and

$$U_k = U_{11} - U_{12}(U_{22} - \rho I_{n-k})^{-1} U_{21}, \tag{2.5}$$

where $\rho \in \mathbb{C}, |\rho| = 1$, and ρ is not an eigenvalue of U. Then U_k is unitary. Let

$$\lambda_1 \leq \cdots \leq \lambda_n$$

be the eigenvalues of U and

$$\mu_1 \leq \cdots \leq \mu_k$$

be those of U_k ordered with respect to ρ. Then

$$\lambda_i \leq \mu_i \leq \lambda_{i+n-k}.$$

U_k is called the *modified kth leading principal submatrix* of U. Furthermore, analogues of the Hoffman–Wielandt theorem and a Weyl-type theorem are derived in [15].

With the help of Theorem 2.4 one can specify for each eigenvalue of a modified leading principal submatrix U_k an arc on the unit circle which contains that eigenvalue. These bounds are fairly rough, especially if k is much smaller than n. This result is of a theoretical nature, because in practice we are more interested in the question of whether the arc (μ_1, μ_2) contains an eigenvalue of U.

The same problem arises in the Hermitian case. Lehmann and Kahan derived inclusion theorems which consider this problem (see, e.g., [28] and the references therein). A special case of their results is the following theorem.

THEOREM 2.5 (see Parlett [28, p. 196]). *Let $X \in \mathbb{C}^{n\times n}$ be a Hermitian matrix. Partition X as*

$$X = \begin{bmatrix} X_k & C^H \\ C & V \end{bmatrix}, \qquad \text{where } X_k \in \mathbb{C}^{k\times k}.$$

Let μ_i be the eigenvalues of $X_k, i = 1, \ldots, k$, $\mu_1 \leq \cdots \leq \mu_k$. If $\operatorname{rank}(C) = 1$, *then each interval $[\mu_i, \mu_{i+1}], i = 0, 1, \ldots, k$, contains an eigenvalue of X, where $\mu_0 = -\infty, \mu_{k+1} = \infty$.*

The following theorem states the analogous result for unitary matrices.

THEOREM 2.6. *Let $U \in \mathbb{C}^{n\times n}$ be a unitary matrix and let U be partitioned as in* (2.4). *Let $\rho \in \mathbb{C}, |\rho| = 1$, be not an eigenvalue of U. Define a unitary matrix U_k as in* (2.5) *with eigenvalues $\mu_1, \ldots, \mu_k$. The eigenvalues are numbered starting at ρ moving counterclockwise along the unit circle. If* $\operatorname{rank}(U_{21}) = 1$, *then each arc (μ_i, μ_{i+1}) on the unit circle, $i = 1, \ldots, k$, contains at least one eigenvalue of U, where $\mu_0 = \mu_{k+1} = \rho$.*

Proof. Since ρ is not an eigenvalue of U, it is not an eigenvalue of U_{22}: assume ρ is an eigenvalue of U_{22}. Then there is a normalized eigenvector $x \in \mathbb{C}^{n-k}$ such that $U_{22}x = \rho x$. Therefore,

$$U \begin{bmatrix} 0 \\ x \end{bmatrix} = \begin{bmatrix} U_{12}x \\ U_{22}x \end{bmatrix} = \begin{bmatrix} U_{12}x \\ \rho x \end{bmatrix}.$$

As $|\rho| = 1$, $\|x\| = 1$, and U is unitary, $U_{12}x$ has to be zero. But this would imply that ρ is an eigenvalue of U in contradiction to our general assumption. Hence ρ is not an eigenvalue of U_{22}.

U_k is defined as

$$U_k = U_{11} - U_{12}(U_{22} - \rho I_{n-k})^{-1}U_{21}.$$

$U_k - \rho I_k$ can be interpreted as the Schur complement of $U_{22} - \rho I_{n-k}$ in $U - \rho I_n$. We can make use of this fact to construct $(U - \rho I_n)^{-1}$ using the following result of Duncan [13] (see Corollary 2.4 in [27]):

Let $A \in \mathbb{C}^{n\times n}$ be partitioned as

$$A = \begin{bmatrix} E & F \\ G & H \end{bmatrix}, \qquad \text{where } H \in \mathbb{C}^{n-k\times n-k}.$$

Let A and H be nonsingular; then $T = E - FH^{-1}G \in \mathbb{C}^{k\times k}$ is the Schur complement of H in A, T is nonsingular, and

$$A^{-1} = \begin{bmatrix} T^{-1} & -T^{-1}FH^{-1} \\ -H^{-1}GT^{-1} & H^{-1} + H^{-1}GT^{-1}FH^{-1} \end{bmatrix}.$$

With $V = (U_k - \rho I_k)^{-1}$ and $W = (U_{22} - \rho I_{n-k})^{-1}$ we obtain

$$(U - \rho I_n)^{-1} = \begin{bmatrix} V & -WU_{21}V \\ VU_{12}W & W[I_{n-k} + U_{21}VU_{12}W] \end{bmatrix}.$$

In particular, $(U_k - \rho I_k)^{-1}$ is the $k \times k$ leading principal submatrix of $(U - \rho I_n)^{-1}$.

Now we consider the Cayley transformation with respect to ρ of U:

$$\begin{aligned} X = C(U) &= \imath(\rho I_n - U)^{-1}(\rho I_n + U) \\ &= -\imath I_n - 2\imath\rho(U - \rho I_n)^{-1}. \end{aligned}$$

This yields

$$X + \imath I_n = -2\imath\rho(U - \rho I_n)^{-1}. \tag{2.6}$$

We partition X as we did U:

$$X = \begin{bmatrix} X_k & X_{12} \\ X_{21} & X_{22} \end{bmatrix},$$

where X_k is the $k \times k$ leading principal submatrix of X. From (2.6) it follows that

$$X_k + \imath I_k = -2\imath\rho(U_k - \rho I_k)^{-1}.$$

Therefore, X_k is the Cayley transformation of U_k. Further, we obtain from (2.6)

$$X_{21} = 2\imath\rho(U_{22} - \rho I_{n-k})^{-1}U_{21}(U_k - \rho I_k)^{-1}.$$

If $\operatorname{rank}(U_{21}) = 1$, then $\operatorname{rank}(X_{21}) = 1$ as the other two matrices in the product have full rank.

Now we can use Theorem 2.5 to obtain that each interval formed by two eigenvalues of X_k contains at least one eigenvalue of X. We have seen in (2.1) that the

eigenvalues of X and X_k can be obtained from those of U and U_k via the Cayley transformation. As the Cayley transformation is monotone, this yields the following: each arc (μ_i, μ_{i+1}) on the unit circle, $i = 1, \ldots, k-1$, contains at least one eigenvalue of U. For the two outer arcs $(\rho, \mu_1), (\mu_k, \rho)$ the statement follows directly from Theorem 2.4. ∎

The last result we mention in this section clarifies the question of how the eigenvalues of a unitary matrix change if the matrix is modified by a unitary differing from I only by rank one. For the Hermitian case, the answer is given, e.g., in [17, Chapter 12.5.3]. For the unitary case we obtain the following.

THEOREM 2.7. *Let $U, S \in \mathbb{C}^{n\times n}$ be unitary matrices and S such that*

$$\operatorname{rank}(I_n - S) = 1.$$

Then the eigenvalues of U and US interlace on the unit circle.

Proof. See [4, section 6]. ∎

3. Unitary upper Hessenberg matrices. It is well known that any (unitary) $n \times n$ matrix can be transformed to an upper Hessenberg matrix H by a unitary similarity transformation Q. If the first column of Q is fixed and H is an unreduced upper Hessenberg matrix with positive subdiagonal elements (that is, $h_{i+1,i} > 0$), then the transformation is unique. Any $n \times n$ unitary upper Hessenberg matrix with nonnegative subdiagonal elements can be uniquely parameterized by $2n-1$ real parameters. This compact form is used in [1, 3, 9, 11, 14, 19, 20, 21, 22, 23, 32] to develop fast algorithms for solving the unitary eigenvalue problem.

Let

$$G_k = G_k(\gamma_k) = \operatorname{diag}\left(I_{k-1}, \begin{bmatrix} -\gamma_k & \sigma_k \\ \sigma_k & \overline{\gamma_k} \end{bmatrix}, I_{n-k-1}\right)$$

with $\gamma_k \in \mathbb{C}, \sigma_k \in \mathbb{R}^+$, and $|\gamma_k|^2 + \sigma_k^2 = 1$, and let

$$\widetilde{G}_n(\gamma_n) = \operatorname{diag}(I_{n-1}, -\gamma_n)$$

with $\gamma_n \in \mathbb{C}, |\gamma_n| = 1$.

The product

$$H = H(\gamma_1, \gamma_2, \ldots, \gamma_n) := G_1(\gamma_1) \cdots G_{n-1}(\gamma_{n-1}) \widetilde{G}_n(\gamma_n)$$

$$= \begin{pmatrix} -\gamma_1 & -\sigma_1\gamma_2 & -\sigma_1\sigma_2\gamma_3 & \cdots & -\sigma_1 \ldots \sigma_{n-1}\gamma_n \\ \sigma_1 & -\overline{\gamma_1}\gamma_2 & -\overline{\gamma - 1}\sigma_2\gamma_3 & \cdots & -\overline{\gamma_1}\sigma_2 \ldots \sigma_{n-1}\gamma_n \\ & \sigma_2 & -\overline{\gamma}_2\gamma_3 & \cdots & -\overline{\gamma}_2\sigma_3 \ldots \sigma_{n-1}\gamma_n \\ & & \ddots & \ddots & \vdots \\ & & & \sigma_{n-1} & -\overline{\gamma_{n-1}}\gamma_n \end{pmatrix}$$

is a unitary upper Hessenberg matrix with positive subdiagonal elements. Conversely, if $H \in \mathbb{C}^{n\times n}$ is a unitary upper Hessenberg matrix with positive subdiagonal elements, then it follows from elementary numerical linear algebra that one can determine matrices $G_1, G_2, \ldots, G_{n-1}, \widetilde{G}_n$ such that $\widetilde{G}_n^H G_{n-1}^H \cdots G_2^H G_1^H H = I$. Since H as a unitary matrix has a unique inverse, this has to be $\widetilde{G}_n^H G_{n-1}^H, \ldots, G_2^H G_1^H$. Thus H has a unique factorization of the form

$$H = H(\gamma_1, \gamma_2, \ldots, \gamma_n) = G_1(\gamma_1) G_2(\gamma_2) \cdots G_{n-1}(\gamma_{n-1}) \widetilde{G}_n(\gamma_n). \tag{3.1}$$

The *Schur parameters* $\{\gamma_k\}_{k=1}^n$ and the *complementary Schur parameters* $\{\sigma_k\}_{k=1}^n$ can be computed from the elements of H by a stable $O(n^2)$ algorithm [19]. In statistics the Schur parameters are referred to as partial correlation coefficients and in signal processing as reflection coefficients [2, 11, 12, 25, 29, 30, 33].

If $\sigma_k = 0$, then $|\gamma_k| = 1$, and we have the direct sum decomposition

$$H = H(\gamma_1, \ldots, \gamma_k) \oplus H(\overline{\gamma}_k\gamma_{k+1}, \ldots, \overline{\gamma}_k\gamma_n).$$

Hence, in general $\sigma_1\sigma_2\cdots\sigma_{n-1} > 0$ is assumed if the factorization (3.1) is used to solve a unitary eigenvalue problem. Such a unitary upper Hessenberg matrix is called *unreduced*. If λ is an eigenvalue of an unreduced Hessenberg matrix, then its geometric multiplicity is one [17, Theorem 7.4.4]. Since unitary matrices are diagonalizable, no eigenvalue of an unreduced unitary upper Hessenberg matrix is defective; that is, the eigenvalues of an unreduced unitary upper Hessenberg matrix are distinct.

We will adapt the general theorems given in the last section to the more specific case of unitary upper Hessenberg matrices. Let $H \in \mathbb{C}^{n\times n}$ be a unitary upper Hessenberg matrix with positive subdiagonal elements, $H = H(\gamma_1, \ldots, \gamma_n)$. Partition H as

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}, \tag{3.2}$$

where H_{11} is the $k \times k$ leading principal submatrix. From Theorem 2.4 we obtain that the modified kth leading principal submatrix of H, $H_k = H_{11} - H_{12}(H_{22} - \rho I)^{-1}H_{21}$, is unitary (if $\rho \in \mathbb{C}, |\rho| = 1, \rho$ not an eigenvalue of H). As H_k is a unitary upper Hessenberg matrix, we can factor $H_k = G_1(\widehat{\gamma}_1)G_2(\widehat{\gamma}_2)\cdots G_{k-1}(\widehat{\gamma}_{k-1})\widetilde{G}_k(\widehat{\gamma}_k)$. Taking a closer look at H_k reveals that H_k differs from H_{11} only in the last column. Hence the modification of H_{11} to H_k is equivalent to a modification of the reflection coefficient γ_k, i.e.,

$$H_k = H_k(\gamma_1, \ldots, \gamma_{k-1}, \zeta_k), \qquad |\zeta_k| = 1.$$

The following theorem by Bunse-Gerstner and He [10] characterizes the correct choice of the parameter ζ_k.

THEOREM 3.1. *Let $H = H(\gamma_1, \ldots, \gamma_n) \in \mathbb{C}^{n\times n}$ be a unitary upper Hessenberg matrix with positive subdiagonal elements. For $k \in \{1, \ldots, n-1\}$ let H be partitioned as in* (3.2). *Let $\rho \in \mathbb{C}, |\rho| = 1$, not be an eigenvalue of H. Define parameters $\zeta_l(\rho), l = 1, \ldots, n$ by*

$$\begin{aligned} \zeta_n(\rho) &= \gamma_n, \\ \zeta_l(\rho) &= \frac{\rho\gamma_l + \zeta_{l+1}(\rho)}{\rho + \overline{\gamma_l}\zeta_{l+1}(\rho)}, \qquad l = n-1, \ldots, 1. \end{aligned} \tag{3.3}$$

Then

$$H_k(\gamma_1, \ldots, \gamma_{k-1}, \zeta_k(\rho)) = H_{11} - H_{12}(H_{22} - \rho I_{n-k})^{-1}H_{21}.$$

Let $\mu_1, \ldots, \mu_k$ and $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of $H_k(\gamma_1, \ldots, \gamma_{k-1}, \zeta_k(\rho))$ and H, respectively, where the eigenvalues are numbered starting at ρ moving counterclockwise along the unit circle. Then for each $i = 1, \ldots, k$ the eigenvalue μ_i lies on the arc $(\lambda_i, \lambda_{i+n-k})$.

Proof. The interlace property follows directly from Theorem 2.4. For the rest of the proof see [10]. ∎

Using Theorem 2.6 and 3.1 we obtain the following theorem.

THEOREM 3.2. *Let $H = H(\gamma_1, \ldots, \gamma_n) \in \mathbb{C}^{n\times n}$ be a unitary upper Hessenberg matrix with positive subdiagonal elements. Let $\rho \in \mathbb{C}, |\rho| = 1$, not be an eigenvalue of H. For $k \in \{1, \ldots, n\}$ let $H_k = H_k(\gamma_1, \ldots, \gamma_{k-1}, \zeta_k(\rho))$, where $\zeta_k(\rho)$ is defined as in* (3.3). *Let $\mu_1, \ldots, \mu_k$ be the eigenvalues of H_k. Then for $i = 0, \ldots, k$ each arc (μ_i, μ_{i+1}) on the unit circle contains at least one eigenvalue of H, where $\mu_0 = \mu_{k+1} = \rho$.*

Moreover, we obtain the following.

THEOREM 3.3. *Let $H = H(\gamma_1, \ldots, \gamma_n) \in \mathbb{C}^{n\times n}$ be a unitary upper Hessenberg matrix with positive subdiagonal elements. Then for any $\zeta \in \mathbb{C}, |\zeta| = 1$ there exists a cutting point $\rho \in \mathbb{C}, |\rho| = 1$ such that the eigenvalues of $H_k = H_k(\gamma_1, \ldots, \gamma_{k-1}, \zeta)$ and H have the interlace properties with respect to ρ on the unit circle given by Theorems* 3.1 *and* 3.2.

Proof. We will show that $\rho \mapsto \zeta_k(\rho)$ is an automorphism on the unit circle; this proves the theorem. Note that for unitary upper Hessenberg matrices with positive subdiagonal elements we have $|\gamma_j| < 1, j = 1, \ldots, n-1$.

Obviously

$$\rho \mapsto \zeta_{n-1}(\rho) = \frac{\rho\gamma_{n-1} + \gamma_n}{\rho + \overline{\gamma_{n-1}}\gamma_n}$$

is bijective on the unit circle. The same is true for the mapping

$$\zeta_j(\rho) \mapsto \zeta_{j-1}(\rho) = \frac{\rho\gamma_{j-1} + \zeta_j(\rho)}{\rho + \overline{\gamma_{j-1}}\zeta_j(\rho)}, \qquad j = n-1, \ldots, 2.$$

Hence $\rho \mapsto \zeta_k(\rho)$ is a one-to-one mapping of the unit circle onto itself. ∎

The statement of the above theorem can be summarized as follows. Any leading principal submatrix of a unitary upper Hessenberg matrix with positive lower subdiagonal elements can be modified to be unitary by replacing the last reflection coefficient with a parameter on the unit circle. No matter how this parameter is chosen, there is always a cutting point ρ on the unit circle such that the eigenvalues of the modified leading principal submatrix and those of the entire matrix satisfy the interlace properties given by Theorems 3.1 and 3.2.

Disregarding the cutting point ρ and the two arcs formed with it, Theorem 3.3 implies the following corollary.

COROLLARY 3.4. *Let $H = H(\gamma_1, \ldots, \gamma_n) \in \mathbb{C}^{n\times n}$ be a unitary upper Hessenberg matrix with positive subdiagonal elements. For $\zeta \in \mathbb{C}, |\zeta| = 1$ let $H_k = H_k(\gamma_1, \ldots, \gamma_{k-1}, \zeta) \in \mathbb{C}^{k\times k}, k \in \{1, \ldots, n\}$. Then every arc on the unit circle formed by two eigenvalues of H_k contains an eigenvalue of H.*

In particular the above theorems show that the eigenvalues of two consecutive modified leading principal submatrices H_k and H_{k+1} of a unitary upper Hessenberg matrix with positive subdiagonal elements interlace on the unit circle. More specifically, consider the modified leading principal submatrices $H_k = H_k(\gamma_1, \ldots, \gamma_{k-1}, \zeta_k)$ and $H_{k+1} = H_{k+1}(\gamma_1, \ldots, \gamma_k, \zeta_{k+1})$, where $|\zeta_k| = |\zeta_{k+1}| = 1$ and $k < n$. The eigenvalues of H_k and H_{k+1} interlace with respect to the cutting point ρ on the unit circle where ρ is given by

$$\rho = \frac{\zeta_{k+1}(1 - \overline{\gamma_k}\zeta_k)}{\zeta_k - \gamma_k}.$$

The remaining question is, How strongly do the eigenvalues of $H_k(\gamma_1, \ldots, \gamma_{k-1}, \zeta)$ depend of the choice of ζ? We present some results on this dependence on the last reflection parameter.

THEOREM 3.5. *Let $H_a = H(\gamma_1, \ldots, \gamma_{n-1}, \zeta_a), H_b = H(\gamma_1, \ldots, \gamma_{n-1}, \zeta_b)$ be unitary upper Hessenberg matrices with positive subdiagonal elements, $|\zeta_a| = |\zeta_b| = 1$.*

1. *The eigenvalues of H_a and H_b interlace on the unit circle.*
2. *$\nu(H_a, H_b) \leq |\zeta_a - \zeta_b|$, where the eigenvalue variation $\nu(U, B)$ is defined by*

$$\nu(U, B) = \min \left\{ \max_{i \in \{1,\ldots,n\}} |\lambda_i - \mu_{\Pi(i)}|, \Pi \text{ permutation of } \{1, \ldots, n\} \right\},$$

   the λ_i's being the eigenvalues of U and the μ_i's those of B.
3. *Let $\lambda_1^a, \ldots, \lambda_n^a$ and $\lambda_1^b, \ldots, \lambda_n^b$ be the eigenvalues of H_a and H_b. Let*

$$S_k^H H_a S_k = \operatorname{diag}(\lambda_1^a, \ldots, \lambda_n^a)$$

   be the Schur decomposition of H_a, $S_k = [s_1, \ldots, s_k] = [s_{ij}]_{i,j=1}^k$. Then for $i = 1, \ldots, k$

$$\begin{aligned} \min_{j \in \{1,\ldots,k\}} |\lambda_i^a - \lambda_j^b| &\leq ||H_b s_i - \lambda_i^a s_i||_2 \\ &\leq |\zeta_a - \zeta_b| \ |s_{ki}|. \end{aligned}$$

Proof.

1. We have

$$H_b = H_a S, \qquad \text{where } S = \operatorname{diag}(I_{n-1}, \overline{\zeta_a}\zeta_b),$$

   S is unitary, and $\operatorname{rank}(I_n - S) = 1$. According to Theorem 2.7, the eigenvalues of H_a and H_b interlace on the unit circle.
2. As the matrices H_a and H_b differ only in the last column we have

$$||H_a - H_b||_2 = ||G_1(\gamma_1) \cdots G_{n-1}(\gamma_{n-1})[\widetilde{G}_n(\zeta_a) - \widetilde{G}_n(\zeta_b)]||_2 = |\zeta_a - \zeta_b|.$$

   Since H_a and H_b are unitary, statement 2 follows from the following theorem of Bhatia and Davis [5]:

   For all constant multiples $U = \alpha Q$ and $B = \beta V$ of two unitary matrices Q and V we have

$$\nu(U, B) \leq ||U - B||_2.$$

   (For a completely different proof and extension of the result to multiples of unitaries see [6]. When U and B are Hermitian, the above inequality is a classical result of Weyl.)
3. H_b is unitary and therefore unitarily diagonalizable. The first inequality follows directly from the following easy-to-prove result [8, Satz 1.8.14]:

   Let $A \in \mathbb{C}^{n \times n}$ be diagonalizable, $A = TDT^{-1}$, where $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ and $T \in \mathbb{C}^{n \times n}$. Let $\mu \in \mathbb{C}, x \in \mathbb{C}^n, ||x||_2 = 1$, and $r = Ax - \mu x$. Then

$$\min_{i \in \{1,\ldots,n\}} |\lambda_i - \mu| \leq ||T||_2 \ ||T^{-1}||_2 \ ||r||_2.$$

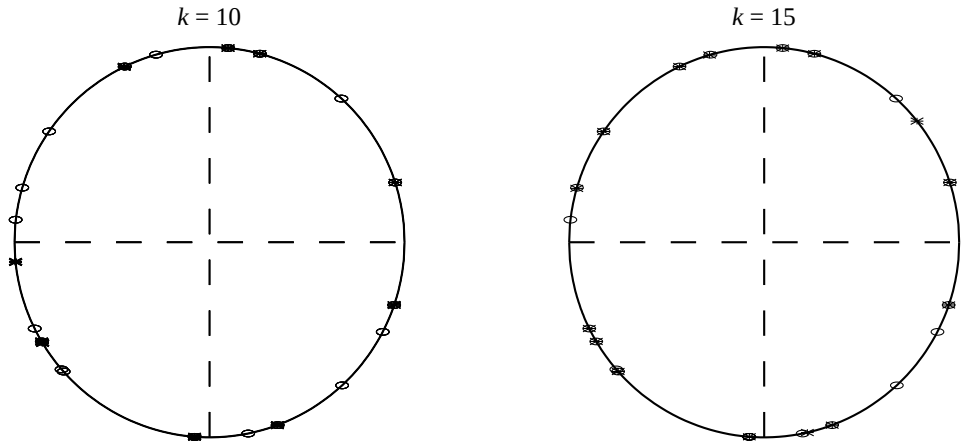


FIG. 4.1.

Furthermore, we obtain

$$\begin{aligned} ||H_b s_i - \lambda_i^a s_i||_2 &= ||(H_b - H_a)s_i||_2 \\ &= ||G_1(\gamma_1)\cdots G_{n-1}(\gamma_{n-1})[\widetilde{G}_n(\zeta_a) - \widetilde{G}_n(\zeta_b)]s_i||_2 \\ &= ||(\zeta_a - \zeta_b)e_k^T s_i||_2 \\ &= |\zeta_a - \zeta_b|\ |s_{ki}|. \qquad \square \end{aligned}$$

Hence, eigenvalues of a unitary upper Hessenberg matrix, whose eigenvectors have a small last component, are not sensitive to changes in the last reflection parameter.

4. Numerical examples. Numerical examples are presented to elucidate the statements of section 3. The eigenvalues of a unitary upper Hessenberg matrix H are compared with the eigenvalues of modified kth leading principal submatrices H_k for different dimensions k. The essential statements of section 3 can be observed clearly:

- Between two eigenvalues of H_k on the unit circle there lies an eigenvalue of H (Corollary 3.4).
- The eigenvalues of unitary upper Hessenberg matrices, whose corresponding eigenvectors have a small last component, are not sensitive against changes of the last reflection coefficient (Theorem 3.5).

All computations were done using MATLAB on a SUN SparcStation 10.

A unitary upper Hessenberg matrix $H = H(\gamma_1, \ldots, \gamma_{20}) \in \mathbb{C}^{20\times 20}$ was constructed from 20 randomly chosen reflection coefficients $\gamma_1, \ldots, \gamma_{20} \in \mathbb{C}$. The eigenvalues λ_j of H lie randomly on the unit circle. The eigenvalue μ_j of the modified kth leading principal submatrices $H_k = H_k(\gamma_1, \ldots, \gamma_{k-1}, \zeta_k)$ were computed for different dimensions $k < 20$.

For the first example $\zeta_k = \frac{\gamma_k}{|\gamma_k|}$ was chosen. The eigenvalues of H and H_k are plotted for $k = 10$ and $k = 15$ in Figure 4.1, a typical figure. The eigenvalues of H are marked by o, the eigenvalues of H_k by '*'.

For the second example a random complex number $\zeta_k, |\zeta_k| = 1$ was chosen. Figure 4.2 displays the same information as before.

Corollary 3.4 states that every arc on the unit circle formed by two eigenvalues of H_k contains an eigenvalue of H. This can be seen in Figures 4.1 and 4.2. Comparing the results of the two examples presented, one observes that independent of the choice of ζ_k the same eigenvalues of H are approximated. In Theorem 3.5 it was proved

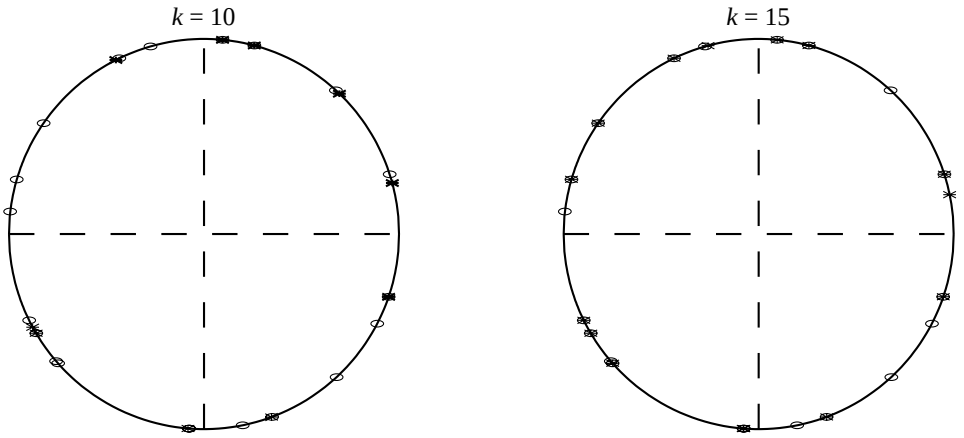


FIG. 4.2.

TABLE 4.1

	$\min_{j\in\{1,\dots,n\}} \lvert\lambda_i^{\gamma_{10}} - \lambda_j^{rand}\rvert$	$\lvert\zeta_a - \zeta_b\rvert\ \lvert s_{ki}\rvert$
$\lambda_1^{\gamma_{10}}$	1.8894e-06	1.2717e-03
$\lambda_2^{\gamma_{10}}$	8.6787e-05	7.2976e-03
$\lambda_3^{\gamma_{10}}$	4.7180e-02	1.4922e-01
$\lambda_4^{\gamma_{10}}$	3.8116e-04	1.9408e-02
$\lambda_5^{\gamma_{10}}$	1.5810e-02	1.7357e-01
$\lambda_6^{\gamma_{10}}$	2.4628e-03	7.1265e-02
$\lambda_7^{\gamma_{10}}$	5.3716e-03	1.2358e-01
$\lambda_8^{\gamma_{10}}$	9.5131e-05	5.8360e-02
$\lambda_9^{\gamma_{10}}$	8.4488e-03	6.0709e-01
$\lambda_{10}^{\gamma_{10}}$	3.9452e-01	1.7845e+00

that if the last component of an eigenvector of a unitary upper Hessenberg matrix is small, then the corresponding eigenvalue is not sensitive against changes in the last reflection coefficient. Individual bounds for the minimal distance of each eigenvalue λ_j^a of $H_a = H_k(\gamma_1, \dots, \gamma_{k-1}, \zeta_a)$ to the eigenvalues λ_j^b of $H_b = H_k(\gamma_1, \dots, \gamma_{k-1}, \zeta_b)$ are given:

$$\min_{j\in\{1,\dots,n\}} |\lambda_i^a - \lambda_j^b| \le |\zeta_a - \zeta_b|\ |s_{ki}|,$$

where $s_{\ell i}, \ell \in \{1, \dots, k\}$ is the ℓth component of the eigenvector for the ith eigenvalue of H_a. This means that if there is an eigenvalue λ_i^a of H_a such that the last component of the corresponding eigenvector is small, then any unitary upper Hessenberg matrix of the form H_b will have an eigenvalue λ_j^b that is close to λ_i^a.

Table 4.1 reports the minimal distance between each eigenvalue $\lambda_i^{\gamma_k}$ of $H_k = H_k(\gamma_1, \dots, \gamma_{k-1}, \gamma_k/|\gamma_k|)$ and the eigenvalues λ_j^{rand} of $H_k = H_k(\gamma_1, \dots, \gamma_{k-1}, \zeta)$ (where $\zeta \in \mathbb{C}, |\zeta| = 1$ is randomly chosen as above) as well as the error bounds for $k = 10$. The absolute difference between ζ_a and ζ_b was $|\zeta_a - \zeta_b| = 1.9052e{+}00$. Comparing the actual minimal distance with the error bound one observes that the approximations are much better than the error bound predicts.

The same results can be observed for larger unitary upper Hessenberg matrices H. Moreover, one can observe that the eigenvalues of the modified leading principal submatrices H'_{k-1} and H'_k interlace on the unit circle with respect to a cutting point ρ.

5. Concluding remarks. In this paper, we have proved that the angles θ_k associated with the eigenvalues λ_j of a unitary matrix U can be characterized by Rayleigh quotients. An inclusion theorem for the eigenvalues of symmetric matrices given by Kahan was adapted to the unitary case. We discussed the special case of unitary Hessenberg matrices, which is important for certain applications. We proved that every arc on the unit circle formed by two eigenvalues of a modified kth leading principal submatrix of a unitary upper Hessenberg matrix contains an eigenvalue of the complete matrix. Results on the dependence of the eigenvalues of unitary upper Hessenberg matrices on the last reflection coefficient are given.

Parts of this paper (section 2 and most of section 3) first appeared in [7]. Bohnhorst analyzed the connection between a unitary matrix U and its Cayley transformation more closely with the help of structure ranks.

Acknowledgment. All three authors owe a special thanks to Ludwig Elsner for stimulating discussions and many helpful suggestions.

REFERENCES

[1] G. S. Ammar, W. B. Gragg, and L. Reichel, *On the eigenproblem for orthogonal matrices*, in Proceedings of the 25th IEEE Conference on Decision and Control, Athens, Greece, 1986, pp. 1963–1966.

[2] G. S. Ammar, W. B. Gragg, and L. Reichel, *Determination of Pisarenko frequency estimates as eigenvalues of an orthogonal matrix*, in Advanced Algorithms and Architectures for Signal Processing II, F. T. Luk, ed., Proc. SPIE 826, Bellingham, WA, 1987, pp. 143–145.

[3] G. S. Ammar, L. Reichel, and D. C. Sorensen, *An implementation of a divide and conquer algorithm for the unitary eigenproblem*, ACM Trans. Math. Software, (1992), pp. 292–307.

[4] P. Arbenz and G. H. Golub, *On the spectral decomposition of Hermitian matrices modified by low rank perturbations with applications*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 40–58.

[5] R. Bhatia and C. Davis, *A bound for the spectral variation of a unitary operator*, Linear and Multilinear Algebra, 15 (1984), pp. 71–76.

[6] R. Bhatia and J. A. R. Holbrook, *Short normal paths and spectral variation*, Proc. Amer. Math. Soc., 94 (1985), pp. 377–382.

[7] B. Bohnhorst, *Beiträge zur numerischen Behandlung des unitären Eigenwertproblems*, Ph.D. thesis, Fakultät für Mathematik, Universität Bielefeld, Bielefeld, Germany, 1993.

[8] W. Bunse and A. Bunse-Gerstner, *Numerische lineare Algebra*, Teubner, Leipzig, Stuttgart, 1985.

[9] A. Bunse-Gerstner and L. Elsner, *Schur parameter pencils for the solution of the unitary eigenproblem*, Linear Algebra Appl., 154/156 (1991), pp. 741–778.

[10] A. Bunse-Gerstner and C. He, *On a Sturm sequence of polynomials for unitary Hessenberg matrices*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1043–1055.

[11] G. Cybenko, *Computing Pisarenko frequency estimates*, in Proceedings of the 18th Annual Princeton Conference on Information Systems and Sciences, Princeton, NJ, 1984, pp. 587–591.

[12] P. Delsarte, Y. Genin, Y. Kamp, and P. V. Dooren, *Speech modelling and the trigonometric moment problem*, Philips J. Res., 37 (1982), pp. 277–292.

[13] W. J. Duncan, *Some devices for the solution of large sets of simultaneous linear equations (with an appendix on the reciprocation of partitioned matrices)*, Philos. Mag. (7), 35 (1944), pp. 660–670.

[14] P. J. Eberlein and C. P. Huang, *Global convergence of the QR algorithm for unitary matrices with some results for normal matrices*, SIAM J. Numer. Anal., 12 (1975), pp. 97–104.

[15] L. Elsner and C. He, *Perturbation and interlace theorems for the unitary Eigenvalue problem*, Linear Algebra Appl., 188/189 (1993), pp. 207–229.

[16] H. Fassbender, *On numerical methods for discrete least-squares approximation for trigonometric polynomials*, Math. Comp., 66 (1997), pp. 719–741.

[17] G. H. Golub and C. F. Van Loan, *Matrix Computation*, 2nd ed., The John Hopkins University Press, Baltimore, MD, 1989.

[18] W. B. Gragg, *Positive definite Toeplitz matrices, the Arnoldi process for isometric operators, and Gaussian quadrature on the unit circle*, in Numerical Methods in Linear Algebra, E. S.

Nikolaev, ed., Moscow University Press, Moscow, 1982, pp. 16–32 (in Russian). See also J. Comput. Appl. Math., 46 (1993), pp. 183–198.

[19] W. B. Gragg, *The QR algorithm for unitary Hessenberg matrices*, J. Comput. Appl. Math., 16 (1986), pp. 1–8.

[20] W. B. Gragg and L. Reichel, *A divide and conquer algorithm for the unitary eigenproblem*, in Hypercube Multiprocessors 1987, M. T. Heath, ed., Proc. Appl. Math. 29, SIAM, Philadelphia, PA, 1987, pp. 639–647.

[21] W. B. Gragg and L. Reichel, *A divide and conquer method for the unitary and orthogonal eigenproblem*, Numer. Math., 57 (1990), pp. 695–718.

[22] W. B. Gragg and T. L. Wang, *Convergence of the Shifted QR Algorithm for Unitary Hessenberg Matrices*, Report NPS-53-90-007, Naval Postgraduate School, Monterey, CA, 1990.

[23] W. B. Gragg and T. L. Wang, *Convergence of the Unitary Hessenberg QR algorithm with Unimodular Shifts*, Report NPS-53-90-008, Naval Postgraduate School, Monterey, CA, 1990.

[24] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 1985.

[25] J. Makhoul, *Linear prediction: A tutorial review*, Proc. IEEE, 63 (1975), pp. 561–580.

[26] M. Marcus and H. Minc, *A Survey of Matrix Theory and Matrix Inequalities*, Allyn and Bacon, Boston, 1964.

[27] D. V. Ouellette, *Schur complements and statistics*, Linear Algebra Appl., 36 (1981), pp. 187–295.

[28] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.

[29] V. F. Pisarenko, *The retrieval of harmonics from a covariance function*, Geophys. J. R. Astr. Soc., 33 (1973), pp. 347–366.

[30] L. Reichel and G. S. Ammar, *Fast approximation of dominant harmonics by solving an orthogonal eigenvalue problem*, in Proceedings of the Second IMA Conference on Mathematics in Signal Processing, Warwick, UK, J. G. McWhirther et al., eds., Oxford University Press, London, 1990.

[31] L. Reichel, G. S. Ammar, and W. B. Gragg, *Discrete least squares approximation by trigonometric polynomials*, Math. Comp., 57 (1991), pp. 273–289.

[32] H. Rutishauser, *Bestimmung der Eigenwerte orthogonaler Matrizen*, Numer. Math., 9 (1966), pp. 104–108.

[33] S. Sagayama and F. Itakura, *Duality theory of composite sinusoidal modelling and linear prediction*, in Proceedings International Acoustics Speech Signal Processing, Tokyo, 1986, pp. 1261–1264.

[34] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 3, pp. 825–839
© 2000 Society for Industrial and Applied Mathematics

THE USE OF *QR* FACTORIZATION IN SPARSE QUADRATIC PROGRAMMING AND BACKWARD ERROR ISSUES*

M. ARIOLI†

Abstract. We present a roundoff error analysis of a null space method for solving quadratic programming minimization problems. This method combines the use of a direct QR factorization of the constraints with an iterative solver on the corresponding null space. Numerical experiments are presented which give evidence of the good performances of the algorithm on sparse matrices.

Key words. augmented systems, sparse matrices, orthogonal factorization, roundoff

AMS subject classifications. 65F05, 65F10, 64F25, 65F50, 65G05

PII. S0895479898338147

1. Introduction.

Let $M \in \mathbb{R}^{n\times n}$ be a symmetric and positive semidefinite matrix, and let $A \in \mathbb{R}^{n\times m}, m \le n$, be a real full rank matrix, $q \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$. The quadratic programming problem with equality constraints

$$\min_{A^T x = b} \frac{1}{2} x^T M x + q^T x \tag{1.1}$$

has a unique solution $\hat{x}$ if and only if $\mathrm{Ker}(A^T) \cap \mathrm{Ker}(M) = \{0\}$. Introducing the vector $u \in \mathbb{R}^m$ of the Lagrangian parameters the problem (1.1) is equivalent to the augmented system

$$\begin{bmatrix} M & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} = \begin{bmatrix} -q \\ b \end{bmatrix}. \tag{1.2}$$

The augmented matrix is invertible, and the solution $[\hat{x}^T, \hat{u}^T]^T$ is composed of the solution of problem (1.1) and the Lagrangian parameters of the gradient of the objective function at $\hat{x}$.

In this paper, we present a roundoff error analysis of a null space method which uses a mixture of direct and iterative solvers. The direct solver is based on the QR factorization of the matrix A. Similar error analyses are discussed in [7, 17] for the least squares problems and in [10] for a different algorithm for quadratic programming problems. The use of iterative algorithms is not taken into account in [7, 10]. In [3, 6, 19] similar algorithms for the constrained least squares problems are proposed; these papers focus on the convergence properties of the iterative methods and use block partitioning of $A^T = [S^T; N^T]$, where S^T is assumed square and nonsingular, to compute a null space basis.

In section 2 we present the details of the algorithm, and in section 3 we analyze its roundoff error propagation. In section 4 we introduce upper bounds for the error affecting the computed solution $\overline{x}$ and the computed Lagrangian parameters $\overline{u}$. It should be pointed out that the augmented system (1.2) also gives the solution to the

*Received by the editors May 1, 1998; accepted for publication (in revised form) by N.J. Higham June 8, 1999; published electronically February 18, 2000.
http://www.siam.org/journals/simax/21-3/33814.html
†Istituto di Analisi Numerica, Consiglio Nazionale delle Ricerche, via Ferrata 1, 27100 Pavia, Italy (arioli@dragon.ian.pv.cnr.it).

(dual) least squares problem

$$\min_u \frac{1}{2}(Au+q)^T M^{-1}(Au+q) + b^T u.$$

Hence, e.g., the results given in section 3 and 4 are equally valid for this problem.

In section 5 we show the results of the numerical tests that we conducted on selected experiments using a sparse QR factorization algorithm illustrated in [29, 26]. We give numerical evidence supporting the theoretical results discussed in [23] and show that the backward error properties of the algorithm explain how the computed solution fulfills some of the requirements of the physical model. Finally, in section 6 we give our conclusions.

In the following, we will denote the augmented matrix by $\mathfrak{A}$, and by E_1 and E_2 the matrices

$$E_1 = \begin{bmatrix} I_m \\ 0_{n-m,m} \end{bmatrix} \qquad \text{and} \qquad E_2 = \begin{bmatrix} 0_{m,n-m} \\ I_{n-m} \end{bmatrix}.$$

Let $fl(\cdot)$ denote the result of a floating point computation. We assume that

$$fl(\alpha \square \beta) = (\alpha \square \beta)(1+\delta(\square,\alpha,\beta)), \quad |\delta(\square,\alpha,\beta)| \le \varepsilon, \tag{1.3}$$

where α and β are floating point numbers, ε is the machine precision, and $\square$ is one of $+ - */$. To a great extent modern computers have arithmetic that satisfies assumption (1.3), with the exception of some CRAY computers (e.g., CRAY2 and CRAY-YMP). Furthermore, we assume that the scalar products are accumulated using either extended precision arithmetic or the Kahan summation formula [20, 8] (for more details about these techniques, we refer to [13, 18]). As a consequence of these assumptions, given x and y real vectors of dimension n, we have

$$fl(x^T y) = x^T y + x^T D y + s, \quad |D| \le 3\varepsilon I, \quad |s| \le \mathcal{O}(n\varepsilon^2 |x|^T |y|).$$

Moreover, given an $n \times m$ matrix B of entries B_{ij} and an n-vector v of entries v_i, we will denote by $\|B\|_2$ the usual spectral norm for a matrix, by $\|v\|_2 = (\sum_{i=1}^n v_i^2)^{1/2}$ the usual 2-norm of a vector, by $\kappa_B = \|B^+\|_2 \|B\|_2$ the 2-norm condition number of the matrix B and, finally, by $\|B\|_F = (\sum_{i=1}^n \sum_{j=1}^m B_{ij}^2)^{1/2}$ the Frobenius norm. Finally for simplicity we will denote by c_i, $i = 0, 1, \ldots$, the constants that will be used in the expressions of the roundoff errors. Their values will be integers less than 174.

2. Algorithms. In this section, we take into account the Null Space classical algorithm, which is described in [12], for the solution of (1.1).

We choose a formulation of this algorithm which is based on the factorization of the augmented matrix, and in the next section we will show how this formulation enables us to give a roundoff error analysis of the algorithm.

The matrix A can be factorized in the following way:

$$A = H \begin{bmatrix} R \\ 0 \end{bmatrix}, \tag{2.1}$$

where $R \in \mathbb{R}^{m \times m}$ is an upper triangular matrix, and $H \in \mathbb{R}^{n \times n}$ is the orthonormal matrix generated by the Householder algorithm applied to A [14](see also [22, 34, 15]). The matrix H is the product of m elementary unitary transformation matrices P_k,

$k = 1, \ldots, m$. Let e_k be the kth column of the unit matrix, and let d_1, d_2 be the subvectors of $d \in \mathbb{R}^n$ made by the first $k-1$ entries of d and by the last $n-k+1$ entries of d, respectively. The Householder unitary transformation $P_k d = \rho \left[\begin{smallmatrix} d_1 \\ e_1 \end{smallmatrix} \right]$ uses $P_k = I - 2ww^T/\|w\|_2^2$, $w = d - e_k\rho$, $\rho = \text{sign}(d_k)\|d_2\|_2$.

Let

$$\tilde{M} = H^T M H.$$

The augmented matrix $\mathfrak{A}$ can be factorized in the following way:

$$\mathfrak{A} = \begin{bmatrix} H & 0 \\ 0 & R^T \end{bmatrix} \begin{bmatrix} \tilde{M}_{11} & \tilde{M}_{12} & I_m \\ \tilde{M}_{12}^T & \tilde{M}_{22} & 0 \\ I_m & 0 & 0 \end{bmatrix} \begin{bmatrix} H^T & 0 \\ 0 & R \end{bmatrix}.$$

Let Y and Z be, respectively, HE_1 and HE_2. The matrix Y is the basis of the range of A and Z is the basis of the kernel of A^T.

On the basis of the previous discussion, we obtain the following algorithm.

Null Space algorithm.

$$\begin{bmatrix} h \\ v \end{bmatrix} = \begin{bmatrix} -H^T q \\ R^{-T} b \end{bmatrix},$$
$$h_1 = E_1^T h,$$
$$h_2 = E_2^T h.$$

Solve the block lower triangular system:

$$\begin{bmatrix} I_m & 0 & 0 \\ \tilde{M}_{12}^T & \tilde{M}_{22} & 0 \\ \tilde{M}_{11} & \tilde{M}_{12} & I_m \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} v \\ h_2 \\ h_1 \end{bmatrix}, \tag{2.2}$$

and let

$$x = H \begin{bmatrix} z_1 \\ z_2 \end{bmatrix},$$
$$u = R^{-1} z_3.$$

It is quite obvious that the product $H^T M H$ cannot be performed directly because the complexity would be too high ($\mathcal{O}(n^3)$) and because the resulting matrix would be fairly dense despite the sparsity of M. Nevertheless, for solving the triangular system, we can perform the product of a submatrix of $\tilde{M}$ by a vector in the following way:

$$\tilde{M}_{ij} z = E_i^T H^T (M(HE_j z)), \quad \text{with } (i,j) = (1,2). \tag{2.3}$$

This approach has the advantage of performing the product of H by a vector using the elementary matrices P_j computed during the Householder process. Even though the matrix H may be structurally full, the single P_j matrices relative to A can be quite sparse (see [11, 23, 29]). Thus, for a special class of matrices, the computational complexity of the product H times a vector and the complexity of $H^T M H z$ can be reduced to $\mathcal{O}(n \log(n))$.

In the Null Space algorithm, we also need to solve $\tilde{M}_{22} z = p$. We have two alternative ways to proceed. If $n - m$ is small (the number of constraints is very close

to the number of unknowns), we can explicitly compute $\tilde{M}_{22}$ using the algorithm (2.3) and then solve the system using the Cholesky factorization. Otherwise, we can solve the linear system $\tilde{M}_{22}z = p$ using a conjugate gradient algorithm (see [16]) without explicitly computing $\tilde{M}_{22}$ and using (2.3) to perform the matrix by vector products required.

Finally, in section 4 we use the following properties. If we denote by $x_{min} = HE_1R^{-T}b$ the minimum norm solution of $A^Ty = b$, by $A^+ = R^{-1}E_1^TH^T$ the Moore-Penrose Pseudoinverse of A [22], and by $\mathcal{P} = I - MZ\tilde{M}_{22}^{-1}Z^T$ the oblique projection onto $span(Y)$ along $span(MZ)$ [21, 31], it is straightforward to verify that

$$\mathfrak{A}^{-1} = \begin{bmatrix} Z\tilde{M}_{22}^{-1}Z^T & \mathcal{P}^TA^{+T} \\ A^+\mathcal{P} & -A^+\mathcal{P}MA^{+T} \end{bmatrix} \tag{2.4}$$

and

$$\begin{bmatrix} x \\ u \end{bmatrix} = \begin{bmatrix} -Z\tilde{M}_{22}^{-1}Z^T(Mx_{min}+q) + x_{min} \\ -A^+(Mx+q) \end{bmatrix}.$$

3. Roundoff error analysis. The roundoff properties of the Householder matrices are very favorable. In [34] it is shown that the computed version of each P_k is very close to the true one:

$$\|fl(P_k) - P_k\|_2 \le c_0\varepsilon + \mathcal{O}(\varepsilon^2)$$

(see also [22, 15, 4]). Hereafter, we will follow the notations used in [22].

Moreover, the computed updates of a vector y with $fl(P_k)$ are close to the exact updates with P_k:

$$\begin{aligned} fl(fl(P_k)y) &= P_k(y+w), \quad \|w\|_2 \le c_1\varepsilon\|y\|_2 + \mathcal{O}(\varepsilon^2), \\ fl(y^Tfl(P_k)) &= (y+w)^TP_k, \quad \|w\|_2 \le c_1\varepsilon\|y\|_2 + \mathcal{O}(\varepsilon^2), \end{aligned}$$

and, in general,

$$\begin{aligned} fl(fl(P_1)\cdots fl(P_j)y) &= P_1\cdots P_j(y+z), \quad \|z\|_2 \le c_1j\varepsilon\|y\|_2 + \mathcal{O}(\varepsilon^2), \\ fl(fl(P_j)\cdots fl(P_1)y) &= P_j\cdots P_1(y+\bar{z}), \quad \|\bar{z}\|_2 \le c_1j\varepsilon\|y\|_2 + \mathcal{O}(\varepsilon^2). \end{aligned} \tag{3.1}$$

In [22] it is proved that, in the floating point computation of the Householder algorithm to the matrix A, there exists an orthonormal matrix $\hat{H}$ and a matrix $F \in \mathbb{R}^{n\times m}$ such that the computed value $\overline{R}$ of the upper triangular matrix R satisfies the equation

$$(A+F) = \hat{H}\begin{bmatrix} \overline{R} \\ 0 \end{bmatrix}, \tag{3.2}$$

with $\|F\|_F \le c_1m\varepsilon\|A\|_F$. We now use the previous properties of P_k to analyze the roundoff in the algorithm (2.3) for the matrix vector product. First of all, we have [15]

$$fl(Mx) = (M+G)x, \quad \text{with } |G| \le 3\varepsilon|M|.$$

Therefore, it follows that for $i, j = 1, 2$,

$$\begin{aligned} fl(\tilde{M}_{ij}z) &= E_i^Tfl(H^Tfl(Mfl(HE_jz))) \\ &= E_i^T\hat{H}^T((M+G)\hat{H}(E_jz+f_1)+f_2), \end{aligned} \tag{3.3}$$

with

$$\|f_1\|_2 \le c_1m\varepsilon\|z\|_2, \qquad \|f_2\|_2 \le c_1m\varepsilon\|M\|_2\|z\|_2.$$

If we introduce in (3.3) the expression

$$\frac{z^T E_j^T \hat{H}^T \hat{H} E_j z}{\|z\|_2^2} = 1,$$

we obtain

$$fl(\tilde{M}_{ij} z) = E_i^T \hat{H}^T (M + \mathcal{G}^{ij}) \hat{H} E_j z, \tag{3.4}$$

where

$$\begin{aligned}
\mathcal{G}^{ij} &= G^{ij} + G_1^{ij} + G_2^{ij} + G_3^{ij}, \\
G_1^{ij} &= \frac{M f_1 z^T E_j^T \hat{H}^T}{\|z\|_2^2}, \quad \|G_1^{ij}\|_2 \le c_1 m \varepsilon \|M\|_2 + \mathcal{O}(\varepsilon^2), \\
G_2^{ij} &= \frac{G f_1 z^T E_j^T \hat{H}^T}{\|z\|_2^2}, \quad \|G_2^{ij}\|_F \le \mathcal{O}(\varepsilon^2), \\
G_3^{ij} &= \frac{f_2 z^T E_j^T \hat{H}^T}{\|z\|_2^2}, \quad \|G_3^{ij}\|_2 \le c_1 m \varepsilon \|M\|_2 + \mathcal{O}(\varepsilon^2), \\
&\quad \|G^{ij}\|_2 \le 3\sqrt{n}\varepsilon \|M\|_2 + \mathcal{O}(\varepsilon^2),
\end{aligned}$$

and

$$\|\mathcal{G}^{ij}\|_2 \le (c_3 m + 3\sqrt{n})\varepsilon \|M\|_2 + \mathcal{O}(\varepsilon^2). \tag{3.5}$$

If we can afford both the complexity and the storage necessary to the explicit computation of $\tilde{M}_{22}$, then we can invert the diagonal block in (2.2) by the Cholesky factorization algorithm, which computes a lower triangular matrix L such that $\tilde{M}_{22} = LL^T$. Thus, following the analysis given by Wilkinson [34] and from (3.4), we prove that $\bar{L}$, the computed value of L, and $fl(\tilde{M}_{22})$, the computed value of $\tilde{M}_{22}$, satisfy the following relations:

$$\begin{aligned}
fl(\tilde{M}_{22}) &= E_2^T \hat{H}^T M \hat{H} E_2 + \mathcal{E}, \\
&\quad \|\mathcal{E}\|_F \le \sqrt{(n-m)}(c_3 m + 3\sqrt{n})\varepsilon \|M\|_2 + \mathcal{O}(\varepsilon^2),
\end{aligned} \tag{3.6}$$

$$\begin{aligned}
\bar{L}\bar{L}^T &= E_2^T \hat{H}^T M \hat{H} E_2 + E_2^T \hat{H}^T \hat{H} E_2 (G_4 + \mathcal{E}) E_2^T \hat{H}^T \hat{H} E_2 \\
&= E_2^T \hat{H}^T (M + \overline{\mathcal{G}^{22}}) \hat{H} E_2,
\end{aligned} \tag{3.7}$$

with

$$\|G_4\|_2 \le (n-m)\varepsilon \|M\|_2, \tag{3.8}$$

$$\|\overline{\mathcal{G}^{22}}\|_F \le \big(\sqrt{(n-m)}(c_3 m + 3\sqrt{n}) + (n-m)\big)\varepsilon \|M\|_2 + \mathcal{O}(\varepsilon^2). \tag{3.9}$$

Alternatively, we can solve the system $\tilde{M}_{22} z_2 = h_2 - \tilde{M}_{12}^T z_1$ using a conjugate gradient method, which incorporates a stopping criterion based on the Rigal–Gaches a posteriori normwise backward error theory [30, 18].

A stopping criterion such as

$$\text{IF} \quad \|\tilde{M}_{22}\tilde{z}_2 - fl(h_2 - \tilde{M}_{12}^T z_1)\|_2 \le \eta \|\tilde{M}_{22}\|_2 \|\tilde{z}_2\|_2, \quad \text{THEN STOP}, \tag{3.10}$$

with $\eta < 1$ an a priori threshold fixed by the user, will guarantee that the computed solution $\tilde{z}_2$ satisfies the perturbed linear system

$$(\tilde{M}_{22} + E_{22})\tilde{z}_2 = fl(h_2 - \tilde{M}_{12}^T z_1), \quad \|E_{22}\|_2 \le \eta \|\tilde{M}_{22}\|_2.$$

Moreover, we can project the error on the null space such that we have

$$\big(E_2^T\hat{H}^T(M+\mathcal{E}_1)\hat{H}E_2\big)\tilde{z}_2 = fl\big(h_2 - \tilde{M}_{12}^T z_1\big), \tag{3.11}$$

with

$$\mathcal{E}_1 = \hat{H}E_2E_{22}E_2^T\hat{H}^T, \quad \|\mathcal{E}_1\|_2 \le \eta(\|M\|_2 + \mathcal{O}(\varepsilon)). \tag{3.12}$$

The choice of η will depend on the properties of the problem that we want to solve, and, in the practical cases, η can be frequently much larger than ε.

Using the previous results, we can now prove the following theorem.

THEOREM 3.1. *Let $\bar{x}$ and $\bar{u}$ be the values of x and u, solutions of* (2)*, computed with the* Null Space algorithm. *If $n\varepsilon \ll 1$, then there exist the matrices δM_1, $\delta M_2 \in \mathbb{R}^{n\times n}$, and F_1, $F_2 \in \mathbb{R}^{n\times m}$, and the vector $\delta q \in \mathbb{R}^n$ such that*

$$\begin{bmatrix} M+\delta M_1+\delta M_2 & A+\delta A_1 \\ (A+\delta A_2)^T & 0 \end{bmatrix}\begin{bmatrix}\bar{x}\\ \bar{u}\end{bmatrix} = \begin{bmatrix}-(q+\delta q)\\ b\end{bmatrix}.$$

Furthermore, we have

$$\begin{aligned}\|\delta A_1\|_F &\le c_2 m\varepsilon\|A\|_F + \mathcal{O}(\varepsilon^2),\\ \|\delta A_2\|_F &\le c_2 m\varepsilon\|A\|_F + \mathcal{O}(\varepsilon^2),\\ \|\delta q\|_2 &\le c_1\varepsilon\|q\|_2 + \mathcal{O}(\varepsilon^2),\end{aligned}$$

and, if we use an iterative solver with (3.10) *and a threshold η,*

$$\begin{aligned}&\|\delta M_1\|_F \le 4(\sqrt{(n-m)}(c_3 m + 3\sqrt{n}) + (n-m))\varepsilon\|M\|_2 + \mathcal{O}(\varepsilon^2),\\ &\|\delta M_2\|_F \le \eta(\|\tilde{M}_{22}\|_2 + \mathcal{O}(\varepsilon)).\end{aligned}$$

Otherwise, if we build $\tilde{M}_{22}$ then factorize and solve the linear system by Cholesky factorization, we have $\delta M_2 = 0$ and

$$\|\delta M_1\|_F \le 5\big(\sqrt{(n-m)}(c_3 m + 3\sqrt{n}) + (n-m)\big)\varepsilon\|M\|_2 + \mathcal{O}(\varepsilon^2).$$

Proof. From [18, Theorem 8.5, p. 154] and (3.1) it follows that the computed values $\bar{x}$ and $\bar{u}$ satisfy the equations

$$(\bar{R}+F_1)\bar{u} = \bar{z}_3, \qquad \bar{x} = \hat{H}\begin{bmatrix}\bar{z}_1+\delta z_1\\ \bar{z}_2+\delta z_2\end{bmatrix}, \tag{3.13}$$

where

$$\begin{aligned}\|F_1\|_F &\le m\varepsilon\|\bar{R}\|_F,\\ \left\|\begin{bmatrix}\delta z_1\\ \delta z_2\end{bmatrix}\right\|_2 &\le c_1 m\varepsilon\left\|\begin{bmatrix}\bar{z}_1\\ \bar{z}_2\end{bmatrix}\right\|_2.\end{aligned} \tag{3.14}$$

Writing

$$\hat{x} = \hat{H}\begin{bmatrix}\bar{z}_1\\ \bar{z}_2\end{bmatrix}, \qquad \delta x = \hat{H}\begin{bmatrix}\delta z_1\\ \delta z_2\end{bmatrix},$$

we have $\overline{x} = \hat{x} + \delta x$ and

$$\|\delta x\|_2 \le m\varepsilon\|\hat{x}\|_2 \le \frac{n\varepsilon}{1-n\varepsilon}\|\overline{x}\|_2.$$

Moreover, the vector $\bar{z}$ satisfies the system

$$\begin{bmatrix} \widehat{M}_{11}+E_{11} & \widehat{M}_{12}+E_{12} & I_m \\ \widehat{M}_{12}^T+E_{21} & \widehat{M}_{22}+E_{22} & 0 \\ I_m & 0 & 0 \end{bmatrix} \begin{bmatrix} \bar{z}_1 \\ \bar{z}_2 \\ \bar{z}_3 \end{bmatrix} = \begin{bmatrix} \bar{h}_1 \\ \bar{h}_2 \\ \bar{v} \end{bmatrix}, \tag{3.15}$$

where $\bar{h} = fl(h)$ and $\bar{v} = fl(v)$, $\widehat{M} = \hat{H}^T M \hat{H}$ and, because of (3.4),

$$\|E_{ij}\|_2 \le (c_3 m + 3\sqrt{n})\varepsilon \|M\|_2 + \mathcal{O}(\varepsilon^2), \quad (i,j) \ne (2,2).$$

If we choose to invert the diagonal block $\tilde{M}_{22}$ by the Cholesky factorization, the $\|E_{22}\|_F$ can be bounded using (3.6), (3.7), (3.8), and (3.9) as follows:

$$\|E_{22}\|_F \le \big(\sqrt{(n-m)}(c_3 m + 3\sqrt{n}) + 3(n-m)\big)\varepsilon \|M\|_F + \mathcal{O}(\varepsilon^2). \tag{3.16}$$

Alternatively, using an iterative solver with (3.10), we deduce from (3.11) and (3.12) that

$$\|E_{22}\|_2 \le \eta(\|M\|_2 + \mathcal{O}(\varepsilon)). \tag{3.17}$$

Once more, from [18, Theorem 8.5, p. 154] and (3.1) it follows that the computed values $\bar{h}$ and $\bar{v}$ satisfy

$$(\bar{R} + F_2)^T \bar{v} = b, \qquad \bar{h} = -\hat{H}^T(q + \delta q), \tag{3.18}$$

where

$$\|F_2\|_F \le m\varepsilon \|\bar{R}\|_F, \qquad \|\delta q\|_2 \le c_1 m \varepsilon \|q\|_2 + \mathcal{O}(\varepsilon^2). \tag{3.19}$$

Finally, combining (3.18), (3.15), and (3.13) with (3.2), we have

$$\begin{bmatrix} M+\delta M_3 & A+\delta A_1 \\ (A+\delta A_3)^T & 0 \end{bmatrix} \begin{bmatrix} \bar{x}-\delta x \\ \bar{u} \end{bmatrix} = \begin{bmatrix} -(q+\delta q) \\ b \end{bmatrix},$$

where

$$\delta M_3 = \sum_{i,j=1,2} \hat{H} E_i E_{ij} E_j^T \hat{H}^T$$

and

$$\delta A_1 = F + F_1, \qquad \delta A_3 = F + F_2.$$

Furthermore, because we have

$$\bar{x} - \delta x = \left(I - \frac{\delta x \bar{x}^T}{\bar{x}^T \bar{x}}\right)\bar{x},$$

it follows that

$$\begin{bmatrix} M+\delta M & A+\delta A_1 \\ (A+\delta A_2)^T & 0 \end{bmatrix} \begin{bmatrix} \bar{x} \\ \bar{u} \end{bmatrix} = \begin{bmatrix} -(q+\delta q) \\ b \end{bmatrix},$$

where

$$\delta M = \delta M_3 \left(I - \frac{\delta x \overline{x}^T}{\overline{x}^T \overline{x}} \right) - M \frac{\delta x \overline{x}^T}{\overline{x}^T \overline{x}}$$

and

$$\delta A_2^T = \delta A_3^T \left(I - \frac{\delta x \overline{x}^T}{\overline{x}^T \overline{x}} \right) - A^T \frac{\delta x \overline{x}^T}{\overline{x}^T \overline{x}}.$$

If we use an iterative solver, the matrix δM can be split into two contributions: $\delta M = \delta M_1 + \delta M_2$, where

$$\delta M_2 = E_2 E_{22} E_2^T \left(I - \frac{\delta x \overline{x}^T}{\overline{x}^T \overline{x}} \right).$$

We can now evaluate the norms of the perturbations by the use of (3.2), (3.5), (3.14), (3.19), and, alternatively, either (3.16) or (3.17):

$$\begin{aligned} \|\delta A_1\|_F &\le c_2 m \varepsilon \|A\|_F + \mathcal{O}(\varepsilon^2), \\ \|\delta A_2\|_F &\le c_2 m \varepsilon \|A\|_F + \mathcal{O}(\varepsilon^2), \\ \|\delta q\|_2 &\le c_1 m \varepsilon \|q\|_2 + \mathcal{O}(\varepsilon^2), \end{aligned}$$

and either

$$\begin{aligned} \|\delta M_1\|_F &\le \big(\|E_{11}\|_F + \|E_{21}\|_F + \|E_{12}\|_F + \|E_{22}\|_F\big)\Big(1 + \frac{n\varepsilon}{1-n\varepsilon}\Big) + \frac{n\varepsilon}{1-n\varepsilon}\|M\|_2 \\ &\le 5\big(\sqrt{(n-m)}(c_3 m + 3\sqrt{n}) + (n-m)\big)\varepsilon \|M\|_2 + \mathcal{O}(\varepsilon^2) \\ \|\delta M_2\|_F &= 0 \end{aligned}$$

or

$$\begin{aligned} \|\delta M_1\|_F &\le (\|E_{11}\|_F + \|E_{21}\|_F + \|E_{12}\|_F)\Big(1 + \frac{n\varepsilon}{1-n\varepsilon}\Big) + \frac{n\varepsilon}{1-n\varepsilon}\|M\|_2 \\ &\le 4\big(\sqrt{(n-m)}(c_3 m + 3\sqrt{n}) + (n-m)\big)\varepsilon \|M\|_2 + \mathcal{O}(\varepsilon^2) \\ \|\delta M_2\|_F &\le \eta(\|M\|_F + \mathcal{O}(\varepsilon)). \qquad \square \end{aligned}$$

4. Forward error. Following the results of section 3, we are able to represent the errors $\delta x = \overline{x} - x$ and $\delta u = \overline{u} - u$ as follows:

$$\begin{bmatrix} \delta x \\ \delta u \end{bmatrix} = \mathfrak{A}^{-1} \begin{bmatrix} -\delta q - \delta A_1 \overline{u} - (\delta M_1 + \delta M_2)\overline{x} \\ -\delta A_2^T \overline{x} \end{bmatrix}.$$

Let us denote:

$$\begin{aligned} \kappa_A &= \|A^+\|_2 \|A\|_2, \\ \mathfrak{K}_A &= \|A^+ \mathcal{P}\|_2 \|A\|_2, \\ \kappa_{\tilde{M}_{22}} &= \|\tilde{M}_{22}^{-1}\|_2 \|\tilde{M}_{22}\|_2, \\ \mathfrak{C}_M &= \|\tilde{M}_{22}^{-1}\|_2 \|M\|_2. \end{aligned}$$

Using the block form of the inverse of $\mathfrak{A}$ (2.4), the results of Theorem 3.1, and denoting $c_2 = 4(\sqrt{(n-m)}(c_3 m + 3\sqrt{n}) + (n-m))$, we obtain the following bounds:

$$\begin{aligned}\frac{\|\delta x\|_2}{\|\overline{x}\|_2} &\le c_2\varepsilon\left[\mathfrak{C}_M\left(1+\frac{\|q\|_2+\|A\|_2\|\overline{u}\|_2}{\|M\|_2\|\overline{x}\|_2}\right)+\mathfrak{K}_A\right]\\ &\quad+\eta\kappa_{\tilde{M}_{22}}+\mathcal{O}((\eta+\varepsilon)\varepsilon),\\ \frac{\|\delta u\|_2}{\|\overline{u}\|_2} &\le c_2\varepsilon\left[\mathfrak{K}_A\left(1+\frac{\|q\|_2+\|M\|_2\|\overline{x}\|_2}{\|A\|_2\|\overline{u}\|_2}\right)+\mathfrak{K}_A\kappa_A\frac{\|M\|_2\|\overline{x}\|_2}{\|A\|_2\|\overline{u}\|_2}\right]\\ &\quad+\eta\mathfrak{K}_A\frac{\|M\|_2\|\overline{x}\|_2}{\|A\|_2\|\overline{u}\|_2}+\mathcal{O}((\eta+\varepsilon)\varepsilon).\end{aligned}$$

We can see from these bounds that the solution x is less sensitive than the Lagrangian parameters u to the perturbations in the data. Moreover, the dependence of δu on the square of the condition number $\mathfrak{K}_A$ is in agreement with the analogous results that we have for the least squares problems (see [14, 7, 9]).

The condition numbers $\kappa_{\tilde{M}_{22}}$ and $\mathfrak{C}_M$ are bounded as follows:

$$\kappa_{\tilde{M}_{22}} \le \mathfrak{C}_M \le \|M\|_2\|M^{-1}\|_2 = \kappa_M.$$

Normally, the value of $\kappa_{\tilde{M}_{22}}$ is very much smaller than the value of κ_M (see [32]); nevertheless, they can coincide if the $\mathrm{Ker}(A^T)$ contains the invariant spaces of M corresponding to the smallest and to the largest eigenvalues. Thus the ratio $\kappa_{\tilde{M}_{22}}/\kappa_M$ is the measure of how much the initial hypothesis $\mathrm{Ker}(A^T)\cap\mathrm{Ker}(M) = \{0\}$ is satisfied.

From the following properties of $\mathcal{P}$,

$$\mathcal{P}^2 = \mathcal{P}, \quad \mathrm{Range}(\mathcal{P}) = \mathrm{Range}(A) = \mathrm{Range}(Y), \quad \mathrm{Ker}(\mathcal{P}) = \mathrm{Range}(MZ),$$

we can straightforwardly prove that

$$\kappa_A \le \mathfrak{K}_A \le \kappa_A\|\mathcal{P}\|_2.$$

The spectral norm of an idempotent matrix (projector) $\mathcal{P}$ can be exactly computed (see [33]).

THEOREM 4.1. *If $\mathcal{P} \in \mathbb{R}^{n\times n}$ is a general idempotent matrix, then*

$$\|I-\mathcal{P}\|_2 = \|\mathcal{P}\|_2 = \frac{1}{c},$$

with c being the cosine of the largest principal angle α between $\mathrm{Range}(\mathcal{P})$ *and* $\mathrm{Ker}(\mathcal{P})^\perp$ *(the angle α is also the largest principal angle between* $\mathrm{Range}(\mathcal{P})^\perp$ *and* $\mathrm{Ker}(\mathcal{P})$*).*

It is important to observe that

$$\frac{1}{c} < \sqrt{\kappa_M}$$

and that the two quantities are close if the invariant space of M corresponding to the smallest eigenvalue is very close to the space generated by Z.

5. Numerical experiments. In our numerical experiments we used an augmented system obtained by a mixed finite element discretization of a magnetostatic problem. Hereafter, we give a short description of the physical problem and its discretization; more details can be found in [27] and in [28].

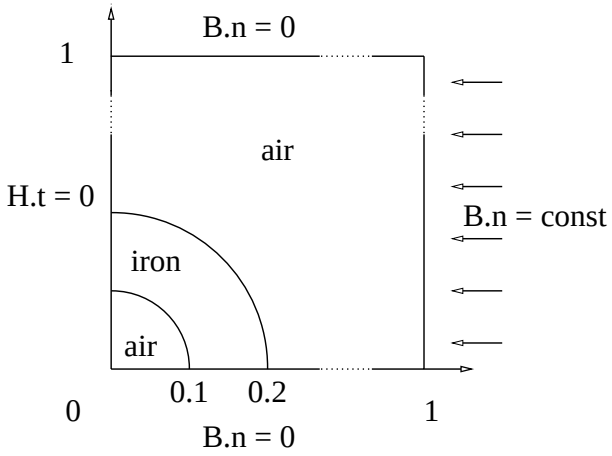


FIG. 5.1. *Computational domain.*

5.1. Model problem. The solution of the three-dimensional magnetostatic problem consists of finding the magnetic displacement field $\underline{B}$ and the magnetic field $\underline{H}$ which satisfy the physical conditions expressed by Maxwell's equations

$$\begin{aligned} &\text{i)} \quad \operatorname{div} \underline{B} = 0, \\ &\text{ii)} \quad \operatorname{curl} \underline{H} = \underline{J}, \end{aligned} \tag{5.1}$$

where $\underline{J}$ is the imposed current density. The relation between $\underline{B}$ and $\underline{H}$ in a material with magnetic permeability $\mu = \mu_0 \mu_r$ is expressed by means of the constitutive law

$$\underline{B} = \mu \underline{H}. \tag{5.2}$$

We consider a convex polyhedron Ω of $\mathbb{R}^3$ as the computational domain, with an outward normal unit vector $\underline{n}$ and a boundary $\Gamma = \Gamma_B \cup \Gamma_H$. We shall assume that $\Gamma_H \neq \{\emptyset\}$. Using a suitable functional norm, we have chosen the following formulation of the problem:

$$\min_{\substack{\operatorname{div} \underline{B}=0 \\ \operatorname{curl} \underline{H}=\underline{J}}} \frac{1}{2} \|\underline{B} - \mu \underline{H}\|^2 \quad \text{in} \quad \Omega,$$

$$\text{with} \quad \underline{B} \cdot \underline{n} = g_1 \quad \text{on} \quad \Gamma_B, \qquad \underline{H} \wedge \underline{n} = \underline{g}_2 \quad \text{on} \quad \Gamma_H.$$

In our experiments we consider a cylindrical magnetostatic screen which consists of a hollow iron cylinder of infinite length having external and internal radii equal to 0.2 m and 0.1 m, respectively. The screen is placed in a uniform induction field of prescribed value, orthogonal to the cylinder axis. Since the field does not vary in the axial direction, only a cross-section of the screen needs be considered, and then we can reduce the analysis of the three-dimensional problem to the analysis of a two-dimensional one. Moreover, owing to the symmetry of the problem, we may restrict the analysis to a quarter of the device surrounded by an air region of suitable extension. Thus, we used the square $[0,1] \times [0,1]$ as the computational domain, with $\Gamma_H = \{0\} \times [0,1]$ and $\Gamma_B = \{[0,1] \times \{0\}\} \cup \{[0,1] \times \{1\}\} \cup \{\{1\} \times [0,1]\}$ (see Figure 5.1).

The constraint $\operatorname{curl} \underline{H} = \underline{J}$ is replaced by $\operatorname{rot} \underline{H} = j$, where $\operatorname{rot} \underline{H} = -\frac{\partial H_1}{\partial y} + \frac{\partial H_2}{\partial x}$ and j is the third component of $\underline{J}$. The boundary condition on $\underline{H}$ is then $\underline{H} \cdot \underline{t} = g_2$ on Γ_H, where $\underline{t}$ is the tangential unit vector (couterclockwise oriented). In our experiments, we have chosen $g_2 = 0$ and $g_1 = 0$ on $\{[0,1] \times \{0\}\} \cup \{[0,1] \times \{1\}\}$ and $g_1 =$ constant on $\{1\} \times [0,1]$. We then give a mixed variational formulation of the problem, introducing the Lagrangian multipliers ξ and ϕ. The problem is discretized by the use of the standard and "rotated" Raviart–Thomas element spaces. We refer to [28, 27] for a complete analysis.

We would like to point out that in the test problem, the physical laws (5.1) have to be satisfied with a larger accuracy compared to that requested for the constitutive law (5.2). Moreover, (5.2) is merely a first-order approximation of a more complex nonlinear condition where the experimental parameter μ becomes a nonlinear function of H.

5.2. Algebraic problem. The algebraic linear system, arising in the discretization of the two-dimensional problem, has the following structure

$$\left[\begin{array}{cc|cc} M_{11} & M_{12} & A_1 & 0 \\ M_{12}^T & M_{22} & 0 & A_2 \\ \hline A_1^T & 0 & 0 & 0 \\ 0 & A_2^T & 0 & 0 \end{array}\right] \left[\begin{array}{c} \mathfrak{b} \\ \mathfrak{h} \\ \xi \\ \varphi \end{array}\right] = \left[\begin{array}{c} -q_1 \\ -q_2 \\ b_1 \\ b_2 \end{array}\right].$$

For a grid of NT triangular elements and NE edges, the system size is 2(NE+NT) (n = 2NT and m = 2NE), with NE unknowns in $\mathfrak{b}$ and $\mathfrak{h}$ and NT unknowns in ξ and φ.

The block M is positive definite and A is full-rank. The submatrices A_1^T and A_2^T applied to the vectors $\mathfrak{b}$ and $\mathfrak{h}$ give the action of div and rot on the finite dimensional space functions. Thus the corresponding residual entries must be satisfied as accurately as possible compared with those of the remaining equations where the constitutive law of the material is confined. Therefore, the use of an iterative solver for the reduce system is appropriate: In the stopping criterion (3.10), we can choose for η a value which is related to the experimental errors that take place when determining the parameter μ and then the accuracy of the constitutive law (5.2). Theorem 3.1 guarantees that the norms of the residuals corresponding to A_1^T and A_2^T are bounded by machine precision times the norms of the matrices times the norms of the solution vectors. Moreover, it guarantees that the remaining residuals corresponding to the constitutive law are bounded by η times the norm of M.

The condition number of the augmented matrix grows with the dimension of the system or, equivalently, with the inverse of the meshsize. In its original form, the coefficient matrix is extremely ill scaled because of the presence in the variational formulation of the factors μ and μ^{-1}, where $\mu = \mu_0 \mu_r$ with $\mu_0 = 4\pi 10^{-7}$, and the magnitude of μ_r can vary from 10 to 10^4 (see [27]). Stability considerations thus call for a diagonal prescaling procedure before any solver is used, so as to mitigate the effect of the magnetic permeability. In view of the structure of the augmented matrix $\mathfrak{A}$, the symmetric diagonal scaling can easily be constructed as follows. Let $D_1 = \operatorname{diag}(M_{11})^{-\frac{1}{2}}$, $D_2 = \operatorname{diag}(M_{22})^{-\frac{1}{2}}$, and D_3 and D_4 be the diagonal matrices of size NT such that $(D_3)_{ii} = \max_j |(D_1 A_1)_{ji}|$ and $(D_4)_{ii} = \max_j |(D_2 A_2)_{ji}|$. The scaling matrix is given by

$$D = \operatorname{diag}\left[D_1, D_2, D_3^{-1}, D_4^{-1}\right].$$

The resulting system is of the type $D\mathfrak{A}Dy = D\left[\begin{smallmatrix} -q \\ b \end{smallmatrix}\right]$, with $\left[\begin{smallmatrix} x \\ u \end{smallmatrix}\right] = Dy$.

5.3. Multifrontal approach. All our experiments were performed on a SUN workstation using Matlab. We performed the QR factorization using the Matlab version of the SQR subroutine produced by P. Matstoms [26] and modified by M. Adlers to cope with the storage of the frontal orthogonal matrices Q_i, the product of which gives the matrix H of the Householder factorization. The QR algorithm implemented in the SQR subroutine is a multifrontal approach to the Householder factorization of sparse matrices: In the symbolic part, the code computes the column ordering of A following the minimum degree reordering of A^TA in order to minimize the fill-in in the upper triangular matrix R, and then a row ordering making A ordered by the leading entries.

Since the matrix A is block diagonal, we applied the multifrontal separately to each diagonal block $A_k \in \mathbb{R}^{n_k \times m_k}$, $k = 1, 2$. Moreover, because each diagonal block satisfies the Strong Hall property [5], the structure of the pattern of each $A_k^TA_k$ yields to a correct prediction of the pattern of R_k.

For each diagonal block A_k, $k = 1, 2$, the multifrontal algorithm builds the *elimination tree* $T(A_k^TA_k)$ based on the upper triangular Cholesky factor R_k of $A_k^TA_k$:

$$R_k^TR_k = A_k^TA_k.$$

The elimination tree $T(A_k^TA_k)$ is a rooted tree with n_k nodes labeled from 1 to n_k, where node p is the *parent* of node j if and only if

$$p = \min_j\{j > i : (R_k)_{ij} \neq 0\}.$$

With each node j of the elimination tree there is associated a frontal matrix $\mathcal{F}_j$, which consists of the set of rows in A_k with the first nonzero entry in location j, together with one updated matrix contributed by each child node of j. The frontal matrix $\mathcal{F}_j$ is stored in a small dense matrix which is partially factorized by the Householder algorithm:

$$\mathcal{F}_j = Q_jU_j.$$

The first row of U_j will be the jth row of R_k, and the remaining rows will contribute to form the updating matrix of the parent node of j.

We can store Q_j either as a full matrix or through the Householder vectors; we will report in the next section on the storage needed by each choice. In [23] it is proved that for matrices having the $\sqrt{n}$-separator property the storage needed for the Householder vectors is of order $\mathcal{O}(n \log n)$, and we have experimentally observed that our test matrices satisfy this property.

5.4. Numerical results. We have considered three different discretizations, by 408, 748, and 1800 triangles, of the model problem with discontinuity of the magnetic permeability μ_r at the air-iron border of 10^3.

For the sake of comprehensiveness and for later use we report the sparsity of all resulting matrices: the 2088×2088 matrix has 15480 nonzeros, the 3810×3810 matrix has 28542 nonzeros, and, finally, the 9102×9102 matrix has 69282 nonzero entries. Hereafter, we will denote by $nz(B)$ the number of nonzero entries in a matrix B. In Table 5.1, we report on the values of $nz(R) = nz(R_1) + nz(R_2)$, $nz_Q = \sum_i nz(Q_i)$, nz_H being the number of entries of the Householder vectors needed for storing the Q_i, and of $nz(U)$, where U is the upper triangular factor of the Cholesky factorization of M. In Table 5.1, we denote by NT and NE, respectively, the number of triangles

TABLE 5.1
Number of floating point entries needed by the Householder factorization of A.

NT	NE	$nz(R)$	nz_Q	nz_H	$nz(U)$
408	636	5971	37172	23694	23704
748	1157	12865	80151	52073	47009
1800	2751	40439	258310	173083	135872

and the number of edges involved in the triangularization of the domain. In our test problems, even if we use minimum degree reordering, nz_H always satisfies the theoretical value of $\mathcal{O}(n \log n)$ indicated in [23] for the nested dissection reordering of a special class of matrices.

The matrix M for the discretization parameter going to zero becomes singular. For the three values of NT, which we chose, the condition number of M increases with NT. Nevertheless, the condition number of the projected scaled Hessian matrix $\kappa_{\tilde{M}_{22}} = \|\tilde{M}_{22}^{-1}\|_2 \|\tilde{M}_{22}\|_2$ is always small ($\kappa_{\tilde{M}_{22}} \leq 1.94$) and independent of the values of NT and NE. On the contrary, the values of κ_A depend on NT and, in particular, for $NT = 408$, the value of κ_A after the scaling is still of order 10^3.

This situation is extremely favorable for the Null Space algorithm because the rate of convergence of the conjugate gradient algorithm becomes independent of NE and constant.

Then we can assume that the complexity of the conjugate gradient algorithm will be of the order of $\mathcal{O}(n \log n)$. More precisely, we can estimate that the conjugate gradient algorithm will take around 20 iterations to reduce the $\tilde{M}_{22}$-norm of the error in the solution to the level of ε by the $\tilde{M}_{22}$-norm of the solution. This conclusion follows from the upper bound of the $\tilde{M}_{22}$-norm of the error [24]:

$$(z - z_{exact})^T \tilde{M}_{22} (z - z_{exact}) \leq \left(\frac{\sqrt{\kappa_{\tilde{M}_{22}}} - 1}{\sqrt{\kappa_{\tilde{M}_{22}}} + 1} \right)^{2k}.$$

6. Conclusions. We have proved the backward stability of the Null Space algorithm for solving augmented systems. Moreover, we have given evidence of the good applicability of the method in the sparse case. However, we must point out that in our test problems it was not necessary to precondition the projected Hessian matrix during the application of the conjugate gradient algorithm because this matrix was already very well conditioned. We reserve for future work the analysis of the preconditioning techniques proposed by Nash and Sofer in [25]. We want to emphasize here that our test problems are significant examples of the more general class of mixed finite element discretization of elliptic problems in saddle point formulation.

Moreover, the method we propose could also be generalized to the solution of nonlinear problems with linear equality constraints, where it would be possible to build a specialized version of the conjugate gradient taking advantage of the sparse QR factorization.

Finally, we want to comment on the alternative approach where the QR factorization is replaced by the LU factorization. In [1, 2] both a roundoff error analysis of the Null Space algorithm is presented (where the QR factorization is substituted by the LU factorization) and a result similar to the one of presented in Theorem 3.1 is proved. In the LU-based algorithm, the matrix A is factorized as follows:

$$A = L \begin{bmatrix} U \\ 0 \end{bmatrix},$$

where $U \in \mathbb{R}^{m\times m}$ is an upper triangular matrix, and $L \in \mathbb{R}^{n\times n}$ is the lower triangular matrix with ones on the main diagonal computed by the Gaussian algorithm with partial pivot. In the matrix $\tilde{M} = H^T M H$ the orthonormal matrix H^T is then substituted by L^{-1} and the matrix $L^{-T}E_2$ is a nonorthogonal basis for the null space. The upper bounds of the perturbations in the backward error depend on the product of the norm of $L^{-1}ML^{-T}$ and the norm of $L^{-T}E_2$.

Because the pivot strategy operates only on the matrix A, it is fairly difficult to control the norm of $L^{-1}ML^{-T}$. Moreover, the norm of the null space basis matrix can also be much larger than 1 (see [18, Chapter 8]). However, in many practical situations this product is usually of the order $n\|M\|_\infty$. In particular, if A is the incidence matrix of a network (i.e., a totally unimodular matrix), the algorithm based on the LU factorization is backward stable and faster than the QR-based algorithm.

Acknowledgments. I would like to thank Dr. Pontus Matstoms, who kindly allowed us to use his Matlab code SQR, and especially Dr. Mikael Adlers, who modified the code to cope with the storage of the frontal orthogonal matrices.

REFERENCES

[1] M. Arioli and L. Baldini, *A Backward Error Analysis of a Null Space Algorithm in Sparse Quadratic Programming*, Tech. report IAN 1150, IAN-CNR, Pavia, Italy, 1999.
[2] L. Baldini, *Propagazione degli errori di arrotondamento in algoritmi per la risoluzione di sistemi aumentati*, Master's thesis, University of Pavia, Pavia, Italy, 1997.
[3] J. L. Barlow, N. K. Nichols, and R. J. Plemmons, *Iterative methods for equality-constrained least squares problems*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 892–906.
[4] Å. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, PA, 1996.
[5] T. F. Coleman, A. Edenbrant, and J. R. Gilbert, *Predicting fill for sparse orthogonal factorization*, J. ACM, 33 (1986), pp. 517–532.
[6] T. F. Coleman and A. Pothen, *The Null space problem.* I. *Complexity*, SIAM J. Algebraic Discrete Methods, 7 (1986), pp. 527–537.
[7] A. J. Cox and N. J. Higham, *Accuracy and stability of the null space method for solving the equality constrained least squares problem*, BIT, 39 (1999), pp. 34–50.
[8] T. J. Dekker, *A floating point technique for extending the available precision*, Numer. Math., 18 (1971), pp. 224–242.
[9] L. Elden, *Perturbation theory for the least squares problem with linear equality constraints*, SIAM J. Numer. Anal., 17 (1980), pp. 338–350.
[10] E. Galligani and L. Zanni, *Error analysis in equality constrained quadratic optimization problems*, Boll. Un. Mat. Ital. A (7), 11 (1997), pp. 595–611.
[11] J. G. Gilbert, E. G. Ng, and B. W. Peyton, *Separators and structure prediction in sparse orthogonal factorization*, Linear Algebra Appl., 262 (1997), pp. 83–97.
[12] P. H. Gill, W. Murray, and M. H. Wright, *Practical Optimization*, Academic Press, London, 1981.
[13] D. Goldberg, *What every computer scientist should know about floating-point arithmetic*, ACM Computing Surveys, 23 (1991), pp. 5–48.
[14] G. H. Golub, *Numerical methods for solving least squares problems*, Numer. Math., 7 (1965), pp. 206–216.
[15] G. H. Golub and C. Van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.
[16] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, PA, 1997.
[17] M. Gulliksson, *Algorithms for Overdetermined System of Equations*, Ph.D. thesis UMINF 93-11, Department of Computing Science, Institute of Information Processing, University of Umeå, Sweden, 1993.
[18] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, PA, 1996.
[19] D. James, *Order-reducing conjugate gradients versus block AOR for constrained least-squares problems*, Linear Algebra Appl., 154–156 (1991), pp. 23-43.
[20] W. Kahan, *Further remarks on reducing truncation errors*, Comm. ACM, 8 (1965), pp. 1–40.
[21] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, 2nd ed., Academic Press, New York, 1985.

[22] C. Lawson and R. Hanson, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, NJ, 1974.
[23] S.-M. Lu and J. L. Barlow, *Multifrontal computation with the orthogonal factors of sparse matrices*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 658–679.
[24] D. G. Luenberger, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1973.
[25] S. G. Nash and A. Sofer, *Preconditioning reduced matrices*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 47–68.
[26] P. Matstoms, *Sparse QR Factorization with Applications to Linear Least Squares Problems*, in Linköping Studies in Science and Technology, Dissertations, No. 337, Linköping Univ., Sweden, 1994.
[27] I. Perugia, V. Simoncini, and M. Arioli, *Linear algebra methods in a mixed approximation of magnetostatic problems*, Tech. report IAN 1060, IAN-CNR, Pavia, Italy, 1997.
[28] I. Perugia, *A field-based mixed formulation for the two-dimensional magnetostatic problem*, SIAM J. Numer. Anal., 34 (1997), pp. 2382–2391.
[29] C. Puglisi, *Factorisation QR de grandes matrices creuses basée sur une méthode multifrontale dans un environnement multiprocesseur*, Thesis INP, Tech. report TH/PA/93/33, CERFACS, 1993.
[30] J. L. Rigal and J. Gaches, *On the compatibility of a given solution with the data of a linear system*, J. ACM, 14 (1967), pp. 543–548.
[31] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.
[32] G. W. Stewart, *Constrained definite Hessian tend to be well conditioned*, Math. Programming, 21 (1981), pp 235–234.
[33] P. Å. Wedin, *Perturbation theory and condition numbers for generalized and constrained linear least squares problems*, Tech. report UMINF-125.85, Institute of Information Processing, University of Umeå, Umeå, Sweden, 1985.
[34] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, UK, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 3, pp. 840–848
© 2000 Society for Industrial and Applied Mathematics

LINEAR-TIME COPOSITIVITY DETECTION FOR TRIDIAGONAL MATRICES AND EXTENSION TO BLOCK-TRIDIAGONALITY*

IMMANUEL M. BOMZE†

Abstract. Determining whether a given symmetric matrix is copositive (i.e., whether it generates a quadratic form taking no negative values on the positive orthant) is an NP-hard problem. However, for diagonal matrices this amounts to the trivial check of signs of the diagonal entries. Here, a linear-time algorithm for tridiagonal matrices is presented which similarly checks only for signs of diagonal entries, but (depending on the sign of an off-diagonal entry) sometimes updates the matrix by an ordinary pivot step. Ramifications for the sign-constrained border and generalizations for the block-tridiagonal case are also specified. As a key tool, we establish a monotonicity result for the Löwner ordering of partial Schur complements for general symmetric matrices with a positive definite principal block.

Key words. copositive matrix, tridiagonal matrix, block pivoting, Schur complement, Löwner ordering

AMS subject classifications. 65F30, 65F50, 15A57, 15A63

PII. S0895479898341487

1. Introduction and basics.

Denote by $\mathbb{R}^n_+ = \{x \in \mathbb{R}^n : x_i \geq 0 \text{ for all } i\}$ the positive orthant in n-dimensional Euclidean space $\mathbb{R}^n$. A symmetric $n \times n$ matrix Q is said to be $\mathbb{R}^n_+$-copositive (or simply copositive) if

$$x'Qx \geq 0 \quad \text{for all } x \in \mathbb{R}^n_+ , \tag{1.1}$$

where $'$ denotes transposition. To find out whether a given symmetric matrix is copositive is an NP-hard problem [7]. On the other hand, for diagonal matrices this amounts to the trivial check of signs of the diagonal entries. The next difficult case is that of tridiagonal matrices. In [8], the authors propose also a polynomial-time algorithm for this problem for which they gave a worst-case complexity of $\mathcal{O}(n^6)$. Here, a different procedure is proposed which—similar to the diagonal case—checks only for signs of diagonal entries, but (depending on the sign of an off-diagonal entry) sometimes updates the matrix by an ordinary pivot step.

First, let us review a recursive characterization of copositivity by dimensional reduction, following the ideas in [1] and [2]. To this end, recall that given a general cone $\Gamma \subseteq \mathbb{R}^n$, a symmetric $n \times n$ matrix Q is said to be Γ-copositive if its quadratic form takes no negative values on Γ, i.e., if

$$x'Qx \geq 0 \quad \text{for all } x \in \Gamma. \tag{1.2}$$

Next, partition Q into the submatrices

$$Q = \left[\begin{array}{c|c} R & b \\ \hline b' & \gamma \end{array}\right] , \tag{1.3}$$

where R is $(n-1) \times (n-1)$, $b \in \mathbb{R}^n$, and $\gamma \in \mathbb{R}$.

*Received by the editors July 6, 1998; accepted for publication (in revised form) by R. Brualdi April 14, 1999; published electronically February 18, 2000.
http://www.siam.org/journals/simax/21-3/34148.html
†Institut für Statistik, Operations Research und Computerverfahren, University of Vienna, Universitätsstraße 5, A-1010 Vienna, Austria (immanuel.bomze@univie.ac.at).

LEMMA 1.1. *Let the $n \times n$ matrix Q be partitioned as in* (1.3) *and put*

$$H_{\pm} = \{y \in \mathbb{R}_+^{n-1} : \pm b'y \geq 0\}\,. \tag{1.4}$$

Then Q is $\mathbb{R}_+^n$-copositive if and only if $\gamma \geq 0$ and

1) *the $(n-1) \times (n-1)$ matrix R is H_+-copositive; and*
2) *the $(n-1) \times (n-1)$ matrix $\gamma R - bb'$ is H_--copositive if $\gamma > 0$.*

Proof. We fix $y = [x_1, \ldots, x_{n-1}]' \in \mathbb{R}_+^{n-1}$ and consider $x'Qx$ as a quadratic function in x_n:

$$f(x_n|y) = x'Qx = \gamma x_n^2 + 2(b'y)x_n + y'Ry\,. \tag{1.5}$$

Clearly, Q can be copositive only if $\gamma \geq 0$ (put $y = o$ and $x_n = 1$) and only if R is copositive (put $x_n = 0$); however, unless $b = o$, these two conditions are not sufficient. Further, if $\gamma = 0$, then copositivity of Q yields $b'y \geq 0$ for all $y \in \mathbb{R}_+^{n-1}$, which is equivalent to $b \in \mathbb{R}_+^{n-1}$. If, on the other hand, $\gamma > 0$, then with $t(y) = \operatorname{argmin}_{x \in \mathbb{R}} f(x|y) = -(b'y)/\gamma$ we get

$$\gamma f(t(y)|y) = \gamma y'Ry - (b'y)^2 = y'[\gamma R - bb']y\,. \tag{1.6}$$

Convexity of $f(.|y)$ now yields for $x_n \geq 0$

$$x'Qx = f(x_n|y) \geq f(\max\{0, t(y)\}|y) \tag{1.7}$$

or, equivalently, in a more compact form,

$$\gamma x'Qx \geq y'[\gamma R - h(-b'y)bb']y, \tag{1.8}$$

with h denoting the Heavyside function: $h(t) = 1$ if $t \geq 0$, and $h(t) = 0$ if $t < 0$. Thus the result follows. ∎

This recursive characterization can be used for general branch-and-bound-like copositivity detection algorithms as in [1, 2], since in every step, dimension is reduced by one, and solving the one-dimensional question is a trivial sign check. Two difficulties arise: First, proceeding this way, one ends up with 2^n of such checks; the second problem consists of the more complicated structure of the cones $H_{\pm}$, which in general have up to $\frac{n^2}{2}$ extremal rays rather than $n-1$ as the positive orthant $\mathbb{R}_+^{n-1}$, cf. [1, p. 260]. Although these rays can be calculated explicitly, retransformation into a positive orthant can result only in $\mathbb{R}_+^d$ with $d \geq n$ in general, which destroys dimensional reduction. An alternative strategy which decomposes H_+ into a series of subcones isomorphic to $\mathbb{R}_+^{n-1}$ is discussed in [4].

However, if Q has a special structure, then one of the cones H_+ or H_- is contained in the other one so that we end up with a problem chain rather than with a problem tree, since $y'[\gamma R - bb']y \leq \gamma y'Ry$ always. This observation is the basis for the ramification described in the following section.

2. Ramification for sign-constrained borders. According to Lemma 1.1, copositivity of Q already implies a sign constraint on the lower-right corner entry: $\gamma \geq 0$. We now show how additional sign constraints on the remaining entries b_i of the last column simplify this reduction result. For the proof and the arguments to follow in section 4, it will be convenient to recall that, given two symmetric matrices A and B of the same size, A is considered to precede B with regard to (w.r.t.) the Löwner ordering, denoted by $A \preceq B$, if and only if $B - A$ is positive semidefinite.

Positive semidefiniteness of A can thus be abbreviated by $O \preceq A$. Here and in what follows, the generically used letter O denotes a (possibly rectangular) zero matrix of appropriate size.

LEMMA 2.1. *Let the $n \times n$ matrix Q be partitioned as in* (1.3) *with $\gamma \geq 0$.*

(a) *If $b \in \mathbb{R}^n_+$, then Q is $\mathbb{R}^n_+$-copositive if and only if R is $\mathbb{R}^{n-1}_+$-copositive.*

(b) *If $-b \in \mathbb{R}^n_+$, then Q is $\mathbb{R}^n_+$-copositive if and only if*

1) *either $\gamma = 0$, $b = o$, and R is $\mathbb{R}^{n-1}_+$-copositive; or*

2) *$\gamma > 0$ and $\gamma R - bb'$ is $\mathbb{R}^{n-1}_+$-copositive.*

Proof. If $\pm b \in \mathbb{R}^{n-1}_+$, then the cones $H_\pm$ from Lemma 1.1 are contained in each other: Indeed for $b \in \mathbb{R}^{n-1}_+$, we get $H_+ = \mathbb{R}^{n-1}_+$ while $H_- = \mathbb{R}^{n-1}_+ \cap b^\perp \subseteq H_+$, whereas for $-b \in \mathbb{R}^{n-1}_+$, we have similarly $H_+ = \mathbb{R}^{n-1}_+ \cap b^\perp \subseteq \mathbb{R}^{n-1}_+ = H_-$.

(a) Now if $\gamma > 0$, then $y'[\gamma R - bb']y = \gamma y' R y$ for all $y \in H_-$. Hence the result follows from Lemma 1.1, because condition 1 implies condition 2 due to $H_- \subseteq H_+$.

(b) As argued in the proof of Lemma 1.1, $\gamma = 0$ entails $b \in \mathbb{R}^{n-1}_+$, which here means $b = o$. On the other hand, if $\gamma > 0$, we obviously get $\gamma R - bb' \preceq \gamma R$, so that condition 2 in Lemma 1.1 yields condition 1 by means of $H_+ \subseteq H_-$. ∎

Using Lemma 2.1, we not only have the benefit of a single smaller successor problem rather than two of these, but we also must deal with the positive orthants rather than with $H_\pm$, which in general are more complicated. So a promising reduction strategy for general symmetric matrices Q could consist of reordering coordinates such that the last row of Q contains only positive (negative) off-diagonal entries and then proceed recursively as in Lemma 2.1; cf. [1, pp. 256f]. Note that case (a) occurs exclusively if Q has no negative entries, in which case the answer is trivial. So this result becomes relevant only if case (b) occurs at least occasionally.

Example 2.2. Consider the data from Example 1 of [2]:

$$Q = \begin{bmatrix} 1 & -1 & 1 & 2 \\ -1 & 2 & -3 & -3 \\ 1 & -3 & 5 & 6 \\ 2 & -3 & 6 & 5 \end{bmatrix},$$

where the second row of Q has only negative off-diagonal entries: $b = [-1, -3, -3]'$ while $\gamma = 2$. Hence we have to deal with the updated matrix $\tilde{Q} = \gamma R - bb'$, which reads

$$\tilde{Q} = \begin{bmatrix} 1 & -1 & 1 \\ -1 & 1 & 3 \\ 1 & 3 & 1 \end{bmatrix}.$$

Now the last row of $\tilde{Q}$ has positive entries only, so that the new $\tilde{R}$ is obtained by extracting the leading 2×2 principal submatrix. It is readily seen that $\tilde{R}$ is positive semidefinite; however, to proceed consistently, consider any of the two principal submatrices of $\gamma\tilde{R} - \tilde{b}\tilde{b}'$, which are now 1×1 and of the form $1 - (-1)^2 = 0$, so that the algorithm stops with the affirmative answer. Compare the minuscule effort here with that in [2, 9]; e.g., in the latter reference, ten tableaux have to be calculated to establish copositivity of Q.

In case that there is no (off-diagonally) sign-constrained column, one can reduce the number of successor problems by choosing that row with the least product of number of positive/negative entries (cf. again [1]). On the other hand, there are matrices which guarantee that all columns are off-diagonally sign-constrained and that

this property is also hereditary to the updates. Important instances are tridiagonal matrices with which we will deal explicitly in the next section.

3. Linear-time copositivity check for tridiagonal matrices. Suppose that Q is tridiagonal, i.e., $q_{ij} = 0$ if $|i - j| \geq 2$. This is a special instance of the sign-constrained border case. For ease of reference, we formulate the result explicitly.

LEMMA 3.1. *If Q is tridiagonal with $b' = [0, \ldots, 0, \rho]$ in the partition* (1.3)*, then Q is $\mathbb{R}^n_+$-copositive if and only if*

(a) *either $\gamma \geq 0$, $\rho \geq 0$, and R is $\mathbb{R}^{n-1}_+$-copositive; or*

(b) *$\gamma > 0$, $\rho < 0$, and $\gamma R - bb'$ is $\mathbb{R}^{n-1}_+$-copositive.*

Proof. Of course, $\pm b \in \mathbb{R}^{n-1}_+$ if and only if $\pm\rho \geq 0$. Note that for the case $\rho = 0$ we get $H_+ = H_-$ and also $bb' = O$, so that again only $\mathbb{R}^{n-1}_+$-copositivity of R is decisive. Hence the result follows immediately from Lemma 2.1. ∎

Finally, observe that the updated matrix $\gamma R - bb'$ is again tridiagonal if Q is so, since the update affects only the entry in the lower-right corner of the R matrix. Summarizing, we arrive at the following algorithm.

COPOSITIVITY DETECTION ALGORITHM.
Input: Tridiagonal $n \times n$ matrix Q.
For $i = n$ downto 1 do
  if $q_{ii} < 0$, stop ("Q is not copositive");
    else if $i = 1$, stop ("Q is copositive");
    else if $q_{i,i-1} < 0$ update $q_{jk} := q_{ii}q_{jk} - q_{ij}q_{ik}$ for all $j, k < i$ with $|j - k| \leq 1$
    endif
  endif
endfor.

THEOREM 3.2. *The above algorithm delivers the exact answer after at most n iterations.*

Proof. In view of Lemma 3.1, it remains to show only that the case $\gamma = q_{ii} = 0$ is treated correctly. Indeed, then necessarily $\rho \geq 0$ if Q is copositive, and in this case we have to proceed with R as in the described algorithm. If, however, $\rho = q_{i,i-1} < 0$ (and hence Q is not copositive), then the update gives for $j = k = i - 1$

$$\tilde{q}_{i-1,i-1} = \gamma q_{i-1,i-1} - \rho^2 = -\rho^2 < 0\,,$$

so violation of copositivity will be detected immediately at the next iteration. ∎

Next let us deal with the following problem: Suppose that Q fails to be $\mathbb{R}^n_+$-copositive. Then one would like to have a violating vector $x \in \mathbb{R}^n_+$ such that $x'Qx < 0$. This is simple, if $\rho \geq 0$ and a vector is found such that $y'Ry < 0$: Just take $x = [y'|0]' \in \mathbb{R}^n_+$. However, if in the other case $\rho < 0$ we obtain $y \in \mathbb{R}^{n-1}_+$ with $y'[\gamma R - bb']y < 0$, then this means $t(y) = -(b'y)/\gamma = (-\rho/\gamma)y_{n-1} \geq 0$ and $f(t(y)|y) < 0$. Then we have to take

$$x = [y'|t(y)]' = [y_1, \ldots, y_{n-1}, -\frac{\rho}{\gamma}y_{n-1}]' \in \mathbb{R}^n_+\,.$$

Proceeding recursively with this backtracking, one gets the desired result, again, after at most n repetitions. Even if some entries of Q are successively overwritten, as in the above description, the necessary information (which branch according to the sign of ρ is followed) can be simply retrieved by $h(-\rho/\gamma) = h(-q_{n,n-1}/q_{nn})$ to get the correct multiplier of the last y-coordinate for the additional entry of x. See the description below, where n is replaced with j in the current cycle. Note that the input $\tilde{Q}$ is

supposed to result from updating the originally given matrix Q with the previously sketched algorithm.

ALGORITHM FOR DETECTING A VIOLATING VECTOR.
Input: Tridiagonal $n \times n$ matrix $\tilde{Q}$ with $\tilde{q}_{ii} < 0$ while $\tilde{q}_{jj} \geq 0$, for all $j \in \{i+1, \ldots, n\}$.
Output: violating vector $x \in \mathbb{R}^n_+$ with $x'Qx < 0$.
Initialize $x_i := 1$ and $x_j := 0$ for all $j \neq i$.
For $j = i + 1$ to n do
  if $\tilde{q}_{j,j-1}\tilde{q}_{jj} < 0$ update $x_j := -\frac{\tilde{q}_{j,j-1}}{\tilde{q}_{jj}} x_{j-1}$;
  endif
endfor.

4. Block versions. In this section we follow the general block-pivoting strategy in order to reduce dimensionality of the problem as described in [2]. To this end choose a principal $k \times k$ submatrix A of Q such that A is either positive definite or negative semidefinite and decompose Q as follows:

$$Q = \left[\begin{array}{c|c} A & B \\ B' & C \end{array}\right] . \tag{4.1}$$

Accordingly, B and C are submatrices of order $k \times (n-k)$ and $(n-k) \times (n-k)$, respectively.

Case (i) A is negative semidefinite. Then the following is a quite straightforward characterization (cf. [2, Remark 8]): Q is copositive if and only if (i) $A = O$; (ii) $b_{ij} \geq 0$ for all i, j; and (iii) C is $\mathbb{R}^{n-k}_+$-copositive.

Case (ii) A is positive definite. Then Theorem 8 of [2] provides the following characterization (for a neat formulation of which we will need some more notation): Let $I \subseteq \{1, \ldots, k\}$ and $J = \{1, \ldots, k\} \setminus I$. Partition A and B according to I and J:

$$A = \left[\begin{array}{c|c} A_I & A_{IJ} \\ A'_{IJ} & A_J \end{array}\right] \quad \text{and} \quad B = \left[\begin{array}{c} B_I \\ B_J \end{array}\right] , \tag{4.2}$$

and form

$$Q_I = C - B'_J A_J^{-1} B_J \tag{4.3}$$

(note that A_J is nonsingular as A is supposed to be positive definite), as well as

$$\Gamma_I = \{z \in \mathbb{R}^{n-k}_+ : A_J^{-1} B_J z \leq o \quad \text{and} \quad (B_I - A_{IJ} A_J^{-1} B_J) z \geq o\} \tag{4.4}$$

(if $I = \{1, \ldots, k\}$, simply put $B_I = B$ and $B_J = O$, while for $I = \emptyset$, let $Q_\emptyset = C - B' A^{-1} B$ and $\Gamma_\emptyset = \{z \in \mathbb{R}^{n-k}_+ : A^{-1} B z \leq o\}$).

THEOREM 4.1. *Suppose that Q is partitioned as in* (4.1) *with A positive definite. Then Q is $\mathbb{R}^n_+$-copositive if and only if the $(n-k) \times (n-k)$ matrices*

$$Q_I \text{ are } \Gamma_I\text{-copositive for all } I \subseteq \{1, \ldots, k\} .$$

Proof. See [2]. □

Note that for general matrices, this characterization involves up to 2^k successor cones, so not much is gained for large k. But we want to deal here with specially structured matrices Q, to be more specific, with simple off-diagonal blocks B. The easiest case here is that of $B = O$. This corresponds to the block-diagonal case and can be treated exactly as the diagonal case ($k = 1$). Indeed, Remark 8 of [2] gives

$\Gamma_I = \mathbb{R}^{n-k}_+$ and similarly $Q_I = C$ for all I. Hence Q is $\mathbb{R}^n_+$-copositive if and only if C is $\mathbb{R}^{n-k}_+$-copositive. Of course, this is directly evident due to the block-diagonal structure of Q.

The next complicated case can be viewed as the block-tridiagonal one, where we assume that only one row of B is not vanishing: $B = [O|c]'$, so that $B_I = O$ whenever $k \in J$, and $B_J = O$ whenever $k \in I$, since c' is the kth row of B. In the following lemma, we deal with both cases separately. For neat formulation let us introduce $s_I = [S_I]_k$, the last (i.e., kth) column of the matrix $S_I = \left[\begin{smallmatrix} A_{IJ}A_J^{-1} \\ A_J^{-1} \end{smallmatrix}\right]$, and the system

$$\mathcal{I}_+ = \{I \subseteq \{1, \ldots, k-1\} : s_I \geq o\}, \tag{4.5}$$

as well as $\sigma_I = [s_I]_k$, the last coordinate of s_I. Note that $\sigma_I > 0$ since this is the last diagonal entry of the positive definite matrix A_J^{-1}.

LEMMA 4.2. *Denote* $H_\pm = \{z \in \mathbb{R}^{n-k}_+ : \pm c'z \geq 0\}$ *and let* $I \subseteq \{1, \ldots, k\}$ *with* $J = \{1, \ldots, k\} \setminus I$.

(a) *If* $k \in I$, *then* $Q_I = C = Q_{\{1,\ldots,k\}}$ *and* $\Gamma_I = H_+ = \Gamma_{\{1,\ldots,k\}}$.

(b) *If* $k \in J$, *then* $Q_I = C - \sigma_I cc'$ *and* $\Gamma_I = H_-$ *if* $I \in \mathcal{I}_+$ *while* $\Gamma_I = H_+ \cap H_- \subseteq H_-$ *if* $I \notin \mathcal{I}_+$.

Proof. (a) If $k \in I$, then $B_J = O$ while $B_I = [O|c]'$, so that (4.3) gives $Q_I = C$ and (4.4) yields

$$\Gamma_I = \{z \in \mathbb{R}^{n-k}_+ : c'z \geq 0\} = H_+ .$$

(b) If $k \in J$, then

$$\Gamma_I = \{z \in \mathbb{R}^{n-k}_+ : (c'z)s_I \leq o\} .$$

Now s_I has at least one positive entry $\sigma_I > 0$, and therefore we arrive at $\Gamma_I \subseteq H_-$ with equality if $s_I \geq o$, i.e., if $I \in \mathcal{I}_+$. If, however, (at least) one entry of s_I is negative (i.e., if $I \notin \mathcal{I}_+$), then one obviously obtains $\Gamma_I = H_+ \cap H_-$. Hence the result. ∎

Interestingly enough, we derive from Theorem 4.1 and Lemma 4.2 the following result on the sign structure of partial inverses of positive definite matrices, which ensures $\mathcal{I}_+ \neq \emptyset$. I am not aware of a more simple and direct proof of this result.

COROLLARY 4.3. *Let* A *be an arbitrary positive definite* $k \times k$ *matrix, partitioned as in* (4.2). *Then for any* $i \in \{1, \ldots, k\}$, *there is at least one* $I \subseteq \{1, \ldots, k\}$ *with* $i \notin I$, *such that the* i*th column of* $\left[\begin{smallmatrix} A_{IJ}A_J^{-1} \\ A_J^{-1} \end{smallmatrix}\right]$ *has no negative entries.*

Proof. Without loss of generality we assume $i = k$. Hence we have to show that the system $\mathcal{I}_+$ defined in (4.5) is nonvoid. To this end we put $n = k + 1$ and enrich A with $B = [0, \ldots, 0, \rho]' \in \mathbb{R}^k$ as well as $C = \gamma > 0$, arriving at an $n \times n$ matrix Q as in (4.1). Now, if $\gamma = a_{kk} > 0$ and $\rho = -2a_{kk}$, then for $x = [0, \ldots, 0, 1, 1]' \in \mathbb{R}^n_+$ we get $x'Qx = -2\gamma < 0$. Hence Q cannot be copositive. On the other hand, C is clearly H_+-copositive and thus $Q_I = C$ is Γ_I-copositive for all I with $k \in I$ due to Lemma 4.2(a). Now suppose $\mathcal{I}_+ = \emptyset$. Then Lemma 4.2(b) would entail $\Gamma_I = H_+ \cap H_-$ and thus $z'Q_Iz = z'Cz - 0 = \gamma z^2 \geq 0$ for all z and all I not containing k. Hence also for $k \notin I$, we get Γ_I-copositivity of Q_I, from which Theorem 4.1 yields copositivity of Q, which, as we have seen above, is absurd. Therefore, $\mathcal{I}_+$ cannot be void. ∎

Summarizing, we arrive at the following theorem.

THEOREM 4.4. *Suppose* Q *is partitioned as in* (4.1) *with* A *a positive definite* $k \times k$ *matrix and* $B = [O|c]'$ *for some* $c \in \mathbb{R}^{n-k}$. *Denote by* $H_\pm = \{z \in \mathbb{R}^{n-k}_+ : \pm c'z \geq 0\}$. *Then* Q *is* $\mathbb{R}^n_+$*-copositive if and only if*

1) *the* $(n-k)\times(n-k)$ *matrix* C *is* H_+*-copositive; and*
2) *the* $(n-k)\times(n-k)$ *matrix* $\gamma_k C - cc'$ *is* H_-*-copositive, where*

$$\gamma_k = \max\{\sigma_I^{-1} : I \in \mathcal{I}_+\}\,.$$

Proof. Choose $I_o \in \mathcal{I}_+$ such that $\gamma_k = 1/\sigma_{I_o}$. Due to Lemma 4.2(b) we derive that Q_I are Γ_I-copositive for all $I \in \mathcal{I}_+$ if and only if Q_{I_o} is H_--copositive. This follows from

$$z'Q_I z = z'Cz - \sigma_I(c'z)^2 \geq z'Cz - \sigma_{I_o}(c'z)^2 = z'Q_{I_o}z\,.$$

Also, Γ_I-copositivity of Q_I follows from H_+-copositivity of C if either $k \in I$—due to Lemma 4.2(a)—or $k \notin I \notin \mathcal{I}_+$, because in the latter case we have $z'Q_I z = z'Cz$ for all $z \in H_+ \cap H_- = \Gamma_I$. The result now follows by the observation that $\gamma_k C - cc' = \gamma_k Q_{I_o}$ from Theorem 4.1. □

Crucial for practical application is the fact that to obtain γ_k, one need not calculate σ_I for all $I \in \mathcal{I}_+$. Rather it is sufficient to consider the subsystem

$$\mathcal{I}_o = \{I \in \mathcal{I}_+ : I \text{ is minimal w.r.t. } \subseteq\}\,.$$

To prove this, we need a general result on the Löwner ordering for partial Schur complements, i.e., for the matrices Q_I as defined in (4.3). Indeed, if A is positive definite we can derive the following monotonicity property.

LEMMA 4.5. *If* $\tilde{I} \subseteq I$, *then* $Q_{\tilde{I}} \preceq Q_I$.

Proof. Since the assertion affects only submatrices of $[A_{\tilde{J}}|B_{\tilde{J}}]$ (cf. the partition (4.2)), we may, and for notational convenience do, assume $\tilde{I} = \emptyset$. Note that following the usual convention for empty matrices, we get for $I = \emptyset$ the Schur complement $Q_\emptyset = C - B'A^{-1}B$, while for $J = \emptyset$ we obtain $Q_{\{1,\ldots,k\}} = C$, the remaining principal block. First, let us establish the relation

$$O \preceq \begin{bmatrix} O & | & O \\ O & | & A_J^{-1} \end{bmatrix} \preceq A^{-1}\,. \tag{4.6}$$

Indeed, using the Schur complement $D = A_I - A_{IJ}A_J^{-1}A'_{IJ}$ of A_J in A and its inverse $E = D^{-1}$, we have the following block-inversion formula for A (see, e.g., [5]):

$$A^{-1} = \begin{bmatrix} A_I & | & A_{IJ} \\ A'_{IJ} & | & A_J \end{bmatrix}^{-1} = \begin{bmatrix} E & | & -EA_{IJ}A_J^{-1} \\ -A_J^{-1}A'_{IJ}E & | & F \end{bmatrix}\,, \tag{4.7}$$

where $F = A_J^{-1} + A_J^{-1}A'_{IJ}EA_{IJ}A_J^{-1}$. Now clearly A^{-1} is positive definite if A is so, and hence F is also positive definite. Moreover, since $E = D^{-1}$ is positive definite as the inverse of the Schur complement [6, 3] and thus $O \preceq A_J^{-1}A_{JI}EA'_{JI}A_J^{-1}$, we even have $O \preceq A_J^{-1} \preceq F$. This relation is generalized by (4.6). Now to prove (4.6), we use (4.7) so that all we have to show is positive semidefiniteness of the difference matrix

$$\begin{bmatrix} E & | & -EA_{IJ}A_J^{-1} \\ -A_J^{-1}A'_{IJ}E & | & A_J^{-1}A'_{IJ}EA_{IJ}A_J^{-1} \end{bmatrix}\,.$$

To this end, observe that this matrix coincides with the symmetric product

$$\begin{bmatrix} I_m & | & O \\ -A_J^{-1}A'_{IJ} & | & O \end{bmatrix} \begin{bmatrix} E & | & O \\ O & | & O \end{bmatrix} \begin{bmatrix} I_m & | & -A_{IJ}A_J^{-1} \\ O & | & O \end{bmatrix}\,,$$

where m is the number of elements in I and the symbol I_m is used for the $m \times m$ identity matrix. Since $O \preceq E$, relation (4.6) is established. But then

$$B'_J A_J^{-1} B_J = B' \begin{bmatrix} O & | & O \\ O & | & A_J^{-1} \end{bmatrix} B \preceq B' A^{-1} B,$$

which entails via (4.3) the result $Q_\emptyset \preceq Q_I$. □

Remark 4.6. By virtually the same arguments, we could have established the identity

$$Q_I = Q_\emptyset + H'_I G_I H_I \,, \tag{4.8}$$

where $G_I = (E_I A^{-1} E'_I)^{-1}$ and $H_I = E_I A^{-1} B - F_I$ for some matrices E_I (of full row rank) and F_I of appropriate sizes. Equation (4.8) has been formulated (but not proved) in Theorem 8 of [2] and immediately yields via $O \preceq G_I$ the relation $Q_\emptyset \preceq Q_I$ of Lemma 4.5.

From the general result of Lemma 4.5, it follows that for all $I \in \mathcal{I}_+$ there is a subset $\tilde{I} \in \mathcal{I}_o$ such that $z' Q_I z \geq z' Q_{\tilde{I}} z$ for all z and therefore $\sigma_I \geq \sigma_{\tilde{I}}$. Hence we can obtain γ_k in Theorem 4.4 from

$$\gamma_k = \max\{\sigma_{\tilde{I}}^{-1} : \tilde{I} \in \mathcal{I}_o\} \,.$$

Applying Theorem 4.4, one gets a significant reduction in successor problems, namely, from (possibly) 2^k down to 2. Two problems remain: (a) This still yields a binary tree of subproblems, albeit with less height than those in the previous section; (b) the resulting subproblems are of different structure, since the cones, as pointed out already, are not isomorphic to the positive orthant in $\mathbb{R}^{n-k}$. The latter problem is not that stringent if $k > n - \sqrt{2n}$, because $H_\pm$ has at most $(n-k)^2/2$ extremal rays, and hence dimensional reduction is still possible. Unfortunately there is no guarantee of finding large positive definite principal submatrices A of Q, even ignoring the structure of B; cf. [2, Example 2].

However, if c is again sign-constrained, we can argue as in section 2 to get only one successor problem. In particular, if $c = [\rho | o']'$ has only one nonvanishing entry, then as in the scalar case presented in the previous section, this structure is hereditary to the subproblem: If, again, C is block-tridiagonal, then so will $\tilde{C} = \gamma_k C - cc'$ be block-tridiagonal, since only one element in the main diagonal will be affected by the update. Furthermore, again $H_\pm = \mathbb{R}^{n-k}_+$ if $\pm\rho > 0$, so virtually the same algorithms as in the previous section apply.

5. Conclusion. Using a recursive dimensional reduction strategy, we developed in section 2 a ramification for matrices with sign-constrained borders and applied this in section 3 to obtain a linear-time algorithm for detecting copositivity of tridiagonal matrices. The procedure proposed consists of simple sign checks of diagonal entries and, depending on the sign of an off-diagonal entry, occasional updates of parts of the matrix by an ordinary pivot step. Section 4 is devoted to an extension for the block-tridiagonal case with a positive definite principal block, where the off-diagonal block has only one nonvanishing row. Here the characterization comes from a general reduction scheme using block-pivoting arguments. A key tool for this approach is the Löwner monotonicity of partial Schur complements. Again, this extension becomes especially effective in the sign-constrained border case, e.g., if only one entry in the off-diagonal block is nonzero. But even if the off-diagonal entries vary in sign, a considerable reduction in effort is guaranteed, provided the positive definite block is large enough.

Acknowledgment. The problem addressed in this paper was brought to my attention by the SSOR group, ITS Faculty, Delft University of Technology. Their kind hospitality is warmly appreciated. I also owe thanks to an anonymous referee for valuable comments which helped to improve the presentation of this paper.

REFERENCES

[1] I. M. Bomze, *Remarks on the recursive structure of copositivity*, J. Inform. Optim. Sci., 8 (1987), pp. 243–260.

[2] I. M. Bomze, *Block pivoting and shortcut strategies for detecting copositivity*, Linear Algebra Appl., 248 (1996), pp. 161–184.

[3] R. W. Cottle, *Manifestations of the Schur complement*, Linear Algebra Appl., 8 (1974), pp. 189–211.

[4] G. Danninger, *A recursive algorithm to detect (strict) copositivity of a symmetric matrix*, in Methods of Operations Research 62, U. Rieder, P. Gessner, A. Peyerimhoff, and F. J. Radermacher, eds., Hain, Meisenheim, 1990, pp. 45–52.

[5] G. Golub and C. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[6] E. V. Haynsworth, *Determination of the inertia of a partitioned Hermitian matrix*, Linear Algebra Appl., 1 (1967), pp. 73–81.

[7] K. G. Murty and S. N. Kabadi, *Some NP-complete problems in quadratic and linear programming*, Math. Programming, 39 (1987), pp. 117–129.

[8] A. J. Quist, E. de Klerk, C. Roos, and T. Terlaky, *Copositive relaxation for general quadratic programming*, Optim. Methods Softw., 9 (1998), pp. 185–209.

[9] H. Väliaho, *Criteria for copositive matrices*, Linear Algebra Appl., 81 (1986), pp. 19–34.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 3, pp. 849–866
© 2000 Society for Industrial and Applied Mathematics

STABILIZATION BY PERTURBATION OF A $4n^2$ TOEPLITZ SOLVER*

PER CHRISTIAN HANSEN† AND PLAMEN Y. YALAMOV‡

Abstract. We consider a $4n^2$-algorithm for computing the Gohberg–Semencul representation of the inverse of a nonsymmetric Toeplitz matrix and for obtaining the solution of a Toeplitz system. The original algorithm is unstable when one or more of the leading principal submatrices are ill-conditioned. In our stabilized version we slightly perturb the matrix when we encounter a division by a small number. As a consequence, the solution is also perturbed, and its accuracy can be improved by taking a small number of iterative refinement steps. We present a roundoff error analysis of the new algorithm as well as numerical results that support our analysis.

Key words. general Toeplitz matrices, $O(n^2)$ algorithms, stabilization

AMS subject classifications. 65F30, 65G05

PII. S0895479895292266

1. Introduction.

Toeplitz matrices and Toeplitz systems appear in many applications, such as signal processing, statistics, integral equations, and Padé approximations; see [5]. A number of specialized algorithms have been proposed for solving Toeplitz systems and for inversion of Toeplitz matrices. A survey of many algorithms in this area is given in [4].

These algorithms are faster than the usual algorithms for general dense matrices because they exploit the structure of the Toeplitz matrix. Many of them require $\mathcal{O}(n^2)$ operations, where n is the order of the matrix, instead of $\mathcal{O}(n^3)$ operations for a general dense solver. But the $\mathcal{O}(n^2)$-algorithms can break down or produce completely wrong results. For example, this happens in Levinson's algorithm [12, section 4.7.3], [17] if one or more of the leading principal submatrices are singular or ill-conditioned. Pivoting techniques are not useful because they spoil the Toeplitz structure.

In order to handle the ill-conditioning of leading principal submatrices in the Levinson algorithm, Chan and Hansen incorporated look-ahead in the algorithm [6]. Their look-ahead algorithm is based on the idea that we can take a block step instead of the usual step in the Levinson algorithm, so that any ill-conditioned leading principal submatrix (which is implicitly to be inverted) is avoided. This algorithm, as well as the related algorithm by Freund and Zha [10], needs some overhead in time and storage which grows with the size of the block step, and a maximum block size is required in a practical implementation. Hence, the look-ahead algorithms are useful only for matrices that require look-ahead steps with small block size.

In this paper we consider a different approach for improving the stability of fast Toeplitz solvers. This stabilization does not require look-ahead, and a fixed size of additional storage, $3n$ locations, is needed. Our approach is based on the idea that

*Received by the editors September 25, 1995; accepted for publication (in revised form) by F. T. Luk May 17, 1999; published electronically February 18, 2000.
http://www.siam.org/journals/simax/21-3/29226.html

†Department of Mathematical Modelling, Building 321, Technical University of Denmark, DK-2800 Lyngby, Denmark (pch@imm.dtu.dk).

‡Center of Applied Mathematics and Informatics, University of Rousse, 7017 Rousse, Bulgaria (yalamov@ami.ru.acad.bg). This author was partially supported by grants MM-707/97 and I-702/97 from the Bulgarian Ministry of Education and Science and by grant 21-0804/26-1821 from the Danish Research Councils.

if we add relatively small perturbations to ill-conditioned submatrices, then we can improve their conditioning. If the whole matrix is well-conditioned, it will not be influenced strongly by these perturbations, and by taking a few steps of iterative refinement we can compute the solution of a system of equations, or a representation of the inverse of the Toeplitz matrix, to satisfactory accuracy.

The same type of stabilization was proposed by Concus and Saylor [7] for indefinite symmetric Toeplitz matrices in connection with Trench's inversion algorithm [12, section 4.7.4].

We apply the stabilization to an algorithm for computing the Gohberg–Semencul representation of the inverse of a nonsymmetric Toeplitz matrix. This algorithm is discussed, for example, by Heinig and Rost [13] and Voevodin and Tyrtyshnikov [21], and it is also the basis algorithm in Argonne's Toeplitz package [2]. It is a modification of Durbin's algorithm [12, section 4.7.2] for solving the Yule–Walker problem. The algorithm requires $4n^2 + \mathcal{O}(n)$ operations, and we shall demonstrate that the computational overhead in our stabilization is $\mathcal{O}(1)$.

When the $4n^2$-algorithm is used to solve a system of equations, we need to multiply the right-hand side with the inverse of the Toeplitz matrix; the complete algorithm is sometimes referred to as Jain's algorithm [16]. The multiplications can be performed in $\mathcal{O}(n \log_2 n)$ operations, and we refer to [1] and [20, section 4.2.4] for details about fast methods for multiplication with Toeplitz matrices. When the stabilized algorithm is used, a few iterative refinement steps are necessary, and these steps can also be performed in $\mathcal{O}(n \log_2 n)$ operations. Hence, the stabilized algorithm needs $4n^2 + \mathcal{O}(n \log_2 n)$ operations to solve a system of equations. Even for moderate n, this is faster than the classical Levinson algorithm, which requires $6n^2 + \mathcal{O}(n)$ operations, as long as the number of refinement steps is moderate.

Consider the nonsymmetric $(n+1) \times (n+1)$ Toeplitz matrix

$$T_n = \begin{pmatrix} t_0 & t_{-1} & \cdots & t_{-n} \\ t_1 & t_0 & \cdots & t_{1-n} \\ \vdots & \vdots & \ddots & \vdots \\ t_n & t_{n-1} & \cdots & t_0 \end{pmatrix},$$

and let T_k, $k = 0, \ldots, n$, denote the $(k+1) \times (k+1)$ leading principal submatrix of T_n. The algorithm computes representations of T_k^{-1} successively. The inverses are represented via their first and last columns $\mathbf{x}^{(k)} = (T_k^{-1})_{:,1}$ and $\mathbf{y}^{(k)} = (T_k^{-1})_{:,k+1}$. In each step of the algorithm we update recursively a scaling factor p_k and the *scaled* vectors $\tilde{\mathbf{x}}^{(k)} = p_k^{-1}\mathbf{x}^{(k)}$ and $\tilde{\mathbf{y}}^{(k)} = p_k^{-1}\mathbf{y}^{(k)}$ to arrive at p_n, $\tilde{\mathbf{x}}^{(n)} = p_n^{-1}(T_n^{-1})_{:,1}$ and $\tilde{\mathbf{y}}^{(n)} = p_n^{-1}(T_n^{-1})_{:,n+1}$.

As we show in the next section, ill-conditioning of T_k leads to division by a small number and growth of the scaling factor p_k. The only dangerous place (i.e., division by a small number) where the error can blow up is in the computation of the scaling factors p_k. Therefore, we need to apply stabilization only to the computation of p_k.

In a practical implementation of the look-ahead Levinson algorithm it is necessary that the number of successive ill-conditioned submatrices is relatively small, while our algorithm does not need this assumption. Moreover, our algorithm needs less storage than the look-ahead algorithm: one array of length $2n-1$ to store the stabilized matrix and one n-array for storage of the residuals in the iterative refinement part.

Although the basic idea in our algorithm is the same as that used in the algorithm by Concus and Saylor [7], there are important differences. Our algorithm works for general Toeplitz matrices, and it computes the Gohberg–Semencul representation

$\tilde{\mathbf{x}}^{(0)} = \tilde{\mathbf{y}}^{(0)} = 1;\ p_0 = 1/t_0;$
for $k = 1, \ldots, n$
$f_k = t_k \tilde{\mathrm{x}}_0^{(k-1)} + \cdots + t_1 \tilde{\mathrm{x}}_{k-1}^{(k-1)};$
$g_k = t_{-1} \tilde{\mathrm{y}}_0^{(k-1)} + \cdots + t_{-k} \tilde{\mathrm{y}}_{k-1}^{(k-1)};$
$\alpha_k = p_{k-1} f_k; \quad \beta_k = p_{k-1} g_k;$
$p_k = p_{k-1}/(1 - \alpha_k \beta_k);$
$\tilde{\mathbf{x}}^{(k)} = \begin{pmatrix} \tilde{\mathbf{x}}^{(k-1)} \\ 0 \end{pmatrix} - \alpha_k \begin{pmatrix} 0 \\ \tilde{\mathbf{y}}^{(k-1)} \end{pmatrix};$
$\tilde{\mathbf{y}}^{(k)} = -\beta_k \begin{pmatrix} \tilde{\mathbf{x}}^{(k-1)} \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \tilde{\mathbf{y}}^{(k-1)} \end{pmatrix};$
end

FIG. 2.1. *The original $4n^2$-algorithm.*

of T_n^{-1}. Moreover, we provide a roundoff error analysis, as well as theory that supports our choice of the necessary perturbation parameter.

The stabilization idea should carry over to other algorithms that also depend on well-conditioned leading submatrices, such as the Levinson and Schur algorithms. The systolic version of the Bareiss algorithm, derived in [4], is another potential application of the stabilization-by-perturbation idea, since systolic arrays require algorithms with very simple structure.

The outline of our paper is as follows. In section 2 we present a brief summary of the $4n^2$-algorithm and we illustrate how unstable the original algorithm can be. In section 3 the new stabilized version of the algorithm is presented. An error analysis of both algorithms is given in section 4. Finally, numerical experiments are presented in section 5.

A few words about our notation: Matrices are denoted by capital letters (such as T_k), vectors are denoted by lower-case bold letters (such as $\mathbf{x}^{(k)}$), and scalars are denoted by lower-case letters (such as p_k). The ith element of a vector is denoted by a subscript i, such as $\mathbf{x}_i^{(k)}$, and note that vector indices start at 0. The machine precision is denoted by ϵ_M.

2. A brief summary of the original algorithm. The $4n^2$-algorithm for computation of the first and last columns $\mathbf{x}^{(n)}$ and $\mathbf{y}^{(n)}$ of T_n^{-1} is summarized in Figure 2.1. Inside the loop this algorithm computes the scaled first and last columns

$$\tilde{\mathbf{x}}^{(k)} = p_k^{-1}\mathbf{x}^{(k)} = \begin{pmatrix} \tilde{\mathrm{x}}_0^{(k)} \\ \vdots \\ \tilde{\mathrm{x}}_k^{(k)} \end{pmatrix}, \qquad \tilde{\mathbf{y}}^{(k)} = p_k^{-1}\mathbf{y}^{(k)} = \begin{pmatrix} \tilde{\mathrm{y}}_0^{(k)} \\ \vdots \\ \tilde{\mathrm{y}}_k^{(k)} \end{pmatrix}$$

of the matrix T_k^{-1} with scaling factor p_k. This is a simplified version of the algorithm proposed in [21], where there are two separate scaling factors for $\tilde{\mathbf{x}}^{(k)}$ and $\tilde{\mathbf{y}}^{(k)}$ and the initial values $\tilde{\mathbf{x}}^{(0)}$ and $\tilde{\mathbf{y}}^{(0)}$ are arbitrarily chosen. We simplify this choice and, thus, simplify the computation of the scaling factors. As a consequence of our choice, we have

$$\tilde{\mathrm{x}}_0^{(k)} = \tilde{\mathrm{y}}_k^{(k)} = 1, \qquad k = 0, \ldots, n, \tag{2.1}$$

which follows from the persymmetry of T_k^{-1}.

$t_{00}^{-1} = p_n$;
for $i = 1, \ldots, \lceil n/2 \rceil$
  $t_{ii}^{-1} = t_{i-1,i-1}^{-1} + p_n(\tilde{\mathrm{x}}_i^{(n)}\tilde{\mathrm{y}}_{n-i}^{(n)} + \tilde{\mathrm{x}}_{n-i+1}^{(n)}\tilde{\mathrm{y}}_{i-1}^{(n)})$;
end

FIG. 2.2. *Computation of the first* $\lceil n/2 \rceil$ *diagonal entries* t_{ii}^{-1} *of* T_n^{-1}.

Once p_n, $\tilde{\mathbf{x}}^{(n)}$ and $\tilde{\mathbf{y}}^{(n)}$ are computed, the explicit inverse can be obtained either via Trench's algorithm [12, section 4.7.4] (which we shall not pursue here) or from the first Gohberg–Semencul formula [11]. Using p_n and the scaled vectors $\tilde{\mathbf{x}}^{(n)}$ and $\tilde{\mathbf{y}}^{(n)}$, we obtain:

$$T_n^{-1} = p_n \left[\begin{pmatrix} \tilde{\mathrm{x}}_0^{(n)} & & & 0 \\ \tilde{\mathrm{x}}_1^{(n)} & \ddots & & \\ \vdots & \ddots & \tilde{\mathrm{x}}_0^{(n)} & \\ \tilde{\mathrm{x}}_n^{(n)} & \cdots & \tilde{\mathrm{x}}_1^{(n)} & \tilde{\mathrm{x}}_0^{(n)} \end{pmatrix} \begin{pmatrix} \tilde{\mathrm{y}}_n^{(n)} & \tilde{\mathrm{y}}_{n-1}^{(n)} & \cdots & \tilde{\mathrm{y}}_0^{(n)} \\ & \tilde{\mathrm{y}}_n^{(n)} & \ddots & \vdots \\ & & \ddots & \tilde{\mathrm{y}}_{n-1}^{(n)} \\ 0 & & & \tilde{\mathrm{y}}_n^{(n)} \end{pmatrix} \right.$$
$$\left. - \begin{pmatrix} 0 & & & 0 \\ \tilde{\mathrm{y}}_0^{(n)} & \ddots & & \\ \vdots & \ddots & 0 & \\ \tilde{\mathrm{y}}_{n-1}^{(n)} & \cdots & \tilde{\mathrm{y}}_0^{(n)} & 0 \end{pmatrix} \begin{pmatrix} 0 & \tilde{\mathrm{x}}_n^{(n)} & \cdots & \tilde{\mathrm{x}}_1^{(n)} \\ & 0 & \ddots & \vdots \\ & & \ddots & \tilde{\mathrm{x}}_n^{(n)} \\ 0 & & & 0 \end{pmatrix} \right].$$

The computation of $\tilde{\mathbf{x}}^{(n)}$ and $\tilde{\mathbf{y}}^{(n)}$ requires $4n^2 + \mathcal{O}(n)$ operations. If we want to solve the system $T_n\,\mathbf{x} = \mathbf{b}$, then we need additional $\mathcal{O}(n \log_2 n)$ operations to compute $T_n^{-1}\mathbf{b}$, because the product of a Toeplitz matrix times a vector can be computed via a few fast Fourier transforms (FFTs) [20], [21]. Hence, the solution of the system $T_n\,\mathbf{x} = \mathbf{b}$ requires $4n^2 + \mathcal{O}(n \log_2 n)$ operations.

In some applications the diagonal entries of T_n^{-1} are needed, but not the rest of T_n^{-1}. From the representation of T_n^{-1} and from the relation $\tilde{\mathrm{x}}_0^{(n)} = \tilde{\mathrm{y}}_n^{(n)} = 1$ (cf. (2.1)), it is not difficult to derive an algorithm for computation of these entries. Due to the persymmetry of T_n^{-1}, we need only to compute the first $\lceil n/2 \rceil$ diagonal entries, and the algorithm requires only $\lceil 5n/2 \rceil$ flops; see Figure 2.2.

The only place where the $4n^2$-algorithm can break down is in the computation of the scaling factors p_k, namely, when $\alpha_k\beta_k = 1$. More frequently, as we show below, it becomes dangerous when the quantity $1 - \alpha_k\beta_k$ is very small compared to p_{k-1}. For better understanding of this situation we derive another expression for $1 - \alpha_k\beta_k$.

THEOREM 2.1. *In the absence of breakdown in the* $4n^2$*-algorithm, i.e.,* $p_k \neq 0$, $k = 0, \ldots, n$, *we have*

$$1 - \alpha_k\beta_k = \mathrm{x}_0^{(k-1)}\gamma_k = p_{k-1}\gamma_k, \quad k = 0, \ldots, n, \tag{2.2}$$

where γ_k *is the Schur complement of* T_{k-1} *in the matrix* T_k.

Proof. To prove (2.2), we write the block LU factorization of the matrix T_k in the following way:

$$T_k = \begin{pmatrix} T_{k-1} & \boldsymbol{\tau}_{-k} \\ \boldsymbol{\tau}_k^T & t_0 \end{pmatrix} = \begin{pmatrix} I_k & \mathbf{0} \\ \boldsymbol{\tau}_k^T T_{k-1}^{-1} & 1 \end{pmatrix} \begin{pmatrix} T_{k-1} & \boldsymbol{\tau}_{-k} \\ \mathbf{0}^T & \gamma_k \end{pmatrix},$$

where

$$\boldsymbol{\tau}_{-k} = \begin{pmatrix} t_{-k} \\ \vdots \\ t_{-1} \end{pmatrix}, \qquad \boldsymbol{\tau}_k = \begin{pmatrix} t_k \\ \vdots \\ t_1 \end{pmatrix}, \tag{2.3}$$

and $\gamma_k = t_0 - \boldsymbol{\tau}_k^T T_{k-1}^{-1} \boldsymbol{\tau}_{-k}$ is the Schur complement of T_{k-1}. Next, we need the last component $\mathrm{y}_k^{(k)}$ of the vector $\mathbf{y}^{(k)}$. After some simple manipulations, we obtain that the last component of the $k+1$ vector $T_k \mathbf{y}^{(k)}$ equals $\gamma_k \mathrm{y}_k^{(k)}$. But this quantity is equal to 1 because $\mathbf{y}^{(k)}$ is the last column of T_k^{-1}. Using (2.1) and taking into account the persymmetry of T_k^{-1}, i.e., $\mathrm{x}_0^{(k)} = \mathrm{y}_k^{(k)}$, we then obtain

$$\gamma_k \mathrm{y}_k^{(k)} = \gamma_k \mathrm{x}_0^{(k)} = \gamma_k p_k.$$

But we have that

$$p_k = p_{k-1}/(1 - \alpha_k \beta_k)$$

and, hence,

$$\gamma_k p_{k-1}/(1 - \alpha_k \beta_k) = 1,$$

which leads to (2.2). ∎

From the above theorem it follows that

$$p_k = p_{k-1}/(1 - \alpha_k \beta_k) = 1/\gamma_k. \tag{2.4}$$

Now from (2.4) we see that the algorithm can break down when $\gamma_k = 0$, which is equivalent to T_k's being singular. Moreover, the quantity $|\gamma_k^{-1}|$ is closely connected to $\|T_k^{-1}\|_2$ and therefore to the condition number of T_k; see [10, section 6] for details. Therefore, in practice, trouble sets in when $|\gamma_k|$ becomes small, which happens when $|1 - \alpha_k \beta_k|$ is small relative to $|p_{k-1}|$. In the next section we propose a stabilized modification of the algorithm that avoids small $|\gamma_k|$.

3. The stabilized algorithm. One strategy for avoiding breakdown or near-breakdown in the $4n^2$-algorithm was proposed by Huckle [15]. His strategy is to switch back and forth between two different iteration schemes, of which the second is more expensive and used only when the first is inaccurate.

3.1. Derivation of our algorithm. Our strategy is different and easier to implement: We stay with one iteration scheme and instead perturb γ_k by $\mathrm{Sign}(\gamma_k)\delta$, when necessary, such that $|\gamma_k|$ increases. Here, δ is a small positive number whose magnitude is discussed in section 4.5, and Sign is defined as

$$\mathrm{Sign}(\gamma_k) = \begin{cases} 1 & \text{for } \gamma_k \geq 0, \\ -1 & \text{for } \gamma_k < 0. \end{cases}$$

This approach is inspired by similar perturbations used to "stabilize" Strassen's recursive matrix inversion algorithm in [3]. In order to analyze our approach, we express the perturbation of γ_k as a perturbation in α_k and β_k. Suppose that the perturbed quantities are

$$\hat{\alpha}_k = \alpha_k + \delta_{\alpha_k}, \qquad \hat{\beta}_k = \beta_k + \delta_{\beta_k}$$

and that $\omega_k = 1 - \alpha_k\beta_k$. Then from the identity

$$1 - \hat{\alpha}_k\hat{\beta}_k = (\gamma_k + \operatorname{Sign}(\gamma_k)\delta)p_{k-1} \tag{3.1}$$

we obtain

$$\alpha_k\delta_{\beta_k} + \beta_k\delta_{\alpha_k} + \delta_{\alpha_k}\delta_{\beta_k} = -\operatorname{Sign}(\gamma_k)p_{k-1}\delta. \tag{3.2}$$

This equation is nonlinear with respect to δ_{α_k} and δ_{β_k}. It has an infinite number of solutions, and we prefer a simple solution. Therefore, we choose either

$$\delta_{\alpha_k} = -\operatorname{Sign}(\gamma_k)\delta/g_k, \qquad \delta_{\beta_k} = 0 \tag{3.3}$$

or

$$\delta_{\alpha_k} = 0, \qquad \delta_{\beta_k} = -\operatorname{Sign}(\gamma_k)\delta/f_k. \tag{3.4}$$

In this way we transfer the perturbation of γ_k to a perturbation of α_k or β_k. Whether to use (3.3) or (3.4) is explained below.

We shall now show that the perturbation of δ_{α_k} in (3.3) cannot change α_k dramatically (a similar analysis holds for β_k). Notice that if β_k is zero or very small, then a perturbation is not necessary. Therefore, we can assume that

$$|\beta_k| \geq c > 0, \tag{3.5}$$

where c is a constant. From (3.1) we obtain

$$|\hat{p}_k| = |p_{k-1}/(1 - \hat{\alpha}_k\hat{\beta}_k)| = |1/(\gamma_k + \operatorname{Sign}(\gamma_k)\delta)| \leq 1/\delta. \tag{3.6}$$

Then it follows that

$$\hat{\alpha}_k = \alpha_k - \operatorname{Sign}(\gamma_k)\delta/g_k = \alpha_k - \operatorname{Sign}(\gamma_k)\delta p_{k-1}/\beta_k,$$

where the quantity p_{k-1} from the previous step is either perturbed or unperturbed. If p_{k-1} is unperturbed, then $|\gamma_{k-1}| \geq \delta$, and it follows from (2.4) that $|p_{k-1}| \leq \delta^{-1}$. If p_{k-1} is perturbed, then (11) implies that $|p_{k-1}| \leq \delta^{-1}$. Hence, in both cases we obtain the bound

$$|\alpha_k| + \delta|p_{k-1}|/c \leq |\alpha_k| + 1/c.$$

We see that $\hat{\alpha}_k$ can increase only mildly as long as $|\beta_k|$ is not too small. This analysis shows that we should always choose to perturb the smaller of $|\alpha_k|$ and $|\beta_k|$.

For the analysis in the next section we need to express the perturbation as a perturbation of T_n. Assume again that α_k is perturbed (a "symmetric" analysis holds for β_k). The perturbation of α_k is achieved by perturbing the matrix element t_k. To derive the size of the perturbation Δt_k, we equate $\hat{\alpha}_k$ with $p_{k-1}\hat{f}_k$, where

$$\hat{f}_k = (t_k + \Delta t_k)\tilde{\mathrm{x}}_0^{(k-1)} + \sum_{j=1}^{k-1} t_{k-j}\tilde{\mathrm{x}}_j^{(k-1)}$$

to obtain

$$\Delta t_k = \frac{\delta_{\alpha_k}}{p_{k-1}\tilde{\mathrm{x}}_0^{(k-1)}} = -\frac{\operatorname{Sign}(\gamma_k)\delta}{p_{k-1}\tilde{\mathrm{x}}_0^{(k-1)}g_k} = -\frac{\operatorname{Sign}(\gamma_k)\delta}{\beta_k}.$$

A perturbation of β_k corresponds to a perturbation of t_{-k} by $\Delta t_{-k} = -\operatorname{Sign}(\gamma_k)\delta/\alpha_k$.

Obviously, the perturbation of T_n at stage k of the algorithm is a Toeplitz matrix with a single nonzero off-diagonal at position k below (or above) the diagonal with elements Δt_k (or Δt_{-k}), and the ∞-norm of this matrix equals $|\Delta t_k|$ (or $|\Delta t_{-k}|$). The total perturbation matrix ΔT_n of T_n, after completion of the algorithm and possibly a number of intermediate perturbations, is the sum of the individual perturbation matrices. From (3.5) we have

$$\|\Delta T_n\|_\infty \le \ell\,\delta/c.$$

Here, $\ell \le n+1$ is the number of perturbation matrices; usually we have $\ell = \mathcal{O}(1)$ and $\ell = 1$ in all our numerical examples.

If, instead of perturbing t_k explicitly, we merely perturb α_k and keep using the original T_n as input (which would make the implementation easier), then it can be shown that we compute the inverse of a Toeplitz matrix $\tilde{T}_n$ that differs from T_n in all elements on and below the kth diagonal. This $\tilde{T}_n$ can differ significantly from T_n, and therefore it is not a useful approach.

We need extra storage for the perturbed matrix, because once an entry has been perturbed it is this perturbed entry that is used in the remaining steps of the algorithm. In principle, we could save only the perturbed entries along with information about which entries were perturbed, but doing so requires some bookkeeping which is not easy to pipeline. Instead, we suggest a much simpler solution, namely, to use an additional array of length $2n-1$ to store the perturbed matrix, of which most of the elements are identical to those of the unperturbed matrix. If iterative refinement is not needed, then the perturbed matrix can overwrite the original matrix, and no additional storage is needed.

3.2. Summary of the stabilized algorithm. The stabilized algorithm—without the refinement step—is summarized in Figure 3.1. Notice that the parameter δ acts both as a threshold for deciding when to add a small perturbation, and as a parameter controlling the size of this perturbation. By adding the perturbation we avoid division by a small number, such that p_k stays reasonably small. The choice of the parameter δ is discussed in the next section.

In practice we do not want to perturb the data unnecessarily. If we add a perturbation at some step of the algorithm, then in subsequent steps the quantity ω_k/p_{k-1} might be slightly smaller than the threshold δ, and therefore we should again add the same perturbation. To avoid such unnecessary perturbations, we reduce the threshold δ by some factor once we have added a perturbation. In our implementation, we use a factor 0.1 (which was also used by Chan and Hansen [6] and found to work well in their algorithm). This approach also works well in our algorithm, and in all our examples the perturbation is added only once.

3.3. Iterative refinement. Our perturbations lead to an inaccurate solution that can be made more accurate by taking a few steps of iterative refinement [12, section 3.5.3], which is relatively cheap for Toeplitz systems. The output of our algorithm is a representation of the matrix $\hat{T}_n^{-1}$ which approximates the inverse T_n^{-1}, and during the iterative refinement we need to compute the products of T_n and $\hat{T}_n^{-1}$ times a vector. This can be done relatively fast because such products can be computed in $\mathcal{O}(n \log_2 n)$ operations using FFTs; see [1], [20, section 4.2.4], [21].

The stabilization overhead is at most $\mathcal{O}(n)$ operations if stabilization is used in every step. In practice, even in the case of successive singular submatrices, only the

if $(|t_0| < \delta)$
  $t_0 = t_0 + \mathrm{Sign}(t_0)\delta$;
  $\delta = 0.1\delta$;
end
$\tilde{\mathbf{x}}^{(0)} = \tilde{\mathbf{y}}^{(0)} = 1$; $p_0 = 1/t_0$;
for $k = 1, \ldots, n$
  $f_k = t_k \tilde{\mathrm{x}}_0^{(k-1)} + \cdots + t_1 \tilde{\mathrm{x}}_{k-1}^{(k-1)}$;
  $g_k = t_{-1} \tilde{\mathrm{y}}_0^{(k-1)} + \cdots + t_{-k} \tilde{\mathrm{y}}_{k-1}^{(k-1)}$;
  $\alpha_k = p_{k-1} f_k$; $\beta_k = p_{k-1} g_k$;
  $\omega_k = 1 - \alpha_k \beta_k$;
  if $(|\omega_k / p_{k-1}| < \delta)$
    if $(|\alpha_k| < |\beta_k|\ )$
      $\alpha_k = \alpha_k - \mathrm{Sign}(\gamma_k)\delta / g_k$;
      $t_k = t_k - \mathrm{Sign}(\gamma_k)\delta / \beta_k$;
    else
      $\beta_k = \beta_k - \mathrm{Sign}(\gamma_k)\delta / f_k$;
      $t_{-k} = t_{-k} - \mathrm{Sign}(\gamma_k)\delta / \alpha_k$;
    end
    $\omega_k = 1 - \alpha_k \beta_k$;
    $\delta = 0.1\delta$;
  end
  $p_k = p_{k-1} / \omega_k$;
  $\tilde{\mathbf{x}}^{(k)} = \begin{pmatrix} \tilde{\mathbf{x}}^{(k-1)} \\ 0 \end{pmatrix} - \alpha_k \begin{pmatrix} 0 \\ \tilde{\mathbf{y}}^{(k-1)} \end{pmatrix}$;
  $\tilde{\mathbf{y}}^{(k)} = -\beta_k \begin{pmatrix} \tilde{\mathbf{x}}^{(k-1)} \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \tilde{\mathbf{y}}^{(k-1)} \end{pmatrix}$;
end

FIG. 3.1. *The stabilized algorithm.*

first singular submatrix is usually perturbed, while the subsequent ones do not need to be perturbed. Thus, in most cases we need only $\mathcal{O}(1)$ operations for the stabilization. The operations count for the whole algorithm including iterative refinement is therefore $4n^2 + \mathcal{O}(n \log_2 n)$, where the constant in front of $n \log_2 n$ depends on the number of refinement steps. Our experience is that two or three iterations are enough in most of the cases.

Note that for symmetric Toeplitz matrices, only the first column is needed for representation of the inverse, and the number of operations is halved. The stabilization part and the error analysis are analogous, and therefore we do not discuss the symmetric case here.

4. Error analysis. Error analysis of Toeplitz algorithms is quite rare in the literature—the classical paper is Cybenko's analysis in [8] of Levinson's algorithm; see also Bunch [5]. One reason for this is probably that the standard "tools" of classical error analysis cannot immediately be applied to the highly structured Toeplitz algorithms; a specialized analysis is required.

Our error analysis is based on a method developed by Voevodin and Yalamov [22], which uses the concept of equivalent perturbations for each piece of data in the algorithm: input, intermediate, and output. Our main motivation for using this approach

is that it works, i.e., it is possible to derive useful results without too many tedious computations (although we do not claim that our analysis is simple).

We perform a detailed analysis of one step of the algorithm for a fixed k from which the overall error in the computed solution can be determined.

4.1. Equivalent perturbations in step k. In the kth step of the algorithm, we write the relations between the rounding errors and the equivalent perturbations in the data in the following linear form, where all second-order term are neglected:

$$B_k \boldsymbol{\epsilon}_k = \boldsymbol{\eta}_k. \tag{4.1}$$

Here B_k is a sparse matrix consisting of the Fréchét derivatives of all the local operations in step k, $\boldsymbol{\epsilon}_k$ is a vector of all the equivalent perturbations in step k, and $\boldsymbol{\eta}_k$ is a vector of all the corresponding rounding errors in step k . The details of this system are given below.

Solving the underdetermined system (4.1), we obtain bounds for the equivalent perturbations in step k. If the equivalent perturbations of all outputs are zero, then we have backward analysis; if the equivalent perturbations of all inputs are zero, we have forward analysis; and if some of the inputs and some of the outputs have nonzero perturbations, the analysis is mixed. In this paper we use mixed analysis to study the roundoff error propagation, and our goal is to bound the equivalent perturbations of both T_n and the solution vectors.

Our analysis is nonstructured, i.e., the entries t_k and t_{-k} are considered as different entries at each step of the algorithm. Throughout our analysis we assume that $p_k \neq 0$ and $1 - \alpha_k\beta_k \neq 0$ for $k = 0, \ldots, n$. These assumptions are valid for our stabilized algorithm because we always perturb the denominators $1-\alpha_k\beta_k$ if they are zero or very small.

The vector $\boldsymbol{\eta}_k$ in (4.1) consists of all the local rounding errors involved in computing the scalars f_k, g_k, α_k, β_k, p_k and the vectors $\tilde{\mathbf{x}}^{(k)}$ and $\tilde{\mathbf{y}}^{(k)}$:

$$\boldsymbol{\eta}_k^T = (\eta_{f_k}, \eta_{g_k}, \eta_{\alpha_k}, \eta_{\beta_k}, \eta_{p_k}, \boldsymbol{\eta}_{\tilde{\mathbf{x}}^{(k)}}^T, \boldsymbol{\eta}_{\tilde{\mathbf{y}}^{(k)}}^T). \tag{4.2}$$

Both $\boldsymbol{\eta}_{\tilde{\mathbf{x}}^{(k)}}$ and $\boldsymbol{\eta}_{\tilde{\mathbf{y}}^{(k)}}$ are vectors of length $k+1$, and thus $\boldsymbol{\eta}_k$ is a vector of length $2k+7$. The vector $\boldsymbol{\epsilon}_k$ in (4.1) consists of the equivalent perturbations of the following data: $\boldsymbol{\tau}_{k-1}$, $\boldsymbol{\tau}_{-(k-1)}$, $\tilde{\mathbf{x}}^{(k-1)}$, $\tilde{\mathbf{y}}^{(k-1)}$, t_k, t_{-k}, f_k, g_k, p_{k-1}, α_k, β_k, p_k, $\tilde{\mathbf{x}}^{(k)}$, and $\tilde{\mathbf{y}}^{(k)}$, where the vectors $\boldsymbol{\tau}_{k-1}$ and $\boldsymbol{\tau}_{-(k-1)}$ are defined in (2.3). Thus, $\boldsymbol{\epsilon}_k$ has length $6k + 8$ and has the form

$$\boldsymbol{\epsilon}_k^T = (\boldsymbol{\epsilon}_{\boldsymbol{\tau}_{k-1}}, \boldsymbol{\epsilon}_{\boldsymbol{\tau}_{-(k-1)}}, \ldots, \boldsymbol{\epsilon}_{\tilde{\boldsymbol{y}}^{(k)}}^T). \tag{4.3}$$

Finally, the matrix B_k is very sparse, and we partition it as follows:

$$B_k = \begin{pmatrix} C_k & D_k & F_k & J_k & 0 \\ 0 & P_k & Q_k & 0 & -I_{2k+2} \end{pmatrix} \in \Re^{(2k+7)\times(6k+8)},$$

where I_{2k+2} is the identity matrix of order $2k+2$, and the other submatrices are

$$C_k = \begin{pmatrix} \tilde{\mathrm{x}}_1^{(k-1)} & \cdots & \tilde{\mathrm{x}}_{k-1}^{(k-1)} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \tilde{\mathrm{y}}_0^{(k-1)} & \cdots & \tilde{\mathrm{y}}_{k-2}^{(k-1)} \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix} \in \Re^{5\times 2(k-1)},$$

$$D_k = \begin{pmatrix} t_k & \cdots & t_1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & t_{-1} & \cdots & t_{-k} \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix} \in \Re^{5\times 2k},$$

$$F_k = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & p_{k-1} & 0 & f_k & -1 & 0 \\ 0 & 0 & 0 & p_{k-1} & g_k & 0 & -1 \\ 0 & 0 & 0 & 0 & \frac{1}{1-\alpha_k\beta_k} & \frac{p_{k-1}\beta_k}{(1-\alpha_k\beta_k)^2} & \frac{p_{k-1}\alpha_k}{(1-\alpha_k\beta_k)^2} \end{pmatrix} \in \Re^{5\times 7},$$

$$J_k = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ -1 \end{pmatrix} \in \Re^{5\times 1},$$

$$P_k = \begin{pmatrix} 1 & & & 0 & \cdots & 0 \\ & \ddots & & -\alpha_k & & \\ & & 1 & & \ddots & \\ 0 & \cdots & 0 & & & -\alpha_k \\ -\beta_k & & & 0 & \cdots & 0 \\ & \ddots & & 1 & & \\ & & -\beta_k & & \ddots & \\ 0 & \cdots & 0 & & & 1 \end{pmatrix} \in \Re^{(2k+2)\times 2k},$$

$$Q_k = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\tilde{\mathbf{y}}_{0:\,k-2}^{(k-1)} & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\tilde{\mathbf{x}}_{1:\,k-1}^{(k-1)} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \Re^{(2k+2)\times 7}.$$

The first three block columns of B_k correspond to the input ($\boldsymbol{\tau}_{k-1}$, $\boldsymbol{\tau}_{-(k-1)}$, $\tilde{\mathbf{x}}^{(k-1)}$, $\tilde{\mathbf{y}}^{(k-1)}$, t_k, t_{-k}, p_{k-1}) and the intermediate quantities (f_k, g_k, α_k, β_k) of step k, and the last two block columns correspond to the output (p_k, $\tilde{\mathbf{x}}^{(k)}$, $\tilde{\mathbf{y}}^{(k)}$) of the step. We have used the fact that $\tilde{\mathrm{x}}_0^{(k)} = \tilde{\mathrm{y}}_k^{(k)} = 1$, and the corresponding entries of B_k are replaced by 1.

4.2. The local rounding errors. Now we consider the local absolute rounding errors in the computation of each output. Let ϵ_M denote the machine precision. Then simple roundoff analysis [14, p. 69] of the computation of f_k, g_k, α_k, β_k, p_k, $\tilde{\mathrm{x}}_i^{(k)}$, and $\tilde{\mathrm{y}}_i^{(k)}$ gives (ignoring terms of higher order in ρ_0)

$$
\begin{aligned}
\eta_{f_k} &= \sum_{i=0}^{k-1} t_{k-i}\tilde{\mathrm{x}}_i^{(k-1)}\sigma_i, \qquad |\sigma_i| \le k\epsilon_M, \quad i = 0, \ldots, k-1, \\
\eta_{g_k} &= \sum_{i=1}^{k} t_{-i}\tilde{\mathrm{y}}_{i-1}^{(k-1)}\mu_i, \qquad |\mu_i| \le k\epsilon_M, \quad i = 0, \ldots, k-1, \\
\eta_{\alpha_k} &= \alpha_k\zeta_1, \quad \eta_{\beta_k} = \beta_k\zeta_2, \qquad |\zeta_1| \le \epsilon_M, \quad |\zeta_2| \le \epsilon_M, \\
\eta_{p_k} &= \frac{p_{k-1}}{1-\alpha_k\beta_k}\nu_1 + \frac{p_{k-1}\alpha_k\beta_k}{(1-\alpha_k\beta_k)^2}\nu_2, \qquad |\nu_1| \le 2\epsilon_M, \quad |\nu_2| \le \epsilon_M,
\end{aligned} \tag{4.4}
$$

$$
\left.
\begin{aligned}
(\boldsymbol{\eta}_{\tilde{\mathbf{x}}^{(k)}})_i &= \tilde{\mathrm{x}}_i^{(k-1)}\rho_1^{(i)} + \alpha_k\tilde{\mathrm{y}}_{i-1}^{(k-1)}\rho_3^{(i)}, \\
(\boldsymbol{\eta}_{\tilde{\mathbf{y}}^{(k)}})_i &= \beta_k\tilde{\mathrm{x}}_i^{(k-1)}\rho_4^{(i)} + \tilde{\mathrm{y}}_{i-1}^{(k-1)}\rho_2^{(i)}, \\
|\rho_1^{(i)}| \le \epsilon_M, \quad |\rho_2^{(i)}| &\le \epsilon_M, \quad |\rho_3^{(i)}| \le 2\epsilon_M, \quad |\rho_4^{(i)}| \le 2\epsilon_M,
\end{aligned}
\right\} \quad i = 0, \ldots, k, \tag{4.5}
$$

where the quantities

$$
\tilde{\mathrm{y}}_{-1}^{(k-1)} = \tilde{\mathrm{x}}_k^{(k-1)} = \rho_1^{(0)} = \rho_2^{(k)} = \rho_3^{(k)} = \rho_4^{(0)} = 0
$$

are introduced to make the above notation more compact.

4.3. Simplification of the system. As already mentioned, the system in (4.1) is underdetermined, and we can simplify the system by fixing some of the equivalent perturbations. In each step k we choose

$$
\boldsymbol{\epsilon}_{\boldsymbol{\tau}_{k-1}} = \boldsymbol{\epsilon}_{\boldsymbol{\tau}_{-(k-1)}} = \mathbf{0}, \qquad \epsilon_{\alpha_k} = \alpha_k\nu_2/2, \qquad \epsilon_{\beta_k} = \beta_k\nu_2/2.
$$

With this simplification, the entries $\boldsymbol{\epsilon}_{\boldsymbol{\tau}_{k-1}}$, $\boldsymbol{\epsilon}_{\boldsymbol{\tau}_{-(k-1)}}$, ϵ_{α_k}, and ϵ_{β_k} are removed from the vector $\boldsymbol{\epsilon}_k$, the block column $\begin{pmatrix} C_k \\ 0 \end{pmatrix}$ is removed from B_k, and the last two columns of F_k are removed. Moreover, a few upper bounds are altered,

$$
|\zeta_1| \le 1.5\epsilon_M, \quad |\zeta_2| \le 1.5\epsilon_M, \quad |\rho_3^{(i)}| \le 2.5\epsilon_M, \quad |\rho_4^{(i)}| \le 2.5\epsilon_M,
$$

and the expression for η_{p_k} becomes

$$
\eta_{p_k} = \frac{p_{k-1}}{1-\alpha_k\beta_k}\nu_1, \quad |\nu_1| \le 2\epsilon_M.
$$

It is this modified system that we use to derive our bounds.

4.4. Derivation of the bounds. From (4.1) we see that if the errors $\epsilon_{\tilde{\mathbf{x}}^{(k-1)}}$ and $\epsilon_{\tilde{\mathbf{y}}^{(k-1)}}$ are given, then we can compute $\epsilon_{\tilde{\mathbf{x}}^{(k)}}$ and $\epsilon_{\tilde{\mathbf{y}}^{(k)}}$ from the second block equation as follows:

$$
\begin{pmatrix} \boldsymbol{\epsilon}_{\tilde{\mathbf{x}}^{(k)}} \\ \boldsymbol{\epsilon}_{\tilde{\mathbf{y}}^{(k)}} \end{pmatrix} = P_k \begin{pmatrix} \boldsymbol{\epsilon}_{\tilde{\mathbf{x}}^{(k-1)}} \\ \boldsymbol{\epsilon}_{\tilde{\mathbf{y}}^{(k-1)}} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\eta}_{\tilde{\mathbf{x}}^{(k)}} \\ \boldsymbol{\eta}_{\tilde{\mathbf{y}}^{(k)}} \end{pmatrix}. \tag{4.6}
$$

This means that we can do forward analysis for the pairs $\tilde{\mathbf{x}}^{(k)}$, $\tilde{\mathbf{y}}^{(k)}$ because their equivalent perturbations are defined independently of the rest of the equivalent perturbations in (4.1). Applying (4.6) recursively, we obtain

$$\begin{pmatrix} \boldsymbol{\epsilon}_{\tilde{\mathbf{x}}^{(k)}} \\ \boldsymbol{\epsilon}_{\tilde{y}^{(k)}} \end{pmatrix} = -\sum_{i=1}^{k} \left(\prod_{j=k}^{i+1} P_j \right) \begin{pmatrix} \boldsymbol{\eta}_{\tilde{\mathbf{x}}^{(i)}} \\ \boldsymbol{\eta}_{\tilde{\mathbf{y}}^{(i)}} \end{pmatrix}, \quad k = 1, \ldots, n. \tag{4.7}$$

Here we assume that $\boldsymbol{\epsilon}_{\tilde{\mathbf{x}}^{(0)}} = \boldsymbol{\epsilon}_{\tilde{\mathbf{y}}^{(0)}} = 0$. Now we bound the local errors $\boldsymbol{\eta}_{\tilde{\mathbf{x}}^{(i)}}$ and $\boldsymbol{\eta}_{\tilde{\mathbf{y}}^{(i)}}$ using standard notation from componentwise analysis (see, e.g., [14]). From (4.5) we have

$$\begin{pmatrix} |\boldsymbol{\eta}_{\tilde{\mathbf{x}}^{(k)}}| \\ |\boldsymbol{\eta}_{\tilde{\mathbf{y}}^{(k)}}| \end{pmatrix} \leq \; |P_k| \begin{pmatrix} |\tilde{\mathbf{x}}^{(k-1)}| \\ |\tilde{\mathbf{y}}^{(k-1)}| \end{pmatrix} 2.5\epsilon_M. \tag{4.8}$$

From the updating equations for $\tilde{\mathbf{x}}^{(k)}$ and $\tilde{\mathbf{y}}^{(k)}$ (the last two statements in the algorithms), we find that

$$\begin{pmatrix} \tilde{\mathbf{x}}^{(k-1)} \\ \tilde{\mathbf{y}}^{(k-1)} \end{pmatrix} = P_{k-1} \begin{pmatrix} \tilde{\mathbf{x}}^{(k-2)} \\ \tilde{\mathbf{y}}^{(k-2)} \end{pmatrix} = P_{k-1} \cdots P_1, \tag{4.9}$$

where

$$P_1 = \begin{pmatrix} \tilde{\mathbf{x}}^{(0)} \\ \tilde{\mathbf{y}}^{(0)} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{4.10}$$

Now (4.7), (4.8), and (4.9) imply that

$$\begin{pmatrix} |\boldsymbol{\epsilon}_{\tilde{\mathbf{x}}^{(k)}}| \\ |\boldsymbol{\epsilon}_{\tilde{\mathbf{y}}^{(k)}}| \end{pmatrix} \leq 2.5\epsilon_M \sum_{i=1}^{k} \left| \prod_{j=k}^{i+1} P_j \right| |P_i| \left| \prod_{j=i-1}^{1} P_j \right|, \tag{4.11}$$

and, hence,

$$\|\boldsymbol{\epsilon}_{\tilde{\mathbf{x}}^{(k)}}\|_\infty \leq 2.5\pi_k k\epsilon_M, \qquad \|\boldsymbol{\epsilon}_{\tilde{\mathbf{y}}^{(k)}}\|_\infty \leq 2.5\pi_k k\epsilon_M, \tag{4.12}$$

where we have defined the scalar

$$\pi_k \equiv \max_{1 \leq i \leq k} \left\| \left| \prod_{j=k}^{i+1} P_j \right| |P_i| \left| \prod_{j=i-1}^{1} P_j \right| \right\|_\infty. \tag{4.13}$$

Hence, we have bounded the equivalent perturbations of $\tilde{\mathbf{x}}^{(k)}$ and $\tilde{\mathbf{y}}^{(k)}$ by means of forward analysis. As we shall see, π_k appears in the final bounds. Unfortunately, there is no simple way to analyze or bound π_k, and we shall therefore resort to numerical experiments; cf. (4.22).

Having obtained the equivalent output perturbations $\boldsymbol{\epsilon}_{\tilde{\mathbf{x}}^{(k-1)}}$ and $\boldsymbol{\epsilon}_{\tilde{\mathbf{y}}^{(k-1)}}$ from the previous step, we obtain the equivalent input perturbations $\epsilon_{p_{k-1}}$, ϵ_{t_k}, and $\epsilon_{t_{-k}}$ by means of backward error analysis. After inverting F_k (which is possible because of our assumption that $p_{k-1} \neq 0$ and $1 - \alpha_k\beta_k \neq 0$), we get

$$\epsilon_{p_{k-1}} = (1 - \alpha_k\beta_k)\epsilon_{p_k} + p_{k-1}\nu_1 = p_{k-1}(\rho_{p_k} + \nu_1),$$

where we have used the fact that $p_k = p_{k-1}/(1-\alpha_k\beta_k)$, and $\rho_{p_k} = \epsilon_{p_k}/p_k$ is the relative equivalent perturbation of p_k. Hence,

$$\rho_{p_{k-1}} = \rho_{p_k} + \nu_1, \qquad |\nu_1| \le 2\epsilon_M,$$

which, when applied recursively to the previous steps, gives

$$|\rho_{p_{k-1}}| \le 2(n-k+1)\epsilon_M. \tag{4.14}$$

Analogously, we obtain

$$|\rho_{f_k}| \le |\zeta_1| + |\rho_{p_{k-1}}| \le (2n-2k+3.5)\epsilon_M, \tag{4.15}$$

$$|\rho_{g_k}| \le |\zeta_2| + |\rho_{p_{k-1}}| \le (2n-2k+3.5)\epsilon_M.$$

Finally, from (4.4), (4.12), and (4.15) we have

$$\begin{aligned}
|\epsilon_{t_k}| &\le |\eta_{f_k}| + |\epsilon_{f_k}| + \sum_{i=1}^{k-1} |t_{k-i}(\boldsymbol{\epsilon}_{\tilde{\mathbf{x}}^{(k-1)}})_i\| \\
&\le k\sum_{i=0}^{k-1} |t_{k-i}\tilde{\mathrm{x}}_i^{(k-1)}|\epsilon_M + (2n-2k+3.5)\sum_{i=0}^{k-1} |t_{k-i}\tilde{\mathrm{x}}_i^{(k)}|\epsilon_M + \sum_{i=1}^{k-1} |t_{k-i}(\boldsymbol{\epsilon}_{\tilde{\mathbf{x}}^{(k-1)}})_i| \\
&\le (2n+3.5)\|\tilde{\mathbf{x}}^{(k-1)}\|_\infty \|T_k\|_\infty \epsilon_M + 2.5\pi_k\|T_k\|_\infty\epsilon_M.
\end{aligned}$$

Here, we need to bound $\|\tilde{\mathbf{x}}^{(k-1)}\|_\infty$. From (4.9) we obtain

$$\left\| \begin{pmatrix} \tilde{\mathbf{x}}^{(k-1)} \\ \tilde{\mathbf{y}}^{(k-1)} \end{pmatrix} \right\|_\infty = \|P_{k-1}\cdots P_1\|_\infty,$$

and it follows from the definition of π_k in (4.13) that π_k is an upper bound for the above norm and therefore also for $\|\tilde{\mathbf{x}}^{(k-1)}\|_\infty$. Hence, we get

$$|\epsilon_{t_k}| \le 2(n+3)\pi_k\|T_k\|_\infty\epsilon_M. \tag{4.16}$$

Analogously, we obtain

$$|\epsilon_{t_{-k}}| \le 2(n+3)\pi_k\|T_k\|_\infty\epsilon_M. \tag{4.17}$$

It remains to analyze the first step of the algorithm, i.e., $p_0 = 1/t_0$. For this operation we have $\epsilon_{t_0}/t_0^2 - \epsilon_{p_0} = \eta_{p_0}$ in the absolute case, and $-\rho_{t_0} - \rho_{p_0} = \nu$, $|\nu| \le \epsilon_M$ in the relative case. Then from (4.14) we get

$$|\rho_{t_0}| \le (2n+1)\epsilon_M, \tag{4.18}$$

showing that the perturbation of t_0 is very small.

If $\pi_k < 1$ for all k, it is clear that all the bounds are small. Now we shall suppose that $\pi_k \ge 1$, which is the interesting case. From (4.16), (4.17), and (4.18) we obtain the following bound for the equavalent perturbation E_{T_n} of the matrix T_n:

$$\|E_{T_n}\|_\infty \le 2(n^2+3n)\pi\|T_n\|_\infty\epsilon_M, \quad \pi = \max_k \pi_k,$$

which is a normwise bound.

The solution x of the system $T_n\mathbf{x} = \mathbf{b}$ is computed from $\tilde{\mathbf{x}}^{(n)}$, $\tilde{\mathbf{y}}^{(n)}$, p_n, and $\mathbf{b}$ via FFTs. It is well-known that the FFT is very stable [14, section 23.1], and therefore we do not analyze these transformations here but instead use the following bound:

$$\|\boldsymbol{\epsilon}_{\mathbf{x}}\|_\infty \le c_n \, \|T_n^{-1}\|_\infty \, \|\mathbf{b}\|_\infty \, \epsilon_M,$$

where c_n is a polynomial in n.

4.5. Main result and discussion. We summarize our analysis as follows.

THEOREM 4.1. *For the $4n^2$-algorithm, we have*

$$(T_n + E_{T_n})(\hat{\mathbf{x}} + \boldsymbol{\epsilon}_{\mathbf{x}}) = \mathbf{b},$$

where $\hat{\mathbf{x}}$ is the computed solution of $T_n\mathbf{x} = \mathbf{b}$ and

$$\begin{aligned} \|E_{T_n}\|_\infty &\le 2(n^2 + 3n)\,\pi \|T_n\|_\infty\, \epsilon_M, \\ \|\boldsymbol{\epsilon}_{\mathbf{x}}\|_\infty &\le c_n\,\pi\, \|T_n^{-1}\|_\infty\, \|\mathbf{b}\|_\infty\, \epsilon_M, \\ \pi &= \max_k \pi_k. \end{aligned}$$

Here, π_k is defined in (4.13)*, ϵ_M is the machine precision, and c_n is a polynomial in n.*

Theorem 4.1 presents a mixed error analysis for both the original and the stabilized algorithms. Now let us consider the stabilized algorithm in more detail. By explicitly perturbing some of the entries t_k or t_{-k}, we change the inputs as noted in section 3. Instead of solving a system with the matrix T_n, we solve a system with the matrix $\hat{T}_n = T_n + \Delta T_n$, where

$$\|\Delta T_n\|_\infty \le (\ell/c)\,\delta. \tag{4.19}$$

For this system we have

$$(\hat{T}_n + E_{\hat{T}_n})(\hat{\mathbf{x}} + \boldsymbol{\epsilon}_{\mathbf{x}}) = \mathbf{b}, \tag{4.20}$$

where $E_{\hat{T}_n}$ and $\boldsymbol{\epsilon}_{\mathbf{x}}$ are bounded in Theorem 4.1, with T_n replaced by $\hat{T}_n$.

Let $\Delta\mathbf{x} = \hat{\mathbf{x}} - \mathbf{x}$, where $\mathbf{x}$ is the exact solution of the unperturbed problem $T_n\mathbf{x} = \mathbf{b}$. Then after some standard manipulation, and ignoring higher order terms, we get from (4.20) that

$$\Delta\mathbf{x} = -T_n^{-1}(\Delta T_n + E_{\hat{T}_n})\mathbf{x} - \boldsymbol{\epsilon}_{\mathbf{x}}.$$

Provided that $\|T_n + \Delta T_n\|_\infty \le 2\,\|T_n\|_\infty$ (which is true for our choice of δ), Theorem 4.1 and (4.19) then lead to the bound

$$\begin{aligned} \|\Delta\mathbf{x}\|_\infty &\le (\ell/c)\,\|T_n^{-1}\|_\infty \|\mathbf{x}\|_\infty \delta + 4(n^2+3n)\pi \|T_n^{-1}\|_\infty \|T_n\|_\infty \|\mathbf{x}\|_\infty \epsilon_M \\ &\quad +\ c_n\,\pi\,\|T_n^{-1}\|_\infty\,\|\mathbf{b}\|_\infty\,\epsilon_M \\ &\le c_n' \|T_n^{-1}\|_\infty \|T_n\|_\infty \|\mathbf{x}\|_\infty (\delta + \pi\epsilon_M), \end{aligned}$$

where c_n' is a new polynomial in n. The relative error is bounded as

$$\frac{\|\Delta\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} \le c_n'\,\kappa(\delta + \pi\epsilon_M), \quad \kappa = \|T_n^{-1}\|_\infty \|T_n\|_\infty. \tag{4.21}$$

The last bound depends on π which, in turn, depends on the growth of the α_k and β_k, and thus on the growth of the scaling factors p_k. Since all p_k are bounded by (3.6) we use a model of the form

$$\pi \le \delta^{-s}, \tag{4.22}$$

where s is a parameter. It is very difficult to find a bound on s. From practical experience (cf. the next section), we find that s usually lies in the range $1 \le s \le 2$.

TABLE 5.1
The forward error $\|\mathbf{x}-\hat{\mathbf{x}}\|_\infty$ and the number of IR steps for the "random" matrices T_n^{CH}.

		$\theta = 10^{-7}$	$\theta = 10^{-10}$	$\theta = 10^{-13}$	$\theta = 0$
Original	Average	$1.26 \cdot 10^{-5}$	$1.34 \cdot 10^{-2}$	6.11	$8.00 \cdot 10^{1}$
before IR	Worst	$2.05 \cdot 10^{-3}$	$9.79 \cdot 10^{-1}$	$8.38 \cdot 10^{2}$	$1.70 \cdot 10^{4}$
Original	Average	$2.00 \cdot 10^{-12}$	$1.79 \cdot 10^{-11}$	$1.06 \cdot 10^{4}$	$4.44 \cdot 10^{4}$
after IR	Worst	$5.66 \cdot 10^{-10}$	$1.62 \cdot 10^{-8}$	$7.21 \cdot 10^{6}$	$2.09 \cdot 10^{7}$
N_{it}	Average	2.12	3.61	14.0	13.9
	Worst	3	10	divergence	divergence
Stabilized	Average	$1.80 \cdot 10^{-5}$	$1.43 \cdot 10^{-4}$	$1.77 \cdot 10^{-4}$	$1.71 \cdot 10^{-4}$
before IR	Worst	$2.05 \cdot 10^{-3}$	$1.56 \cdot 10^{-2}$	$3.02 \cdot 10^{-2}$	$5.18 \cdot 10^{-2}$
Stabilized	Average	$1.81 \cdot 10^{-12}$	$1.92 \cdot 10^{-12}$	$1.92 \cdot 10^{-12}$	$1.85 \cdot 10^{-12}$
after IR	Worst	$5.66 \cdot 10^{-10}$	$5.05 \cdot 10^{-10}$	$5.08 \cdot 10^{-10}$	$6.18 \cdot 10^{-10}$
N_{it}	Average	2.21	2.60	2.59	2.60
	Worst	3	5	5	5

Now from (4.21) and (4.22) we can derive the recommended value of δ. The right-hand side of (4.21) is minimal when

$$\delta = (s\epsilon_M)^{1/(s+1)},$$

and if $1 \le s \le 2$, then

$$\epsilon_M^{\frac{1}{3}} \le \delta \le \epsilon_M^{\frac{1}{2}}.$$

For example, if $\epsilon_M \approx 10^{-16}$ (double precision), then $10^{-5} \le \delta \le 10^{-8}$. In all the experiments, we used the value $\delta = 10^{-6}$, which turns out to be suitable.

This choice of δ assures us that the computed solution of the perturbed problem is not far from the exact solution of the unperturbed problem for well-conditioned matrices T_n (because the bound (4.21) depends on the condition number κ). Hence, the computed solution $\hat{\mathbf{x}}$ is a good starting vector for the iterative procedure which converges very fast (usually in two or three iterations).

5. Numerical experiments. The numerical experiments are done in MATLAB, the roundoff unit is $\epsilon_M \approx 2.22 \cdot 10^{-16}$, and all matrix-vector multiplications are performed with MATLAB's FFT routine. The exact solution in all the examples is $\mathbf{x} = (1, \ldots, 1)^T$, and we measure the forward error $\|\mathbf{x}-\hat{\mathbf{x}}\|_\infty$, where $\hat{\mathbf{x}}$ is the solution computed with rounding errors. We use the standard iterative refinement (IR) procedure from [12, section 3.5.3] with stopping criterion $\|T_n\hat{\mathbf{x}} - \mathbf{b}\|_\infty / \|\mathbf{b}\|_\infty \le 1000\epsilon_M$.

Table 5.1 presents the results for "random" test matrices T_n^{CH} from [6]. For each row of this table we generated 1,000 random matrices of order $n = 500$ in such a way that at least one leading principal submatrix is ill-conditioned. The order k of this ill-conditioned submatrix is chosen randomly in the range $1 \le k \le 250$. The conditioning of the submatrix is controlled by a real parameter θ, which is given in the first row of Table 5.1. In order to generate these test matrices, we first generate a random nonsymmetric Toeplitz matrix T_n with positive entries. Let λ denote the numerically smallest real eigenvalue of the leading principal matrix T_k of T_n (due to Perron's theorem [18], a positive matrix always has at least one real eigenvalue, which is identical to the spectral radius of the matrix). Then we set our test matrix to

$$T_n^{\mathrm{CH}} = T_n - (\lambda - \theta)I_n,$$

TABLE 5.2
The forward error $\|\mathbf{x} - \hat{\mathbf{x}}\|_\infty$ and the number of IR steps for the shifted KMS matrices.

n	Original before IR	Original after IR	N_{it}	Stabilized before IR	Stabilized after IR	N_{it}	Condition number
15	$2.39 \cdot 10^{-3}$	$1.55 \cdot 10^{-15}$	5	$1.50 \cdot 10^{-6}$	$2.22 \cdot 10^{-15}$	3	$2.55 \cdot 10^{1}$
30	$2.82 \cdot 10^{-3}$	$6.77 \cdot 10^{-14}$	6	$2.00 \cdot 10^{-6}$	$2.66 \cdot 10^{-15}$	3	$5.13 \cdot 10^{1}$
60	$2.85 \cdot 10^{-3}$	$2.35 \cdot 10^{-13}$	7	$2.00 \cdot 10^{-6}$	$3.77 \cdot 10^{-15}$	3	$1.01 \cdot 10^{2}$
120	$2.92 \cdot 10^{-3}$	$3.18 \cdot 10^{-13}$	9	$2.00 \cdot 10^{-6}$	$1.63 \cdot 10^{-14}$	3	$2.01 \cdot 10^{2}$
1200	$7.94 \cdot 10^{-3}$	divergence	–	$2.00 \cdot 10^{-6}$	$5.95 \cdot 10^{-14}$	3	$1.98 \cdot 10^{3}$

and we know that the leading principal submatrix T_k^{CH} has an eigenvalue equal to θ. Since the numerically smallest eigenvalue is bounded below by the smallest singular value [12, section 7.1.6], we are guaranteed that the smallest singular value of T_k^{CH} is bounded by θ. In this way we can control the conditioning of T_k^{CH} by means of θ.

Table 5.1 presents average and worst-case errors for the two versions of the algorithm with and without iterative refinement. It is clear that the stabilized algorithm can produce many more correct solutions than the original algorithm. The average number of IR steps is small and almost independent of θ. For $\theta = 0$, the original algorithm breaks down in 12 out of 1,000 cases. Therefore, the results for the original algorithm in the last column of Table 5.1 take into account only the remaining cases where the algorithm did not break down.

The entries N_{it} in Table 5.1 show the average and maximum number of IR steps for the two algorithms. We see that the stabilized algorithm never needs more IR steps than the original algorithm, and usually it needs much fewer IR steps.

Next, we applied the algorithms to three symmetric test matrices. The first is the shifted Kac–Murdock–Szegö matrix T_n^{KMS} from [19] with entries given by

$$t_0 = 10^{-14}, \quad t_i = t_{-i} = (1/2)^{i-1}, \quad i = 1, \ldots, n,$$

for which every third leading principal submatrix T_k^{KMS}, $k = 0, 3, 6, \ldots$, is very ill-conditioned. The results are given in Table 5.2. After IR, the stabilized algorithm produces results which are as accurate as can be expected, and the number of IR steps is constant. In contrast to this, the solutions computed by the original algorithm with IR are less accurate, and the number of IR steps increases with n. For large n, exemplified by the case $n = 1,200$ in Table 5.2, the condition number and the initial error are so large that IR diverges. We conclude that the perturbation in our algorithm is essential for the stability and the robustness of the algorithm.

The second symmetric test matrix T_{311}^{F} is taken from [9]. The entries $t_i = t_{-i}$, $i = 0, \ldots, 311$, of T_{311}^{F} are given as the coefficients of the Taylor expansion

$$f(z) = \sum_{i=0}^{\infty} t_i z^i$$

of the function

$$f(z) = 18 + \frac{1}{1-z} - \frac{3}{1-z^3} + \frac{6}{1-z^6} - \frac{24}{1-z^{24}} + \frac{48}{1-z^{48}} - \frac{96}{1-z^{96}}.$$

All leading principal submatrices are well-conditioned with the exception of the submatrices T_k^{F}, $k = 50, \ldots, 56$, which are exactly singular. The forward error $\|\mathbf{x} - \hat{\mathbf{x}}\|_\infty$ is 0.82 for the original algorithm before IR (it does not break down in the presence

TABLE 5.3
The forward error $\|\mathbf{x} - \hat{\mathbf{x}}\|_\infty$ *and the number of IR steps for the matrix* tridiag $(1, 10^{-13}, 1)$.

n	Original before IR	Original after IR	N_{it}	Stabilized before IR	Stabilized after IR	N_{it}
100	2.27	divergence	–	$8.10 \cdot 10^{-5}$	$6.43 \cdot 10^{-15}$	3
200	2.27	divergence	–	$7.90 \cdot 10^{-5}$	$9.99 \cdot 10^{-15}$	3
400	2.27	divergence	–	$7.87 \cdot 10^{-5}$	$3.14 \cdot 10^{-14}$	3

TABLE 5.4
Number of iterative refinement steps for different choices of δ.

δ	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}
KMS, $n = 120$	10	5	3	3	2
tridiag $(1, 0, 1)$, $n = 400$	10	5	3	3	7
T_{311}, $n = 311$	4	3	3	2	3

of roundoff errors), and the IR process diverges. The forward error for the stabilized algorithm is $3.14 \cdot 10^{-7}$ before IR and $5.24 \cdot 10^{-14}$ after two IR steps.

The third symmetric test matrix is the tridiagonal matrix tridiag $(1, 10^{-13}, 1)$ which is well-conditioned for odd n and almost singular for even n, i.e., each second principal submatrix is almost singular. The results are given in Table 5.3. The original algorithm is not capable of solving this system, while the stabilized algorithm computes solutions to almost full precision after three IR steps.

These numerical tests demonstrate that our stabilized algorithm is always capable of computing a solution to a satisfactory accuracy. In most cases, two or three steps of iterative refinement are enough, and only one perturbation is needed in all the examples independent of the number of singular or ill-conditioned submatrices.

Finally, in Table 5.4 we give some numerical evidence that the range of s is probably $1 \le s \le 2$. The first row shows different choices of δ, and the other rows show the number of IR steps for three test matrices. We see that the minimum number of iterative steps is achieved for δ in the range $10^{-7} \le \delta \le 10^{-5}$. This means that δ probably should be chosen in this range, which is equivalent to s being in the range $1 \le s \le 2$. We did not find examples in which δ is outside the above-mentioned range. We thus recommend to use the value $\delta = 10^{-6}$.

Acknowledgments. We wish to thank Prof. Thomas Huckle for sharing his insights about Toeplitz algorithms, as well as his MATLAB routines, with us. We also thank the referees for many detailed comments.

REFERENCES

[1] G. S. AMMAR AND W. B. GRAGG, *Superfast solution of real positive definite Toeplitz systems*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 61–76.

[2] O. B. ARUSHANIAN, M. K. SAMARIN, V. V. VOEVODIN, E. E. TYRTYSHNIKOV, B. S. GARBOW, J. M. BOYLE, W. R. COWELL, AND K. W. DRITZ, *The Toeplitz Package Users' Guide*, Technical report ANL-83-16, Argonne National Laboratory, October 1983.

[3] S. M. BALLE, P. C. HANSEN AND N. J. HIGHAM, *A Strassen-Type Matrix Inversion Algorithm for the Connection Machine*, Technical report CNC/1993/028, Centre for Novel Computing, University of Manchester, UK, October 1993.

[4] R. BRENT, *Parallel algorithms for Toeplitz systems*, in Numerical Linear Algebra, Digital Signal Processing and Parallel Algorithms, G. H. Golub and P. Van Dooren, eds., Springer-Verlag, Berlin, 1991, pp. 75–92.

[5] J. R. Bunch, *Stability of methods for solving Toeplitz systems of equations*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 349–364.
[6] T. F. Chan and P. C. Hansen, *A look-ahead Levinson algorithm for indefinite Toeplitz systems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 490–506.
[7] P. Concus and P. Saylor, *A modified direct preconditioner for indefinite symmetric Toeplitz systems*, Numer. Linear Algebra Appl., 2 (1995), pp. 415–429.
[8] G. Cybenko, *The numerical stability of the Levinson-Durbin algorithm for Toeplitz systems of equations*, SIAM J. Sci. Statist. Comput., 1 (1980), pp. 303–319.
[9] R. W. Freund, *A look-ahead Bareiss algorithm for general Toeplitz matrices*, Numer. Math., 68 (1994), pp. 35–69.
[10] R. W. Freund and H. Zha, *Formally biorthogonal polynomials and a look-ahead Levinson algorithm for general Toeplitz matrices*, Linear Algebra Appl., 188 (1993), pp. 255–303.
[11] I. Gohberg and A. Semencul, *On the inversion of finite Toeplitz matrices and their continual analogues*, Mat. Issled., 7 (1972), pp. 201–233.
[12] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MO, 1996.
[13] G. Heinig and F. Rost, *Algebraic Methods for Toeplitz and Toeplitz-Like Operators*, Birkhäuser, Cambridge, MA, 1984.
[14] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.
[15] T. Huckle, *Computations with Gohberg–Semencul formulas for Toeplitz Matrices*, Linear Algebra Appl., 273 (1998), pp. 169–198.
[16] J. R. Jain, *An efficient algorithm for a large Toeplitz set of linear equations*, IEEE Trans. Acoust. Speech, Signal Process, ASSP-27 (1979), pp. 612–615.
[17] N. Levinson, *The Wiener RMS (root mean square) error criterion in filter design and prediction*, J. Math. Phys. Mass. Inst. Tech., 25 (1947), pp. 261–278.
[18] J. M. Ortega, *Matrix Theory*, Plenum, New York, 1987.
[19] W. F. Trench, *Numerical solution of the eigenvalue problem for Hermitian Toeplitz matrices*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 135–146.
[20] C. F. Van Loan, *Computational Frameworks for the Fast Fourier Transform*, SIAM, Philadelphia, 1992.
[21] V. V. Voevodin and E. E. Tyrtyshnikov, *Computational Processes with Toeplitz Matrices*, Nauka, Moscow, 1987 (in Russian).
[22] V. V. Voevodin and P. Y. Yalamov, *A new method of roundoff error estimation*, in Parallel and Distributed Processing, K. Boyanoved, ed., Elsevier, Amsterdam, 1990, pp. 315–333.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 3, pp. 867–881
© 2000 Society for Industrial and Applied Mathematics

ASYMPTOTIC SPECTRA OF HERMITIAN BLOCK TOEPLITZ MATRICES AND PRECONDITIONING RESULTS*

M. MIRANDA† AND P. TILLI†

Abstract. We study the asymptotic behavior of the eigenvalues of Hermitian $n \times n$ block Toeplitz matrices T_n, with $k \times k$ blocks, as n tends to infinity. No hypothesis is made concerning the structure of the blocks. Such matrices $\{T_n\}$ are generated by the Fourier coefficients of a Hermitian matrix-valued function $f \in L^2$, and we study the distribution of their eigenvalues for large n, relating their behavior to some properties of f as a function; in particular, we show that the distribution of the eigenvalues converges to a limit μ_f, and we explicitly compute μ_f in terms of f, showing that $\int F\, d\mu_f = 1/k \int \operatorname{Tr} F(f)$. Some consequences of this distribution and some localization results for the eigenvalues of T_n are discussed. We also study the eigenvalues of the preconditioned matrices $\{P_n^{-1}T_n\}$, where the sequence $\{P_n\}$ is generated by a positive definite matrix-valued function p. We show that the spectrum of any $P_n^{-1}T_n$ is contained in the interval $[r, R]$, where r is the smallest and R the largest eigenvalue of $p^{-1}f$. We also prove that the first m eigenvalues of $P_n^{-1}T_n$ tend to r and the last m tend to R, for any fixed m. Finally, the exact limit value of the condition number of the preconditioned matrices is computed.

Key words. block Toeplitz matrix, distribution of eigenvalues, conjugate gradient, preconditioning

AMS subject classification. 65D32

PII. S0895479896313036

1. Introduction.

In this paper we investigate the asymptotic distribution of the eigenvalues of Hermitian block Toeplitz (BT) matrices and we study how the spectra of these matrices are affected if a BT preconditioner is applied. A Hermitian matrix T_n has an $n \times n$ BT structure with $k \times k$ blocks if

$$T_n = \begin{pmatrix} A_0 & A_1 & \dots & A_{n-1} \\ A_{-1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & A_1 \\ A_{-n+1} & \dots & A_{-1} & A_0 \end{pmatrix} \in \mathbf{C}^{nk \times nk}, \tag{1}$$

with

$$A_j \in \mathbf{C}^{k \times k}, \quad A_j = A_{-j}^*, \quad j = -n+1, \dots, n-1. \tag{2}$$

In this paper we consider the case where the blocks A_j are the Fourier coefficients of a $k \times k$ Hermitian matrix-valued function $f : Q \to H^k$, which is integrable on $Q = (-\pi, \pi)$, that is, $f \in L^1(Q, H^k)$ and

$$A_j = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-\hat{\imath} j x}\, dx, \quad j = 0, \pm 1, \pm 2, \dots . \tag{3}$$

For every natural number n, we associate the BT matrix (1) with the Hermitian matrix-valued function f, and we say that $\{T_n\}$ is the set of BT matrices *generated*

*Received by the editors November 27, 1996; accepted for publication (in revised form) September 8, 1998; published electronically February 18, 2000.
http://www.siam.org/journals/simax/21-3/31303.html
†Scuola Normale Superiore, Piazza Cavalieri 7, 56100 Pisa, Italy (miranda@cibs.sns.it, tilli@cibs.sns.it).

by f. Each matrix T_n has an $n \times n$ BT structure, and each block (3) is a $k \times k$ matrix with complex entries (we remark that no hypothesis is made concerning the structure of these blocks); one easily checks that $T_n \in \mathbf{C}^{nk \times nk}$ is Hermitian, since f is Hermitian-valued and hence (3) implies $A_j = A_{-j}^*$ (throughout, the symbol * denotes conjugate transpose).

If A is an $n \times n$ Hermitian matrix, we indicate by $\lambda_j(A)$, $j = 1, \ldots, n$, all the eigenvalues of A, counted with their multiplicities, numbered in nondecreasing order.

Our goal here is to study the asymptotic spectral distribution of the matrices $\{T_n\}$ generated by f when n tends to infinity. Well-known is the case where the generating function f takes scalar values (that is, $k = 1$ in (1) and (2)): in this particular case, the matrices $\{T_n\}$ associated with f are $n \times n$ plain Toeplitz matrices, and the following classical theorem holds (see [6]).

THEOREM 1.1 (see Grenander and Szegő [6]). *Suppose $f \in L^\infty(Q, \mathbf{R})$ (that is, f is scalar-valued, measurable, and bounded), and let $\{T_n\}$ be the sequence of Hermitian Toeplitz matrices generated by f; then for any function F, continuous on the interval $[\inf_Q f, \sup_Q f]$, it holds that*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{\lambda \in \sigma(T_n)} F(\lambda) = \frac{1}{2\pi} \int_{-\pi}^{\pi} F\big(f(x)\big)\, dx. \tag{4}$$

(Here $\sigma(T_n)$ indicates the spectrum of T_n: it is understood that each eigenvalue is counted according to its multiplicity.)

In a recent paper [11], Tyrtyshnikov has shown that it is not necessary that $f \in L^\infty(Q, \mathbf{R})$ for (4) to hold, but it suffices $f \in L^2(Q, \mathbf{R})$, provided the test function F has compact support. In this paper, we extend Tyrtyshnikov's result to the case where the generating function f is $k \times k$ Hermitian matrix-valued.

Our main result (Theorem 3.4) states that if $f \in L^2(Q, H^k)$ (that is, f is $k \times k$ Hermitian matrix-valued and square-integrable over Q) then, for every continuous function F with compact support, there holds

$$\lim_{n\to\infty} \frac{1}{nk} \sum_{\lambda \in \sigma(T_n)} F(\lambda) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1}{k} \sum_{\lambda \in \sigma(f(x))} F(\lambda)\, dx. \tag{5}$$

(The factor $1/k$ on both sides has a normalization purpose, since T_n has order nk and $f(x)$ has order k.) We remark that, in the particular case where $k = 1$ (that is, f is scalar-valued), each T_n is a Toeplitz matrix of order n and (5) reduces to the Szegő–Tyrtyshnikov theorem for the eigenvalues of Toeplitz matrices [11].

It is clear that the main formula (5) provides much information about the eigenvalues of the matrices T_n, for large n. For example, we obtain (Corollary 3.5) that for any positive integer m it holds that

$$\lim_{n\to\infty} \lambda_m(T_n) = \inf_Q f, \qquad \lim_{n\to\infty} \lambda_{nk+1-m}(T_n) = \sup_Q f, \tag{6}$$

that is, for any fixed m the first m eigenvalues of T_n tend to $\inf_Q f$, and the last m eigenvalues of T_n tend to $\sup_Q f$, as n approaches infinity (the precise meaning of $\inf_Q f$ and $\sup_Q f$ is that given by (10) and (11): roughly speaking, they are, respectively, the smallest and the largest eigenvalue of $f(x)$, for $x \in Q$).

Observe that it might hold $\sup_Q f = \infty$ as well as $\inf_Q f = -\infty$, since f is not assumed to be bounded. In the particular case where f is bounded, anyway, the limit

relation (5) makes sense and holds also for any continuous function F defined only on the compact set $[\inf_Q f, \sup_Q f]$ (see Theorem 3.3). In fact, any such function can be extended outside $[\inf_Q f, \sup_Q f]$ to a continuous function with compact support; this extension, of course, is not unique, but neither the right-hand side nor the left-hand side of (5) is affected by the extension, since the spectra of $f(x)$ and T_n lie inside $[\inf_Q, f \sup_Q, f]$, as proved in Theorem 3.1.

In applications it is often necessary to solve linear systems of the kind $T_n x = b$ for large n, where the matrices $\{T_n\}$ are generated by some f which is positive definite for almost all $x \in Q$, but satisfies $\inf_Q f = 0$. Then all the T_n are positive definite (Corollary 3.2) and the conjugate gradient (CG) method can be applied to solve the system $T_n x = b$; unfortunately, by virtue of (6), the matrix T_n is always ill-conditioned for large n and conjugate gradients will converge very slowly. Therefore, the linear system $T_n x = b$ needs preconditioning [3].

A natural way to build a positive definite preconditioner P_n for the system $T_n x = b$ is to let P_n be generated by a positive definite $k \times k$ matrix-valued function p in such a way that the eigenvalues of $P_n^{-1} T_n$ have nicer properties than those of the unpreconditioned T_n (see [2, 3]). In particular, it is desirable that the eigenvalues of the preconditioned matrices have a strictly positive lower bound r, independent of n. In Theorems 4.1 and 4.3 we prove that, under reasonable assumptions on f and p, for any n the spectrum of $P_n^{-1} T_n$ lies in the (not necessarily bounded) interval $[r, R]$, where

$$r = \inf_{Q_+} p(x)^{-1} f(x), \qquad R = \sup_{Q_+} p(x)^{-1} f(x),$$

and $Q_+ \subset Q$ is the set, where $p(x)$ is positive definite. Moreover, we show that, for any integer $m \geq 1$, it holds that

$$\lim_{n\to\infty} \lambda_m(P_n^{-1} T_n) = r, \qquad \lim_{n\to\infty} \lambda_{nk+1-m}(P_n^{-1} T_n) = R; \tag{7}$$

that is, for fixed m, the first m eigenvalues of $P_n^{-1} T_n$ tend to the smallest eigenvalue of $p^{-1} f$, whereas the last m tend to the largest eigenvalue of $p^{-1} f$.

From (7) with $m = 1$ it follows that the condition number μ_n of the matrix $P_n^{-1/2} T_n P_n^{-1/2}$ satisfies $\lim_{n\to\infty} \mu_n = R/r$ (see Corollary 4.4); moreover, since (7) holds for arbitrary m, the convergence rate of the CG method [1, 12] crucially depends on the ratio R/r (the smaller the better). Since the ratio R/r depends only on the generating functions f and p, the above results provide a simple but very general tool to study BT preconditioners: indeed, given f and $\{T_n\}$, a sequence of BT preconditioners $\{P_n\}$, generated by some positive semidefinite p, will be suitable for $\{T_n\}$ if the two following conditions are satisfied:

- the ratio R/r is small,
- the matrices P_n are easily invertible.

In section 5 we clarify all these considerations by a numerical example, which shows how suitable BT preconditioners may remove the ill-conditioning of a bad sequence of matrices $\{T_n\}$, for which even the block circulant preconditioners fail.

Finally, these results on preconditioners can be considered the analog (in the case where the generating function is matrix valued) of the results obtained by Serra and Tilli in the scalar case (see [8, 9]).

2. Notation and preliminary considerations. Given complex vectors x, y we indicate by $\langle y, x\rangle$ their Hermitian product and by $\|x\|$ the induced Euclidean norm;

sometimes we write $x \perp y$ instead of $\langle x, y\rangle = 0$. If A is a square matrix, we denote by $\|A\|_F$ its Frobenius norm and by $\sigma(A)$ its spectrum: it is understood that each eigenvalue is counted according to its multiplicity, in such a way that $\sum_{\lambda\in\sigma(A)} \lambda$ is the trace of A. Finally, we indicate by H^k the space of all Hermitian matrices of order k.

If $A \in \mathbf{C}^{k\times k}$ has real eigenvalues, we indicate by $\lambda_j(A)$, $j = 1, \dots, k$, the eigenvalues of A, numbered in nondecreasing order, counted with their multiplicities; moreover, we let $\mathrm{cond}(A)$ denote the ratio $\lambda_k(A)/\lambda_1(A)$ (if A is positive definite, this is the usual condition number).

REMARK 2.1. *If $T, P \in H^k$ are Hermitian matrices such that P is positive definite, then the eigenvalues of $P^{-1}T$ are all real, since they coincide with those of $P^{-\frac{1}{2}}TP^{-\frac{1}{2}}$; in particular, the number* $\mathrm{cond}(P^{-1}T)$ *is the usual condition number (in the spectral norm) of $P^{-\frac{1}{2}}TP^{-\frac{1}{2}}$. Therefore, if P is meant as a preconditioner for T, we will refer to the number* $\mathrm{cond}(P^{-1}T)$ *as the "condition number of the preconditioned matrix."*

In the following, we will deal with Lebesgue-integrable matrix-valued functions over the interval $Q = (-\pi, \pi)$; when we say that a set $Q' \subset Q$ has measure zero, we always refer to Lebesgue measure on Q (also expressions such as "almost everywhere" refer to Lebesgue measure). Most of the functions we deal with are Hermitian matrix-valued functions, defined (almost everywhere) on the interval Q. If $p \geq 1$, we indicate by $L^p(Q, H^k)$ the Banach space of all $k \times k$ Hermitian matrix-valued functions, which are p-integrable on Q, that is,

$$f \in L^p(Q, H^k) \quad \Longleftrightarrow \quad f(x) \in H^k \text{ and } \|f\|_{L^p}^p = \frac{1}{2\pi}\int_Q \|f(x)\|_F^p \, dx < +\infty. \tag{8}$$

If $f \in L^2(Q, H^k)$, we refer to the $k \times k$ matrices A_j's defined by (3) as the matrix Fourier coefficients of f.

The main reason to use the Frobenius norm in (8) is that $L^2(Q, H^k)$ is a Hilbert space, with the scalar product given by

$$\langle f, g\rangle_{L^2} = \frac{1}{2\pi}\int_{-\pi}^{\pi} [\mathrm{Tr}\, g(x)^* f(x)]\, dx,$$

and Parseval's relation in $L^2(Q, H^k)$ reads $\|f\|_{L^2}^2 = \sum_{j=-\infty}^{+\infty} \|A_j\|_F^2$.

There is a connection between the Frobenius norm of T_n and the L^2 norm of the generating function f, namely,

$$\|T_n\|_F \leq \sqrt{n}\, \|f\|_{L^2}. \tag{9}$$

This inequality easily follows from (1) and Parseval's relation in $L^2(Q, H^k)$: in fact, $\|T_n\|_F^2$ equals the sum of the square of the Frobenius norm of its blocks, and each block A_j appears in T_n at most n times.

When we state a pointwise property of a matrix-valued function f which is only supposed to be integrable, we will always mean that such a property holds *almost everywhere* on Q, that is, that it holds at every point of Q except for a set of zero Lebesgue measure. For example, if $f \in L^1(Q, H^k)$ by saying that f is (almost everywhere) positive definite, we mean that the matrix $f(x)$ is positive definite for all $x \in Q$, except for a set of measure zero.

If $f : Q \to \mathbf{C}^{k\times k}$ is a matrix-valued measurable function such that $f(x)$ has real eigenvalues, for $Q' \subseteq Q$, we set

$$\inf_{Q'} f = \sup\{y \in \mathbf{R} : \lambda_1\big(f(x)\big) > y \text{ for almost every } x \in Q'\} \tag{10}$$

and, similarly, we set

$$\sup_{Q'} f = \inf\{y \in \mathbf{R} : \lambda_k\big(f(x)\big) < y \text{ for almost every } x \in Q'\}. \tag{11}$$

Roughly speaking, $\inf_{Q'} f$ is the smallest eigenvalue of $f(x)$ and $\sup_{Q'} f$ is the greatest eigenvalue of $f(x)$ as x ranges over Q', disregarding sets of measure zero. It is immediate to see that, in the scalar case $k = 1$ (that is, f takes scalar values), the above two notions coincide with the usual notions of essential infimum and, respectively, essential supremum of f over Q'. Sometimes we will simply write $\inf f$ for $\inf_Q f$, and $\sup f$ for $\sup_Q f$.

Relevant to our purposes is also the Banach space $L^\infty(Q, H^k)$ of Hermitian matrix-valued functions which are *essentially bounded* over Q, that is,

$$f \in L^\infty(Q, H^k) \iff \begin{cases} f \text{ is measurable,} \\ \|f\|_{L^\infty} = \inf\{y \in \mathbf{R} : \|f(x)\|_F < y \text{ for a.e. } x \in Q\} < \infty. \end{cases}$$

Finally, we will use the following perturbation theorem for the eigenvalues of a Hermitian matrix, whose original version is due to Hoffman and Wielandt (for a proof, see [7]).

THEOREM 2.1. *For $T, \tilde{T} \in H^m$, it holds that*

$$\left(\sum_{j=1}^{m} \left|\lambda_j(T) - \lambda_j(\tilde{T})\right|^2\right)^{\frac{1}{2}} \le \left\|T - \tilde{T}\right\|_F.$$

3. Spectral theorems for block Toeplitz matrices. The next theorem provides a sharp bound for the eigenvalues of the sequence of BT matrices $\{T_n\}$ generated by f. It should be noted that such bounds are independent of n (they only depend on the spectrum of f).

THEOREM 3.1. *Suppose that $f \in L^1(Q, H^k)$, and let $\{T_n\}$ be the set of Hermitian BT matrices generated by f. Then, for any natural number n, if λ is an eigenvalue of T_n, it holds that*

$$\inf_Q f \le \lambda \le \sup_Q f. \tag{12}$$

Moreover, if $\lambda = \inf f$, then the smallest eigenvalue of $f(x)$ is constant and it equals $\inf f$ for almost all $x \in Q$; conversely, if $\lambda = \sup f$, then the largest eigenvalue of $f(x)$ is constant and it equals $\sup f$ for almost all $x \in Q$.

Proof. Suppose that, for some n, λ is an eigenvalue of T_n and $u \in \mathbf{C}^{nk}$ is the corresponding eigenvector. Partition u into blocks as follows

$$u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}, \quad u_j \in \mathbf{C}^k, \quad j = 1, \dots, n, \tag{13}$$

and define $v(x) = \sum_{h=1}^{n} e^{-\hat{\imath}hx} u_h$. Then an elementary computation, using (1), (13), and (3), yields

$$\lambda \|u\|^2 = \langle T_n u, u \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} \langle f(x)v(x), v(x) \rangle \, dx.$$

Now, since for almost every x it holds that $\langle f(x)v(x), v(x) \rangle \leq (\sup f) \|v(x)\|^2$, we obtain

$$\lambda \|u\|^2 \leq (\sup f) \frac{1}{2\pi} \int_{-\pi}^{\pi} \|v(x)\|^2 \, dx = (\sup f) \sum_{j=1}^{n} \|u_j\|^2 = (\sup f) \|u\|^2 .$$

Dividing by $\|u\|^2 \neq 0$, we have $\lambda \leq \sup f$. Now suppose that the equality holds; then it holds that $\langle f(x)v(x), v(x) \rangle = \sup f \|v(x)\|^2$ for almost every x. Then $\sup f$ (and hence λ) is the largest eigenvalue of $f(x)$ for almost all x, since the Rayleigh quotient $\langle f(x)v(x), v(x) \rangle / \langle v(x), v(x) \rangle$, almost everywhere, cannot exceed $\sup f$. Replacing f by $-f$ and repeating the above argument we conclude the proof. ∎

COROLLARY 3.2. *Suppose that $f \in L^1(Q, H^k)$ and that $f(x)$ is positive semidefinite; moreover, suppose that the set $Q' \subset Q$, made of all $x \in Q$ such that $f(x)$ is positive definite, has strictly positive Lebesgue measure. Then the matrices $\{T_n\}$ generated by f are all positive definite.*

Proof. By Theorem 3.1, if λ is an eigenvalue of some T_n, then $\lambda \geq 0$ (in fact, f is positive semidefinite almost everywhere if and only if $\inf f \geq 0$). Therefore, T_n is positive semidefinite. Now suppose that $\lambda = 0$; then it holds that $\inf f = 0$ and, again by Theorem 3.1, the smallest eigenvalue of $f(x)$ is zero for almost every x, and so f cannot be positive definite on a set of positive measure. ∎

The next two theorems, respectively, extend the classic Szegő theorem [6] and the Szegő–Tyrtyshnikov theorem for Toeplitz matrices [11] to the case where f is matrix-valued.

THEOREM 3.3. *Suppose that $f \in L^\infty(Q, H^k)$, and let $\{T_n\}$ be the set of BT matrices generated by f; then for any function F, continuous on the interval $[\inf f, \sup f]$, there holds*

$$\lim_{n\to\infty} \frac{1}{nk} \sum_{\lambda \in \sigma(T_n)} F(\lambda) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1}{k} \sum_{\lambda \in \sigma(f(x))} F(\lambda) \, dx. \tag{14}$$

THEOREM 3.4. *Suppose that $f \in L^2(Q, H^k)$, and let $\{T_n\}$ be the set of BT matrices generated by f; then for any continuous function $F(\lambda)$ with compact support in $\mathbf{R}$, it holds that*

$$\lim_{n\to\infty} \frac{1}{nk} \sum_{\lambda \in \sigma(T_n)} F(\lambda) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1}{k} \sum_{\lambda \in \sigma(f(x))} F(\lambda) \, dx. \tag{15}$$

As we observed in the Introduction, Theorem 3.3 follows from Theorem 3.4 by extending F outside $[\inf f, \sup f]$ so that the new function is continuous and has bounded support. Therefore, it suffices to prove Theorem 3.4.

Proof. First suppose that f is a $k \times k$ matrix-valued trigonometric polynomial, that is, there exists a natural number d such that

$$f(x) = \sum_{j=-d}^{d} A_j e^{\hat{\imath}jx}, \quad A_j \in \mathbf{C}^{k\times k}, \quad j = -d, \ldots, d. \tag{16}$$

If $n > 2d+1$, then T_n is a block banded matrix whose block bandwidth equals $2d+1$. Consider the block circulant matrices associated with f [5]

$$S_n = \mathrm{BCirc}\,(A_0, A_1, \ldots, A_d, 0, \ldots, 0, A_{-d}, A_{-d+1}, \ldots, A_{-1})_{n\times n};$$

it is immediate to check that $\mathrm{rank}(T_n - S_n) \le 2dk$ (indeed, T_n and S_n differ only in the upper-right and lower-left corners by submatrices of order dk). By a result of Tyrtyshnikov (see [11, Theorem 6.1]) the eigenvalues of T_n and those of S_n are equally distributed as $n \to \infty$: therefore, in order to prove (15), it suffices to establish it with S_n in place of T_n (see [11, Section 6]). Since the spectrum of the block circulant S_n is the union of the spectra of $f(x)$ as x ranges over the uniform grid $M_n = \{-\pi/2 + 2\pi i/n\}_{i=0}^{n-1}$ (see [5]), for any F we have

$$\frac{1}{nk}\sum_{\lambda\in\sigma(S_n)} F(\lambda) = \frac{1}{nk}\sum_{i=0}^{n-1}\sum_{\lambda\in\sigma\left(f(-\frac{\pi}{2}+\frac{2\pi i}{n})\right)} F(\lambda).$$

Observe that right-hand side is the nth Riemann sum of the continuous function

$$G(x) = \frac{1}{k}\sum_{\lambda\in\sigma\left(f(x)\right)} F(\lambda)$$

over the grid M_n. Taking the limit as $n \to \infty$, the Riemann sums converge to the Lebesgue integral of G over $[-\pi, \pi]$, and hence (15) is established with S_n in place of T_n: this proves the theorem when f is a matrix-valued trigonometric polynomial of the kind (16).

The general case $f \in L^2(Q, H^k)$ is obtained by approximating f, in the norm of L^2, by its Fourier sums, as follows. By a standard density argument [6, 10], it suffices to establish (15) when F has compact support and is continuously differentiable; therefore, let such an F be given, and choose an arbitrary $\epsilon > 0$. Denote $D = \max_{x\in\mathbf{R}} |F'(x)|$, and choose d so large that, letting $\tilde{f}(x) = \sum_{j=-d}^{d} A_j e^{\hat{\imath} jx}$, it holds that

$$D\left\|f - \tilde{f}\right\|_{L^2} \le \epsilon. \tag{17}$$

Denote $\{\tilde{T}_n\}$, the sequence of BT matrices generated by $\tilde{f}$; letting

$$D_n = \left|\frac{1}{nk}\sum_{\lambda\in\sigma(T_n)} F(\lambda) - \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{k}\sum_{\lambda\in\sigma(f(x))} F(\lambda)\,dx\right|,$$

from the triangle inequality we obtain $D_n \le S_1(n) + S_2(n) + S_3(n)$, where

$$S_1(n) = \left|\frac{1}{nk}\sum_{\lambda\in\sigma(T_n)} F(\lambda) - \frac{1}{nk}\sum_{\lambda\in\sigma(\tilde{T}_n)} F(\lambda)\right|,$$

$$S_2(n) = \left|\frac{1}{nk}\sum_{\lambda\in\sigma(\tilde{T}_n)} F(\lambda) - \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{k}\sum_{\lambda\in\sigma(\tilde{f}(x))} F(\lambda)\,dx\right|,$$

$$S_3 = \left|\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{k}\sum_{\lambda\in\sigma(\tilde{f}(x))} F(\lambda)\,dx - \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{k}\sum_{\lambda\in\sigma(f(x))} F(\lambda)\,dx\right|.$$

From the mean value theorem applied to F and the Cauchy inequality, we obtain

$$S_1(n) \le \frac{D}{nk} \sum_{j=1}^{nk} \left| \lambda_j(T_n) - \lambda_j(\tilde{T}_n) \right| \le \frac{D}{\sqrt{nk}} \left(\sum_{j=1}^{nk} \left| \lambda_j(T_n) - \lambda_j(\tilde{T}_n) \right|^2 \right)^{\frac{1}{2}};$$

from Theorem 2.1 and inequality (17) we obtain

$$S_1(n) \le \frac{D}{\sqrt{nk}} \left\| T_n - \tilde{T}_n \right\|_F \le \frac{D}{\sqrt{k}} \left\| f - \tilde{f} \right\|_{L^2} \le \epsilon. \tag{18}$$

A similar reasoning, involving the mean value theorem, the Cauchy inequality, and Theorem 2.1, leads to

$$\begin{aligned} S_3(n) &\le \frac{D}{2\pi} \int_{-\pi}^{\pi} \frac{1}{\sqrt{k}} \left\| f(x) - \tilde{f}(x) \right\|_F \, dx \\ &\le \frac{D}{\sqrt{k}} \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \left\| f(x) - \tilde{f}(x) \right\|_F^2 \, dx \right)^{\frac{1}{2}} \\ &= \frac{D}{\sqrt{k}} \left\| f - \tilde{f} \right\|_{L_2} \le \epsilon. \end{aligned}$$

Therefore, it holds that $D_n \le 2\epsilon + S_2(n)$ for all n; on the other hand, $\tilde{f}$ is a trigonometric polynomial and, according to the first part of the proof, $\lim_{n\to\infty} S_2(n) = 0$. In the limit we obtain $\limsup_{n\to\infty} D_n \le 2\epsilon$; since $\epsilon > 0$ was arbitrary, it holds that $\limsup_{n\to\infty} D_n = 0$, which, combined with $D_n \ge 0$, yields $\lim_{n\to\infty} D_n = 0$, and the proof is completed. □

Recalling Theorem 3.1, by a suitable choice of F in (15) one easily proves the following corollary.

COROLLARY 3.5. *Under the same hypothesis as in Theorem* 3.4*, for any positive integer m it holds*

$$\lim_{n\to\infty} \lambda_m(T_n) = \inf_Q f, \qquad \lim_{n\to\infty} \lambda_{nk+1-m}(T_n) = \sup_Q f; \tag{19}$$

that is, for any fixed m, the first m eigenvalues of T_n tend to $\inf_Q f$, and the last m eigenvalues of T_n tend to $\sup_Q f$, as n approaches infinity.

4. Preconditioning block Toeplitz matrices. We begin this section with a result concerning the localization of the eigenvalues of BT matrices preconditioned with BT preconditioners. Throughout, if $Q' \subseteq Q$ is a measurable set, we indicate by $\mathcal{L}(Q')$ the Lebesgue measure of Q'.

THEOREM 4.1. *Suppose $f, p \in L^1(Q, H^k)$, with $p(x)$ positive definite for almost every $x \in Q$, that is,*

$$\mathcal{L}(Q \setminus Q_+) = 0, \qquad \text{where} \qquad Q_+ = \{x \in Q : p(x) \text{ is positive definite}\}.$$

Let $\{T_n\}$ and $\{P_n\}$ be the sets of BT matrices generated by f and, respectively, by p. Then, for any $n \in$ bf N, all the eigenvalues of $P_n^{-1} T_n$ lie in the (not necessarily bounded) interval $[r, R]$, where

$$r = \inf_{Q_+} p(x)^{-1} f(x), \qquad R = \sup_{Q_+} p(x)^{-1} f(x). \tag{20}$$

Proof. We observe that, according to Corollary 3.2, the matrices $\{P_n\}$ generated by p are all positive definite. Let λ be an eigenvalue of some $P_n^{-1}T_n$, and let $u \in \mathbf{C}^{nk}$ be the corresponding eigenvector. Then it clearly holds that $T_n u = \lambda P_n u$, and hence $\langle T_n u, u\rangle = \lambda \langle P_n u, u\rangle$. Partitioning u into blocks as in (13) and arguing as in the proof of Theorem 3.1, $\langle (T_n - \lambda P_n)u, u\rangle = 0$ can be written as

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \langle (f(x) - \lambda p(x))v(x), v(x)\rangle \, dx = 0, \tag{21}$$

where $v(x) = \sum_{h=1}^{n} e^{-\hat{\imath}hx}u_h$. Since $v(x)$ is a (vector-valued) trigonometric polynomial which is not identically zero, it holds that $v(x) \neq 0$ for almost every $x \in Q$, and hence (21) cannot hold if $f(x) - \lambda p(x)$ is positive definite (or negative definite) for almost every $x \in Q$. On the other hand, it is easy to see that if $\lambda < r$ (respectively, $\lambda > R$), then $f(x) - \lambda p(x)$ is positive definite (respectively, negative definite) almost everywhere on Q, since the eigenvalues of $f - \lambda p$ coincide with those of $p^{-1}f$ (recall that p is positive definite almost everywhere). Therefore, neither $\lambda < r$ nor $\lambda > R$ can hold, and hence $\lambda \in [r, R]$. □

From now on, we adopt the following notation: if k and n are natural numbers, we indicate by $\mathcal{T}_n(Q, \mathbf{C}^k)$ the vector space of $\mathbf{C}^k$-valued trigonometric polynomials of degree at most n, that is, the space of functions of the kind

$$u(x) = \sum_{j=-n}^{n} u_j e^{\hat{\imath}jx}, \quad u_j \in \mathbf{C}^k, \quad j = -n, \ldots, n.$$

LEMMA 4.2. *Suppose $f, p \in L^\infty(Q, H^k)$, with p continuous on Q and positive definite for almost every $x \in Q$, that is,*

$$\mathcal{L}(Q \setminus Q_+) = 0, \qquad \text{where} \qquad Q_+ = \{x \in Q : p(x) \text{ is positive definite}\}.$$

If we set

$$r = \inf_{Q_+} p(x)^{-1} f(x),$$

then for any real number $r' > r$ and any integer $N \geq 0$, there exists a natural number n such that, for every N-tuple of polynomials $h_i \in \mathcal{T}_n(Q, \mathbf{C}^k)$ $i = 1, \ldots, N$, a not identically zero polynomial $u \in \mathcal{T}_n(Q, \mathbf{C}^k)$ can be found, such that

$$\int_{-\pi}^{\pi} \langle f(x)u(x), u(x)\rangle \, dx \leq r' \int_{-\pi}^{\pi} \langle p(x)u(x), u(x)\rangle \, dx, \tag{22}$$

and

$$\int_{-\pi}^{\pi} \langle u(x), h_i(x)\rangle \, dx = 0, \quad i = 1, \ldots, N. \tag{23}$$

In order to simplify the notation, in the proof of this lemma the symbol $\int$ should read as $\int_{-\pi}^{\pi}$.

Proof. Choose any real number r'' such that $r < r'' < r'$; then the set

$$S = \{x \in Q_+ : \lambda_1(p(x)^{-1}f(x)) < r''\} \tag{24}$$

has strictly positive Lebesgue measure, since $r'' > r$. By elementary measure theory, there exist $N+1$ pairwise disjoint compact sets $K_j \subset S$, $j = 0, \dots, N$ such that each K_j has strictly positive Lebesgue measure. We let $\delta = \min\{\mathcal{L}(K_0), \dots, \mathcal{L}(K_N)\} > 0$; letting $K = \bigcup_{j=0}^{N} K_j$, it follows that $p(x) \geq \gamma \mathrm{I}_k$ for some $\gamma > 0$ and for every $x \in K$, since p is continuous and positive definite on K, and K is compact (I_k is the $k \times k$ identity matrix). We now define the vectors $\phi_j(x) \in \mathbf{C}^k$ by

$$\|\phi_j(x)\| = 1, \quad f(x)\phi(x)_j = \lambda_1\big(p(x)^{-1} f(x)\big) p(x)\phi_j(x) \quad \forall x \in K_j, \quad j = 0, \dots, N; \tag{25}$$

Moreover, we let $\phi_j(x) = 0$ for $x \in Q \setminus K_j$. It is obvious that $\phi_j \in L^\infty(Q, \mathbf{C}^k)$, for $j = 0, \dots, N$. It is well-known that, for every $\epsilon > 0$, we can find trigonometric polynomials $g_j^\epsilon \in \mathcal{T}_{n_j}(Q, \mathbf{C}^k)$, $j = 0, \dots, N$, such that

$$\big\|\phi_j(x) - g_j^\epsilon(x)\big\|_{L^2} < \epsilon, \quad j = 0, \dots, N. \tag{26}$$

We let $n(\epsilon) = \max\{n_0, \dots, n_N\}$, so that $g_j^\epsilon \in \mathcal{T}_{n(\epsilon)}(Q, \mathbf{C}^k)$, for $j = 0, \dots, N$; now consider any polynomial $u \in \mathcal{T}_{n(\epsilon)}(Q, \mathbf{C}^k)$ of the form

$$u(x) = \sum_{j=0}^{N} c_j g_j^\epsilon(x), \tag{27}$$

where the coefficients c_j are complex numbers such that $|c_j| \leq 1$, and at least one of them, say c_{j^*}, satisfies $c_{j^*} = 1$. It then holds that

$$\int \langle f(x)u(x), u(x)\rangle \, dx = \sum_{i,j=0}^{N} c_i \bar{c}_j \int \big\langle f(x) g_i^\epsilon(x), g_j^\epsilon(x)\big\rangle \, dx. \tag{28}$$

Since $f(x)$ is bounded, it is easy to see (by polarization and Cauchy–Schwartz inequality) that the sesquilinear form

$$(v(x), w(x)) \to \int \langle f(x)v(x), w(x)\rangle \, dx$$

is continuous, from $L^2(Q, \mathbf{C}^k) \times L^2(Q, \mathbf{C}^k)$ to $\mathbf{C}$. For this reason, from (26), (27), and (28) it follows that

$$\int \langle f(x)u(x), u(x)\rangle \, dx = \sum_{i,j=0}^{N} c_i \bar{c}_j \int \langle f(x)\phi_i(x), \phi_j(x)\rangle \, dx + \omega(\epsilon), \tag{29}$$

where $\omega(\epsilon) = \omega(\epsilon, c_0, \dots, c_N)$ stands for a quantity which is infinitesimal when $\epsilon \to 0$: it is easy to see that $\omega(\epsilon)$ is $O(\epsilon)$, but we do not need this here; what is fundamental, instead, is that $\omega(\epsilon)$ tends to zero uniformly with respect to $c_0, \dots, c_N$, since $|c_j| \leq 1$, $j = 0, 1, \dots, N$. This is the reason why we simply write $\omega(\epsilon)$, omitting the dependence on the c_j's.

Moreover, since ϕ_j and ϕ_i have disjoint support when $i \neq j$ (because $K_i \cap K_j = \emptyset$), the previous relation reduces to

$$\int \langle f(x)u(x), u(x)\rangle \, dx = \sum_{j=0}^{N} |c_j|^2 \int_{K_j} \langle f(x)\phi_j(x), \phi_j(x)\rangle \, dx + \omega(\epsilon). \tag{30}$$

Similarly, replacing $f(x)$ by $p(x)$, we obtain

$$\int \langle p(x)u(x), u(x)\rangle \, dx = \sum_{j=0}^{N} |c_j|^2 \int_{K_j} \langle p(x)\phi_j(x), \phi_j(x)\rangle \, dx + \omega'(\epsilon) \tag{31}$$

(again, $\omega'(\epsilon) \to 0$ when $\epsilon \to 0$, uniformly with respect to the c_j's). Now, since $c_{j^*} = 1$, $p(x)$ is positive definite for $x \in Q_+$, and $p(x) \geq \gamma I_k$ for $x \in K \supset K_{j^*}$, we obtain

$$\begin{aligned}\sum_{j=0}^{N} |c_j|^2 \int_{K_j} \langle p(x)\phi_j(x), \phi_j(x)\rangle \, dx &\geq \int_{K_{j^*}} \langle p(x)\phi_{j^*}(x), \phi_{j^*}(x)\rangle \, dx \\ &\geq \gamma \int_{K_{j^*}} \|\phi_{j^*}(x)\|^2 \, dx = \gamma \mathcal{L}(K_{j^*}) \geq \gamma\delta > 0,\end{aligned}$$

and therefore the left side of (31) is positive for small ϵ (and uniformly with respect to the c_j's). From (30), (25), and (24) (recall that $K_j \subset S$) it follows that

$$\begin{aligned}\int \langle f(x)u(x), u(x)\rangle \, dx &= \sum_{j=0}^{N} |c_j|^2 \int_{K_j} \lambda_1(p(x)^{-1}f(x)) \langle p(x)\phi_j(x), \phi_j(x)\rangle \, dx + \omega(\epsilon) \\ &\leq r'' \sum_{j=0}^{N} |c_j|^2 \int_{K_j} \langle p(x)\phi_j(x), \phi_j(x)\rangle \, dx + \omega(\epsilon); \qquad (32)\end{aligned}$$

dividing by the left side of (31) (which is positive for small ϵ), we obtain

$$\frac{\int \langle f(x)u(x), u(x)\rangle \, dx}{\int \langle p(x)u(x), u(x)\rangle \, dx} \leq \frac{r'' \sum_{j=0}^{N} |c_j|^2 \int_{K_j} \langle p(x)\phi_j(x), \phi_j(x)\rangle \, dx + \omega(\epsilon)}{\sum_{j=0}^{N} |c_j|^2 \int_{K_j} \langle p(x)\phi_j(x), \phi_j(x)\rangle \, dx + \omega'(\epsilon)}.$$

It is clear that, since $r'' < r'$, for $\epsilon > 0$ sufficiently small the left side of the last inequality is smaller than r', independently of the coefficients c_j's. Choosing such an ϵ, we have proved that every polynomial $u \in \mathcal{T}_{n(\epsilon)}(Q, \mathbf{C}^k)$ of the kind (27) verifies (22), provided the coefficients c_j verify $|c_j| \leq 1$ and one of them verifies $c_{j^*} = 1$.

Now suppose that N trigonometric polynomials $h_i \in \mathcal{T}_{n(\epsilon)}(Q, \mathbf{C}^k)$, $i = 1, \ldots, N$ are given; then, seeking $u \in \mathcal{T}_{n(\epsilon)}(Q, \mathbf{C}^k)$ of the kind (27), (23) reduces to

$$\sum_{j=0}^{N} c_j \int \langle g_j^{\epsilon}(x), h_i(x)\rangle \, dx = 0, \quad i = 1, \ldots, N,$$

that is, a homogeneous linear system of N equations in $N+1$ indeterminates c_j, $j = 0, \ldots, N$. Such a system always has a nontrivial solution, which can be supposed normalized such that $|c_j| \leq 1$ and $c_{j^*} = 1$ for a suitable j^*, and the lemma thus is proved. ∎

The following theorem summarizes our main results on BT preconditioners.

THEOREM 4.3. *Suppose $f, p \in L^{\infty}(Q, H^k)$, with p continuous on Q and positive definite almost everywhere, that is,*

$$\mathcal{L}(Q \setminus Q_+) = 0, \qquad \textit{where} \qquad Q_+ = \{x \in Q : p(x) \textit{ is positive definite}\}.$$

Let $\{T_n\}$ and $\{P_n\}$ be the sets of BT matrices generated by f and, respectively, by p. Let r, R be defined by (20); then, for any $n \in$ bf N, all the eigenvalues of $P_n^{-1}T_n$ lie in the interval $[r, R]$, and for any integer $m \geq 1$ it holds that

$$\lim_{n\to\infty} \lambda_m(P_n^{-1}T_n) = r, \qquad \lim_{n\to\infty} \lambda_{nk+1-m}(P_n^{-1}T_n) = R, \tag{33}$$

that is, for fixed m, the first m eigenvalues of $P_n^{-1}T_n$ tend to the smallest eigenvalue of $p^{-1}f$, while the last m tend to the largest eigenvalue of $p^{-1}f$. Moreover, both limits are monotone.

Proof. Let $m \geq 1$ be given; each matrix $P_n^{-1}T_n$ has nk eigenvalues and, if $nk \geq m$, its mth eigenvalue can be characterized, according to the minimax theorem of Courant and Fischer, as follows:

$$\lambda_m^{(n)} := \lambda_m(P_n^{-1}T_n) = \sup_{s_1,\dots,s_{m-1}} \inf_{\substack{u\neq 0 \\ \langle u, s_j\rangle = 0 \\ j=1,\dots,m-1}} \frac{\langle T_n u, u\rangle}{\langle P_n u, u\rangle},$$

where u and $s_1, \dots, s_{m-1}$ all belong to $\mathbf{C}^{nk}$, and the sup symbol is not present if $m = 1$.

We observe that $\lambda_m^{(n)}$ is nonincreasing with respect to n, since T_n and P_n, respectively, are principal submatrices of T_{n+1} and P_{n+1}: therefore, the first limit in (33) surely exists and is monotone. From Theorem 4.1 we already know that $\lambda_m^{(n)} \in [r, R]$ and hence, in order to show the first limit relation in (33), it suffices to show that, for every $r' > r$, there exists $\tilde{n}$ such that $\lambda_m^{(\tilde{n})} \leq r'$. Apply Lemma 4.2 with $N = m - 1$ and let $\tilde{n} = 2n + 1$, where n is that given by Lemma 4.2. We want to show that

$$\lambda_m^{(\tilde{n})} = \sup_{s_1,\dots,s_{m-1}} \inf_{\substack{u\neq 0 \\ \langle u, s_j\rangle = 0 \\ j=1,\dots,m-1}} \frac{\langle T_{\tilde{n}} u, u\rangle}{\langle P_{\tilde{n}} u, u\rangle} \leq r', \tag{34}$$

where u and $s_1, \dots, s_{m-1}$ all belong to $\mathbf{C}^{\tilde{n}k}$. We observe that if

$$v(x), w(x) \in \mathcal{T}_n(Q, \mathbf{C}^k),$$

that is,

$$u(x) = \sum_{j=-n}^{n} e^{\hat{\imath} jx} u_j, \qquad v(x) = \sum_{j=-n}^{n} e^{\hat{\imath} jx} v_j, \quad u_j, v_j \in \mathbf{C}^k, \quad j = 0, \pm 1, \dots, \pm n,$$

and we build the vectors $u, v \in \mathbf{C}^{\tilde{n}k}$, partitioned into blocks

$$u = \begin{pmatrix} u_{-n} \\ u_{-n+1} \\ \vdots \\ u_n \end{pmatrix}, \quad \begin{pmatrix} v_{-n} \\ v_{-n+1} \\ \vdots \\ v_n \end{pmatrix},$$

then it holds that

$$\langle T_{\tilde{n}} u, u\rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} \langle f(x)u(x), u(x)\rangle \, dx, \qquad \langle P_{\tilde{n}} u, u\rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} \langle p(x)u(x), u(x)\rangle \, dx; \tag{35}$$

TABLE 1
Condition numbers of unpreconditioned, BT and BC preconditioned matrices.

	$n = 10$	$n = 25$	$n = 50$	$n = 100$	$n = 200$	bound
$\mathrm{cond}(T_n)$	102.94	631.68	2514.38	10023.90	40046.49	∞
$\mathrm{cond}(P_n^{-1}T_n)$	2.21	2.35	2.41	2.45	2.46	2.48
$\mathrm{cond}(C_n^{-1}T_n)$	11.19	36.39	94.08	252.90	699.08	∞

moreover, $u \perp v$ (as vectors) if and only if $u(x) \perp v(x)$ (as polynomials). This follows from

$$\langle u, v\rangle = \sum_{j=-n}^{n} \langle u_j, v_j\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} \langle u(x), v(x)\rangle \, dx. \tag{36}$$

It is clear that, by virtue of (35) and (36), the inequality (34) comes straightforward from Lemma 4.2. □

COROLLARY 4.4. *Under the same hypotheses and notations as in Theorem* 4.3*, suppose that f is positive semidefinite almost everywhere on Q, with f positive definite on a set of positive measure. Then it holds that*

$$\mathrm{cond}(P_n^{-1}T_n) < \frac{R}{r} \quad \forall n \geq 1, \qquad \textit{and} \qquad \lim_{n\to\infty} \mathrm{cond}(P_n^{-1}T_n) = \frac{R}{r},$$

where it is understood that $R/r = +\infty$ if $r = 0$. Moreover, the limit is nondecreasing with respect to n.

5. A numerical example. Consider the generating function

$$f(x) = \begin{bmatrix} 8x^2 & (\sin x)^4 \\ (\sin x)^4 & 8x^2 \end{bmatrix}, \quad x \in [0, 2\pi].$$

It is easy to check that f is everywhere positive semidefinite and is singular if and only if $x = 0$; it holds that $\inf f = 0$, $M = \sup f = 8\pi^2$. According to Theorem 3.1 and Corollary 3.2, the BT matrices $\{T_n\}$ generated by f are positive definite, their spectra lie inside the interval $(0, M)$, and by Corollary 3.5 the condition number of T_n tends to infinity as $n \to \infty$. Therefore, the matrices $\{T_n\}$ need preconditioning (see Table 1, first row).

Observe that, near the point $x = 0$, there holds

$$f(x) = \begin{bmatrix} 8x^2 & 0 \\ 0 & 8x^2 \end{bmatrix} \begin{bmatrix} 1 & O(x^2) \\ O(x^2) & 1 \end{bmatrix} \tag{37}$$

and, letting

$$p(x) = \begin{bmatrix} 32(1-\cos x) & 0 \\ 0 & 32(1-\cos x) \end{bmatrix},$$

the function $p^{-1}f$ has uniformly positive eigenvalues near $x = 0$. The spectrum of $p(x)^{-1}f(x)$, as a function of x, is plotted in Figure 1, where the bounds $r \approx 0.49$ and $R \approx 1.23$ are pointed out (the numbers r and R are defined by (20)).

The knowledge of the spectrum of $p^{-1}f$, together with the theory developed in section 4, provides detailed information on the spectra of the preconditioned matrices $\{P_n^{-1}T_n\}$, where the P_n's are the BT matrices generated by p. The spectra of $P_n^{-1}T_n$

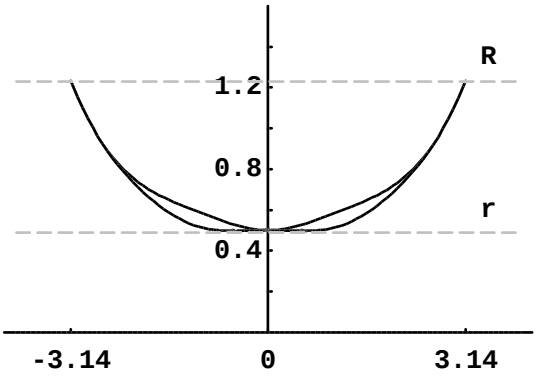


FIG. 1. *Spectrum of $p(x)^{-1}f(x)$ as a function of x.*

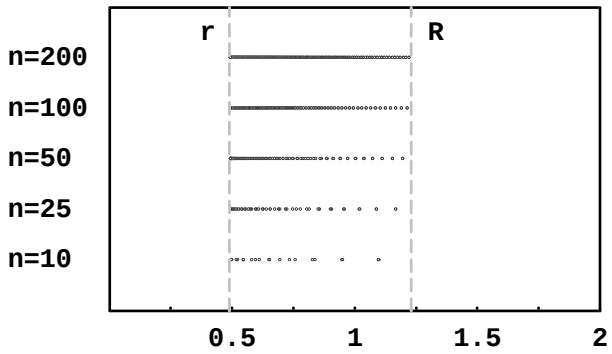


FIG. 2. *Spectra after block Toeplitz preconditioning.*

are plotted in Figure 2, where the interval $[r, R]$ is highlighted; according to Theorem 4.1, the spectrum of each $P_n^{-1}T_n$ lies inside $[r, R]$ and, by Corollary 4.4, the ratio $R/r \approx 2.48$ is an upper bound (attained in the limit) to the condition number of the preconditioned matrices $P_n^{-1}T_n$ (see Table 1, second row).

It is worth observing that $P_n = A_n \otimes I_2$ holds, where

$$A_n = \text{Tridiag}\,(-16, 32, -16)_{n\times n}$$

and I_2 is the identity matrix of order 2. As a consequence, from a computational point of view, the matrices P_n can be efficiently handled.

It is interesting to compare the spectra of $\{P_n^{-1}T_n\}$ (see Figure 2) with those of $\{C_n^{-1}T_n\}$ (see Figure 3), where C_n is the T. Chan block circulant (BC) preconditioner associated with T_n [3, 4] (among all $n \times n$ BC matrices with 2×2 blocks, C_n is the one that minimizes the distance from T_n in the Frobenius norm; see [3]). It turns out that the matrices $\{C_n^{-1}T_n\}$ are still ill-conditioned (see Table 1, third row); indeed (see Figure 3), although a bunch of eigenvalues of $C_n^{-1}T_n$ are close to 1, there are also a few large eigenvalues and, what is the main drawback, many eigenvalues approach zero as n increases. The failure of the BC preconditioners C_n when f is singular at some point should not surprise, since it is a well-known feature when f is scalar-valued and $\inf f = 0$ (see [3, p. 447]).

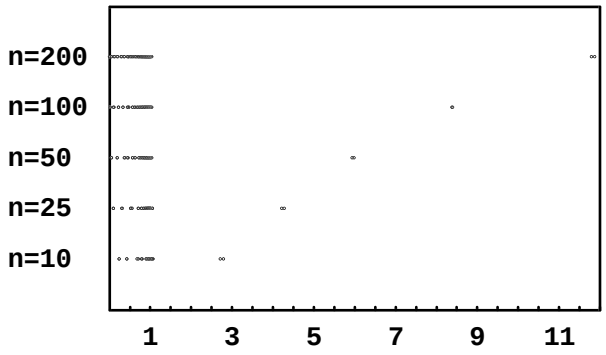


FIG. 3. *Spectra after block circulant preconditioning.*

The ideas underlying BT preconditioners and T. Chan BC preconditioners are different in nature: indeed, C_n is the best approximation to T_n (with respect to the Frobenius norm) in the BC class, whereas P_n may be far from T_n but is a suitable preconditioner, since $p^{-1}f$ is uniformly positive definite for $x \in [0, 2\pi]$. Moreover, while C_n is built from the (block) entries of T_n [3], $\{P_n\}$ is a standalone BT sequence generated by p; therefore, the construction of a suitable sequence of BT preconditioners is reduced to finding a generating function p such that $p^{-1}f$ is uniformly positive definite.

REFERENCES

[1] O. AXELSSON AND G. LINDSKOG, *On the rate of convergence of the preconditioned conjugate gradient method*, Numer. Math., 48 (1986), pp. 499–523.
[2] R. H. CHAN, *Toeplitz preconditioners for Toeplitz systems with nonnegative generating functions*, IMA J. Numer. Anal., 11 (1991), pp. 333–345.
[3] R. H. CHAN AND M. K. NG, *Conjugate gradient methods for Toeplitz systems*, SIAM Rev., 38 (1996), pp. 427–482.
[4] T. CHAN AND J. OLKIN, *Circulant preconditioners for Toeplitz-block matrices*, Numer. Algorithms, 6 (1994), pp. 89–101.
[5] P. J. DAVIS, *Circulant Matrices*, John Wiley and Sons, New York, 1979.
[6] U. GRENANDER AND G. SZEGŐ, *Toeplitz Forms and Their Applications*, 2nd edition, Chelsea, New York, 1984.
[7] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, New York, 1985.
[8] S. SERRA, *Preconditioning strategies for asymptotically ill-conditioned block Toeplitz systems*, BIT, 34 (1994), pp. 579–594.
[9] P. TILLI, *On the asymptotic spectrum of Hermitian block Toeplitz matrices with Toeplitz blocks*, Math. Comp., 66 (1997), pp. 1147–1159.
[10] P. TILLI, *Singular values and eigenvalues of non-Hermitian block Toeplitz matrices*, Linear Algebra Appl., 272 (1998), pp. 59–89.
[11] E. E. TYRTYSHNIKOV, *A unifying approach to some old and new theorems on distribution and clustering*, Linear Algebra Appl., 232 (1996), pp. 1–43.
[12] A. VAN DER SLUIS AND H. A. VAN DER VORST, *The rate of convergence of conjugate gradients*, Numer. Math., 48 (1986), pp. 543–560.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 3, pp. 882–903
© 2000 Society for Industrial and Applied Mathematics

COMPUTABLE CONVERGENCE BOUNDS FOR GMRES*

JÖRG LIESEN†

Abstract. The purpose of this paper is to derive new computable convergence bounds for GMRES. The new bounds depend on the initial guess and are thus conceptually different from standard "worst-case" bounds. Most importantly, approximations to the new bounds can be computed from information generated during the run of a certain GMRES implementation. The approximations allow predictions of how the algorithm will perform. Heuristics for such predictions are given. Numerical experiments illustrate the behavior of the new bounds as well as the use of the heuristics.

Key words. linear systems, convergence analysis, GMRES method, Krylov subspace methods, iterative methods

AMS subject classification. 65F10

PII. S0895479898341669

1. Introduction.

The GMRES algorithm by Saad and Schultz [28] is a popular iterative method for solving systems of linear equations:

$$Ax = b,\ A \in \mathbf{C}^{N\times N},\ b \in \mathbf{C}^N. \tag{1.1}$$

Given an initial guess x_0 for the solution of (1.1), GMRES yields iterates x_n so that

$$\|r_n\| := \|b - Ax_n\| = \min_{p_n\in\pi_n} \|p_n(A)r_0\|, \tag{1.2}$$

where $\|\cdot\|$ denotes the Euclidean vector and corresponding matrix norm, $r_0 := b - Ax_0$ is the initial residual, and π_n denotes the set of polynomials of degree at most n with value 1 at the origin.

The convergence analysis of GMRES has been an active field of research in recent years. Among others, Saad and Schultz [28], Nachtigal, Reddy, and Trefethen [23], Eiermann [8], Campbell et al. [5], and Starke [30] derived upper bounds on the residual norms. An overview of GMRES convergence results is given in [21, section 5.2].

These bounds are typically derived from the "ideal GMRES" approximation problem [16]: Find $p_n^* \in \pi_n$ so that

$$\|p_n^*(A)\| = \min_{p_n\in\pi_n} \|p_n(A)\|.$$

Being independent of the initial guess, "ideal GMRES" gives an upper bound for all possible GMRES convergence curves for the given matrix A. Hence, bounds on the GMRES residual norms derived from ideal GMRES intend to describe the algorithm's worst-case behavior. This approach sometimes leads to sharp bounds, for example, when A is normal [13, 17]. But as demonstrated in an example by Toh [33], ideal GMRES can overestimate even the worst-case behavior of GMRES arbitrarily high. While the practical implication of Toh's results and the similar results of others (see,

*Received by the editors June 25, 1998; accepted for publication (in revised form) by A. Greenbaum June 19, 1999; published electronically February 18, 2000. This work was supported by the Studienstiftung des deutschen Volkes, Bonn, Germany.

http://www.siam.org/journals/simax/21-3/34166.html

†Fakultät für Mathematik, Universität Bielefeld, 33501 Bielefeld, Germany (liesen@mathematik.uni-bielefeld.de).

e.g., [10]) is not clear yet, none of the bounds derived from ideal GMRES can completely characterize the algorithm's convergence behavior for any matrix A. Moreover, while some of the known bounds are only applicable to a restricted class of nonsingular linear systems, others include factors that are not computable. It is generally agreed that the convergence analysis of GMRES needs further work (see, e.g., [12, p. 55], [31, p. 187]).

In this paper we take an alternative approach to deriving convergence bounds for GMRES. Using characterizations of matrices that generate the same GMRES residuals—so-called GMRES-equivalent matrices—we derive bounds that depend on the initial guess. Our bounds are thus a posteriori and do not give new insight into how the convergence of GMRES depends on properties of A. However, all factors in our bounds can be computed from information generated during the run of WZ-GMRES, the GMRES implementation by Walker and Zhou [35]. Moreover, approximations to the factors in our bounds that are obtained early in the iteration often allow us to *predict how GMRES will perform* in later stages of the iteration. Our theory is applicable to any nonsingular linear system and can even be used in the singular case if a certain condition is satisfied.

The paper is organized as follows. In section 2 we introduce the GMRES algorithm and its mathematical properties. We generalize characterizations of GMRES-equivalent matrices by Greenbaum and Strakoš [15] in section 3. Section 4 includes the derivation of our new convergence bounds. In section 5 we explain how to obtain approximations to the factors in our bounds from information generated during the run of WZ-GMRES. We also devise heuristics for the behavior of these approximations and for the prediction of the performance of GMRES. Numerical examples are presented in section 6, and section 7 contains our concluding remarks.

By $\Lambda(M)$ we denote the spectrum and by M^H the hermitian transpose of the matrix $M \in \mathbf{C}^{N\times N}$, respectively. If M is nonsingular, we define $\kappa(M) := \|M\|\|M^{-1}\|$, and $\kappa(M) := \infty$, otherwise. We define $K_n(M, v) := \text{span}\{v, Mv, \ldots, M^{n-1}v\}$, $n = 1, 2, \ldots$, the nth Krylov space generated by the matrix M and the vector $v \in \mathbf{C}^N$. For technical reasons we also define $K_0(A, v) := \{0\}$.

2. Mathematical properties of GMRES. In this section we introduce the GMRES algorithm and state several of its properties. Note that the linear system (1.1) might be singular, if not assumed otherwise, and that we are not concerned with implementational details.

DEFINITION 2.1. *Suppose that x_0 is an initial guess for the solution of a linear system* (1.1) *and define the initial residual $r_0 := b - Ax_0$. For $n = 1, 2, \ldots$, let*

$$x_n \in x_0 + K_n(A, r_0) \tag{2.1}$$

be a vector, such that

$$r_n := b - Ax_n \in r_0 + AK_n(A, r_0) \tag{2.2}$$

satisfies

$$\|r_n\| = \min_{r\in r_0+AK_n(A,r_0)} \|r\|; \tag{2.3}$$

then the vectors x_n and r_n are called nth GMRES iterate and residual, respectively.

As a comparison with [28, formulas (4) and (8)] shows, our definition of the GMRES iterates and residuals is mathematically equivalent to their original derivation

by Saad and Schultz. The following well-known result (see, e.g., [6, p. 8]) justifies Definition 2.1.

LEMMA 2.2. *Suppose that* $A \in \mathbf{C}^{N\times N}$ *and* $r_0 \in \mathbf{C}^N$. *Then for every* $n = 1, 2, \ldots$, *there exists a unique vector* $r_n \in r_0 + AK_n(A, r_0)$ *that satisfies* (2.3).

Thus, for $n = 1, 2, \ldots$, the GMRES residuals are uniquely defined. In case of a nonsingular linear system, uniqueness of the residuals does result in uniqueness of the iterates. Definition 2.1 leads to the following algorithm, which does not, however, require uniqueness of the iterates.

ALGORITHM 2.3 (GMRES).
Input: A linear system (1.1) *and an initial guess* x_0.
Initialize: $r_0 := b - Ax_0$, $n = 0$
While $r_n \neq 0$
$n = n + 1$
Construct $x_n \in x_0 + K_n(A, r_0)$ *so that*

$$\|r_n\| = \|b - Ax_n\| = \min_{r\in r_0+AK_n(A,r_0)} \|r\|$$

End While

If $r_n = 0$ for some $n \geq 0$, or, equivalently, x_n is a solution of (1.1), we say that GMRES terminates at step n. Before analyzing the termination properties of the algorithm, we point out an equivalent way to define the GMRES residuals.

LEMMA 2.4. *Let* $S \subset \mathbf{C}^N$ *be a subspace and suppose that* $y_0 \in \mathbf{C}^N$ *and* $y_n \in y_0+S$. *Then* $\|y_n\| = \min_{y\in y_0+S} \|y\|$ *if and only if* $y_n \perp S$.

From this classical result (see, e.g., [6, p. 9]) it follows that our definition of the GMRES residuals via (2.2) and (2.3) is equivalent to defining them by (2.2) combined with

$$r_n \perp AK_n(A, r_0). \tag{2.4}$$

The following lemma is a direct consequence of (2.3).

LEMMA 2.5. *The GMRES algorithm applied to a linear system* (1.1) *and initial guess* x_0 *terminates at step* $d \geq 0$ *if and only if* d *is the smallest integer for which* $r_0 \in AK_d(A, r_0)$.

To further analyze this situation, we state an important property of the Krylov spaces, which is easy to prove.

LEMMA 2.6. *If* $A \in \mathbf{C}^{N\times N}$ *and* $r_0 \in \mathbf{C}^N$, *then*

$$\dim AK_N(A, r_0) = d \ \text{ for some } \ d, \ \ 0 \leq d \leq N,$$

is equivalent to

$$\begin{aligned} &\dim AK_j(A, r_0) = j \ \text{ for } \ 0 \leq j \leq d \ \text{ and} \\ &\dim AK_j(A, r_0) = d \ \text{ for all } \ j \geq d + 1. \end{aligned}$$

Since the spaces $AK_n(A, r_0)$ coincide for all $n \geq \dim AK_N(A, r_0)$, the following is an immediate consequence.

COROLLARY 2.7. *Suppose that a linear system* (1.1) *and an initial guess* x_0 *are given and define* $d := \dim AK_N(A, r_0)$. *Then for the GMRES residuals* r_n, $n \geq d$, $r_n = r_d$ *follows.*

Thus, either GMRES terminates at some step $n \leq d := \dim AK_N(A, r_0) \leq N$, which requires $r_0 \in AK_N(A, r_0) = AK_d(A, r_0)$, or the algorithm fails to terminate, but stagnates after step d. The next theorem can be shown easily using Lemma 2.6.

THEOREM 2.8 (cf. [28, Proposition 2]). *Suppose that the matrix A in* (1.1) *is nonsingular. Then for any initial guess x_0, GMRES applied to* (1.1) *and x_0 terminates at step $d \geq 0$ if and only if* $\dim AK_N(A, r_0) = d$.

While GMRES must fail to terminate for inconsistent linear systems, for consistent singular systems the question of termination depends on the initial guess. To derive a reasonable bound on the residual norms, one has to assume that GMRES will terminate, which is equivalent to assuming $r_0 \in AK_d(A, r_0) = AK_N(A, r_0)$. In the derivation of our new bounds we will always make this assumption. For further analysis of the application of GMRES to singular systems, in particular the case $r_0 \notin AK_d(A, r_0)$, we refer to [4].

3. GMRES-equivalent matrices. For a given matrix $A \in \mathbf{C}^{N\times N}$ and vector $r_0 \in \mathbf{C}^N$, for which $r_0 \in AK_N(A, r_0)$, we now characterize the matrices $B \in \mathbf{C}^{N\times N}$ that satisfy $AK_n(A, r_0) = BK_n(B, r_0)$, for $1 \leq n \leq \dim AK_N(A, r_0)$.

Since A might be singular and we allow $\dim AK_N(A, r_0) < N$, we obtain generalizations of results of Greenbaum and Strakoš [15, section 2] (also cf. [1, 14]). In the following, by $r_n^{M,r}$ we denote the nth residual, when GMRES is applied to a linear system with the matrix M and the initial residual is r.

General assumptions. A linear system (1.1) and initial guess x_0 are given such that $2 \leq \dim AK_N(A, r_0) =: d \leq N$ and $r_0 \in AK_d(A, r_0)$. The latter assumption is always satisfied when A is nonsingular (cf. Theorem 2.8). Assuming $d \geq 2$, we exclude the trivial cases of GMRES termination at steps zero and one. Furthermore, let $w_1, w_2, \ldots, w_d$ be orthonormal vectors that satisfy

$$\operatorname{span}\{w_1, w_2, \ldots, w_n\} = AK_n(A, r_0) \text{ for } 1 \leq n \leq d. \tag{3.1}$$

Denoting $W_d := [w_1, w_2, \ldots, w_d] \in \mathbf{C}^{N\times d}$, we have

$$r_0 = W_d h, \text{ where } h := [\eta_1, \ldots, \eta_d]^T, \text{ and } \eta_n := w_n^H r_0 \text{ for } 1 \leq n \leq d. \tag{3.2}$$

It is easy to see that

$$AW_d = W_d H, \tag{3.3}$$

where $H \in \mathbf{C}^{d\times d}$ is a nonsingular unreduced upper Hessenberg matrix. If H was not unreduced, then for some n, $1 \leq n \leq d-1$, Aw_n would be a linear combination of $w_1, \ldots, w_n$. This contradicts the dimension assumption on $AK_N(A, r_0)$. Also note that if we define

$$\hat{H} := \begin{bmatrix} 0 & \ldots & 0 & 1/\eta_d \\ 1 & & & -\eta_1/\eta_d \\ & \ddots & & \vdots \\ & & 1 & -\eta_{d-1}/\eta_d \end{bmatrix}, \tag{3.4}$$

then

$$H = \tilde{R}\hat{H}, \tag{3.5}$$

where $\tilde{R} \in \mathbf{C}^{d\times d}$ is some nonsingular and upper triangular matrix.

DEFINITION 3.1. *Suppose that $A \in \mathbf{C}^{N\times N}$ and $r_0 \in \mathbf{C}^N$. Then $B \in \mathbf{C}^{N\times N}$ is called GMRES-equivalent to A with respect to r_0 if*

$$\dim BK_N(B, r_0) = d \text{ and } AK_n(A, r_0) = BK_n(B, r_0) \text{ for } 1 \leq n \leq d,$$

where $d := \dim AK_N(A, r_0)$.

Clearly, if B is GMRES-equivalent to A with respect to r_0, then

$$r_n^{A,r_0} = r_n^{B,r_0} \text{ for } 1 \le n \le d.$$

The following result is a generalization of [15, Theorems 2.1 and 2.2].

THEOREM 3.2. *Under our general assumptions, for a matrix* $B \in \mathbf{C}^{N\times N}$ *the following three statements are equivalent:*

(a) *B is GMRES-equivalent to A with respect to r_0.*

(b) *B satisfies $BW_d = W_d\tilde{R}\hat{H}$, where $\hat{H}$ is defined in* (3.4) *and $\tilde{R} \in \mathbf{C}^{d\times d}$ is a nonsingular upper triangular matrix.*

(c) *B satisfies $BW_d = W_d\bar{R}H$, where H is defined in* (3.3) *and $\bar{R} \in \mathbf{C}^{d\times d}$ is a nonsingular upper triangular matrix.*

Proof. The equivalence of (b) and (c) is obvious from (3.5). We thus have to show only that (a) and (b) are equivalent. Suppose that (a) holds. Since $r_0 = W_d h$, we have

$$B[r_0, W_{d-1}] = BW_d\hat{H}^{-1}. \tag{3.6}$$

Furthermore, from (3.1) and the GMRES-equivalence of A and B, it follows that

$$B[r_0, W_{d-1}] = W_d\tilde{R} \tag{3.7}$$

for some upper triangular matrix $\tilde{R}$. It is easy to show that $\tilde{R}$ must be nonsingular. Now (3.6) and (3.7) yield $BW_d = W_d\tilde{R}\hat{H}$.

Conversely, if (b) holds, then $B^n r_0 = B^n W_d h = r_{11} W_d(\tilde{R}\hat{H})^{n-1}e_1$. Since $\tilde{R}$ is nonsingular and upper triangular, $AK_n(A, r_0) = \operatorname{span}\{w_1, w_2, \ldots, w_n\} = BK_n(B, r_0)$ for $1 \le n \le d$. Finally we note that $B^{d+1}r_0 \in BK_d(B, r_0)$, so that $\dim BK_N(B, r_0) = d$. ∎

By varying $\tilde{R}$ and $\bar{R}$ in Theorem 3.2, we obtain a whole class of matrices which are GMRES-equivalent to the given matrix A with respect to the initial residual $r_0 = W_d h$. We obtain a useful corollary.

COROLLARY 3.3. *Suppose that, in the notation of Theorem* 3.2,

$$B = W_d\bar{R}HW_d^H \text{ and } \hat{B} = W_d\tilde{R}\hat{H}W_d^H,$$

where $\bar{R}$ and $\tilde{R}$ are arbitrary nonsingular upper triangular matrices. Then B and $\hat{B}$ are both GMRES-equivalent to A with respect to r_0 and, in particular,

$$\|r_n^{B,r_0}\| = \|r_n^{\hat{B},r_0}\| = \|r_n^{A,r_0}\| \text{ for } 1 \le n \le d.$$

We remark that if $d < N$, then the matrices B and $\hat{B}$ in Corollary 3.3 must be singular, regardless of whether A is singular.

4. New bounds on GMRES residual norms. The results of this section are derived under the general assumptions stated in section 3. Using Theorem 3.2, we can relate the convergence of GMRES for the given linear system to the convergence for some other system, *which is easier to analyze.* Our starting point for the derivation of the new bounds therefore is the construction of suitable matrices that are GMRES-equivalent to A with respect to the given r_0.

As stated in Corollary 3.3, for arbitrary nonsingular upper triangular matrices $\bar{R}, \tilde{R} \in \mathbf{C}^{d\times d}$, the matrices

$$W_d\bar{R}HW_d^H \text{ and } W_d\tilde{R}\hat{H}W_d^H \tag{4.1}$$

are both GMRES-equivalent to A with respect to r_0. We consider the following decompositions of H and $\hat{H}$:

$$H = QR \quad \text{and} \quad \hat{H} = \hat{R}\hat{Q}. \tag{4.2}$$

The matrices Q and $\hat{Q}$ are unitary and the matrices R and $\hat{R}$ are nonsingular and upper triangular. All four are of size $d \times d$. These two decompositions always exist due to nonsingularity of H and $\hat{H}$ (see, e.g., [11]).

Two other similar decompositions are possible: $H = R_1 Q_1$ and $\hat{H} = Q_2 R_2$. We are particularly interested in the spectra of the Q-factors of these decompositions. Since Q_2 is simply a unitary shift matrix and the results derived for $\Lambda(\hat{Q})$ also hold for $\Lambda(Q_1)$ (cf. (3.5)), we do not consider these two other decompositions.

For later use we note that $W_d^H A W_d = QR$ yields

$$\kappa(R) \le \kappa(A) \tag{4.3}$$

with equality if $d = N$. Often $d = N$ is a reasonable assumption, in particular when the "generic" case of an initial residual without special properties is to be analyzed.

Choosing $\bar{R} = R^{-1}$ and $\tilde{R} = \hat{R}^{-1}$ in (4.1) and using (4.2), we define the matrices

$$B := W_d R^{-1} Q R W_d^H \quad \text{and} \quad \hat{B} := W_d \hat{Q} W_d^H, \tag{4.4}$$

which are both GMRES-equivalent to A with respect to r_0.

Remark 4.1. Greenbaum and Strakoš [15, section 3.1] consider the RQ-decomposition of the matrix H. They point out that if $d = N$, then the matrix $\hat{B} = W_N \hat{Q} W_N^H$ is unitary. Thus, whenever A and r_0 are such that $d = N$, there exists a unitary matrix which is GMRES-equivalent to A with respect to r_0.

THEOREM 4.2. *Suppose that B and $\hat{B}$ are defined as in* (4.4). *Then*

$$\frac{\|r_n\|}{\|r_0\|} \le \kappa(R) \min_{p_n \in \pi_n} \max_{\lambda \in \Lambda(Q)} |p_n(\lambda)| \tag{4.5}$$

and

$$\frac{\|r_n\|}{\|r_0\|} \le \min_{p_n \in \pi_n} \max_{\lambda \in \Lambda(\hat{Q})} |p_n(\lambda)| \tag{4.6}$$

for $1 \le n \le d$.

Proof. Corollary 3.3 yields for $1 \le n \le d$,

$$\begin{aligned} \|r_n\| \equiv \|r_n^{A,r_0}\| = \|r_n^{B,r_0}\| &= \min_{p_n \in \pi_n} \|W_d R^{-1} p_n(Q) R W_d^H r_0\| \\ &\le \|R^{-1}\| \, \|R\| \min_{p_n \in \pi_n} \|p_n(Q)\| \, \|r_0\|, \end{aligned}$$

from which (4.5) follows. The proof of (4.6) is similar. ∎

To study the approximation problem

$$\min_{p_n \in \pi_n} \max_{\lambda \in \Lambda(U)} |p_n(\lambda)|, \quad U \text{ unitary}, \tag{4.7}$$

we need the following definition.

DEFINITION 4.3. *Suppose that the spectrum of the $d \times d$ unitary matrix U is given by $\Lambda(U) := \{e^{i\beta_j} \ : \ 0 \le \beta_1 \le \beta_2 \le \cdots \le \beta_d < 2\pi\}$ and let $\beta_{d+1} := 2\pi + \beta_1$. We define the d gaps in $\Lambda(U)$ by*

$$g_j := \beta_{j+1} - \beta_j \ \text{ for } \ 1 \le j \le d. \tag{4.8}$$

For $1 \le j \le d$, the complex number $e^{i(\beta_{j+1}+\beta_j)/2}$ is called the center of the gap g_j. $\phi := \max\{g_1, g_2, \ldots, g_d\}$ denotes the largest gap in $\Lambda(U)$.

Note that for every unitary matrix U and every $\alpha \in \mathbf{R}$, we have

$$\min_{p_n \in \pi_n} \max_{\lambda \in \Lambda(U)} |p_n(\lambda)| = \min_{p_n \in \pi_n} \max_{\lambda \in \Lambda(e^{i\alpha}U)} |p_n(\lambda)|. \tag{4.9}$$

Therefore, in the study of (4.7) we can assume, without loss of generality, that 1 is the center of the largest gap ϕ in $\Lambda(U)$. Let us define

$$\Omega_\phi := \left\{ e^{i\alpha} \ : \ \frac{\phi}{2} \le \alpha \le 2\pi - \frac{\phi}{2} \right\}. \tag{4.10}$$

Then (4.9) shows that

$$\min_{p_n \in \pi_n} \max_{\lambda \in \Lambda(U)} |p_n(\lambda)| \le \min_{p_n \in \pi_n} \max_{z \in \Omega_\phi} |p_n(z)|. \tag{4.11}$$

Suppose that $0 < \phi < 2\pi$, or, equivalently, that $\Lambda(U)$ contains at least two distinct points. Then for $n \ge 1$, we can introduce the polynomials

$$p_n(z) := \prod_{k=0}^{n-1} \left(1 - z \frac{(1 - \gamma e^{i\psi_k})}{e^{i\psi_k}\,(\gamma - e^{i\psi_k})} \right), \text{ where} \tag{4.12}$$

$$\psi_k := \frac{2\pi k}{n} \ \text{ for } \ 0 \le k \le n-1, \tag{4.13}$$

$$\gamma := \left(\cos \frac{\phi}{4} \right)^{-1}. \tag{4.14}$$

By assumption, $1 < \gamma < \infty$ and hence the polynomials (4.12) are well defined for every $n \ge 1$. Furthermore, we have $p_n \in \pi_n$.

Remark 4.4. The polynomials p_n in (4.12) are constructed using a conformal mapping technique from [19] (also cf. [21]). As shown there, $\Psi(z) := \frac{z(\gamma - z)}{1 - \gamma z}$ maps the exterior of the unit circle conformally onto the exterior of Ω_ϕ. Thus, the zeros of p_n are Fejér points (see, e.g., [29, Chapter 1]) associated with Ω_ϕ.

LEMMA 4.5. *The polynomials introduced in* (4.12) *satisfy*

$$\max_{z \in \Omega_\phi} |p_n(z)| \le \frac{4}{\gamma^n - 1} \ \textit{ for } \ n \ge 1. \tag{4.15}$$

Proof. Suppose that $z = e^{i\alpha} \in \Omega_\phi$, i.e., $\alpha \in \left[\frac{\phi}{2}, 2\pi - \frac{\phi}{2} \right]$. A simple manipulation yields

$$\begin{aligned} \left| p_n(e^{i\alpha}) \right| &= \prod_{k=0}^{n-1} \left| 1 - e^{i\alpha} \frac{(1 - \gamma e^{i\psi_k})}{e^{i\psi_k}\,(\gamma - e^{i\psi_k})} \right| \\ &= \prod_{k=0}^{n-1} \left| \frac{e^{i\frac{\alpha}{2}} \left(\gamma \left(e^{i\frac{\alpha}{2}} + e^{-i\frac{\alpha}{2}} \right) - \left(e^{i(\psi_k - \frac{\alpha}{2})} + e^{-i(\psi_k - \frac{\alpha}{2})} \right) \right)}{\gamma - e^{i\psi_k}} \right| \\ &= \frac{2^n}{\gamma^n - 1} \prod_{k=0}^{n-1} \left| \gamma \cos \frac{\alpha}{2} - \cos\left(\psi_k - \frac{\alpha}{2} \right) \right|. \end{aligned} \tag{4.16}$$

In the last equation, we utilized that $\prod_{k=0}^{n-1}\left(z-e^{i\psi_k}\right)=z^n-1$ for all $z\in\mathbf{C}$. Now note that $-1\le\gamma\cos\frac{\alpha}{2}\le 1$. We define $\beta:=\arccos\left(\gamma\cos\frac{\alpha}{2}\right)$. Then (4.16) yields

$$\begin{aligned}
\left|p_n(e^{i\alpha})\right| &= \frac{2^n}{\gamma^n-1}\prod_{k=0}^{n-1}\left|\cos\beta-\cos\left(\psi_k-\frac{\alpha}{2}\right)\right| \\
&= \frac{4^n}{\gamma^n-1}\prod_{k=0}^{n-1}\left|\sin\frac{\beta+\psi_k-\frac{\alpha}{2}}{2}\quad\sin\frac{\beta-\psi_k+\frac{\alpha}{2}}{2}\right| \\
&= \frac{2^n}{\gamma^n-1}\prod_{k=0}^{n-1}\left|\sqrt{1-\cos\left(\beta+\psi_k-\frac{\alpha}{2}\right)}\sqrt{1-\cos\left(\beta-\psi_k+\frac{\alpha}{2}\right)}\right| \\
&= \frac{1}{\gamma^n-1}\prod_{k=0}^{n-1}\left|e^{i\left(\frac{\alpha}{2}-\beta\right)}-e^{i\psi_k}\right|\left|e^{i\left(\frac{\alpha}{2}+\beta\right)}-e^{i\psi_k}\right| \\
&= \frac{1}{\gamma^n-1}\left|e^{i\left(\frac{\alpha}{2}-\beta\right)n}-1\right|\left|e^{i\left(\frac{\alpha}{2}+\beta\right)n}-1\right| \;\le\; \frac{4}{\gamma^n-1}. \qquad \square
\end{aligned}$$

As Lemma 4.5 shows, the polynomials (4.12) converge to zero on Ω_ϕ, with the speed of convergence being directly related to the gap ϕ: the larger ϕ, the faster the convergence, and vice versa. Since the polynomials (4.12) are constructed using Fejér points, they converge to zero on Ω_ϕ with the largest geometric degree of convergence (cf. [36, Chapter 4.7]). Because of the moderate constant 4, Lemma 4.5 is useful not only asymptotically, but also for small n.

We now combine (4.3), (4.11), Theorem 4.2, and Lemma 4.5 for our main convergence result. Note that from our general assumption $d\ge 2$, it follows that $\Lambda(Q)$ and $\Lambda(\hat{Q})$ both contain at least two distinct points, so that Lemma 4.5 is applicable.

THEOREM 4.6. *Suppose that a linear system* (1.1) *and an initial guess* x_0 *are given, such that* $2\le d:=\dim AK_N(A,r_0)\le N$ *and* $r_0\in AK_d(A,r_0)$. *Define the matrices* Q, R, *and* $\hat{Q}$ *as in* (4.2) *and let* ϕ *and* $\hat{\phi}$ *denote the largest gaps in* $\Lambda(Q)$ *and* $\Lambda(\hat{Q})$, *respectively. Define* $\gamma:=(\cos\frac{\phi}{4})^{-1}$ *and* $\hat{\gamma}:=(\cos\frac{\hat{\phi}}{4})^{-1}$; *then the GMRES residuals for* (1.1) *and* x_0 *satisfy*

$$\frac{\|r_n\|}{\|r_0\|}\;\le\;4\frac{\kappa(R)}{\gamma^n-1} \tag{4.17}$$

and

$$\frac{\|r_n\|}{\|r_0\|}\;\le\;\frac{4}{\hat{\gamma}^n-1} \tag{4.18}$$

for $1\le n\le d$. *Furthermore,* $\kappa(R)\le\kappa(A)$ *with equality if* $d=N$.

It is important to note that we cannot expect (4.17) and (4.18) to be sharp, except for special cases. There are several reasons for this: First, the proof of Lemma 4.5 uses the submultiplicative property of the Euclidean norm, which generally does not lead to sharp bounds. Second, while the polynomials used in Lemma 4.5 are asymptotically optimal, the GMRES algorithm actually constructs polynomials that are optimal in each step. Third, both our bounds decrease with constant rates, each being derived from only one largest spectral gap. But there might be several considerable gaps in $\Lambda(Q)$ and $\Lambda(\hat{Q})$ that have an impact on the convergence behavior.

As shown in the remainder of this paper, despite all these theoretical objections, (4.17) and (4.18) are valuable for describing and predicting the behavior of GMRES.

We first point out that by construction γ^{-1} and $\hat{\gamma}^{-1}$, the rates of reduction of our bounds, depend on the initial guess. Both rates lie strictly between zero and one. We have not found an algebraic relation between them.

At first sight, (4.18) seems to be more useful than (4.17), since it does not include the potentially large constant $\kappa(R)$. However, it is possible to implement GMRES in a way that approximations to (4.17) are easily computable in each step of the algorithm. On the other hand, computing approximations to (4.18) during the GMRES run is sometimes not possible and is always more expensive.

Furthermore, the appearance of the constant $\kappa(R)$ leads to the following observation: Suppose that GMRES stagnates early in the iteration. Then $\hat{\phi}$ must be small, because (4.18) involves no large constant. But because of $\kappa(R)$ in (4.17), early stagnation of GMRES does not generally imply that ϕ is small. Hence, γ^{-1} and $\hat{\gamma}^{-1}$ might be of different sizes. In particular, when $\kappa(R)$ is large, the rate γ^{-1} has the potential to describe the convergence of GMRES after an initial phase of stagnation. Numerical experiments confirming this observation are shown in section 6.

5. Approximate bounds predict the behavior of GMRES. Note that if close approximations to the right-hand sides of (4.17) and (4.18) are available in early stages of the computation, then these potentially *predict* the behavior of the algorithm in later stages. It is therefore useful to compute such approximations during the GMRES run.

5.1. WZ-GMRES. To obtain the approximations, we consider an implementation of GMRES first proposed by Walker and Zhou [35]. In the following we will refer to this implementation as WZ-GMRES. For a description of WZ-GMRES, we need to introduce the Arnoldi process [2]:

Suppose that $M \in \mathbf{C}^{N\times N}$ and a unit vector $w_1 \in \mathbf{C}^N$ satisfy $\dim K_N(M, w_1) = d \leq N$. Then the Arnoldi process applied to M and w_1 is the construction of an orthonormal basis, $w_1, w_2, \ldots, w_d$, of $K_d(M, w_1)$ by means of any orthogonalization method. In many references (see, e.g., [11]), the orthogonalization performed by the Arnoldi process is implemented using the classical or modified Gram–Schmidt procedure. As pointed out by Walker [34], it is also possible to use Householder transformations for this purpose. For a collection of different implementations of the Arnoldi process, we refer to [7] (see also [24]). Our development is independent of the particular implementation, so that any (numerically stable) one will serve our purpose.

In practice, d is not known, and the basis vectors are constructed recursively. After $n < d$ steps of the Arnoldi process, we have

$$MW_n = W_{n+1}H_n^e, \quad W_n := [w_1, \ldots, w_n], \tag{5.1}$$

where

$$H_n^e = \begin{bmatrix} h_{11} & \ldots & h_{1n} \\ h_{21} & \ddots & \vdots \\ & \ddots & h_{nn} \\ & & h_{n+1,n} \end{bmatrix} \tag{5.2}$$

is an $(n+1)\times n$ unreduced upper Hessenberg matrix. In exact arithmetic, the process has to terminate at step d with the final matrix relation

$$MW_d = W_dH_d, \quad W_d \in \mathbf{C}^{N\times d}, \quad H_d := W_d^H AW_d \in \mathbf{C}^{d\times d}. \tag{5.3}$$

The original implementation of GMRES [28] is based on the Arnoldi process applied to the matrix A and the vector $r_0/\|r_0\|$. WZ-GMRES uses

$$w_1 := Ar_0/\|Ar_0\| \tag{5.4}$$

as the starting vector for the Arnoldi process. Recall that for our given linear system and initial guess, we have assumed $\dim AK_N(A, r_0) = d$. Then, since

$$K_n(A, w_1) = AK_n(A, r_0) \text{ for } 1 \le n \le d,$$

the application of the Arnoldi process to A and w_1 yields orthonormal vectors that satisfy (3.1). Using these vectors, (2.1) can be rewritten as

$$x_n = x_0 + [r_0/\|r_0\|, W_{n-1}]y_n \tag{5.5}$$

for some $y_n \in \mathbf{C}^n$, $1 \le n \le d$. To implement GMRES, we use the orthogonality relation (2.4) in the computation of y_n. Together with (5.1) and (5.5), this yields

$$0 = W_n^H r_n = W_n^H r_0 - W_n^H A[r_0/\|r_0\|, W_{n-1}]y_n,$$

which is equivalent to

$$[(\|Ar_0\|/\|r_0\|)e_1, H_{n-1}^e]y_n = W_n^H r_0. \tag{5.6}$$

Thus, y_n is the solution of an upper triangular system. Since H_{n-1}^e is unreduced, y_n is well defined. It is important to note that r_n can be obtained without explicitly computing x_n [25],

$$r_n = r_{n-1} - (w_n^H r_0)w_n. \tag{5.7}$$

The upper triangular solve for y_n therefore has to be performed only if the residual norm computed from (5.7) is small enough. Clearly, in exact arithmetic, the termination of the Arnoldi process at step d is equivalent to the termination of WZ-GMRES at step d. We remark that from (5.5) and (5.6) it is easy to see that WZ-GMRES is (mathematically) independent of the orthogonalization method used for the Arnoldi process. For more details about WZ-GMRES, we refer to [21, 25, 35].

5.2. Computing approximations to (4.17). If (5.4) holds, then the vectors $w_1, w_2, \ldots, w_d$ produced by the Arnoldi process satisfy (3.1). Consequently, the matrix H_d in (5.3) can be identified with the matrix H in (4.2). We can therefore compute the numbers ϕ and $\kappa(R)$ used in (4.17) from the QR-decomposition of H_d. Thus, the right-hand side of (4.17) is computable from information generated during the run of WZ-GMRES. Moreover, because of the recursive nature of the Arnoldi process, the matrix H_n^e is the leading $(n+1) \times n$ submatrix of H_d. For $n = 1, 2, \ldots, d-1$, we can therefore decompose

$$H_n^e = Q_n \begin{bmatrix} R_n \\ 0 \end{bmatrix}, \quad Q_n \in \mathbf{C}^{(n+1)\times(n+1)}, \quad R_n \in \mathbf{C}^{n\times n}, \tag{5.8}$$

and use ϕ_n, the largest gap in $\Lambda(Q_n)$, as an approximation to ϕ and $\kappa(R_n)$ as an approximation to $\kappa(R)$. Note that because H_n^e is unreduced, R_n is nonsingular and (5.8) is well defined.

The decomposition (5.8) is particularly easy to compute. It is well known that because of the special structure of H_n^e, the matrix Q_n can be derived as a product of n Givens transformations, $Q_n = G_1 \cdots G_n$, where

$$G_j = \begin{bmatrix} I_{j-1} & & & \\ & c_j & -\overline{s}_j & \\ & s_j & c_j & \\ & & & I_{n-j} \end{bmatrix} \in \mathbf{C}^{(n+1)\times(n+1)}, \quad 1 \le j \le n. \tag{5.9}$$

Suppose that for some n, $2 \le n \le d-1$,

$$Q_{n-1}^H H_{n-1}^e = \begin{bmatrix} R_{n-1} \\ 0 \end{bmatrix}, \quad Q_{n-1}^H = G_{n-1}^H \cdots G_1^H. \tag{5.10}$$

We define

$$h^n := [h_{1n}, \ldots, h_{nn}]^T, [\rho_1, \ldots, \rho_n]^T := Q_{n-1}^H h^n, \text{ and } \nu_n := [\rho_1, \ldots, \rho_{n-1}]^T.$$

Then

$$\begin{aligned} Q_n^H H_n^e &= G_n^H \begin{bmatrix} Q_{n-1}^H & \\ & 1 \end{bmatrix} \begin{bmatrix} H_{n-1}^e & h^n \\ & h_{n+1,n} \end{bmatrix} \\ &= G_n^H \begin{bmatrix} R_{n-1} & \nu_n \\ & \rho_n \\ & h_{n+1,n} \end{bmatrix}. \end{aligned} \tag{5.11}$$

If $\rho_n = |\rho_n| e^{i\delta_n}$, we define $\omega_n := (|\rho_n|^2 + |h_{n+1,n}|^2)^{1/2}$ and choose

$$c_n = \frac{|\rho_n|}{\omega_n}, \tag{5.12}$$

$$s_n = \frac{e^{-i\delta_n} h_{n+1,n}}{\omega_n} \tag{5.13}$$

to give (5.11) the required upper triangular form. Hence, to compute Q_n and R_n from Q_{n-1} and R_{n-1}, respectively, only one matrix-vector product and a few flops to compute c_n and s_n are required.

To compute ϕ_n we have to solve an $n \times n$ unitary eigenproblem. For the efficient computation of $\kappa(R_n)$ we can use the method of incremental condition estimation [3]. With this method, the computation of $\kappa(R_n)$ costs only $\mathcal{O}(n)$ operations, provided $\kappa(R_{n-1})$ is known.

5.3. Computing approximations to (4.18). Suppose that we apply WZ-GMRES to (1.1) and x_0 and use the Arnoldi vectors $w_1, w_2, \ldots, w_d$ to set up the matrix $\hat{H}$ as defined in (3.4). Note that $\hat{H}$ can be obtained at no additional cost if the updated residuals are computed using (5.7). The bound (4.18) can then be computed easily because it depends only on the largest gap $\hat{\phi}$ in $\Lambda(\hat{Q})$, where $\hat{Q}$ is defined by the RQ-decomposition $\hat{H} = \hat{R}\hat{Q}$ (cf. (4.2)).

For approximations to (4.18), we consider (5.7), which yields

$$|\eta_n| = |w_n^H r_0| = \|r_{n-1} - r_n\| \text{ for } 1 \le n \le d.$$

Hence, if GMRES does not stagnate in step n, $2 \le n \le d-1$, then $\eta_n \neq 0$ and the decomposition

$$\hat{H}_n := \begin{bmatrix} 0 & \dots & 0 & 1/\eta_n \\ 1 & & & -\eta_1/\eta_n \\ & \ddots & & \vdots \\ & & 1 & -\eta_{n-1}/\eta_n \end{bmatrix} = \hat{R}_n \hat{Q}_n \in \mathbf{C}^{n\times n} \tag{5.14}$$

is defined. In this case, $\hat{\phi}_n$, the largest gap in $\Lambda(\hat{Q}_n)$, can be used as an approximation to $\hat{\phi}$. However, we have no easy way to update the decomposition (5.14). $\hat{R}_n$ and $\hat{Q}_n$ have to be recomputed from scratch in each step. As shown above, the situation is drastically different for the decomposition (5.8).

An approximation $\hat{\phi}_n$ can also be computed from the RQ-decomposition of the matrix $H_n := W_n^H A W_n$ (cf. (5.1)). In exact arithmetic, stagnation of WZ-GMRES in step n leads to singularity of H_n. Hence $\hat{\phi}_n$ is not well defined in this case. But on the contrary to (5.14), $\hat{\phi}_n$ might still be computable. This is particularly true in finite precision arithmetic.

5.4. Behavior of the approximations and prediction of the performance of GMRES. Q_n in (5.8) is obtained from Q_{n-1} by a low-rank modification. If $q_1, \dots, q_n$ denote the columns of Q_{n-1}, then

$$Q_n = \begin{bmatrix} q_1 & \dots & q_{n-1} & c_n q_n & -\overline{s}_n q_n \\ 0 & \dots & 0 & s_n & c_n \end{bmatrix}.$$

By (5.12), $c_n \neq -1$, so that we can readily apply results of Elsner and He [9, section 5]. We partition

$$Q_n = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix}, \quad U_{11} = [q_1, \dots, q_{n-1}, c_n q_n] \in \mathbf{C}^{n\times n}, \quad U_{22} = c_n$$

to get, in the notation of [9, Theorem 5.3]: $U_n := U_{11} - U_{12}(I + U_{22})^{-1} U_{21} = Q_{n-1}$. From [9, Theorem 5.3] it follows that if

$$\begin{aligned} \Lambda(Q_{n-1}) = \Lambda(U_n) &= \{e^{i\beta_j} \ : \ 0 \le \beta_1 \le \dots \le \beta_n < 2\pi\}, \\ \Lambda(Q_n) &= \{e^{i\tau_j} \ : \ 0 \le \tau_1 \le \dots \le \tau_{n+1} < 2\pi\}, \end{aligned}$$

then $\tau_j \le \beta_j \le \tau_{j+1}$ for $1 \le j \le n$.

Because of this interlacing property, we can expect a "smooth" behavior of the eigenvalues of the matrices Q_n and, in particular, of their largest gaps. This is not necessarily the case for the spectra of the matrices $\hat{Q}_n$ in (5.14), since we know of no easy algebraic relation between them.

The matrix R_n in (5.8) is the leading principal submatrix R in (4.2). Thus, we have $\kappa(R_{n-1}) \le \kappa(R_n) \le \kappa(R)$ for $n = 2, \dots, d-1$. Monitoring the numbers $\kappa(R_n)$ is a useful check on the numerical accuracy of the WZ-GMRES iterates. As (5.6) shows, the conditioning of R_{n-1} determines the accuracy of the vector y_n. In the derivation of our bounds we have assumed $r_0 \in AK_d(A, r_0)$ for some $d \le N$. From this assumption follows that $\kappa(R_n) \le \kappa(R) < \infty$. But for ill-conditioned (or singular) A, $\kappa(R)$ and thus $\kappa(R_n)$ might still be very large. In such a case it might be useful to terminate WZ-GMRES early.

In numerical experiments we often observed that for small n, ϕ_n, $\kappa(R_n)$, and $\hat{\phi}_n$ were close to ϕ, $\kappa(R)$, and $\hat{\phi}$, respectively. Such observations become plausible following this heuristic reasoning: In many cases even small leading submatrices of H convey much of the information obtained by the Arnoldi process. For example, close approximations to $\kappa(A)$ and $\Lambda(A)$ can typically be obtained from H_n for small n (see, e.g., [26]). In the same fashion, information about properties of the factors in the QR- and RQ-factorizations of H should be obtainable from its leading submatrices. Hence, close approximations to quantities associated with Q, R, and $\hat{Q}$ should for small n typically be obtainable from Q_n, R_n, and $\hat{Q}_n$, respectively.

With close approximations to ϕ, $\kappa(R)$, and $\hat{\phi}$, predictions can be made on how GMRES will perform in later stages of the iteration. Theorem 4.6 implies the following heuristic for such predictions.

HEURISTIC 5.1. *When the approximations ϕ_n, $\kappa(R_n)$, and $\hat{\phi}_n$ are close to ϕ, $\kappa(R)$, and $\hat{\phi}$, respectively, their sizes predict the behavior of GMRES as follows:*

1. *A large ϕ_n predicts fast GMRES convergence, possibly after an initial phase of stagnation in case $\kappa(R_n)$ is large. A small ϕ_n predicts slow convergence over the whole course of the iteration.*
2. *A large $\hat{\phi}_n$ predicts fast GMRES convergence over the whole course of the iteration. A small $\hat{\phi}_n$ predicts (at least early) stagnation of GMRES.*

Our experiments indicate that stagnation of the approximations for several steps typically indicates that they are close to their final values. This is not, however, a reliable method of checking the accuracy of approximations. For additional information it is possible to check the convergence of the Arnoldi process itself, for example, by estimating how closely $\Lambda(H_n)$ approximates $\Lambda(A)$. Finding direct bounds on the errors $\|\phi - \phi_n\|$, etc., is an interesting open problem which we plan to study in the future.

6. Numerical experiments. We present numerical experiments run in MATLAB 5.0 [32] using five different nonsymmetric (A199: nonhermitian) test matrices:

A	N	$\kappa(A)$	Data type
A199	199	153.5	complex
PDE225	225	39.1	real
FS1836	183	1.7e+11	real
STEAM1	240	2.8e+07	real
ISING100	100	1.0	real

The matrices PDE225, FS1836, and STEAM1 can be obtained from the MatrixMarket website at http://math.nist.gov/MatrixMarket. PDE225 comes from the finite difference discretization of a two-dimensional elliptic partial differential operator, FS1836 represents a problem in chemical kinetics, and STEAM1 comes from an oil recovery problem. More information on these matrices can be obtained from the MatrixMarket website. More information on A199 and ISING100 is given below.

For our experiments we use a Householder version [34] of WZ-GMRES. The approximations ϕ_n are computed as described in section 5.2 and the $\hat{\phi}_n$ are derived from the RQ-decomposition of $H_n = W_n^H A W_n$ (cf. section 5.3). To check the numerical accuracy of our computations, we compare the norms of the directly computed residuals $(b - Ax_n)$ and the updated residuals (cf. (5.7)).

Before presenting our results we point out that the bounds (4.17) and (4.18), in particular the approximations ϕ_n and $\hat{\phi}_n$, are sensitive to the scaling of the diagonal entries of the R-factors in the QR- and RQ-decompositions. The results of this section

are derived with the following choices:

- The QR-decompositions (5.8) are computed using Givens transformations with nonnegative cosines, i.e., $c_n \geq 0$ for all n (cf. (5.12)).
- The RQ-decompositions $H_n = \hat{R}_n \hat{Q}_n$ are computed so that the $\hat{R}_n$ have positive diagonal entries.

Of all the scalings we tested, these choices typically led to the largest gaps ϕ and $\hat{\phi}$ and thus yielded the smallest upper bounds (4.17) and (4.18).

Experiment 6.1. We use the matrices A199, PDE225, and FS1836 to show typical types of behavior of the quantities in our bounds and of the computed approximations. A199 is a complex matrix of the form

$$A199 = \begin{bmatrix} 1 & & & \\ & 1.5 & & \\ & & \ddots & \\ & & & 100 \end{bmatrix} + iB \in \mathbf{C}^{199\times 199},$$

where B is a random matrix generated with the MATLAB command `B=rand(199)`.

For A = A199 and A = PDE225, we use $b = A[1,\ldots,1]^T$ and $x_0 = 0$. We terminate WZ-GMRES at step d when the updated scaled residual norm satisfies $\|r_d\|/\|r_0\| < 10^{-14}$. Since the residual norm has decreased almost to machine precision, $\mathbf{u} = 10^{-16}$, we identify d with $\dim AK_N(A, r_0)$. Hence, we consider last-computed approximations as equal to the quantities used in Theorem 4.6:

	d	ϕ	$\hat{\phi}$	$\kappa(R)$
A199	94	2.1325	2.1601	150.7
PDE225	93	1.9675	1.4250	34.9

Figure 6.1 shows the convergence curve of WZ-GMRES for A199 as well as our convergence bounds. We stress that the actual right-hand sides of (4.17) and (4.18) are plotted and not just the "rates of convergence" γ^{-1} and $\hat{\gamma}^{-1}$. Both bounds describe the actual rate of convergence well. Due to the constant $\kappa(R)$, (4.17) is slightly weaker than (4.18). Figures 6.2 and 6.4 show that the approximations $\kappa(R_n)$ and ϕ_n, $\hat{\phi}_n$ converge "smoothly" and reach their final levels after approximately 40 steps.

Note that, as shown in Figure 6.3, there is only one significant gap in $\Lambda(Q)$ (the same holds for $\Lambda(\hat{Q})$). The number of significant gaps in $\Lambda(Q)$ and $\Lambda(\hat{Q})$ has an important influence on our bounds: The rates of reduction of both bounds depend only on the largest gap in the two respective spectra. Thus, the occurrence of several large gaps might cause our bounds to overestimate the actual convergence behavior. As mentioned in section 4, this is one reason we do not generally expect our bounds to be sharp. However, our bounds, as well as Heuristic 5.1, will not lose their predictive value in cases of several large gaps, as the occurrence of such gaps will still correspond to and predict fast GMRES convergence. Experiment 6.3 gives an example of our bounds in the context of two large gaps.

Results for PDE225 are shown in Figures 6.5–6.8, which have the same notation as Figures 6.1–6.4, respectively. Note the considerable difference between ϕ and $\hat{\phi}$. As Theorem 4.6 implies, $\hat{\phi}$ must be small due to slow initial GMRES convergence. Because of the constant $\kappa(R)$ in (4.17), ϕ might be—and actually is—larger than $\hat{\phi}$.

For A = FS1836 (see Figures 6.9–6.12) we use two different right-hand sides,

$$b^{(1)} = [1,1,\ldots,1]^T, \quad b^{(2)} = A[1,1,\ldots,1]^T,$$

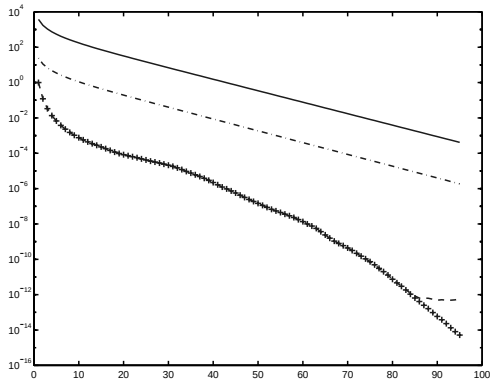


FIG. 6.1. *A*199: *Computed (dashed) and updated (pluses) residual norms, bounds* (4.17) *(solid) and* (4.18) *(dashdot).*

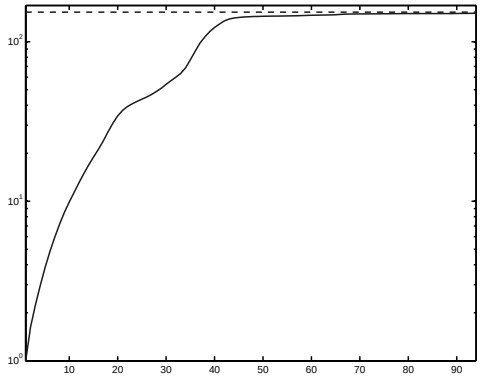


FIG. 6.2. *A*199: $\kappa(R_n)$.

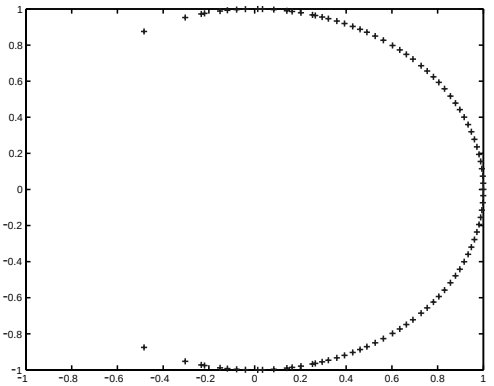


FIG. 6.3. *A*199: $\Lambda(Q)$.

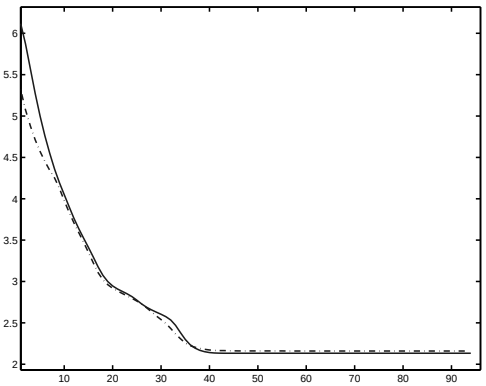


FIG. 6.4. *A*199: ϕ_n *(solid) and* $\hat{\phi}_n$ *(dashdot).*

and $x_0 = 0$ in both cases. FS1836 is ill conditioned, and the performance of GMRES for this matrix is highly sensitive to the initial residual. Furthermore, due to the ill-conditioning, for both right-hand sides a finite precision accuracy of little more than $\mathbf{u} * \kappa(A)$ is achievable by WZ-GMRES. We terminate WZ-GMRES at step n shortly after we notice a considerable difference between the computed and updated residuals. After termination of WZ-GMRES we continue the Arnoldi process to compute ϕ and $\hat{\phi}$. For both right-hand sides these are close to the computed approximations at step n:

	n	ϕ_n	ϕ	$\hat{\phi}_n$	$\hat{\phi}$	$\kappa(R_n)$
FS1836 ($b^{(1)}$)	50	2.2045	2.2045	0.2497	0.2497	1.2e+11
FS1836 ($b^{(2)}$)	23	2.0876	2.0421	1.4332	1.2793	3.4e+09

Our bounds, in particular (4.18), largely overestimate the actual convergence curves. We therefore have not included them in Figure 6.9, which shows the GMRES convergence curves for both right-hand sides. Still, our theory provides useful information because of the different largest gaps in $\Lambda(Q)$ and $\Lambda(\hat{Q})$. GMRES stagnates for $b^{(1)}$, so $\hat{\phi}$ must be small. On the other hand, ϕ for $b^{(1)}$ is large and the behavior of the approximations (cf. Figure 6.11) predicts for small n that GMRES will eventually

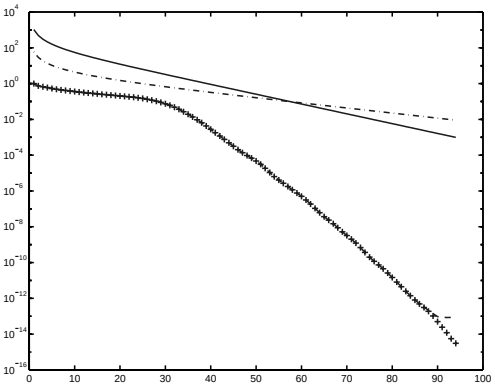


FIG. 6.5. *PDE*225: *Computed (dashed) and updated (pluses) residual norms, bounds* (4.17) *(solid) and* (4.18) *(dashdot).*

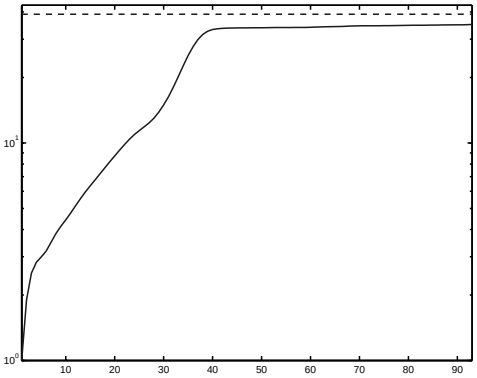


FIG. 6.6. *PDE*225: $\kappa(R_n)$.

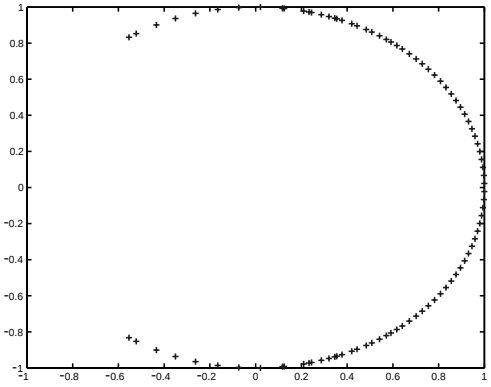


FIG. 6.7. *PDE*225: $\Lambda(Q)$.

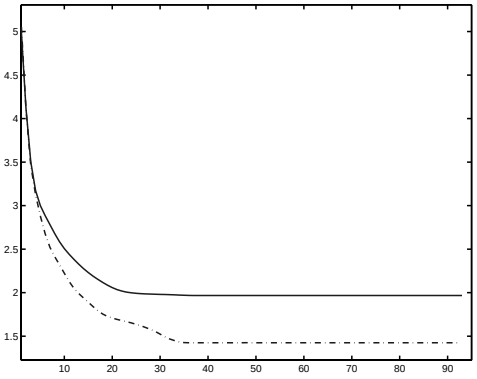


FIG. 6.8. *PDE*225: ϕ_n *(solid) and* $\hat{\phi}_n$ *(dashdot).*

stop to stagnate. If interpreted as "rates of convergence," $\hat{\gamma}^{-1} \approx 0.99$ describes the first phase of the iteration and $\gamma^{-1} \approx 0.85$ describes the second phase. For $b^{(2)}$, there is no stagnation and ϕ and $\hat{\phi}$ are closer to each other (cf. Figure 6.12).

Experiment 6.2. We show the application of our heuristics in the context of preconditioning. We also identify situations in which our heuristics work well and situations in which they yield unsatisfactory results.

We use $A =$ STEAM1 and $A =$ PDE225 and apply WZ-GMRES with $x_0 = 0$ to preconditioned linear systems

$$M^{-1}A\hat{x} = M^{-1}A[1, \ldots, 1]^T.$$

For the preconditioner M we use

- the diagonal of A (Jacobi preconditioner),
- the lower triangular part of A (Gauss–Seidel preconditioner), and
- an incomplete LU factorization of A (ILU(0) preconditioner).

See, e.g., [27, Chapter 10] for more information on these preconditioners. We terminate WZ-GMRES at step d when the updated scaled residual norm is less than 10^{-13}. We identify d with $\dim AK_N(A, r_0)$ and thus consider the last computed approximations as equal to the quantities used in Theorem 4.6.

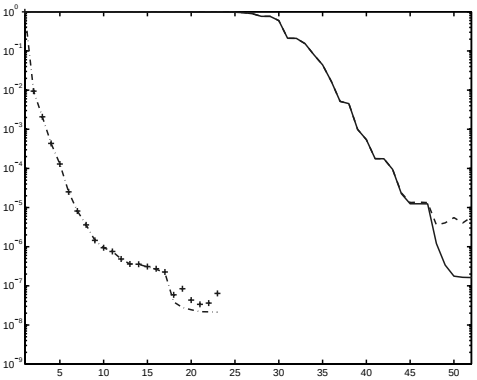


FIG. 6.9. *FS*1836: *Computed (dashed) and updated (solid) residual norms for* $b^{(1)}$*, computed (pluses) and updated (dashdot) residual norms for* $b^{(2)}$.

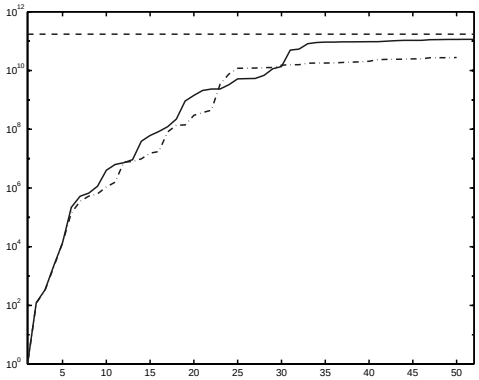


FIG. 6.10. *FS*1836: $\kappa(R_n)$ *for* $b^{(1)}$ *(solid) and* $b^{(2)}$ *(dashdot).*

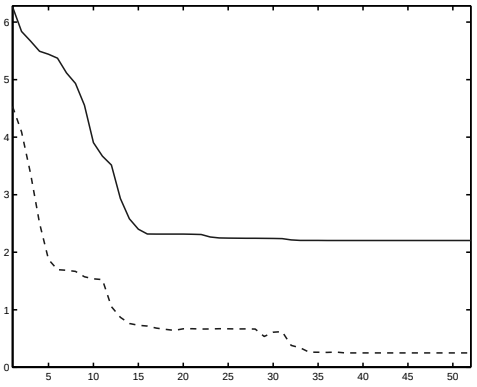


FIG. 6.11. *FS*1836: ϕ_n *(solid) and* $\hat{\phi}_n$ *(dashed) for* $b^{(1)}$.

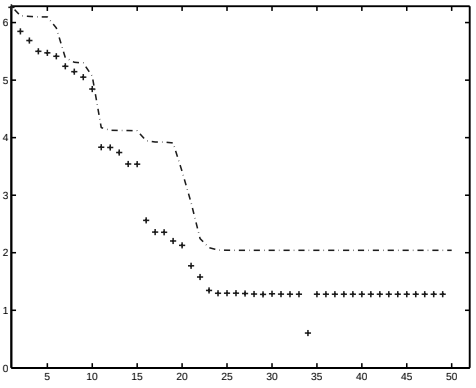


FIG. 6.12. *FS*1836: ϕ_n *(dashdot) and* $\hat{\phi}_n$ *(pluses) for* $b^{(2)}$.

First, consider the results for STEAM1 (see Figures 6.13–6.15):

M	d	ϕ	$\hat{\phi}$	$\kappa(R)$	$\kappa(M^{-1}A)$
Jacobi	22	4.1638	2.5861	2.4e+06	3.4e+06
Gauss–Seidel	15	5.4312	5.4452	2.16	2.25
ILU(0)	7	6.2691	6.2692	1.02	1.02

As a comparison of Figures 6.13 and 6.14 shows, the difference in the three convergence curves is reflected very well by the computed approximations ϕ_n. Heuristic 5.1 correctly predicts the differences in the GMRES convergence behavior for the three preconditioned systems. Of particular interest is the large difference between the ϕ_n for Gauss–Seidel and ILU(0): It implies after only 6 steps that GMRES will converge faster for ILU(0). In contrast, after 10 steps the actual residual norms for these two preconditioners are still approximately equal. Figure 6.15 shows that for Gauss–Seidel and ILU(0) the approximations $\hat{\phi}_n$ behave similar to the ϕ_n. For Jacobi preconditioning we note that ϕ is much larger than $\hat{\phi}$, which is reflected early by the approximations. This difference occurs simultaneously with the appearance of the

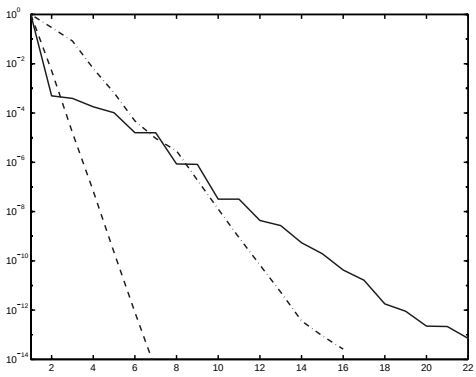


FIG. 6.13. *STEAM*1: *GMRES convergence with Jacobi (solid), Gauss–Seidel (dashdot), and ILU*(0) *(dashed) preconditioning.*

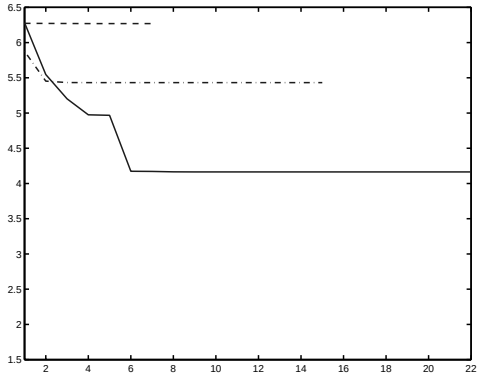


FIG. 6.14. *STEAM*1: ϕ_n *for Jacobi (solid), Gauss–Seidel (dashdot), and ILU*(0) *(dashed) preconditioning.*

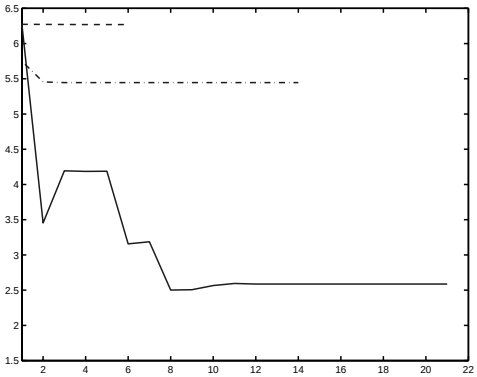


FIG. 6.15. *STEAM*1: $\hat{\phi}_n$ *for Jacobi (solid), Gauss–Seidel (dashdot), and ILU*(0) *(dashed) preconditioning.*

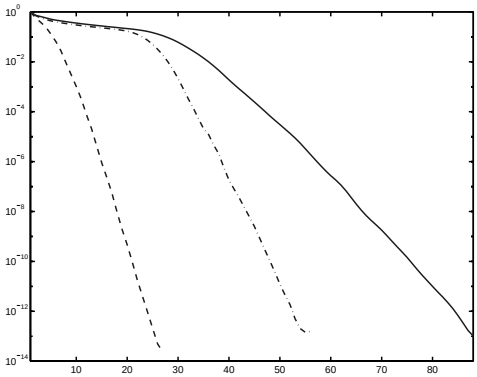


FIG. 6.16. *PDE*225: *GMRES convergence with Jacobi (solid), Gauss–Seidel (dashdot), and ILU*(0) *(dashed) preconditioning.*

large constant $\kappa(R)$ in (4.17) (cf. Experiment 6.1). Figure 6.15 also shows that, while due to interlacing the ϕ_n usually decrease monotonically, the $\hat{\phi}_n$ might behave more erratically.

Experiments with PDE225 (see Figures 6.16–6.18) show limitations of the predictive value of our bounds. As we see in the following table and in Figures 6.17 and 6.18, the difference in the GMRES behavior for Jacobi and Gauss–Seidel preconditioning (cf. Figure 6.16) is neither reflected in nor predicted by our bounds.

M	d	ϕ	$\hat{\phi}$	$\kappa(R)$	$\kappa(M^{-1}A)$
Jacobi	88	1.9289	1.3852	29.8	33.0
Gauss–Seidel	55	2.0605	1.4552	21.7	23.7
ILU(0)	26	4.0565	3.9351	3.7	4.3

In agreement with our earlier observations, the outcome of these experiments can be explained as follows:

In case of ILU(0), our bounds work very well. We have not plotted them, but the reader might check that they both closely approximate the actual convergence curve.

For Jacobi and Gauss–Seidel, GMRES has an initial phase of stagnation. Thus, $\hat{\phi}$

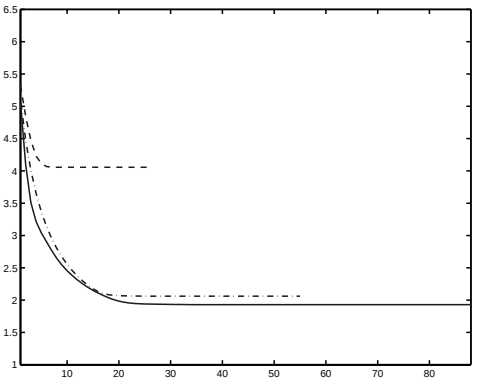


FIG. 6.17. *PDE*225: ϕ_n *for Jacobi (solid), Gauss–Seidel (dashdot), and ILU*(0) *(dashed) preconditioning.*

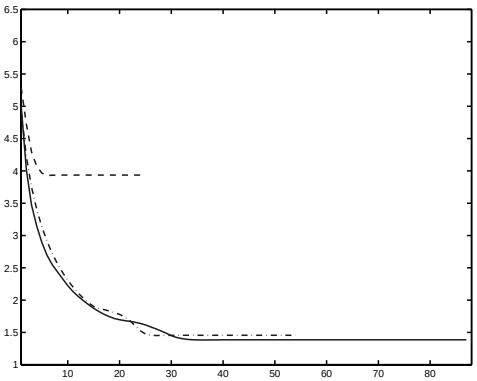


FIG. 6.18. *PDE*225: $\hat{\phi}_n$ *for Jacobi (solid), Gauss–Seidel (dashdot), and ILU*(0) *(dashed) preconditioning.*

must in both cases be relatively small and cannot reflect later phase of faster convergence. Because of the constant $\kappa(R)$, ϕ in both cases might be—and actually is—larger than $\hat{\phi}$. But, because of the initial stagnation, the values of ϕ cannot possibly differ much from each other, because the two respective constants $\kappa(R)$ are only of a similar and moderate size.

Experiment 6.3. We study our bounds in the case of two significantly large gaps in the spectra of the matrices Q and $\hat{Q}$. It was difficult to find a suitable matrix A: With the imposed diagonal scalings of the R-factors, no nonunitary (or nonorthogonal) matrix A we tested led to matrices Q or $\hat{Q}$ having more than one significant gap in their spectra. We therefore had to construct an orthogonal matrix A to start with.

One source of orthogonal matrices is the well-known Ising model (see, e.g., [22, p. 26]). The Ising matrices are products of two matrices K and L of the form

$$K = \begin{bmatrix} E(\alpha) & & \\ & \ddots & \\ & & E(\alpha) \end{bmatrix} \in \mathbf{R}^{2s\times 2s}, \quad E(\alpha) = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix}, \text{ and}$$

$$L = \begin{bmatrix} \cos\beta & & & & -\sin\beta \\ & E(\beta) & & & \\ \vdots & & \ddots & & \vdots \\ & & & E(\beta) & \\ \sin\beta & & & & \cos\beta \end{bmatrix} \in \mathbf{R}^{2s\times 2s}.$$

We choose $A = KL$ with $N = 2s = 100$, $\alpha = \pi/4$, $\beta = \pi/6$ and call this matrix ISING100. We use a random right-hand side generated with the MATLAB call `b=rand(100,1)`, and $x_0 = 0$. We terminate WZ-GMRES at step d when the updated scaled residual norm is less than 10^{-13}. As above, we consider the last computed approximations as equal to the quantities used in Theorem 4.6:

	d	ϕ	$\hat{\phi}$	$\kappa(R)$
ISING100	52	3.6652	3.6652	1.0

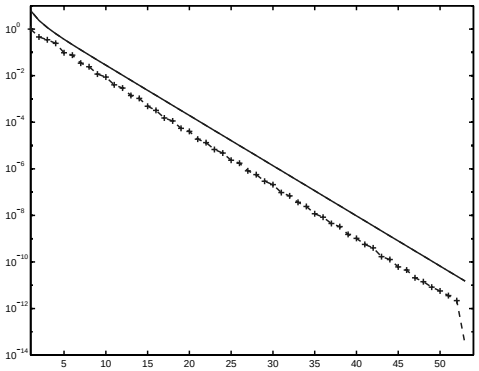


FIG. 6.19. *ISING*100: *Computed (dashed) and updated (pluses) residual norms, bounds* (4.17) *(solid) and* (4.18) *(dashdot).*

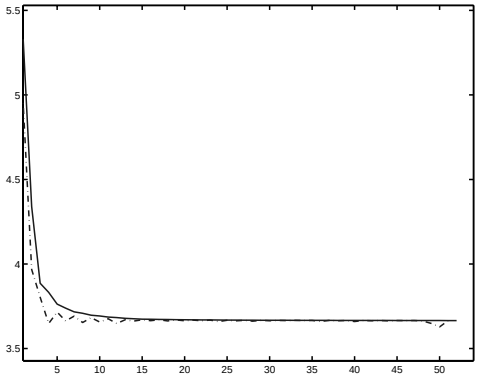


FIG. 6.20. *ISING*100: ϕ_n *(solid) and* $\hat{\phi}_n$ *(dashdot).*

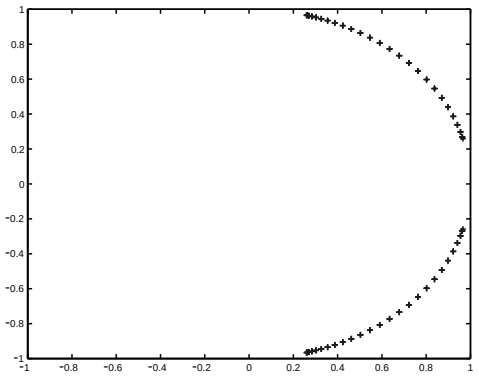


FIG. 6.21. *ISING*100: $\Lambda(A)$.

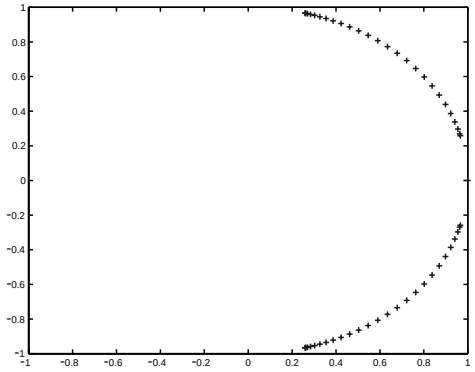


FIG. 6.22. *ISING*100: $\Lambda(Q)$.

Figure 6.19 shows the computed (pluses) and updated (dashed) scaled residual norms produced by WZ-GMRES. These two curves are identical, which is typical for well-conditioned matrices. We also plot our bounds (4.17) and (4.18). The bounds coincide as well, because $\phi = \hat{\phi}$ and $\kappa(R) = 1$. The experiment shows that, despite the occurrence of two significant spectral gaps, our bounds describe the actual convergence curve very well. Figure 6.20 shows the approximations ϕ_n and $\hat{\phi}_n$. Note that, due to interlacing, the curve of the ϕ_n is smoother than the curve of the $\hat{\phi}_n$. Finally, Figures 6.21 and 6.22 show that the spectra of A and Q almost coincide.

7. Conclusions. We have derived two new bounds on the residual norms of GMRES, which are conceptually different from the standard worst-case analysis. Our bounds depend on the initial guess and can be approximated from information generated during the run of a certain GMRES implementation. The approximations allow predictions of how the algorithm will perform in the iteration. We have presented heuristics for such predictions, which are often confirmed in numerical experiments. We have not studied the use of predictions rigorously in the context of large-scale applications. However, we believe that the tools presented here might prove useful in some areas.

Numerical experiments have clearly demonstrated the dependence of our bounds on A *and* r_0. Hence, it will be generally difficult to derive a priori estimates on ϕ and $\hat{\phi}$ based solely on properties of A. Because information about the convergence behavior of GMRES seems to be "distilled" into ϕ and $\hat{\phi}$, a priori estimates for these quantities are of great interest and should be the subject of further investigations.

Interesting related work was recently done by Knizhnerman [18]. He gives bounds for the largest gap in $\Lambda(\hat{Q})$, which depend on the GMRES convergence curve for the given A and r_0. In particular, the bounds show that fast GMRES convergence without steps close to stagnation implies large gaps. Our results, which show that large gaps in $\Lambda(\hat{Q})$ imply fast GMRES convergence, are therefore in some sense dual to Knizhnerman's. It is also interesting to note that, in accordance with our observations, the matrices $\hat{Q}$ in Knizhnerman's examples have only one significant gap in their spectra (cf. [18, Figures 2.1 and 2.2]).

Acknowledgments. Parts of this paper first appeared in [20] and my Ph.D. thesis [21]. Thanks to my advisor Ludwig Elsner for his careful reading of the manuscript and several suggestions that improved the paper. I am happy to acknowledge many fruitful discussions with Zdeněk Strakoš, who made very helpful comments and pointed out several references. Part of the work was performed while I was visiting the Institute of Computer Science of the Academy of Sciences of the Czech Republic. Thanks to Leonid Knizhnerman and Paul Saylor for careful reading of the manuscript and many comments, and to Tino Koch, Reinhard Nabben, and Miro Rozložník for helpful discussions.

REFERENCES

[1] M. Arioli, V. Pták, and Z. Strakoš, *Krylov sequences of maximal length and convergence of GMRES*, BIT, 38 (1998), pp. 636–643.
[2] W. Arnoldi, *The principle of minimized iterations in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.
[3] C. H. Bischof, *Incremental condition estimation*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 312–322.
[4] P. N. Brown and H. F. Walker, *GMRES on (nearly) singular systems*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 37–51.
[5] S. L. Campbell, I. C. F. Ipsen, C. T. Kelley, and C. D. Meyer, *GMRES and the minimal polynomial*, BIT, 36 (1996), pp. 664–675.
[6] J. B. Conway, *A Course in Functional Analysis*, 2nd ed., Springer-Verlag, New York, 1990.
[7] J. Drkošová, A. Greenbaum, M. Rozložník, and Z. Strakoš, *Numerical stability of GMRES*, BIT, 35 (1995), pp. 309–330.
[8] M. Eiermann, *Field of Values and Iterative Methods*, manuscript, 1996.
[9] L. Elsner and C. He, *Perturbation and interlace theorems for the unitary eigenvalue problem*, Linear Algebra Appl., 188 (1993), pp. 207–229.
[10] V. Faber, W. Joubert, E. Knill, and T. Manteuffel, *Minimal residual method stronger than polynomial preconditioning*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 707–729.
[11] G. H. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.
[12] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, 1997.
[13] A. Greenbaum and L. Gurvits, *MAX-MIN properties of matrix factor norms*, SIAM J. Sci. Comput., 15 (1994), pp. 348–358.
[14] A. Greenbaum, V. Pták, and Z. Strakoš, *Any nonincreasing convergence curve is possible for GMRES*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 465–469.
[15] A. Greenbaum and Z. Strakoš, *Matrices that generate the same Krylov residual spaces*, in Recent Advances in Iterative Methods, G. H. Golub, A. Greenbaum, and M. Luskin, eds., IMA Vol. Math. Appl. 60, Springer-Verlag, New York, 1994, pp. 95–118.
[16] A. Greenbaum and L. N. Trefethen, *GMRES/CR and Arnoldi/Lanczos as matrix approximation problems*, SIAM J. Sci. Comput., 15 (1994), pp. 359–368.

[17] W. Joubert, *A robust GMRES-based adaptive polynomial preconditioning algorithm for nonsymmetric linear systems*, SIAM J. Sci. Comput., 15 (1994), pp. 427–439.
[18] L. Knizhnerman, *On GMRES-Equivalent Bounded Operators*, preprint, 1998.
[19] T. Koch and J. Liesen, *The conformal bratwurst maps and associated Faber polynomials*, Numer. Math., to appear.
[20] J. Liesen, *Computable Convergence Bounds for GMRES*, Preprint 98-043, Sonderforschungsbereich 343, Universität Bielefeld, Bielefeld, Germany, 1998.
[21] J. Liesen, *Construction and Analysis of Polynomial Iterative Methods for Non-Hermitian Systems of Linear Equations*, Ph.D. thesis, Fakultät für Mathematik, Universität Bielefeld, Bielefeld, Germany, 1998; also available from http://archiv.ub.uni-bielefeld.de/disshabi/mathe.htm.
[22] M. Marcus and H. Minc, *A Survey of Matrix Theory and Matrix Inequalities*, Dover, New York, 1992.
[23] N. M. Nachtigal, S. Reddy, and L. N. Trefethen, *How fast are nonsymmetric matrix iterations?*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 778–795.
[24] M. Rozložník, *Numerical Stability of the GMRES Method*, Ph.D. thesis, Institute of Computer Science, Academy of Sciences of the Czech Republic, 1997.
[25] M. Rozložník and Z. Strakoš, *Variants of the residual minimizing Krylov space methods*, in Proceedings of the Eleventh Summer School on Software and Algorithms of Numerical Mathematics, I. Marek, ed., 1995, pp. 208–225.
[26] Y. Saad, *Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices*, Linear Algebra Appl., 34 (1980), pp. 269–295.
[27] Y. Saad, *Iterative Methods for Solving Sparse Linear Systems*, PWS Publishing Company, Boston, MA, 1996.
[28] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
[29] V. Smirnov and N. Lebedev, *Functions of a Complex Variable—Constructive Theory*, MIT Press, Cambridge, MA, 1968.
[30] G. Starke, *Field-of-values analysis of preconditioned iterative methods for nonsymmetric elliptic problems*, Numer. Math., 78 (1997), pp. 103–117.
[31] Z. Strakoš, *Convergence and numerical behavior of the Krylov space methods*, in Proceedings of the NATO ASI Institute "Algorithms for Large Sparse Linear Algebraic Systems: The State of the Art and Applications in Science and Engineering" (invited lectures), Kluwer Academic Publishers, 1998, pp. 175–196.
[32] The MathWorks, *Using MATLAB (December* 1996*)*, The MathWorks Inc., Natick, MA, 1996.
[33] K.-C. Toh, *GMRES vs. ideal GMRES*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 30–36.
[34] H. F. Walker, *Implementation of the GMRES method using Householder transformations*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 152–163.
[35] H. F. Walker and L. Zhou, *A simpler GMRES*, Numer. Linear Algebra Appl., 1 (1994), pp. 571–581.
[36] J. Walsh, *Interpolation and Approximation by Rational Functions in the Complex Domain*, 5th ed., AMS, Providence, RI, 1969.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 3, pp. 904–912
© 2000 Society for Industrial and Applied Mathematics

SOME RESULTS ON OPPENHEIM'S INEQUALITIES FOR M-MATRICES*

ZHONGPENG YANG† AND JIANZHOU LIU‡

Abstract. We offer a generalization of Oppenheim's inequalities for the Hadamard product of two M-matrices. We show that our results strengthen known results of Lynn [*Proc. Cambridge Philos. Soc.*, 60 (1994), pp. 425–431], Ando [*Linear and Multilinear Algebra*, 8 (1980), pp. 291–316], and Liu and Zhu [*SIAM J. Matrix Anal. Appl.*, 18 (1997), pp. 305–311].

Key words. Hadamard product, M-matrix, determinant, inequality

AMS subject classifications. 15A15, 15A48

PII. S0895479898332424

1. Introduction.

We denote by M_n the class of all $n \times n$ M-matrices, i.e., matrices all of whose off-diagonal entries are nonpositive (thus belonging to Z_n) and whose inverse exists and is nonnegative [4]. Let $R^{m\times n}$ denote the set of $m \times n$ real matrices. The Hadamard product of two $m \times n$ matrices $A = (a_{ij})$ and $B = (b_{ij})$ is defined by $A \circ B = (a_{ij}b_{ij})$. We write $A \geq B$ if $a_{ij} \geq b_{ij}$ for all i, j.

In 1964, Lynn [3] proved that Oppenheim's inequality on the lower bound of the determinant of Hadamard products of two real positive definite matrices holds for M-matrices, i.e.,

$$\det(A \circ B) \geq \det A \prod_{i=1}^{n} b_{ii}, \qquad A, B \in M_n. \tag{1}$$

In [2], Ando shows that Oppenheim's inequality (1) can be strengthened to

$$\det(A \circ B) + \det A \det B \geq \det A \prod_{i=1}^{n} b_{ii} + \det B \prod_{i=1}^{n} a_{ii}, \qquad A, B \in M_n. \tag{2}$$

In a recent paper Jianzhou Liu and Li Zhu have given the following result (see [1, Theorem 1]):

$$\begin{aligned} \det(A \circ B) &\geq a_{11}b_{11} \prod_{k=2}^{n} \left[b_{kk} \frac{\det A_k}{\det A_{k-1}} + \frac{\det B_k}{\det B_{k-1}} \left(\sum_{i=1}^{k} \frac{a_{ik}a_{ki}}{a_{ii}} \right) \right] \\ &= L_n(A, B), \qquad A, B \in M_n. \end{aligned} \tag{3}$$

In (3) it appears that the summation should run from $i = 1$ to $k - 1$ and not from $i = 1$ to k, where A_k and B_k are the $k \times k$ leading principal submatrix of A and B,

*Received by the editors January 12, 1998; accepted for publication (in revised form) by R. Brualdi December 5, 1998; published electronically February 29, 2000.
http://www.siam.org/journals/simax/21-3/33242.html
†Science Institute of Jilin Teacher's College, Jilin, 132013, People's Republic of China.
‡Department of Mathematices, Xiangtan University, Xiangtan, Hunan, 411105, People's Republic of China (icam@xtu.edu.cn).

respectively. [1] gives the simplified form of (3) (see [1, Corollary 2.2]),

$$\begin{aligned}\det(A\circ B) &\geq \det A\prod_{i=1}^{n} b_{ii}+\det B\prod_{i=1}^{n} a_{ii}\left(\prod_{k=2}^{n}\sum_{i=1}^{k-1}\frac{a_{ik}a_{ki}}{a_{ii}a_{kk}}\right)\\ &=\widetilde{L_n}(A,B),\quad A,B\in M_n,\end{aligned}\tag{4}$$

and points out that inequality (2) is better than (4), i.e.,

$$\begin{aligned}\det(A\circ B) &\geq \det A\prod_{i=1}^{n} b_{ii}+\det B\prod_{i=1}^{n} a_{ii}-\det A\det B\\ &=\varepsilon_n(A,B)\geq\widetilde{L_n}(A,B),\quad A,B\in M_n.\end{aligned}$$

This paper will provide some new lower bound on determinants for the Hadamard products of two M-matrices, which strengthen the results in [1, 2, 3].

For $A=(a_{ij})$, $B=(b_{ij})\in R^{n\times n}$, we shall use the following notations:

$$\omega_n(A,B,k)=\det A\left[\frac{\prod\limits_{\substack{i=1\\ i\neq k}}^{n} a_{ii}}{\det A(k)}-1\right][b_{kk}\det B(k)-\det B]$$

$$+\det B\left[\frac{\prod\limits_{\substack{i=1\\ i\neq k}}^{n} b_{ii}}{\det B(k)}-1\right][a_{kk}\det A(k)-\det A],$$

$\delta_n(A,B,k)=\varepsilon(A,B)+\omega_n(A,B,k)$, if $\det A(k)\neq 0$ and $\det B(k)\neq 0$, where $A(k)\in R^{(n-1)\times(n-1)}$ denotes the submatrix of A obtained by deleting row and column k of A;

$$\beta(A)=\begin{cases}\frac{\det A}{\det A(n)} & \text{if } \det A(n)\neq 0;\\ 0 & \text{otherwise.}\end{cases}$$

Suppose $A\in R^{n\times n}$, the comparison matrix $\mu(A)=(u_{ij})$ is defined by

$$u_{ij}=\begin{cases}-|a_{ij}|, & i\neq j;\\ |a_{ij}|, & i\neq j.\end{cases}$$

2. Some lemmas. In this section we give some lemmas which will be used in the proof of the main results.

LEMMA 1 (see [5, p. 134, Theorem 2.3 (A_1), (E_{17})]). *Let $A\in Z_n$; then each of the following conditions is equivalent to the statement "A is an M-matrix":*

(i) *All of the principal minors of A are positive.*

(ii) *All the leading principal minors of A are positive, i.e.,* $\det A_k>0$, $k=1,2,\ldots,n$.

LEMMA 2 (see [6]). *If $A,B\in M_n$, then $\mu(A\circ B)\in M_n$.*

LEMMA 3. *Let $A=(a_{ij})\in M_n$; then*

$$\prod_{i=1}^{n} a_{ii}\geq a_{kk}\det A(k)\geq\det A>0,\qquad k=1,2,\ldots,n.\tag{5}$$

Proof. Since A is a τ-matrix (see [5, (5.11), p. 160]), from the Hadamard–Fischer inequality (see [7]), we may obtain (5).

Lemma 4 can be found in [8, p. 629].

LEMMA 4. *If D, E, F, and G are diagonal matrices, then*

$$(DAE) \circ (FBG) = DF(A \circ B)EG, \qquad A, B \in R^{n\times n}. \tag{6}$$

LEMMA 5. *If $A, B \in M_n$, then*

$$\det(A \circ B) \geq \det \mu(A \circ B) > 0. \tag{7}$$

Proof. By [5, Theorem 2.3 (I_{29}), p. 134] there exist two positive diagonal matrices $E = \operatorname{diag}(e_1, \ldots, e_n)$ and $G = \operatorname{diag}(g_1, \ldots, g_n)$ such that AE and BG are strictly diagonally dominant matrices, respectively, i.e.,

$$\sum_{j\neq i} |a_{ij}|e_j < a_{ii}e_i, \qquad \sum_{j\neq i} |b_{ij}|g_j < b_{ii}g_i,$$

therefore

$$\sum_{j\neq i} |a_{ij}b_{ij}|e_jg_j < a_{ii}b_{ii}e_ig_i, \qquad i = 1, 2, \ldots, n,$$

from (6), $(AE) \circ (BG) = (A \circ B)EG$ is a strictly diagonally dominant matrix, and $\mu(AE \circ BG) = [\mu(A \circ B)]EG$. Thus, by [1, Lemma 4], we have

$$\det(A \circ B) \det EG = \det(AE \circ BG) \geq \det \mu(A \circ B) \det EG > 0,$$

which proves the inequality (7)

LEMMA 6. *Let $A \in M_n$, $E_{nn} = \operatorname{diag}(0, 0, \ldots, 0, 1) \in R^{n\times n}$. Then $A - tE_{nn} \in M_n$ if and only if $\beta(A) > t (\in R)$.*

Proof. Since $A - tE_{nn} \in Z_n$ and $(A - tE_{nn})_k = A_k$ $(k = 1, 2, \ldots, n-1)$, by Lemma 1 and the equality

$$\det(A - tE_{nn}) = \det A - t \det A_{n-1} = [\beta(A) - t] \det A(n), \tag{8}$$

we know that $A - tE_{nn} \in M_n$ if and only if $\beta(A) > t (\in R)$.

LEMMA 7. *If $A = \left[\begin{smallmatrix} A(n) & A_{12} \\ A_{21} & a_{nn} \end{smallmatrix}\right], B = \left[\begin{smallmatrix} B(n) & B_{12} \\ B_{21} & b_{nn} \end{smallmatrix}\right] \in M_n$, then*

$$\det \mu(A \circ B) \geq [\beta(A)b_{nn} + \beta(B)a_{nn} - \beta(A)\beta(B)] \det \mu(A(n) \circ B(n)). \tag{9}$$

Proof. Let $\varepsilon > 0$, since $A - [\beta(A) - \varepsilon]E_{nn} = \left[\begin{smallmatrix} A(n) & A_{12} \\ A_{21} & a_{nn} - \beta(A) + \varepsilon \end{smallmatrix}\right]$ and $B - [\beta(B) - \varepsilon]E_{nn} = \left[\begin{smallmatrix} B(n) & B_{12} \\ B_{21} & b_{nn} - \beta(B) + \varepsilon \end{smallmatrix}\right]$ are M-matrices by Lemma 6; from the proof of [1, Theorem 2] and Lemma 2 we have

$$\begin{aligned} &\mu((A - (\beta(A) - \varepsilon)E_{nn}) \circ (B - (\beta(B) - \varepsilon)E_{nn})) \\ &\qquad = \begin{bmatrix} \mu(A(n) \circ B(n)) & -A_{12} \circ B_{12} \\ -A_{21} \circ B_{21} & (a_{nn} - \beta(A) + \varepsilon)(b_{nn} - \beta(B) + \varepsilon) \end{bmatrix} \\ &\qquad = \mu(A \circ B - v(\varepsilon)E_{nn}) \in M_n, \end{aligned}$$

where $v(\varepsilon) = [\beta(A)b_{nn} + \beta(B)a_{nn} - \beta(A)\beta(B)] + \varepsilon[\beta(A) + \beta(B) - a_{nn} - b_{nn} - \varepsilon] \in R$. It follows that

$$\begin{aligned}\beta(A)b_{nn} + \beta(B)a_{nn} - \beta(A)\beta(B) &= \beta(A)b_{nn} + \beta(B)\left[\frac{a_{nn}\det A(n) - \det A}{\det A(n)}\right] \\ &\geq \beta(A)b_{nn} > 0\end{aligned}$$

by Lemma 3. This means that if $\varepsilon > 0$ is sufficiently small, then $v(\varepsilon) > 0$ and $a_{nn}b_{nn} - v(\varepsilon) > 0$, i.e., $a_{nn}b_{nn} - v(\varepsilon) = |a_{nn}b_{nn} - v(\varepsilon)|$, hence $\mu(A \circ B - v(\varepsilon)E_{nn}) = \mu(A \circ B) - v(\varepsilon)E_{nn} \in M_n$. Thus, from Lemmas 6 and 2, $\beta(\mu(A \circ B)) > v(\varepsilon)$, letting $\varepsilon \to 0$, we have

$$\beta(\mu(A \circ B)) \geq \beta(A)b_{nn} + \beta(B)a_{nn} - \beta(A)\beta(B) > 0, \qquad A, B \in M_n. \tag{10}$$

It follows that

$$\begin{aligned}0 &= \det(\mu(A \circ B) - \beta(\mu(A \circ B))E_{nn}) \\ &\leq \det \mu(A \circ B) - (\beta(A)b_{nn} + \beta(B)a_{nn} - \beta(A)\beta(B))\det((A \circ B)(n)) \\ &= \det \mu(A \circ B) - (\beta(A)b_{nn} + \beta(B)a_{nn} - \beta(A)\beta(B))\det \mu(A(n) \circ B(n))\end{aligned}$$

by (8) and (10).

LEMMA 8. *Let $A, B \in M_n$; then*

$$a_{kk}\frac{\det B_k}{\det B_{k-1}} - \frac{\det A_k}{\det A_{k-1}}\frac{\det B_k}{\det B_{k-1}} \geq \frac{\det B_k}{\det B_{k-1}}\sum_{i=1}^{k-1}\frac{a_{ik}a_{ki}}{a_{ii}} \geq 0, \quad k = 2, \ldots, n.$$

Proof. Since $A_k = \begin{bmatrix} A_{k-1} & A_{12}^{(k-1)} \\ A_{21}^{(k-1)} & a_{kk} \end{bmatrix}$ and $B_k = \begin{bmatrix} B_{k-1} & B_{12}^{(k-1)} \\ B_{21}^{(k-1)} & b_{kk} \end{bmatrix}$ are $k \times k$ M-matrices, $\operatorname{diag}(a_{11}, \ldots, a_{k-1,k-1}) \geq A_{k-1} (\in M_{k-1})$ and $A_{k-1}^{-1} \geq \operatorname{diag}(a_{11}^{-1}, \ldots, a_{k-1,k-1}^{-1}) > 0$ by [1, Lemma 3] and [9, Definition 1.1]; therefore,

$$A_{21}^{(k-1)}A_{k-1}^{-1}A_{12}^{(k-1)} \geq A_{21}^{(k-1)}\operatorname{diag}(a_{11}^{-1}, \ldots, a_{k-1,k-1}^{-1})A_{12}^{(k-1)} = \sum_{i=1}^{k-1}\frac{a_{ik}a_{ki}}{a_{ii}} \geq 0. \tag{11}$$

Note that, from [10],

$$\det A_k = \left(a_{kk} - A_{21}^{(k-1)}A_{k-1}^{-1}A_{12}^{(k-1)}\right)\det A_{k-1}.$$

Thus, by (5) and (11), we have

$$\begin{aligned}0 &\leq \frac{\det B_k}{\det B_{k-1}}\sum_{i=1}^{k-1}\frac{a_{ik}a_{ki}}{a_{ii}} \\ &\leq \frac{\det B_k}{\det B_{k-1}}A_{21}^{(k-1)}A_{k-1}^{-1}A_{12}^{(k-1)} \\ &= \frac{\det B_k}{\det B_{k-1}}\left(\frac{a_{kk}\det A_{k-1} - \det A_k}{\det A_{k-1}}\right) \\ &= a_{kk}\frac{\det B_k}{\det B_{k-1}} - \frac{\det A_k}{\det A_{k-1}}\frac{\det B_k}{\det B_{k-1}}.\end{aligned}$$

LEMMA 9. *Let $A, B \in M_n$. If P is a permutation matrix, then $PAP^{-1}, PBP^{-1} \in M_n$, and*

$$\begin{aligned} \det(PAP^{-1}) \circ (PBP^{-1}) &= \det(A \circ B), \\ \det \mu((PAP^{-1}) \circ (PBP^{-1})) &= \det \mu(A \circ B). \end{aligned} \tag{12}$$

Proof. Lemma 1 tells us $PAP^{-1}, PBP^{-1} \in M_n$. From the proof of [8, Observations 4 and 5], $(PAP^{-1}) \circ (PBP^{-1}) = P(A \circ B)P^{-1} = P(A \circ B)P^{-1}, \mu((PAP^{-1}) \circ (PBP^{-1})) = P(\mu(A \circ B))P^{-1}$. Thus, we obtain (12).

3. Main results. In this section, we study some improvement of Oppenheim's inequalities (1) and (2) for M-matrices.

Let vector $e_i (\in R^n)$ have a 1 as its ith component, 0's elsewhere and permutation $P_{\alpha_t} = [e_t, e_1, \ldots, e_{t-1}, e_{t+1}, \ldots, e_n]$. If $A, B \in M_n$, then $A_{\alpha_t} = P_{\alpha_t} A P_{\alpha_t}^{-1}$ and $B_{\alpha_t} = P_{\alpha_t} B P_{\alpha_t}^{-1}$ are M-matrices by Lemma 1.

THEOREM 1. *Let $A = (a_{ij}), B = (b_{ij}) \in M_n$ if*

$$Y_n(A, B) = a_{11} b_{11} \prod_{k=2}^{n} [\beta(A_k) b_{kk} + a_{kk} \beta(B_k) - \beta(A_k)\beta(B_k)]$$

and

$$Y_n(A, B, t) = Y_n(A_{\alpha_t}, B_{\alpha_t});$$

then

$$\det \mu(A \circ B) \geq \mu(A \circ B) \geq \max_{1 \leq t \leq n} Y_n(A, B, t). \tag{13}$$

Proof. First, by induction on n we prove that

$$\det \mu(A \circ B) \geq Y_n(A, B), \qquad A, B \in M_n. \tag{14}$$

One can easily verify that (14) holds with equality for $n = 2$. We assume (14) holds for all $m \times m$ ($m < n$ and $n > 2$) M-matrices.

Let any $n \times n$ M-matrices A and B be partitioned

$$A = \begin{bmatrix} A(n) & A_{12} \\ A_{21} & a_{nn} \end{bmatrix} = A_n \quad \text{and} \quad B = \begin{bmatrix} B(n) & A_{12} \\ B_{21} & b_{nn} \end{bmatrix} = B_n;$$

then, from [5] or [9] and Lemma 1, we know that $A(n) = A_{n-1}$ and $B(n) = B_{n-1}$ are $(n-1) \times (n-1)$ M-matrices; by Lemma 3,

$$\begin{aligned} &b_{kk}\beta(A_k) + a_{kk}\beta(B_k) - \beta(A_k)\beta(B_k) \\ &\quad = b_{kk} \frac{\det A_k}{\det A_{k-1}} + a_{kk} \frac{\det B_k}{\det B_{k-1}} - \frac{\det A_k}{\det A_{k-1}} \frac{\det B_k}{\det B_{k-1}} > 0, \quad k = 1, 2, \ldots, n. \end{aligned} \tag{15}$$

Combining (9), (15), and the inductive hypothesis, we obtain

$$\begin{aligned} \det \mu(A \circ B) &\geq [b_{nn}\beta(A) + a_{nn}\beta(B) - \beta(A)\beta(B)] \det \mu\big(A(n) \circ B(n)\big) \\ &\geq [b_{nn}\beta(A_n) + a_{nn}\beta(B) - \beta(A_n)\beta(B_n)] Y_{n-1}(A_{n-1}, B_{n-1}) \\ &= Y_n(A, B), \end{aligned}$$

which completes the proof of (14). □

Thus, the inequality (13) follows from (12), (7), and (14).

By (14) and Lemma 8, we have the following corollary.

COROLLARY 1. *If $A, B \in M_n$, then*

$$\det(A \circ B) \geq \det \mu(A \circ B) \geq Y_n(A, B) \geq L_n(A, B) \geq \det A \prod_{i=1}^{n} b_{ii}.$$

THEOREM 2. *If $A, B \in M_n$, then*

(16)
$$\det(A \circ B) \geq \det \mu(A \circ B) \geq \max_{1 \leq k \leq n} \delta_n(A, B, k) = \varepsilon_n(A, B) + \max_{1 \leq k \leq n} \omega_n(A, B, k).$$

Proof. Since $\omega_n(A, B, k) \geq 0$ by (5), we know that

$$\delta_n(A, B, k) = \varepsilon_n(A, B) + \omega_n(A, B, k) \geq \varepsilon_n(A, B), \quad A, B \in M_n. \tag{17}$$

First using induction on n, we show that

$$Y_n(A, B) \geq \delta_n(A, B, n), \quad A, B \in M_n. \tag{18}$$

Note that $Y_2(A, B) = \delta_2(A, B, 2) = \varepsilon_2(A, B)$ for $A, B \in M_2$; hence, (18) is immediate for $n = 2$. If $A, B \in M_3$, then

$$\begin{aligned}
Y_3(A, B) &= Y_2(A_2, B_2)[b_{33}\beta(A_3) + a_{33}\beta(B_3) - \beta(A_3)\beta(B_3)] \\
&= \left(\det A_2 \prod_{i=1}^{2} b_{ii} + \det B_2 \prod_{i=1}^{2} a_{ii} - \det A_2 \det B_2 \right) \\
&\quad \times \left(b_{33} \frac{\det A_3}{\det A_2} + a_{33} \frac{\det B_3}{\det B_2} - \frac{\det A_3}{\det A_2} \frac{\det B_3}{\det B_2} \right) \\
&= \det A_3 \prod_{i=1}^{3} b_{ii} - \prod_{i=1}^{2} b_{ii} \frac{\det B_3}{\det B_2} (a_{33} \det A_2 - \det A_3) \\
&\quad + \det B_3 \prod_{i=1}^{3} a_{ii} + \prod_{i=1}^{2} a_{ii} \frac{\det A_3}{\det A_2} (b_{33} \det B_2 - \det B_3) \\
&\quad - b_{33} \det B_2 \det A_3 - a_{33} \det A_2 \det B_3 + \det A_3 \det B_3,
\end{aligned}$$

i.e.,

$$Y_3(A, B) = \varepsilon_3(A, B) + \omega_3(A, B, 3), \quad A, B \in M_n. \tag{19}$$

For $\left[\begin{smallmatrix} A_3 & a_{14} \\ a_{41} & a_{44} \end{smallmatrix} \right], B = \left[\begin{smallmatrix} B_3 & b_{14} \\ b_{41} & b_{44} \end{smallmatrix} \right] \in M_4$, from $A_3, B_3 \in M_3$ and (17), (19),

$$\begin{aligned}
Y_4(A, B) &= Y_3(A_3, B_3)[b_{44}\beta(A_4) + a_{44}\beta(B_4) - \beta(A_4)\beta(B_4)] \\
&\geq \varepsilon_3(A_3, B_3)[b_{44}\beta(A_4) + a_{44}\beta(B_4) - \beta(A_4)\beta(B_4)] \\
&= \varepsilon_4(A_4, B_4) + \omega_4(A_4, B_4, 4) \\
&= \delta_4(A_4, B_4, 4).
\end{aligned}$$

If (18) holds for $n-1$, then it follows from the induction hypothesis that

$$\begin{aligned} Y_{n-1}(A_{n-1}, B_{n-1}) &= \varepsilon_{n-1}(A_{n-1}, B_{n-1}) + \omega_{n-1}(A_{n-1}, B_{n-1}, n-1) \\ &= \delta_{n-1}(A_{n-1}, B_{n-1}, n-1), \end{aligned}$$

for $A = \left[\begin{smallmatrix} A_{n-1} & A_{12} \\ A_{21} & a_{nn} \end{smallmatrix}\right], B = \left[\begin{smallmatrix} B_{n-1} & B_{12} \\ B_{21} & b_{nn} \end{smallmatrix}\right] \in M_n$; hence from (17)

$$\begin{aligned} Y_n(A,B) &= Y_{n-1}(A_{n-1}, B_{n-1})[b_{nn}\beta(A_n) + a_{nn}\beta(B_n) - \beta(A_n)\beta(B_n)] \\ &\geq \varepsilon_{n-1}(A_{n-1}, B_{n-1})[b_{nn}\beta(A) + a_{nn}\beta(B) - \beta(A)\beta(B)] \\ &= \left(\det A_{n-1} \prod_{i=1}^{n-1} b_{ii} + \det B_{n-1} \prod_{i=1}^{n-1} a_{ii} - \det A_{n-1} \det B_{n-1} \right) \\ &\quad \left(b_{nn} \frac{\det A}{\det A_{n-1}} + a_{nn} \frac{\det B}{\det B_{n-1}} - \frac{\det A}{\det A_{n-1}} \frac{\det B}{\det B_{n-1}} \right) \\ &= \varepsilon_n(A,B) + \omega_n(A,B,n) \\ &= \delta_n(A,B,n). \end{aligned}$$

This implies that (18) holds. So, from (14) and (18),

(20)
$$\det(A \circ B) \geq \det \mu(A \circ B) \geq \varepsilon_n(A,B) + \omega_n(A,B,n) = \delta_n(A,B,n), \quad A, B \in M_n.$$

Suppose P is a permutation matrix such that $(PAP^{-1})_{n-1} = (PAP^{-1})(n) = A(k)$ and $(PBP^{-1})_{n-1} = (PBP^{-1})(n) = B(k)$. Since $\varepsilon_n(PAP^{-1}, PBP^{-1}) = \varepsilon_n(A,B)$ and $\omega_n(PAP^{-1}, PBP^{-1}, n) = \omega_n(A,B,k)$, from (20), (12), and Lemma 9, we obtain (16).

By (5), (16), and (17), the next corollary follows.

COROLLARY 2. *If $A, B \in M_n$, then*

$$\begin{aligned} \det(A \circ B) \geq \det \mu(A \circ B) &\geq \max_{1 \leq k \leq n} \delta_n(A,B,k) \\ &\geq \varepsilon_n(A,B) \geq \widetilde{L_n}(A,B) \geq (\det A) \left(\prod_{i=1}^{n} b_{ii} \right). \end{aligned}$$

Corollary 2 shows that inequality (2) can be extended further to yield the following corollary.

COROLLARY 3. *If $A, B \in M_n$, then*

$$\begin{aligned} \det(A \circ B) + \det A \det B &\geq \det \mu(A \circ B) + \det A \det B \\ &\geq (\det A) \left(\prod_{i=1}^{n} b_{ii} \right) + (\det B) \left(\prod_{i=1}^{n} a_{ii} \right) + \max_{1 \leq k \leq n} \omega_n(A,B,k) \\ &\geq (\det A) \prod_{i=1}^{n} b_{ii} + (\det B) \prod_{i=1}^{n} a_{ii}. \end{aligned}$$

If there are diagonal matrices D and E such that $DAE \in M_n$, we say A in an M-matrix under diagonal equivalence.

EXAMPLE. Let

$$A = \begin{bmatrix} -1 & 0 & -2 & 2 \\ 0 & -2 & 0 & 0 \\ \frac{1}{2} & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix};$$

then A is not an M-matrix, but from [11, Example 2], we know that

$$DAE = \begin{bmatrix} 2 & 0 & -1 & -1 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \in M_4,$$

where $D = \operatorname{diag}(1, -1, 1, -1)$, $E = \operatorname{diag}(-2, \frac{1}{2}, \frac{1}{2}, -\frac{1}{2})$.

Theorem 3 below shows that Theorem 1 and 2 hold for A and B, which are M-matrices under diagonal equivalence.

THEOREM 3. *Let $A, B \in R^{n\times n}$, if there are diagonal matrices $D = \operatorname{diag}(d_1, \ldots, d_n)$, $E = \operatorname{diag}(e_1, \ldots, e_n)$, $F = \operatorname{diag}(f_1, \ldots, f_n)$, and $G = \operatorname{diag}(g_1, \ldots, g_n)$ such that $DAE, FBG \in M_n$ and $\det(DEFG) > 0$; then* (13) *and* (16) *hold.*

Proof. From (16)

$$\begin{aligned} [\det(A \circ B)][\det(DEFG)] &= \det[(DAE) \circ (FBG)], \\ [\det \mu(A \circ B)][\det(DEFG)] &= \det \mu[(DAE) \circ (FBG)]. \end{aligned} \tag{21}$$

One can easily verify that

$$\begin{aligned} \beta(D_k A_k E_k)(f_k b_{kk} g_k) + (d_k a_{kk} e_k)\beta(F_k B_k G_k) - \beta(D_k A_k E_k)\beta(F_k B_k G_k) \\ = (d_k e_k f_k g_k)[b_{kk}\beta(A_k) + a_{kk}\beta(B_k) - \beta(A_k)\beta(B_k)], \end{aligned}$$

hence, $Y_n(DAE, FBG)Y_n(A, B)[\det(DEFG)]$ and $Y_n(DAE, FBG, t) = Y_n[(DAE)_{\alpha_t}, (FBG)_{\alpha_t}] = Y_n(A_{\alpha_t}, B_{\alpha_t})[\det(DEFG)] = Y_n(A, B, t)\big[\det(DEFG)\big]$. So, from (21) and Theorem 1, we obtain (13).

Since $\varepsilon_n(DAE, FBG) = [\det(DAE)](\prod_{i=1}^n f_i b_{ii} g_i) + [\det(FBG)](\prod_{i=1}^n d_i a_{ii} e_i) - [\det(DAE)][\det(FBG)] = \varepsilon_n(A, B)[\det(DEFG)]$ and

$$\begin{aligned} \omega_n(DAE, FBG, k) = [\det(DAE)] &\left[\frac{\prod_{\substack{i=1 \\ i\neq k}}^n d_i a_{ii} e_i}{\det(D(k)A(k)E(k))} - 1 \right] \\ &\times [f_k b_{kk} g_k \det(F(k)B(k)G(k)) - \det(FBG)] \\ + [\det(FBG)] &\left[\frac{\prod_{\substack{i=1 \\ i\neq k}}^n f_i b_{ii} g_i}{\det(F(k)B(k)G(k))} - 1 \right] \\ &\times [d_k a_{kk} e_k \det(D(k)A(k)E(k)) - \det(DAE)] \\ = \omega_n(A, B, k)&[\det(DEFG)], \end{aligned}$$

i.e., $\delta_n(DAE, FBG, k) = \delta_n(A, B, k)[\det(DEFG)]$; by (21) and Theorem 2 we obtain (16).

REFERENCES

[1] J. Liu and L. Zhu, *Some improvement of Oppenheim's inequality for M-matrices*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 305–311.
[2] T. Ando, *Inequalities for M-matrices*, Linear and Multilinear Algebra, 8 (1980), pp. 291–316.
[3] M. S. Lynn, *On the Schur product of the H-matrices and non-negative matrices, and related inequalities*, Proc. Cambridge Philos. Soc., 60 (1964), pp. 425–431.
[4] M. Fiedler and T. L. Markham, *A characterization of the closure of inverse M-matrices*, Linear Algebra Appl., 105 (1988), pp. 209–223.
[5] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.
[6] K. Fan, *Inequalities for M-matrices*, Indag Math., 26 (1964), pp. 602–610.
[7] G. Engel and H. Schneider, *The Hadamard-Fischer inequality for a class of matrices defined by eignvalue monotonicity*, Linear and Multilinear Algebra, 4 (1977), pp. 155–176.
[8] C. R. Johnson and H. M. Shapiro, *Mathematical aspects of the relative gain array* $A \circ A^{-T}$, SIAM J. Algebraic Discrete Methods, 7 (1986), pp. 627–644.
[9] J. M. Pena, *M-matrices whose inverses are totally positive*, Linear Algebra Appl., 221 (1995), pp. 189–193.
[10] D. Carlson, *What are Schur complements, anyway?* Linear Algebra Appl., 74 (1986), pp. 257–275.
[11] I. N. Iman, *The Schur complement and the inverse M-matrix problem*, Linear Algebra Appl., 62 (1984), pp. 235–240.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 3, pp. 913–930
© 2000 Society for Industrial and Applied Mathematics

CONDITIONING AND PADÉ APPROXIMATION OF THE LOGARITHM OF A MATRIX*

LUCA DIECI† AND ALESSANDRA PAPINI‡

Abstract. In this work we (i) use the theory of piecewise analytic functions to represent the Fréchet derivative of any primary matrix function, in particular of primary logarithms; (ii) propose an indicator to assess inherent difficulties to compute a logarithm; and (iii) revisit Padé approximation techniques for the principal logarithm of block triangular matrices.

Key words. logarithm, conditioning, Fréchet derivative, Padé approximation

AMS subject classifications. 65F30, 65F35, 65F99, 15A24

PII. S089547989834531X

1. Introduction.

Given a matrix $A \in \mathbb{R}^{n\times n}$ or $\mathbb{C}^{n\times n}$, we call logarithm of A any matrix L such that $e^L = A$. It is well known that A has a logarithm L if and only if A is invertible, which we will henceforth assume. Among the (infinitely many) logarithms of A, some are *primary* matrix functions and some are not (see [6]). Primary functions are those which are true functions on $\Lambda(A)$; in this work, $\Lambda(A) = \{\lambda_i(A),\ i = 1,\dots,n\}$ indicates the spectrum of a matrix A. Thus, if L is a logarithm of A, $\mu_i \in \Lambda(L)$, and $\lambda_i \in \Lambda(A)$, then L is a primary function if there exists a (piecewise analytic) function log for which $\mu_i = \log(\lambda_i)$, $i = 1,\dots,n$; we will nonetheless always write $L = \log(A)$, even though "log" may fail to be a function on the spectrum of A.

In applications, it is often important to characterize under which conditions on real A we also have a real logarithm L. The following is well known (e.g., see [6]):

Given $A \in \mathbb{R}^{n\times n}$, a real logarithm L exists if and only if A has an even number of Jordan blocks of each size relative to every negative eigenvalue. If A has any negative eigenvalue, such L cannot be a primary matrix function.

Among the real primary logarithms of A, attention has been almost invariably restricted to the so-called *principal* logarithm, whose eigenvalues have imaginary parts in $(-\pi,\ \pi)$. Such principal logarithms enjoy a very useful integral representation:

$$\log(A) = \int_0^1 (A-I)((A-I)t+I)^{-1}dt\,. \tag{1.1}$$

Recently, there has been some interest in computation of logarithms of matrices; see [2], [8], [10], and the references therein. However, there are important questions still unanswered. In this work, we address the following three issues:

1. All studies so far have only characterized sensitivity of the principal logarithm. In section 2, we characterize the sensitivity of any primary logarithm. We base our approach on the Fréchet derivative of piecewise analytic functions.

*Received by the editors September 25, 1998; accepted for publication (in revised form) by N. J. Higham June 11, 1999; published electronically February 29, 2000. This work was supported in part by NSF grant DMS-9625813 and NATO grant CRG 950865.
http://www.siam.org/journals/simax/21-3/34531.html
†School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332 (dieci@math.gatech.edu).
‡Dep. Energetica S. Stecco, University of Florence, via C. Lombroso 6-17, 50134 Florence, Italy (papini@de.unifi.it).

2. Arguably, a logarithm L has computational meaning only if one can reliably verify to what extent L satisfies $e^L = A$. Thus, in section 3 we propose a criterion to assess the inherent difficulty in computing the log of a matrix by trying to quantify how reliably we can verify whether $e^L = A$. This is similar in spirit to what happens with the square root of a matrix, $A^{1/2}$; see [1] and [4]. Indeed, we elucidate the relation between our indicator and the "square roots" indicator of [1] and [4] in the context of the *inverse scaling and squaring* procedure.

3. Recently, there has been some criticism of the inverse scaling and squaring technique coupled with Padé approximation (the so-called *Briggs method*) for computing principal logarithms of block triangular matrices; see [10]. In section 4, we derive improved error bounds which show precisely the interplay between inverse scaling and squaring and accuracy of Padé approximants. As a result, we show that the appropriate implementation of the Briggs method is still a very viable technique for computing the logarithm of a block triangular matrix.

2. Fréchet derivative and conditioning. Classical sensitivity considerations are statements about the variational problem. In the context of functions of a matrix, this means studying the Fréchet derivative of the function. Therefore, this type of sensitivity analysis is restricted to primary matrix functions. For the logarithm, this study has thus far been restricted to principal logarithms; see [8] and [2]. However, as was made clear in [3], there are many important applications where this restriction just cannot be made and it is desirable to extend these considerations to any primary logarithm. To this end, it is convenient to use the Cauchy integral form to represent a logarithm in the form put forward in [18] for piecewise analytic functions. Below, for the sake of generality, we will not at first restrict f to be the "log" function.

Let $G_k,\ k = 1, \ldots, q$, be disjoint domains[1] of the complex plane and let G be the union of these G_k; for example, the G_k may be disjoint disks. Let f be a single-valued analytic function in the interior of the G_k's. Consider all matrices whose spectrums lie in G and let A be one of those spectrums which is fixed. Then, one can define the primary matrix function $f(A)$ as (see [18, sections I.2.2–2.6])

$$f(A) \;=\; \frac{1}{2\pi i}\sum_{k=1}^{q}\int_{\Gamma_k} f(z)(zI - A)^{-1}dz\,, \tag{2.1}$$

where $\Gamma_k,\ k = 1, \ldots, q$, are rectifiable curves enclosing each and all eigenvalues $\lambda_j \in G_k$ only once, lying entirely in G_k, and where integration along each Γ_k is performed in the positive direction.

We can now use (2.1) to obtain a formula for the Fréchet derivative: (2.2) below. This formula may be known, but we did not find it elsewhere. Attempts of a similar nature for functions analytic on a domain enclosing all eigenvalues were made in [16], [15], and [6, equation 6.6.15]. See also [13] for a different approach.

As usual, the Fréchet derivative of f at A in the direction of the matrix E is the linear map $f'(A)$ for which

$$\lim_{h\to 0} \left\| \tfrac{1}{h}(f(A + hE) - f(A)) - f'(A)E \right\| = 0$$

(h is real). In this paper, $\|\cdot\|$ is the 2-norm; at times, we also use the Frobenius norm $\|\cdot\|_f$.

[1] Recall that a domain is open and simply connected.

LEMMA 2.1. *With previous notation, the Fréchet derivative of f at A is given by the map*

$$f'(A) : E \to \frac{1}{2\pi i} \sum_{k=1}^{q} \int_{\Gamma_k} f(z)(zI - A)^{-1} E (zI - A)^{-1} dz \,. \tag{2.2}$$

Proof. Choose h sufficiently small, so that $\Lambda(A + hE) \subseteq \cup_k G_k$. Then, form the difference $f(A + hE) - f(A)$ and expand

$$(zI - (A + hE))^{-1} - (zI - A)^{-1} = (zI - A)^{-1}[(I + hE(zI - A)^{-1} + O(h^2)) - I] \,. \qquad \square$$

Remark 2.2. If p is the number of distinct eigenvalues of A, in formulas (2.1) and (2.2) it is convenient to take $q = p$. That is, each Γ_k encloses precisely one of the distinct eigenvalues of A.

The norm of the Fréchet derivative is $\|f'(A)\| = \max_{\|Z\|=1} \|f'(A)Z\|$. This is a useful indicator of numerical conditioning of the evaluation of $f(A)$ (see [2], [4], [8], [11], [12]). In fact, let $X = f(A)$ and suppose $X \neq 0$, and let $X + \Delta X = f(A + E)$, where $\|E\|$ is small. Then, at first order we have

$$\frac{\|\Delta X\|}{\|X\|} \leq \|f'(A)\| \, \frac{\|A\|}{\|X\|} \frac{\|E\|}{\|A\|} \,. \tag{2.3}$$

Thus, $\|f'(A)\|$ acts as an absolute error magnification factor, whereas the quantity $\|f'(A)\| \, \frac{\|A\|}{\|X\|}$ acts as a relative error magnification factor, and it is called the condition number of f at A. Of course, this is a measure of conditioning based on worse-type behavior, and more refined estimates are possible if A has a particular structure and one restricts to matrices E for which $A + E$ has this same structure.

Unfortunately, obtaining a sharp estimate of $\|f'(A)\|$ from (2.2) is generally nontrivial. A typical line of action follows. According to Remark 2.2, consider the case $q = p$ in (2.2). Let $V^{-1}AV = J = \left[\begin{smallmatrix} J_1 & & \\ & \ddots & \\ & & J_p \end{smallmatrix}\right]$ be a grouping of the Jordan form of A, so that J_k comprises all Jordan blocks relative to the eigenvalue λ_k. Thus, we have

$$f'(A)E \;=\; V \frac{1}{2\pi i} \sum_{k=1}^{p} \int_{\Gamma_k} f(z)(zI - J)^{-1}(V^{-1}EV)(zI - J)^{-1} dz V^{-1} \,. \tag{2.4}$$

Next, since $(zI - J)^{-1}$ are easy to write down explicitly, repeated use of the Cauchy formula for analytic functions eventually gives a closed expression for the Fréchet derivative. However, it is not a particularly useful one for computational purposes, since one typically gets stuck with the condition number of V. A great simplification occurs if A is normal, because then V is unitary and J is diagonal. This is the content of the following lemma, whose proof is a simple application of the Cauchy formula on (2.4). A similar result was shown to be true for f analytic (see [8] and [12, Equation (1.4)]).

LEMMA 2.3. *If A is a normal matrix, for the Fréchet derivative* (2.2) *of the primary matrix function $f(A)$ in* (2.1) *we have*

$$\|f'(A)\|_f \;=\; \max\left(\max_{1 \leq j \leq p} |f'(\lambda_j)| \;,\; \max_{1 \leq j,k \leq p,\; j \neq k} \frac{|f(\lambda_j) - f(\lambda_k)|}{|\lambda_j - \lambda_k|} \right) . \tag{2.5}$$

Next, we restrict ourselves to the case of $f(A)$ being $\log(A)$ and specialize the study of the Fréchet derivative to this case. Of course, we must understand that $\log(z)$ makes sense in the complex plane slit along a ray going from zero to infinity, and naturally such a ray must not contain any eigenvalue λ_i of A. So, (2.1) is specialized to read as follows:

$$\log(A) \;=\; \frac{1}{2\pi i}\sum_{k=1}^{p}\int_{\Gamma_k}\log(z)(zI-A)^{-1}dz\,, \tag{2.6}$$

where the Γ_k do not intersect one another nor the ray along which we had slit the complex plane. It has to be stressed that $\log(z)$ is a single-valued analytic function within each Γ_k, but it is not required to be so in a domain enclosing $\cup_k\Gamma_k$. To be precise, suppose the plane is slit along the ray $\arg z = \alpha$ and we fix $\alpha < \arg z < \alpha+2\pi$ for all $z \in \cup_k\Gamma_k$. Then

$$\log(z) = \log|z| + i(\arg z + 2\pi m),\;\; z \in \Gamma_k\;, \tag{2.7}$$

for some $m = 0, \pm1, \pm2, \ldots$, but the values of m are not required to be the same for different Γ_k. With this in mind, by construction, (2.6)–(2.7) characterize all primary logarithms. If A is real with no negative eigenvalues, then it has (in general, infinitely many) real primary $\log(A)$. Equation (2.6) is too general to characterize these. First, we slit along the negative real axis, so that $-\pi < \arg z < \pi$. Then, relative to a complex conjugate pair of eigenvalues of A, in (2.7) we must take $\log(\rho e^{\pm i\phi}) = \rho \pm i(\phi + 2\pi m)$. These choices characterize all real primary logarithms.

One can use Lemma 2.3 in the case of $\log(A)$ to classify primary logarithms of normal matrices to determine which ones minimize $\|f'(A)\|_f$. For example, one easily gets:

(a) "If A and $\log(A)$ are complex, and we have chosen the same value of m in (2.7) in each Γ_k, then all logarithms give identical value of $\|f'(A)\|_f$."

(b) "If A is 2×2 (real or complex), then the logarithm giving the smallest value of $\|f'(A)\|_f$ is the one which minimizes the distance between the imaginary parts of its eigenvalues." Therefore, if A is real, normal, with real primary $\log(A)$, it is not necessarily true that the principal logarithm gives a smaller value of $\|f'(A)\|_f$ than a complex logarithm.

3. Inherent difficulties and scaling procedures. Sensitivity assessment based on the Fréchet derivative is of course restricted to primary matrix functions. In essence, this type of sensitivity study tells us if we may (or may not) expect that the computed logarithm $\tilde{L} = L + F$ is close to the exact logarithm $L = \log(A)$. See (2.3) and also [2], [3], [8], [12]. For the sake of the present section, let us henceforth assume that F is of small norm. However, L in general has only an indirect meaning, as solution of the matrix equation $e^L = A$; therefore, if evaluation of the exponential of L is sensitive, calculation of L is intrinsically hard, since we cannot reliably verify to what extent we have solved $e^L = A$. For this reason, in this section we propose a simple indicator to assess the inherent difficulty of computing the logarithm, which attempts to measure how reliably we can verify whether we have solved $e^L = A$. Our indicator is not restricted to primary logarithms and is similar in spirit to the square root indicator of [1] and [4]; we take this similarity further by elucidating the interplay between our indicator and the square root indicator in the context of the popular inverse scaling and squaring procedure.

So, we let $e^{\tilde{L}} = \tilde{A} = A + E$ and ask when we can expect E to be of small norm. Or, in a somewhat backward manner, we can ask the question, Is there a reasonable

hope that we have verifiably found the logarithm of a matrix close to the original matrix A? To answer this question, we are thus led to consider the equation

$$\frac{\|E\|}{\|A\|} \;=\; \frac{\left\|e^{L+F} - e^{L}\right\|}{\|A\|}, \tag{3.1}$$

and the issue has shifted to finding a way to obtain sharp bounds for (3.1). Several different bounds for (3.1) have been obtained by Kågström in [7] and Van Loan in [11]. We are interested in bounds which take the form

$$\frac{\|E\|}{\|A\|} \;\le\; g(\nu\,\|F\|)\;\beta(L)\,, \quad \text{where } \nu \text{ is a positive constant,} \tag{3.2}$$

and we require that

(i) g is monotone increasing and $g(x) = x + O(x^2)$ for $0 < x \ll 1$;
(ii) $\beta:\ L \in \mathbb{R}^{n\times n} \to \mathbb{R}^{+}$, $\beta(0) = 1$, and $\beta(L) \ge 1$.

Our idea is to regard β as an indicator of intrinsic difficulties in computing $\log(A)$. In practice, the success of this approach will depend on how tight the bound is in (3.2). Generally, g can be expressed as

$$g(\nu\,\|F\|) \;=\; e^{\nu\|F\|} - 1 \tag{3.3}$$

(see [7, Equation (4.1)]), and we have experimented with the choices (3.4), (3.5), and (3.8) below for β and ν in (3.2)–(3.3). From [6, Corollary 6.2.32], we get

$$\nu = 1\ ,\ \text{ and }\ \beta(L) = \frac{e^{\|L\|}}{\|A\|}\,. \tag{3.4}$$

From [7, Equation (4.9)], the following is obtained (a small refinement of [11, Equation (3.2)]):

$$\nu = 1\ ,\ \ \beta(L) = \frac{e^{\mu(L)}}{\|A\|}\,, \tag{3.5}$$

where $\mu(L)$ is the logarithmic norm of L. Recall that, for a matrix B, the logarithmic norm is

$$\mu(B) \;=\; \lim_{h\to 0^+} \frac{\|I + hB\| - 1}{h} \;=\; \lim_{h\to 0^+} \frac{\log(\left\|e^{hB}\right\|)}{h} \tag{3.6}$$

(see [17]) and, in the 2–norm, $\mu(B)$ is the largest eigenvalue of $(B + B^*)/2$. For later purposes, it is useful to notice that the second equality of (3.6) gives

$$\mu(B) \;=\; \lim_{k\to\infty} 2^k\ \log(\left\|e^{B/2^k}\right\|)\,. \tag{3.7}$$

Finally, from [7, Equation (4.15)] (a slight refinement of [11, Equation (3.4)]) we also get

$$\nu = \sum_{k=0}^{n-1} \frac{\|N\|^k}{k!}\ ,\ \ \beta(L) = \frac{e^{a(L)}\nu}{\|A\|}\,, \tag{3.8}$$

where $a(L)$ is the spectral abscissa of L (i.e., the largest real part of the eigenvalues of L) and N is the off-diagonal part of a Schur form of L. That is, if $Q:\ Q^*LQ = D + N$

is a Schur form of L, then $D = \mathrm{diag}(\lambda_i)$ and N is the strictly upper triangular part. To be precise, in (3.8) ν is $\nu(L)$, and hence g also depends on L in this case; for simplicity, we omit this dependence. Notice that β in (3.8) is trivial to obtain if computation of a primary logarithm is done after Schur reduction of A.

LEMMA 3.1. *For all choices of β in* (3.4), (3.5), *and* (3.8), *we always have $\beta \geq 1$ and $\beta(0) = 1$. For β given in* (3.5) *and* (3.8), *we also have $\beta(L + cI) = \beta(L)$ for any real number c, and if L is normal, $\beta(L) = 1$.*

Proof. That $\beta(L) \geq 1$ in (3.4) is obvious. For (3.5), it follows from $e^{\mu(L)} \geq \left\|e^L\right\|$, and for (3.8) is similar (see [11]). When L is normal, a simple computation shows that $e^{\mu(L)} = e^{a(L)} = \left\|e^L\right\|$, and thus $\beta(L) = 1$ in (3.5) and (3.8). The statement about $L + cI$ is also an immediate verification. ∎

Remark 3.2. If the original matrix A is highly structured and the algorithm used to compute L exploits this structure, it may be possible to provide refinement of the bound (3.2). An obvious case is if A is block diagonal: $A = \mathrm{diag}\, A_{ii},\ i = 1, \ldots, p$, and one computes $L_{ii} = \log(A_{ii})$; then, in this case, it is more appropriate to look at the factors β relative to each diagonal block. A less trivial instance of this occurrence will be seen in section 4.

Remark 3.3. Our numerical experiments showed that, in general, the bound (3.4) is not satisfactory and (3.5) is often only a modest improvement over (3.4), whereas (3.8) gives the most satisfactory results. This is in agreement with the results of [11]. Nonetheless, the β values in (3.4) and (3.5) are so inexpensive to compute that we have included them in this study.

Next, we restrict to principal logarithms of real matrices. In such cases, a popular procedure for computing the logarithm exploits the relation

$$\log A = 2^k \log(A^{1/2^k}), \tag{3.9}$$

where the progressive square roots have been taken so that the arguments of their eigenvalues are in the strip $(-\pi/2^k, \pi/2^k)$. Since $A^{1/2^k}$ approaches the identity, for k sufficiently large it should be easy to compute $\log(A^{1/2^k})$ to full precision and then obtain $\log A$. This approach, known as inverse scaling and squaring, was first put forward in [8]. The crux of the technique is the computation of square roots of matrices. In [1] and [4], the authors propose an indicator to quantify the inherent difficulty of computation of a square root of A; this indicator is defined as

$$\alpha(A^{1/2}) \ = \ \frac{\left\|A^{1/2}\right\|^2}{\|A\|}\,, \tag{3.10}$$

and clearly $\alpha(A^{1/2}) \geq 1$. Also, $\alpha(A^{1/2}) = 1$ if A is normal and $A^{1/2}$ is a primary square root (see [4]). Next, we want to understand whether or not the inverse scaling and squaring procedure may have introduced additional difficulties (as detected by the factor α) with respect to the intrinsic difficulty of computation of the logarithm (as detected by the factor β).

LEMMA 3.4. *With respect to the choices of $\beta(L)$ in* (3.4) *and* (3.5), *we have*

$$\beta(L) = (\beta(L/2))^2 \alpha(A^{1/2})\,, \tag{3.11}$$

whereas with respect to (3.8) *we have*

$$\beta(L) = (\beta(L/2))^2 \alpha(A^{1/2})\ \frac{\nu(L)}{\nu^2(L/2)}\,. \tag{3.12}$$

Proof. Proving (3.11) for β in (3.4) is simple:

$$\left(\frac{e^{\|L/2\|}}{\|A^{1/2}\|}\right)^2 = \frac{e^{\|L\|}}{\|A\|}\,\frac{\|A\|}{\|A^{1/2}\|^2}\,.$$

In a similar way, for β in (3.5), (3.11) is true, since $\mu(L/2) = \mu(L)/2$. To show (3.12), we have

$$\big(\beta(L/2)\big)^2 \;=\; \frac{e^{a(L)}\nu(L)}{\|A\|}\,\frac{\|A\|}{\|A^{1/2}\|^2}\,\frac{[\nu(L/2)]^2}{\nu(L)}\,. \qquad \square$$

COROLLARY 3.5. *Let k be a positive integer. For $\beta(L)$ in* (3.4) *and* (3.5)*, we have*

$$\beta(L) = (\beta(L/2^k))^{2^k}\,\prod_{j=1}^{k}\big(\alpha(A^{1/2^j})\big)^{2^{j-1}}\,. \tag{3.13}$$

For β in (3.8)*, we have instead*

$$\beta(L) = (\beta(L/2^k))^{2^k}\,\prod_{j=1}^{k}\big(\alpha(A^{1/2^j})\big)^{2^{j-1}}\,\frac{\nu(L)}{(\nu(L/2^k))^{2^k}}\,. \tag{3.14}$$

COROLLARY 3.6. *For β given in* (3.8)*, in* (3.12) *we have*

$$\frac{[\nu(L/2)]^2}{\nu(L)} \ge 1\,, \qquad (\beta(L/2))^2\alpha(A^{1/2}) \ge \beta(L)\,. \tag{3.15}$$

Proof. Some algebra gives

$$\nu^2(L/2) \;=\; \nu(L) + S, \quad S := \sum_{k=n}^{2n-2}\frac{\|N\|^k}{2^k}\frac{1}{k!}\left[2^k - 2\sum_{i=0}^{k-n}\binom{k}{i}\right]\,.$$

Therefore, we have

$$\big(\beta(L/2)\big)^2 \;=\; \frac{\beta(L)}{\alpha(A^{1/2})}\left(1 + \frac{S}{\nu(L)}\right)\,. \qquad \square$$

We are now ready to answer whether computation of $\log(A)$ through the inverse scaling and squaring procedure has introduced additional difficulties with respect to the intrinsic difficulty of computation of $\log(A)$. The general situation is already clear after one square root, that is, when we use $\log(A) = 2\log(A^{1/2})$. We have to compare $\beta(L)$ with the product $\beta(L/2)\ \alpha(A^{1/2})$. By putting together the results of Lemma 3.4 and Corollary 3.6, it is easy to obtain the following:

With respect to β in (3.4) *and* (3.5)*, taking square roots does not lead to a harder computational task; that is,*

$$\beta(L/2)\alpha(A^{1/2}) \le \beta(L)\,. \tag{3.16}$$

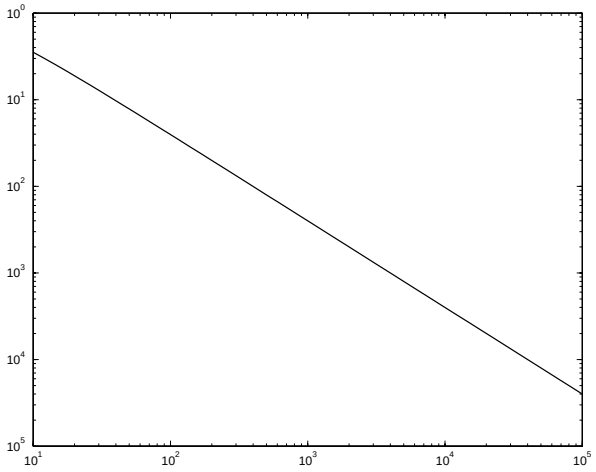


FIG. 1. *A "log-log" plot of the ratio* $\frac{\beta(L)}{\beta(L/2)\alpha(A^{1/2})}$ *for* $b \in [10, 10^5]$.

With respect to β *in* (3.8)*, we instead have*

$$\beta(L) = \beta(L/2)\alpha(A^{1/2}) \left[\frac{e^{a(L/2)}}{\left\|A^{1/2}\right\|} \frac{\nu(L)}{\nu(L/2)} \right]. \tag{3.17}$$

The correct interpretation of (3.16) is more a statement about the inadequacy of the factors β in (3.4)–(3.5) to reveal computational difficulties than it is a statement about real simplifications in computing $\log(A)$ which occur when using inverse scaling and squaring. Instead, (3.17) lends itself to a more insightful and honest interpretation. In general, taking square roots may ease the computational task or increase it, depending on whether the quantity in bracket in (3.17) is greater than or less than 1. A complete characterization of matrices for which the quantity in brackets in (3.17) is greater than 1 seems out of reach, but we observe that for normal matrices, it is 1; in fact, since both β and α equal 1, for normal matrices, taking square roots does not change the inherent difficulties of computing $\log(A)$. Otherwise, the following examples show that computational difficulties, as measured by the ratio $\frac{\beta(L)}{\beta(L/2)\alpha(A^{1/2})}$ of (3.17), can be either increased or decreased by inverse scaling and squaring.

Example 3.7. Consider $A = \left[\begin{smallmatrix} e^c & be^c \\ 0 & e^c \end{smallmatrix}\right]$, with logarithm $L = \left[\begin{smallmatrix} c & b \\ 0 & c \end{smallmatrix}\right]$. For this example, taking square roots may make matters worse. In Figure 1 we show a "log-log" plot of the ratio $\frac{\beta(L)}{\beta(L/2)\alpha(A^{1/2})}$ for $b \in [10, 10^5]$.

Example 3.8. Consider now the matrix $A = \left[\begin{smallmatrix} 0.68 & 0.21 & 0.45 & 0.39 & 0.61 \\ 0.21 & 0.61 & 0.04 & 0.68 & 0.02 \\ 0.84 & 0.63 & 0.03 & 0.09 & 0.02 \\ 0.63 & 0.37 & 0.32 & 0.04 & 0.19 \\ 0.13 & 0.58 & 0.01 & 0.61 & 0.59 \end{smallmatrix}\right]$. In this case, we have $\frac{\beta(L)}{\beta(L/2)\alpha(A^{1/2})} \approx 6.24$, and hence the computational task is eased by taking square roots.

Clearly, regardless of which factor β we are using, since $\beta(0) = 1$, and $\beta(L/2^k) \geq 1$, then $\lim_{k\to\infty} \beta(L/2^k) = 1$. This tells us, if needed, that it is easy to compute the principal logarithm of a matrix close to the identity. However, it is of practical interest to study the rate at which $\beta(L/2^k) \to 1$. We need the following result on the logarithmic norm.

LEMMA 3.9. *Let $A \in \mathbb{R}^{n\times n}$ be invertible with no eigenvalues on the negative real axis and let L be its principal logarithm. For each $k \in \mathbb{Z}^+$, let $A^{1/2^k} := e^{L/2^k}$ be the kth root of A whose eigenvalues have arguments in $(-\pi/2^k,\ \pi/2^k)$, and let*

$$\alpha(A^{1/2^k}) = \frac{\left\|A^{1/2^k}\right\|^2}{\left\|A^{1/2^{k-1}}\right\|}.$$

Then

$$e^{\mu(L)} \;=\; \lim_{k\to\infty} \left\|A^{1/2^k}\right\|^{2^k} \tag{3.18}$$

and

$$e^{\mu(L)} \;=\; \|A\| \lim_{k\to\infty} \prod_{j=1}^{k} \left(\alpha(A^{1/2^j})\right)^{2^{j-1}}. \tag{3.19}$$

Proof. Using (3.7) to characterize $\mu(L)$, (3.18) is obvious:

$$e^{\mu(L)} \;=\; e^{\lim_{k\to\infty} \log(\|e^{L/2^k}\|^{2^k})} \;=\; \lim_{k\to\infty} \left\|e^{L/2^k}\right\|^{2^k}.$$

Now, (3.19) follows from (3.18) and the identity

$$\frac{\left\|A^{1/2^k}\right\|^{2^k}}{\|A\|} = \alpha(A^{1/2})(\alpha(A^{1/4}))^2 \cdots (\alpha(A^{1/2^k}))^{2^{k-1}}. \qquad \square$$

LEMMA 3.10. *Under the same assumptions as those of Lemma* 3.9*, we have:*

(i) *for β in* (3.4),

$$\lim_{k\to\infty} \left(\beta(L/2^k)\right)^{2^k} \;=\; e^{\|L\|-\mu(L)}; \tag{3.20}$$

(ii) *for β in* (3.5),

$$\lim_{k\to\infty} \left(\beta(L/2^k)\right)^{2^k} \;=\; 1; \tag{3.21}$$

(iii) *for β in* (3.8),

$$\lim_{k\to\infty} \left(\beta(L/2^k)\right)^{2^k} \;=\; e^{\|N\|+a(L)-\mu(L)}, \tag{3.22}$$

where N is the strictly upper triangular part in a Schur form of L. In particular, asymptotically,

$$\beta(L/2^k) \;\approx e^{(a(L)-\mu(L))/2^k} \left(1 + \frac{\|N\|}{2^k}\right). \tag{3.23}$$

Proof. Let us begin verifying (3.21). Upon using (3.19) and (3.13), we have

$$\beta(L) = \lim_{k\to\infty} \left(\beta(L/2^k)\right)^{2^k} \beta(L),$$

which is (3.21). Equation (3.20) follows similarly from (3.19) and (3.13). To show (3.22), we pass to the limit as $k \to \infty$ in (3.14) and use (3.19) to obtain

$$\lim_{k\to\infty} \big(\beta(L/2^k)\big)^{2^k} = e^{a(L)}\ e^{-\mu(L)} \lim_{k\to\infty} \big(\nu(L/2^k)\big)^{2^k}\,.$$

Recalling that $\nu(L/2^k) = \sum_{j=0}^{n-1} \frac{\|N/2^k\|^j}{j!}$, we have

$$1 + \frac{\|N\|}{2^k} \le \nu(L/2^k) \le e^{\|N\|/2^k}\,,$$

and therefore $\lim_{k\to\infty} \big(\nu(L/2^k)\big)^{2^k} = e^{\|N\|}$, and (3.22) and (3.23) follow. ∎

4. Block triangular matrices and the Schur–Fréchet method. In this section, we revisit the familiar Briggs method (inverse scaling and squaring with Padé approximants) for computing the principal logarithm L of a (real) block triangular matrix R: $L = \log(R)$. (Hereafter, the notation "log" will always refer to the principal logarithm.) We can think of the triangular matrix R as the result of Schur reduction of a matrix A. Thus, we have

$$R \;=\; \begin{bmatrix} R_{11} & R_{12} & \dots & R_{1p} \\ 0 & R_{22} & \dots & R_{2p} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & R_{pp} \end{bmatrix}, \quad L = \log(R) = \begin{bmatrix} L_{11} & L_{12} & \dots & L_{1p} \\ 0 & L_{22} & \dots & L_{2p} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & L_{pp} \end{bmatrix}. \tag{4.1}$$

Clearly, in (4.1) $L_{ii} = \log(R_{ii})$. The issue becomes finding a way to determine the off-diagonal blocks L_{ij}. We restrict ourselves to the case of $p = 2$ in (4.1):

$$R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}\,,\ \ L = \log(R) = \begin{bmatrix} L_{11} & L_{12} \\ 0 & L_{22} \end{bmatrix}\,. \tag{4.2}$$

Before discussing ways to obtain L_{12}, we must determine whether computation of L_{12} presents some inherent difficulties. We proceed along the lines of section 3.

Let R and L be as in (4.2). In agreement with Remark 3.2, we assume that the chosen algorithm of solution has respected this block structure, and thus we may look at relative errors in a block sense. That is, with $F = \begin{bmatrix} F_{11} & F_{12} \\ 0 & F_{22} \end{bmatrix}$, where $L+F$ represents the computed logarithm, we look at

$$\frac{\big\|e^{L_{11}+F_{11}} - R_{11}\big\|}{\|R_{11}\|}\,,\ \frac{\big\|e^{L_{22}+F_{22}} - R_{22}\big\|}{\|R_{22}\|}\,,\ \frac{\big\|(e^{L+F})_{12} - R_{12}\big\|}{\|R_{12}\|}\,. \tag{4.3}$$

Then, there are three distinct inherent difficulties in computing L: one relative to L_{11}, one relative to L_{22}, and one relative to L_{12}. The first two we treated in section 3; the third one we will deal with here. We will need the familiar formula (e.g., see [11])

$$e^L \;=\; \begin{bmatrix} e^{L_{11}} & X \\ 0 & e^{L_{22}} \end{bmatrix}\,,\ \ X = \int_0^1 e^{L_{11}(1-s)} L_{12} e^{L_{22}s} ds\,. \tag{4.4}$$

We let $(e^{L+F})_{12} = Y$, and with some algebra get

$$\begin{aligned} \frac{\|Y - X\|}{\|X\|} \;\le\; & \frac{\|F_{12}\|}{\|X\|} \int_0^1 h(s)ds \;+ \frac{\|L_{12}\| + \|F_{12}\|}{\|X\|} \\ & \times \int_0^1 h(s)\big[g(\nu_1(1-s)\,\|F_{11}\|)g(\nu_2 s\,\|F_{22}\|)\beta((1-s)L_{11})\beta(sL_{22}) \\ & + g(\nu_1(1-s)\,\|F_{11}\|)\beta((1-s)L_{11}) + g(\nu_2 s\,\|F_{22}\|)\beta(sL_{22})\big]ds\,, \end{aligned} \tag{4.5}$$

where $h(s) := \left\|e^{(1-s)L_{11}}\right\| \left\|e^{sL_{22}}\right\|$ and g and β are the functions in (3.2) and (3.3). Recall that we assume that $g(x) = x + O(x^2)$ for $0 < x \ll 1$ and that the values of β and ν in which we are interested are given in (3.8). Now, we assume that $\|\nu_i F_{ii}\| \le \eta$, $i = 1, 2$, with $\eta \ll 1$, and therefore from (4.5) we get

$$\frac{\|Y - X\|}{\|X\|} \le \frac{\|F_{12}\|}{\|X\|} \int_0^1 h(s)(1 + \eta b(s))ds + \eta \frac{\|L_{12}\|}{\|X\|} \int_0^1 h(s)b(s)ds + O(\eta^2) \,, \tag{4.6}$$

where $b(s) := \beta((1-s)L_{11}) + \beta(sL_{22})$. Since obviously $h(0) = \|e^{L_{11}}\|$ and $h(1) = \|e^{L_{22}}\|$, we arrive at the interesting conclusion that obtaining L_{12} may be an intrinsically hard computational task whenever $h(s)$, $0 < s < 1$, is much larger than the maximum of $\|e^{L_{11}}\|$ and $\|e^{L_{22}}\|$. This cannot happen if L_{ii} are normal (that is, if R_{ii} are normal). This claim can be verified as follows. Suppose that R_{ii}, $i = 1, 2$, are normal, and recall that for normal matrices $\beta(\cdot) = 1$. Then, with notation and results from Lemma 3.1, we have

$$\left\|e^{(1-s)L_{11}}\right\| = e^{(1-s)a(L_{11})} \,, \qquad \left\|e^{sL_{22}}\right\| = e^{sa(L_{22})} \,,$$

where $a(L_{ii})$ indicates the spectral abscissa of the L_{ii}, $i = 1, 2$. Therefore, in this case, $h(s) = e^{(1-s)a(L_{11}) + sa(L_{22})}$, which is a monotone function, and the claim follows.

Example 4.1. Consider again Example 3.7. Obviously, $R_{11} = R_{22}$ are normal, and there is no intrinsic difficulty in computing $L_{12} = b$ which is not already reflected in (and by) the computation of L_{11} and L_{22}.

Next, consider possible ways to approximate L_{12} in (4.2).

(a) *Parlett's method.* This well-known general procedure rests on the identity $RL = LR$ (see [14]). From this, if $\Lambda(R_{11}) \cap \Lambda(R_{22}) = \emptyset$, one can uniquely find L_{12} by solving

$$R_{11}L_{12} \;-\; L_{12}R_{22} \;=\; L_{11}R_{12} \;-\; R_{12}L_{22} \,. \tag{4.7}$$

Remark 4.2. If the blocks R_{11} and R_{22} are sufficiently separated so that the Sylvester equation (4.7) is well conditioned (see [5]), then it is hard for us to think of a better method to determine L_{12} than this one. The difficulty of this approach is to determine a priori if the Sylvester equation is well conditioned. In [2], we used a simple criterion based on the spectra of the diagonal blocks and found it to be empirically reliable. However, if use of the identity (4.7) is not computationally feasible (say, the equation is singular), then some other way to obtain L_{12} must be found.

(b) *Padé approximation.* Here, one uses Padé rational functions to approximate $\log(R) = \log((R - I) + I)$. Therefore, L_{12} is obtained at once along with L_{11} and L_{22}. Good choices are diagonal Padé approximants.

Since Padé approximations are accurate only when R is "close to the identity," a standard practice is to exploit the relation (3.9) (with R there) and use Padé approximants for $\log(R^{1/2^k})$. The resulting approach is known as the Briggs method; it was first put forward in [8] and has enjoyed good success.

From a practical point of view, because of Remark 4.2, we think that the Briggs method is of greater appeal when solution of the system (4.7) is not advisable. Since our scope in this section is to revisit Briggs' technique, we will assume that "*the matrix* $R \in \mathbb{R}^{n\times n}$ *in* (4.2) *does not have a blocking with* R_{11} *and* R_{22} *well separated.*" In particular, if R is in real Schur form, and

we label its eigenvalues $\lambda_i,\ i=1,\dots,k,\ k\le n$, writing only once a complex conjugate pair but repeating multiple eigenvalues, then for each fixed λ_i there exists a $\lambda_j,\ j\neq i$, such that $|\lambda_i-\lambda_j|\le\delta$.

(c) *Fréchet technique.* This is also a general procedure, which for the logarithm consists in exploiting the Fréchet identity

$$L \;=\; \begin{bmatrix} L_{11} & 0 \\ 0 & L_{22} \end{bmatrix} + \log'\left(\begin{bmatrix} R_{11} & 0 \\ 0 & R_{22} \end{bmatrix}\right)\begin{bmatrix} 0 & R_{12} \\ 0 & 0 \end{bmatrix}, \tag{4.8}$$

where $\log'(B)Z$ is the Fréchet derivative of the logarithm at B in the direction Z. Therefore, one needs to get

$$L_{12} = \left(\log'\left(\begin{bmatrix} R_{11} & 0 \\ 0 & R_{22} \end{bmatrix}\right)\begin{bmatrix} 0 & R_{12} \\ 0 & 0 \end{bmatrix}\right)_{12}. \tag{4.9}$$

The *Fréchet method* proposed in [10] is based on approximating the Fréchet derivative of the logarithm in (4.9) by means of the hyperbolic tangent function.

Naturally, no matter what approach we use to obtain L_{12}, we must be at the same time solving (4.7) and satisfying (4.9).

THEOREM 4.3. *Let R and L be as in* (4.2). *Then we have*

$$\begin{aligned} L_{12} \;&=\; \int_0^1 ((R_{11}-I)t+I)^{-1}R_{12}((R_{22}-I)t+I)^{-1}\,dt \\ &=:\; \int_0^1 L_1(t)R_{12}L_2(t)dt \;=:\; \int_0^1 F(t,R_{12})dt\,, \end{aligned} \tag{4.10}$$

where we have set $L_i(t) := ((R_{ii}-I)t+I)^{-1},\ i=1,2$. L_{12} in (4.10) *is the only solution of* (4.7) *which eventually gives a principal logarithm L. Also, the expressions* (4.10) *and* (4.9) *are identical.*

Proof. To show (4.10), just use (1.1) and the block structure. To check that (4.10) solves (4.7) regardless of whether the spectra of R_{11} and R_{22} are disjoint, substitute L_{12} from (4.10) and use (1.1) for both L_{11} and L_{22}. That (4.10) and (4.9) give the same expression for L_{12} is verified using the analytic expression of the Fréchet derivative given in [2, Equation (3.13)]. ∎

Theorem 4.3 suggests that to approximate the Fréchet derivative of the logarithm in (4.9), one might approximate the integral in (4.10) with a quadrature rule. For argument's sake, suppose we do so, say, with any of the standard polynomial rules. Then, it is simple to observe that in order to obtain good bounds for the quadrature error, we generally need R_{11} and R_{22} close to the identity. However, R_{12} does not necessarily have to be close to zero if we are interested in relative accuracy for L_{12}. This observation is at the basis of our revisiting the Briggs method.

We must appreciate that in previous works on computation of the logarithm, in particular in [2], [9], [8], Padé approximations were used to approximate L in (4.2) by attempting to control the *absolute error*. That is, if $\hat{L}$ was the computed approximation, one tried to ensure that

$$\left\|L-\hat{L}\right\| \;\le\; \eta\,, \tag{4.11}$$

where η is a small number, say, the machine precision EPS. Since the original error estimates in [9], it has been clear that Padé approximations cannot obtain a small

TABLE 1
The Briggs–Padé method on Example 3.7: I.

c	k	ω	$\omega_1 = \omega_2$	`abs`	`rel`
0.1	0	1.105E6	0.105	$\begin{bmatrix} 8E{-}17 & 1E{-}10 \\ 0 & 8E{-}17 \end{bmatrix}$	$\begin{bmatrix} 8E{-}16 & 1E{-}16 \\ 0 & 8E{-}16 \end{bmatrix}$
0.1	22	0.22	1+2.4E-8	$\begin{bmatrix} 3E{-}10 & 6E{-}10 \\ 0 & 3E{-}10 \end{bmatrix}$	$\begin{bmatrix} 3E{-}9 & 6E{-}16 \\ 0 & 3E{-}9 \end{bmatrix}$
0.3	0	1.35E6	0.3498	$\begin{bmatrix} 5.5E{-}17 & 8E{-}10 \\ 0 & 5.5E{-}17 \end{bmatrix}$	$\begin{bmatrix} 2E{-}16 & 8E{-}16 \\ 0 & 2E{-}16 \end{bmatrix}$
0.9	2	3E5	0.25	$\begin{bmatrix} 2.2E{-}16 & 1.2E{-}10 \\ 0 & 2.2E{-}16 \end{bmatrix}$	$\begin{bmatrix} 2.5E{-}16 & 1.2E{-}16 \\ 0 & 2.5E{-}16 \end{bmatrix}$

absolute error in (4.11) unless $\|I - R\|$ is sufficiently small. For this reason, people have been using (3.9) to work with $R^{1/2^k}$, where k was chosen so that $\|I - R^{1/2^k}\|$ was sufficiently small. For example, in [2], we chose k so that $\|I - R^{1/2^k}\| \leq 0.35$ and then used the $(9,9)$ diagonal Padé approximant, which guarantees—in absence of other errors—a value of η below EPS (see the first estimate in (4.14)). We also remarked that use of (3.9) may lead to undesired flattening (towards the identity) of the spectra of $R_{11}^{1/2^k}$ and $R_{22}^{1/2^k}$, and observed that (in some ill-conditioned problems) this eventually produced loss of precision. In [10], the authors attributed this loss of precision to the possible loss of significance incurred when forming $I - R^{1/2^k}$. Indeed, in [10], the declared motivation for the Fréchet method and the *direct* approximation of the Fréchet derivative by the hyperbolic tangent was the intent to avoid forming $I - R^{1/2^k}$. Kenney and Laub claimed that this technique allowed them to obtain better accuracy than the standard Padé method. However, we believe that the real reason behind the reported success of the approach in [10] has nothing to do with avoiding the subtraction $I - R^{1/2^k}$ or with the Padé method. To substantiate our claim, let us begin by looking at a simple example.

Example 4.4. In Table 1 we report on some results relative to Example 3.7 (see also [10, Example 2]) for several values of c and constant b, $b = 10^6$. The results have been obtained exploiting the relation $\log(R) = 2^k \log(R^{1/2^k})$, and we used the $(9,9)$ diagonal Padé approximation to obtain the logarithm of $R^{1/2^k}$.

In Table 1, k is the number of square roots taken before using the Padé approximation for the logarithm, $\omega = \|I - R^{1/2^k}\|_\infty$, $\omega_i = \|I - R_{ii}^{1/2^k}\|_\infty$, $i = 1, 2$, "`abs`" refers to the matrix of absolute errors: $\texttt{abs}_{ij} = |L_{ij} - \hat{L}_{ij}|$, and "`rel`" is the matrix of relative errors: $\texttt{rel}_{ij} = \texttt{abs}_{ij}/\,|L_{ij}|$. These numerical experiments were performed in `Matlab`, with machine precision EPS $\approx 2.2E-16$ (exponential notation is used throughout). In Table 2, we instead report on the absolute and relative error matrices after exponentiating $\hat{L}$, "`ABS`," and "`REL`," where $\texttt{ABS}_{ij} = |(R - e^{\hat{L}})_{ij}|$ and $\texttt{REL}_{ij} = \texttt{ABS}_{ij}/|R_{ij}|$ (see section 3 and (4.3)) for the same values of c and k as in Table 1.

The results highlight some interesting features: (i) Taking as many square roots as would have been needed to obtain $\|I - R^{1/2^k}\| \leq 0.35$, as in the second row of Table 1, we eventually lose seven digits along the diagonal of L, but the $(1,2)$ entry has full accuracy (look at the relative errors in Table 1); (ii) taking k so to bring only the diagonal of R to within ω (rows 1, 3, and 4 of Table 1) gives full precision on the diagonal of L and maintains accuracy for the $(1,2)$ entry. The loss of precision observed for $c = 0.1$ and $k = 22$ is easily explained as follows: Write $(e^c)^{1/2^{22}} = 1 + x$. After 22 square roots the number e^c has the finite precision

TABLE 2
Briggs–Padé method on Example 3.7: II.

c	k	ABS	REL
0.1	0	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$
0.1	22	$\begin{bmatrix} 3.3E{-}10 & 3.3E{-}4 \\ 0 & 3.3E{-}10 \end{bmatrix}$	$\begin{bmatrix} 3E{-}10 & 3E{-}10 \\ 0 & 3E{-}10 \end{bmatrix}$
0.3	0	$\begin{bmatrix} 0 & 9.3E{-}10 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 6.9E{-}16 \\ 0 & 0 \end{bmatrix}$
0.9	2	$\begin{bmatrix} 4.4E{-}16 & 9.3E{-}10 \\ 0 & 4.4E{-}16 \end{bmatrix}$	$\begin{bmatrix} 1.8E{-}16 & 3.8E{-}16 \\ 0 & 1.8E{-}16 \end{bmatrix}$

representation 1.000000023841858, so that only eight digits of x are retained. Now, since $\log(1+x) = x - x^2/2 + x^3/3 \ - \ \cdots$, in the double precision representation of the number $\log(1+x)$ we should not expect anything better than eight-digit accuracy (seemingly, we got nine). It is important to stress that the observed loss of digits is unavoidable given the finite precision representation of x, and no algorithm to approximate the logarithm can avoid it, whether or not 1 is subtracted. Apparently, the only way to avoid this is to take fewer square roots, so that x has more (say, sixteen) significant digits. To explain why the $(1,2)$ entry is always fully accurate, we need to wait until Theorem 4.6 below. Finally, the results of Table 2 are in striking agreement with the bound (4.6); in particular, for the second row of Table 2, with the notation from (4.6), $\eta \approx 10^{-10}$, $\frac{\|L_{12}\|}{\|X\|} \approx 1$, and $h(s) \approx 1$.

Example 4.4 makes it evident that the approximation goal expressed by (4.11) (with $\eta \approx$ EPS) is, in general, not attainable. Moreover, the legitimate suspicion is that problems are caused by taking square roots. We quantify this next.

LEMMA 4.5. *Let $R^{1/2^k}$ and $\hat{R}^{1/2^k}$ be the exact and computed kth square roots of R, respectively, where k has been chosen so that $\|I - \hat{R}^{1/2^k}\| = \omega < 1$. Let $\hat{L}^{(k)}$ be the approximation obtained for $\log \hat{R}^{1/2^k}$, for example, with the (N,N) diagonal Padé. Let $\|\log \hat{R}^{1/2^k} - \hat{L}^{(k)}\| \le \eta$ be the absolute error in approximating the logarithm and let $\frac{\|R^{1/2^k} - \hat{R}^{1/2^k}\|}{\|R^{1/2^k}\|} \le \delta$ be the relative error in the computed kth root. Then, we have*

$$\left\|\log R^{1/2^k} - \hat{L}^{(k)}\right\| \le \eta + \frac{\delta}{1-\omega}\left\|R^{1/2^k}\right\| + O(\delta^2)\,, \tag{4.12}$$

and therefore with $\hat{L} = 2^k \hat{L}^{(k)}$ one has

$$\left\|\log R - \hat{L}\right\| \le 2^k\left(\eta + \frac{\delta}{1-\omega}\left\|R^{1/2^k}\right\| + O(\delta^2)\right)\,. \tag{4.13}$$

Proof. We have

$$\begin{aligned}\left\|\log R^{1/2^k} - \hat{L}^{(k)}\right\| &\le \left\|\log R^{1/2^k} - \log \hat{R}^{1/2^k}\right\| + \left\|\log \hat{R}^{1/2^k} - \hat{L}^{(k)}\right\| \\ &\le \left\|\log'(\hat{R}^{1/2^k})\right\| \ \left\|R^{1/2^k} - \hat{R}^{1/2^k}\right\| + \left\|\log \hat{R}^{1/2^k} - \hat{L}^{(k)}\right\| \\ &\quad + O\left(\left\|R^{1/2^k} - \hat{R}^{1/2^k}\right\|^2\right)\,.\end{aligned}$$

Now we use [2, estimate after (3.2)] to get $\|\log'(\hat{R}^{1/2^k})\| \le \frac{1}{1-\omega}$, and the proof is complete. ∎

Suppose we use a very accurate formula to approximate $\log \hat{R}^{1/2^k}$, so that (theoretically) $\eta \leq$ EPS. Then, the overall computational error is controlled by the relative accuracy in the kth root, δ, and the magnification factor 2^k. In the best case, $\delta \approx$ EPS, an approximate equality which can be achieved in case R is normal, but even in this case, these $O(\text{EPS})$ errors may be magnified by 2^k.

Another important feature exhibited by Example 4.4 is that L_{12} is accurate. This is a consequence of the following result.

THEOREM 4.6. *Let R be partitioned as in (4.2) with $R_{11} \in \mathbb{R}^{n_1 \times n_1}$, $R_{22} \in \mathbb{R}^{n_2 \times n_2}$, and $R_{12} \in \mathbb{R}^{n_1 \times n_2}$. Let $R^{(k)} := R^{1/2^k}$ and write $R^{(k)} = \begin{bmatrix} R_{11}^{(k)} & R_{12}^{(k)} \\ 0 & R_{22}^{(k)} \end{bmatrix}$, where k is the smallest integer such that $\|I - R_{11}^{(k)}\| \leq \omega_1 < 1$ and $\|I - R_{22}^{(k)}\| \leq \omega_2 < 1$ for preassigned values of ω_1 and ω_2. Let $\omega = \max(\omega_1, \omega_2)$ and let $P = \begin{bmatrix} P_{11} & P_{12} \\ 0 & P_{22} \end{bmatrix}$ be the approximation obtained with the (N, N) diagonal Padé approximant for $L^{(k)} = \log(R^{(k)})$. Then, we have the following error bounds:*

$$\begin{aligned} \left\| L_{ii}^{(k)} - P_{ii} \right\| &\leq c(N)(2N)! \left(\frac{\omega_i}{1-\omega_i} \right)^{2N+1} , \ i = 1, 2 , \\ \left\| L_{12}^{(k)} - P_{12} \right\| &\leq c(N) \left\| R_{12}^{(k)} \right\| \frac{(2N+1)!}{(1-\omega)^2} \left(\frac{\omega}{1-\omega} \right)^{2N} , \end{aligned} \tag{4.14}$$

where $c(N) = \frac{(N!)^4}{(2N+1)((2N)!)^3}$. Further, we have the following bounds for $\|R_{12}^{(k)}\|$:

$$(1-\omega) \left\| L_{12}^{(k)} \right\| \leq \left\| R_{12}^{(k)} \right\| \leq \frac{1}{1-\omega} \left\| L_{12}^{(k)} \right\| , \tag{4.15}$$

which lead to a computable relative error bound in the second of (4.14).

Proof. The starting point of the proof is to recall that the (N, N) Padé approximant is the same as the N-point Gauss–Legendre quadrature rule for the integral in (1.1); therefore, with $M = I - R^{(k)}$, we have the following error estimate (see [2, Theorem 4.3 and Corollary 4.4]):

$$L^{(k)} - P = c(N) \sum_{j=0}^{\infty} (2N + j) \cdots (j + 1) M^{2N+j+1} \eta_j^j , \tag{4.16}$$

where $0 \leq \eta_j \leq 1$. Now, partition $M = \begin{bmatrix} A & C \\ 0 & B \end{bmatrix}$, and notice that $M^p = \begin{bmatrix} A^p & \sum_{j=0}^{p-1} A^{p-j} C B^j \\ 0 & B^p \end{bmatrix}$. By assumption we have $\|A\| \leq \omega_1 < 1$ and $\|B\| \leq \omega_2 < 1$, so that (taking norms of sub-blocks on the right-hand side of (4.16)) we immediately get

$$\begin{aligned} \left\| L_{ii}^{(k)} - P_{ii} \right\| &\leq c(N) \sum_{j=0}^{\infty} (2N + j) \cdots (j + 1) \omega_i^{2N+j+1} , \ i = 1, 2 , \\ \left\| L_{12}^{(k)} - P_{12} \right\| &\leq c(N) \|C\| \sum_{j=0}^{\infty} (2N + j) \cdots (j + 1) \sum_{l=0}^{2N+j} \omega_1^{2N+j-l} \omega_2^l \\ &\leq c(N) \|C\| \sum_{j=0}^{\infty} (2N + j + 1) \cdots (j + 1) \omega^{2N+j} . \end{aligned}$$

To complete the proof of (4.14), we observe that

$$\sum_{j=0}^{\infty} (2N + j) \cdots (j + 1) x^j = \frac{(2N)!}{(1-x)^{2N+1}} ,$$

from which (4.14) follows. To obtain (4.15), we first use (4.10) to get

$$L_{12}^{(k)} = \int_0^1 ((R_{11}^{(k)} - I)t + I)^{-1} R_{12}^{(k)} ((R_{22}^{(k)} - I)t + I)^{-1} dt\,,$$

and we notice that if $\|R_{ii}^{(k)} - I\| \le \omega_i$, then $\|((R_{ii}^{(k)} - I)t + I)^{-1}\| \le \frac{1}{1-\omega_i t}$, $i = 1, 2$. Therefore,

$$\left\|L_{12}^{(k)}\right\| \le \left\|R_{12}^{(k)}\right\| \int_0^1 \frac{dt}{(1-\omega_1 t)(1-\omega_2 t)}\,.$$

On the other hand, from (4.4), we also have

$$\left\|R_{12}^{(k)}\right\| \le \left\|L_{12}^{(k)}\right\| \int_0^1 \left\|e^{L_{11}^{(k)}(1-s)}\right\| \left\|e^{L_{22}^{(k)} s}\right\| ds\,.$$

Now, $\|e^{L_{ii}^{(k)} t}\| \le e^{\|L_{ii}^{(k)}\| t}$ and $\|L_{ii}^{(k)}\| \le \omega_i \int_0^1 \frac{dt}{1-\omega_i t} = -\log(1-\omega_i)$. Thus, $\|e^{L_{ii}^{(k)} t}\| \le \frac{1}{(1-\omega_i)^t}$; from these, we get

$$\left\|R_{12}^{(k)}\right\| \le \left\|L_{12}^{(k)}\right\| \int_0^1 \frac{dt}{(1-\omega_1)^{1-t}(1-\omega_2)^t}\ .$$

Finally, to get (4.15), it is enough to observe that $(1-\omega_1)^{1-t}(1-\omega_2)^t \ge (1-\omega)$, and also $\frac{1}{(1-\omega_1 t)(1-\omega_2 t)} \le \frac{1}{(1-\omega t)^2}$, for $0 \le t \le 1$. ∎

Remark 4.7. From (4.14) and (4.15), if $\omega = \max(\omega_1, \omega_2) \le 0.30$, we should get a fully accurate approximation for $L_{12}^{(k)}$ in agreement with the results of Example 4.4.

The fundamental implication of Theorem 4.6 is that a (diagonal) Padé approximation for L produces an approximation for L_{12} accurate in a relative-error sense, if the resulting approximations for L_{11} and L_{22} are accurate in an absolute-error sense. To be precise, the estimates (4.15) and (4.14) predict a loss of precision: For $\omega = 0.3$, the relative error bound for the approximation to L_{12} is about 50 times larger than the absolute error bound in the approximations of L_{11} and L_{22}. The situation is quite similar to what was obtained in [10]. However, we carry out the approximation of L_{12} *indirectly* by using Padé approximations for all of L at once, *rather than directly*, as done in [10], without sacrificing accuracy. The need for the kth square root in Theorem 4.6 is also present in the Schur–Fréchet algorithm of Kenney and Laub; this is what Step 1 of the algorithm in [10, p. 651] accomplishes. However, it should be realized that in Theorem 4.6 the value of k is not chosen so that $\|I - R^{1/2^k}\| < 1$, as in [10, Lemma 3.1 and later], but only so that $\|I - R_{ii}^{1/2^k}\| < 1$, $i = 1, 2$. Finally, as we had already remarked, there is no true issue associated with forming the subtraction $\|I - R^{1/2^k}\|$; regardless, the shift of emphasis from a global error bound, as in (4.11), to such an estimate only for the diagonal blocks makes it quite likely that fewer square roots of R must be taken and, thus, less likely that accuracy gets lost (see Lemma 4.5).

It is of course possible to extend the above considerations hinged on Theorem 4.6 to matrices as in (4.1). Although there are many subtle algorithmic issues which arise when we increase the number of blocks, Theorem 4.6 continues to hold for *all* possible choices of block-(2×2) submatrices. For example, suppose we have a block-(3×3) matrix: $\begin{bmatrix} R_{11} & R_{12} & R_{13} \\ 0 & R_{22} & R_{23} \\ 0 & 0 & R_{33} \end{bmatrix}$. Now, assume that $\|I - R_{ii}\| \le \omega_i < 1$, $i = 1, 2, 3$, and that

$\|I - \left[\begin{smallmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{smallmatrix}\right]\| \le \omega_{12} < 1$, or $\|I - \left[\begin{smallmatrix} R_{22} & R_{23} \\ 0 & R_{33} \end{smallmatrix}\right]\| \le \omega_{23} < 1$ (or both). Finally, take the (N, N) Padé approximant for $\log R$ (all of it). Theorem 4.6 applies and we can use the most favorable error estimates predicted by the theorem. We will think of the computed L_{12} and L_{23} as coming from the approximation to the logarithm of $\left[\begin{smallmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} R_{22} & R_{23} \\ 0 & R_{33} \end{smallmatrix}\right]$, respectively; the approximation for L_{13}, instead, can be thought of as coming from two different block partitionings: (i) $R = \left[\begin{smallmatrix} R_{11} & [\,R_{12} & R_{13}\,] \\ 0 & \left[\begin{smallmatrix} R_{22} & R_{23} \\ 0 & R_{33} \end{smallmatrix}\right] \end{smallmatrix}\right]$, or (ii) $R = \left[\begin{smallmatrix} \left[\begin{smallmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{smallmatrix}\right] & \left[\begin{smallmatrix} R_{13} \\ R_{23} \end{smallmatrix}\right] \\ 0 & R_{33} \end{smallmatrix}\right]$. We are free to pick whichever block partitioning is more convenient for us, which, in particular, gives us better error estimates from (4.14).

5. Conclusions. In this work we have

(i) classified and obtained conditioning information for primary logarithms of matrices by using piecewise analytic functions theory;

(ii) proposed an indicator of the inherent difficulty to compute the logarithm of a matrix;

(iii) revisited Padé approximation techniques to compute principal logarithms of block triangular matrices.

From the practical point of view, the most important outcome of this work is that we have shown that the familiar Padé approximation technique with inverse scaling and squaring is still a viable way for computing the logarithm L of a triangular matrix R as in (4.2) with close eigenvalues; however, there are situations where one should not try to obtain an absolute error bound for the approximate L, but only for its diagonal blocks, while a relative error criterion is more appropriate for the off-diagonal block. As a consequence, the number of square roots which we need to perform is not determined by the distance of the whole matrix from the identity, but rather by the distance from the identity of its diagonal blocks.

We believe that point (iii) above has far reaching theoretical and practical implications, and we anticipate some work along similar lines also for the exponential of a matrix.

REFERENCES

[1] A. Bjorck and S. Hammarling, *A Schur method for the square root of a matrix*, Linear Algebra Appl., 52-53 (1983), pp. 127–140.

[2] L. Dieci, B. Morini, and A. Papini, *Computational techniques for real logarithms of matrices*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 570–593.

[3] L. Dieci, B. Morini, A. Papini, and A. Pasquali, *On real logarithms of nearby matrices and structured matrix interpolation*, Appl. Numer. Math., 29 (1999), pp. 145–165.

[4] N. J. Higham, *Computing real square roots of a real matrix*, Linear Algebra Appl., 88-89 (1987), pp. 405–430.

[5] N. J. Higham, *Perturbation theory and backward error for $AX - XB = C$*, BIT, 33 (1993), pp. 124–136.

[6] R. A. Horn and C. R. Johnson, *Topics in Matrix analysis*, Cambridge University Press, New York, 1991.

[7] B. Kagström, *Bounds and perturbation bounds for the matrix exponential*, BIT, 17 (1977), pp. 39–57.

[8] C. Kenney and A. J. Laub, *Condition estimates for matrix functions*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 191–209.

[9] C. Kenney and A. J. Laub, *Padé error estimates for the logarithm of a matrix*, Internat. J. Control, 50 (1989), pp. 707–730.

[10] C. Kenney and A. J. Laub, *A Schur-Fréchet algorithm for computing the logarithm and exponential of a matrix*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 640–663.

[11] C. Van Loan, *The sensitivity of the matrix exponential*, SIAM J. Numer. Anal., 14 (1977), pp. 971–981.
[12] R. Mathias, *Condition estimation for matrix functions via the Schur decomposition*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 565–578.
[13] R. Mathias, *A chain rule for matrix functions and applications*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 610–620.
[14] B. N. Parlett, *A recurrence among the elements of functions of triangular matrices*, Linear Algebra Appl., 14 (1976), pp. 117–121.
[15] D. L. Powers, *The Fréchet differential of a primary matrix function*, Canad. J. Math., 25 (1973), pp. 554–559.
[16] E. Stickel, *On the Fréchet derivative of matrix functions*, Linear Algebra Appl., 91 (1987), pp. 83–88.
[17] T. Ström, *On logarithmic norms*, SIAM J. Numer. Anal., 12 (1975), pp. 741–753.
[18] V. A. Yakubovich and V. M. Starzhinskii, *Linear Differential Equations with Periodic Coefficients*, Vols. 1 and 2, John Wiley, New York, 1975.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 3, pp. 931–951
© 2000 Society for Industrial and Applied Mathematics

UPPER PERTURBATION BOUNDS OF WEIGHTED PROJECTIONS, WEIGHTED AND CONSTRAINED LEAST SQUARES PROBLEMS*

MUSHENG WEI† AND ALVARO R. DE PIERRO‡

Abstract. At each iteration step for solving mathematical programming and constrained optimization problems by using interior-point methods, one often needs to solve the weighted least squares (WLS) problem $\min_{x\in R^n} \|W^{\frac{1}{2}}(Ax-b)\|$, or the weighted and constrained least squares (WLSE) problem $\min_{x\in R^n} \|W^{\frac{1}{2}}(Kx-g)\|$ subject to $Lx=h$, where $W=\mathrm{diag}(w_1,\ldots,w_l)>0$ in which some $w_i\to+\infty$ and some $w_i\to 0$.

In this paper we will derive upper perturbation bounds of weighted projections associated with the WLS and WLSE problems when W ranges over the set $\mathcal{D}$ of positive diagonal matrices. We then apply these bounds to deduce upper perturbation bounds of solutions of WLS and WLSE problems when W ranges over $\mathcal{D}$. We also extend the estimates to the cases when W ranges over a subset of real symmetric positive semidefinite matrices.

Key words. weight, projection, constrained, least squares, upper perturbation bound

AMS subject classifications. 15A012, 65F35, 65K, 90C31

PII. S0895479898336306

1. Introduction.

In this paper we will use the following notation: $C^{m\times n}$ ($R^{m\times n}$) is the set of m by n matrices with complex (real) entries, $C_r^{m\times n}$ is a subset of $C^{m\times n}$ in which any matrix has rank r, $C^m(R^m)=C^{m\times 1}$ ($R^{m\times 1}$). I_m denotes the identity matrix of order m, $0_{m\times n}$ is the m by n matrix with zero entries (if no confusion occurs, we will omit the subscript), and $\|\cdot\|\equiv\|\cdot\|_2$ is the Euclidean matrix or vector norm. For any matrix $A\in C_r^{m\times n}$, $A^\dagger$ is the Moore–Penrose pseudoinverse [2] of A, A^T is the transpose of A, A^H is the conjugate transpose of A, $\mathrm{rank}(A)$ is the rank of A, and $R(A)$ is the range of A. $\mathcal{P}(A)$ is a set of real symmetric positive semidefinite matrices of order m such that for any $W\in\mathcal{P}(A)$, $\mathrm{rank}(WA)=\mathrm{rank}(A)$. $\sigma_j(A)$ denotes the jth largest singular value of A and $\inf_+(A)$ denotes the smallest nonzero singular value of A. $\mathcal{D}$ denotes the set of positive diagonal matrices of appropriate order. Given a vector of r increasing integers $J=\{i_1,\ldots,i_r\}\subset\{1,\ldots,m\}$, $A\in C_r^{m\times n}$, and $D=\mathrm{diag}(d_1,\ldots,d_m)$, A_J denotes a submatrix of A formed with rows $i_1,\ldots,i_r\in J$ and $D_{JJ}=\mathrm{diag}(d_{i_1},\ldots,d_{i_r})$. $J(A)$ is a set of indexes defined by

$$J(A)=\{J=\{i_1,\ldots,i_r\}: 1\le i_1<\cdots<i_r\le m, \mathrm{rank}(A_J)=\mathrm{rank}(A)=r\}. \tag{1.1}$$

Let $A\in C^{m\times n}$, $b\in C^m$, and $W\in\mathcal{P}(A)$ be given, which satisfy $\mathrm{rank}(W^{\frac{1}{2}}A)=\mathrm{rank}(A)$. Then the weighted least squares (WLS) problem

$$\min_{x\in C^n} \|W^{\frac{1}{2}}(Ax-b)\| \tag{1.2}$$

*Received by the editors March 26, 1998; accepted for publication (in revised form) by S. Van Huffel July 2, 1999; published electronically February 29, 2000.

http://www.siam.org/journals/simax/21-3/33630.html

†Department of Mathematics, East China Normal University, Shanghai 200062, China (mwei@math.ecnu.edu.cn). This author was partially supported by the National Natural Science Foundation, People's Republic of China, grant 19871029, and FAPESP, Brazil, grant 1996/2758–0.

‡Department of Applied Mathematics, IMECC, State University of Campinas, SP, Brazil (alvaro@ime.unicamp.br). This author was partially supported by FAPESP, Brazil, grant 1996/2758–0 and CNPq, Brazil, grant 301699–81.

is equivalent to the least squares (LS) problem [4]

$$\min_{x\in C^n} \|A(W)x - b\|, \tag{1.3}$$

in which [27]

$$A(W) = WA(WA)^\dagger A, \qquad A(W)^\dagger = (W^{\frac{1}{2}}A)^\dagger W^{\frac{1}{2}}. \tag{1.4}$$

Similarly, suppose that $L \in C_p^{m_1\times n}$, $K \in C^{m_2\times n}$, and $A = \binom{L}{K} \in C_r^{m\times n}$, $h \in C^{m_1}$, $g \in C^{m_2}$, and $W \in \mathcal{P}(KP)$ (see (1.8)). Then the weighted and constrained least squares (WLSE) problem

$$\min_{x\in C^n} \|W^{\frac{1}{2}}(Kx - g)\| \quad \text{subject to } Lx = h \tag{1.5}$$

is equivalent to the LS problem [4]

$$\min_{x\in C^n} \|A(W)x - b\|, \tag{1.6}$$

in which [27]

(1.7)

$$A(W) = \begin{pmatrix} L \\ WKP(WKP)^\dagger K \end{pmatrix},\ A(W)^\dagger = \left(L^\dagger_{W^{\frac{1}{2}}K}, (W^{\frac{1}{2}}KP)^\dagger W^{\frac{1}{2}}\right),\ b = \begin{pmatrix} h \\ g \end{pmatrix}$$

with

$$P = I - L^\dagger L \ \text{ and } L^\dagger_{W^{\frac{1}{2}}K} = (I - (W^{\frac{1}{2}}KP)^\dagger W^{\frac{1}{2}}K)L^\dagger. \tag{1.8}$$

According to (1.3), (1.6), and [4], any WLS or WLSE solution x of (1.2) or (1.5) has the form

$$x = A(W)^\dagger b + (I - A(W)^\dagger A(W))z = A(W)^\dagger b + (I - A^\dagger A)z, \tag{1.9}$$

where $z \in C^n$ is an arbitrary vector, and we have [27, 4]

$$A(W)^\dagger A(W) = A(W)^\dagger A = A^\dagger A. \tag{1.10}$$

Also, let the QRD of A be $A = QR$, then with $A(W)^\dagger$ defined in (1.4), we have

$$\begin{aligned} (I - AA(W)^\dagger)AA^\dagger &= (I - Q(W^{\frac{1}{2}}Q)^\dagger W^{\frac{1}{2}})QQ^H = 0, \\ I - AA(W)^\dagger &= (I - AA(W)^\dagger)(I - AA^\dagger). \end{aligned} \tag{1.11}$$

The WLS and WLSE problems have been studied extensively for fixed $W \in \mathcal{D}$, including computational algorithms, structures, and perturbation analysis [1, 4, 6, 7, 11, 13, 18, 20, 23, 24, 25].

On the other hand, when solving the mathematical programming and nonlinear constrained optimization problem

$$\min_{y\in R^n} f(y) \ \text{ subject to } y \geq 0 \text{ and } Cy = d$$

by interior-point methods, one will obtain the WLS or WLSE problems of the forms found in (1.2) or (1.5), where $W = \operatorname{diag}(w_1, \ldots, w_l) > 0$ in which some $w_i \to +\infty$, some $w_i \to 0$.

In general, when the matrix W ranges over $\mathcal{D}$, the condition number of the matrix $W^{\frac{1}{2}}A$ or $W^{\frac{1}{2}}K$ can be unbounded, and in this case the WLS problem (1.2) or the WLSE problem (1.5) become "ill-conditioned." From (1.3)–(1.4) or (1.6)–(1.8), the boundedness of the WLS or WLSE solution x depends upon the upper bounds of $\|A(W)^\dagger\|$ when W ranges over a set $\mathcal{P} \subseteq \mathcal{P}(A)$.

By applying the Binet–Cauchy formula for determinants and Cramer's rule for solving a system of linear equations [15], the authors of [3, 5, 14, 17] discussed the geometry of the minimum norm solution of the WLS problem and linear equations. The authors of [5, 9, 10, 12, 21, 22, 28, 29, 30] proposed several algorithms and discussed stability for solving constrained optimization problem.

When W ranges over $\mathcal{D}$ and A is of full column rank, Stewart [19] and O'Leary [16] obtain the supremum of scaled projections $\|AA(W)^\dagger\|$ and an upper bound of $\|A(W)^\dagger\|$. Based on their results, Vavasis [22] derives a perturbation bound of an equilibrium system when the matrix A is of full column rank with no perturbation and the matrix W has small (diagonal) perturbations. Vavasis also proposes some stable algorithms to evaluate the solution of the equilibrium system.

Wei [26] obtains the supremum of scaled pseudoinverses $\|A(W)^\dagger\|$ in the form of (1.4) for general A. When W ranges over the set $\mathcal{P}$ of real symmetric diagonally dominant positive semidefinite matrices, Forsgren [8] derives the supremum of weighted pseudoinverses $\|A(W)^\dagger\|$ in the forms of (1.4) as well as (1.7), and Wei [27] extends his results to derives equivalent formulas of the supremum of weighted pseudoinverses $\|A(W)^\dagger\|$. Notice that when $\operatorname{rank}(A) < m$ and W ranges over the set $\mathcal{P}(A)$ of all real symmetric positive matrices, the supremum of weighted pseudoinverses $\|A(W)^\dagger\|$ is infinity, as shown by Stewart [19] in an example.

Wei [26, 27] also discusses stability properties of these weighted pseudoinverses $\|A(W)^\dagger\|$, that is, under what conditions the supremum of weighted pseudoinverses $\|A(W)^\dagger\|$ depends continuously on A.

Based on the results in [26, 8, 27], we derive upper perturbation bounds of weighted projections, WLS and WLSE solutions of (1.2) and (1.5) when W ranges over $\mathcal{D}$ and $\mathcal{P}$, where both the matrices A and W may have small perturbations.

The paper is organized as follows. Section 2 will derive upper perturbation bounds of weighted projections when W ranges over $\mathcal{D}$. Section 3 will derive upper perturbation bounds of solutions of the WLS and WLSE problems when W ranges over $\mathcal{D}$. Section 4 will extend the results of sections 2 and 3 to the case that W ranges over a subset of real symmetric semipositive matrices. Section 5 will conclude the paper with some remarks.

The following fact is used frequently in the paper.

LEMMA 1.1. *Let* $\alpha_1 > 0,\ \alpha_2 > 0,\ \beta_1 > 0$, *and* $\beta_2 > 0$. *If* $\alpha_1\beta_1 \geq \alpha_2\beta_2$, *then*

(1.12)
$$0 \leq \frac{\alpha_1\beta_1 - \alpha_2\beta_2}{\alpha_1\beta_1 + \alpha_2\beta_2} \leq \max\left\{\left|\frac{\alpha_1}{\alpha_2} - 1\right|, \left|\frac{\beta_2}{\beta_1} - 1\right|\right\},\ 0 \leq \frac{\alpha_1\beta_1 - \alpha_2\beta_2}{\alpha_1\beta_1 + \alpha_2\beta_2} \leq 1 - \frac{\alpha_2\beta_2}{\alpha_1\beta_1} < 1.$$

Proof. It can be verified that

$$0 \leq \frac{\alpha_1\beta_1 - \alpha_2\beta_2}{\alpha_1\beta_1 + \alpha_2\beta_2} = \frac{\alpha_1\beta_1(1 - \frac{\beta_2}{\beta_1}) + \alpha_2\beta_2(\frac{\alpha_1}{\alpha_2} - 1)}{\alpha_1\beta_1 + \alpha_2\beta_2} \leq \max\left\{\left|\frac{\alpha_1}{\alpha_2} - 1\right|,\ \left|\frac{\beta_2}{\beta_1} - 1\right|\right\},$$

proving the first inequality of (1.12). Also, we have

$$0 \leq \frac{\alpha_1\beta_1 - \alpha_2\beta_2}{\alpha_1\beta_1 + \alpha_2\beta_2} = 1 - \frac{2\alpha_2\beta_2}{\alpha_1\beta_1 + \alpha_2\beta_2} \leq 1 - \frac{2\alpha_2\beta_2}{2\alpha_1\beta_1} < 1,$$

proving the second inequality of (1.12). ∎

2. Upper perturbation bounds of the weighted projections when weight matrix ranges over $\mathcal{D}$. In this section, we derive upper perturbation bounds of weighted projections of a matrix $A \in C_r^{m\times n}$ when weight matrix W ranges over $\mathcal{D}$. For given $A \in C_r^{m\times n}$ and $W \in \mathcal{D}$, the weighted projection is defined by $A(W^{\frac{1}{2}}A)^{\dagger}W^{\frac{1}{2}}$ [19]. The following lemma generalizes the result in [5].

LEMMA 2.1. *Let $A \in C_r^{m\times n}$ and $W \in \mathcal{D}$ be given. Then*

$$(W^{\frac{1}{2}}A)^{\dagger}W^{\frac{1}{2}} = \frac{\sum_{J\in J(A)} a_J d_J A_J^{\dagger} I_J}{\sum_{K\in J(A)} a_K d_K}, \tag{2.1}$$

in which

$$I_J \equiv (I_m)_J,\ a_J = \det(A_J(A_J)^H),\ \text{and } d_J = \det(W_{JJ}). \tag{2.2}$$

Proof. Let a unitary decomposition of A be

$$A = QR, \quad \text{where } Q^HQ = I_r \ \text{ and } R \ \text{ has full row rank } r. \tag{2.3}$$

Applying the Binet–Cauchy formula for determinants and Cramer's rule for solving a system of linear equations [15], for given vector $g \in C^m$ we have that

$$\begin{aligned}\left[(W^{\frac{1}{2}}Q)^{\dagger}W^{\frac{1}{2}}g\right]_i &= \left[(Q^HWQ)^{-1}Q^HWg\right]_i = \frac{\det(Q^HW(Q+(g-Qe_i)e_i^T))}{\det(Q^HWQ)}\\ &= \sum_{J\in J(A)} \frac{|\det(Q_J)|^2\cdot\det(W_{JJ})}{\sum_{K\in J(A)}|\det(Q_K)|^2\cdot\det(W_{KK})}\left[(Q_J)^{-1}g_J\right]_i.\end{aligned}$$

So we deduce that

$$(W^{\frac{1}{2}}A)^{\dagger}W^{\frac{1}{2}}g = R^{\dagger}(W^{\frac{1}{2}}Q)^{\dagger}W^{\frac{1}{2}}g = \frac{\sum_{J\in J(A)}|\det(Q_J)|^2\cdot\det(W_{JJ})}{\sum_{K\in J(A)}|\det(Q_K)|^2\cdot\det(W_{KK})}A_J^{\dagger}g_J.$$

Using the fact that $a_J = \det(Q_JRR^HQ_J^H) = \det(RR^H)\cdot|\det(Q_J)|^2$ and taking $g = e_1,\dots,e_m$ in the above identity we obtain that

$$(W^{\frac{1}{2}}A)^{\dagger}W^{\frac{1}{2}} = (W^{\frac{1}{2}}A)^{\dagger}W^{\frac{1}{2}}(e_1,\dots,e_m) = \frac{\det(RR^H)\sum_{J\in J(A)}|\det(Q_J)|^2d_J}{\det(RR^H)\sum_{K\in J(A)}|\det(Q_K)|^2d_K}A_J^{\dagger}I_J,$$

proving (2.1). ∎

LEMMA 2.2. *Let $A,\ A' \in C_r^{m\times n}$, and $W,\ W' \in \mathcal{D}$ be given. Then*

$$(W'^{\frac{1}{2}}A')^{\dagger}W'^{\frac{1}{2}}(I - A(W^{\frac{1}{2}}A)^{\dagger}W^{\frac{1}{2}}) = \frac{\sum_{J\in J(A')}\sum_{M\in J(A)} a'_Jd'_Ja_Md_MA_J'^{\dagger}(I_J - A_JA_M^{\dagger}I_M)}{\sum_{K\in J(A')}\sum_{N\in J(A)} a'_Kd'_Ka_Nd_N} \tag{2.4}$$

$$= \frac{\sum_{J\in J(A')}\sum_{M\in J(A)} a'_Jd'_Ja_Md_MA_J'^{\dagger}(I_J - A_JA_M^{\dagger}I_M) + \sum_{M\in J(A')}\sum_{J\in J(A)} a'_Md'_Ma_Jd_JA_M'^{\dagger}(I_M - A_MA_J^{\dagger}I_J)}{\sum_{K\in J(A')}\sum_{N\in J(A)} a'_Kd'_Ka_Nd_N + \sum_{N\in J(A')}\sum_{K\in J(A)} a'_Nd'_Na_Kd_K},$$

in which

$$a'_J = \det(A'_J(A'_J)^H) \quad \textit{and } d'_J = \det(W'_{JJ}). \tag{2.2a}$$

Proof. By applying Lemma 2.1 we obtain the first identity of (2.4):

$$\begin{aligned}
&(W'^{\frac{1}{2}}A')^\dagger W'^{\frac{1}{2}}(I - A(W^{\frac{1}{2}}A)^\dagger W^{\frac{1}{2}}) \\
&= \frac{\sum_{J\in J(A')} a'_J d'_J A'^{\dagger}_J I_J}{\sum_{K\in J(A')} a'_K d'_K}\left(I - A\frac{\sum_{M\in(A)} a_M d_M A^\dagger_M I_M}{\sum_{N\in J(A)} a_N d_N}\right) \\
&= \frac{\sum_{J\in J(A')}\sum_{M\in J(A)} a'_J d'_J a_M d_M A'^{\dagger}_J (I_J - A_J A^\dagger_M I_M)}{\sum_{K\in J(A')}\sum_{N\in J(A)} a'_K d'_K a_N d_N},
\end{aligned}$$

using the fact that $I_J A = A_J$. Because of the fact that changing subindexes does not change the values of summations, we immediately obtain the second identity of (2.4). ∎

We are now ready to present the main results of this section. In order to better understand the effect of perturbations on the upper bound of $\|(W'^{\frac{1}{2}}A')^\dagger W'^{\frac{1}{2}}(I - A(W^{\frac{1}{2}}A)^\dagger W^{\frac{1}{2}})\|$, we first consider two simple cases.

Case 1. $A' = A \in C_r^{m\times n}$ and $W = \operatorname{diag}(w_1, \ldots, w_m)$, $W' = \operatorname{diag}(w_1 + \delta w_1, \ldots, w_m + \delta w_m) \in \mathcal{D}$. We then have the following theorem.

THEOREM 2.1. *Suppose $A \in C_r^{m\times n}$ with $0 < r$. Then we have the following estimate:*

$$\begin{gathered}
\sup_{W,\ W'\in\mathcal{D},\eta\le\Delta} \|(W'^{\frac{1}{2}}A)^\dagger W'^{\frac{1}{2}}(I - A(W^{\frac{1}{2}}A)^\dagger W^{\frac{1}{2}})\| \le \tfrac{2e_1}{\rho}, \\
e_1 = \min\{(1+\Delta)^{r_m} - 1,\ 1 - (1-\Delta)^{2r_m}\},
\end{gathered} \tag{2.5}$$

in which $\eta = \max\limits_{1\le i\le m} \frac{|\delta w_i|}{w_i}$, $1 > \Delta \ge 0$ *is a constant, and* [26]

$$\rho = \min_{J\in J(A)} \inf\nolimits_+(A_J) = \left(\sup_{W\in\mathcal{D}} \|(W^{\frac{1}{2}}A)^\dagger W^{\frac{1}{2}}\|\right)^{-1},\ r_m = \min\{r, m-r\}. \tag{2.6}$$

Proof. From (2.3), for any $J,\ M \in J(A)$, we have that

$$A^\dagger_J A_J = (Q_J R)^\dagger (Q_J R) = R^\dagger Q_J^{-1} Q_J R = R^\dagger R = A^\dagger_M A_M \quad \text{and} \quad A_J A^\dagger_J = I_r. \tag{2.7}$$

By substituting (2.7) into (2.4) with $A' = A$, so that $J(A') = J(A)$ and $a'_J = a_J$ for all $J \in J(A)$, we obtain

$$\begin{aligned}
&(W'^{\frac{1}{2}}A)^\dagger W'^{\frac{1}{2}}(I - A(W^{\frac{1}{2}}A)^\dagger W^{\frac{1}{2}}) \\
&\quad = \frac{\sum_{J\in J(A)}\sum_{M\in J(A)} a_J a_M (d'_J d_M - d_J d'_M)(A^\dagger_J I_J - A^\dagger_M I_M)}{\sum_{K\in J(A)}\sum_{N\in J(A)} a_K a_N (d'_K d_N + d_K d'_N)},
\end{aligned}$$

from which we obtain

$$\|(W'^{\frac{1}{2}}A)^\dagger W'^{\frac{1}{2}}(I - A(W^{\frac{1}{2}}A)^\dagger W^{\frac{1}{2}})\| \le \max_{J,M} \frac{|d'_J d_M - d_J d'_M|}{d'_J d_M + d_J d'_M} \|A^\dagger_J I_J - A^\dagger_M I_M\|. \tag{2.8}$$

For given $J,\ M \in J(A)$ denote

$$d_J = d_{J\cap M}\cdot d_{J/M},\ d_M = d_{J\cap M}\cdot d_{M/J},\ d'_J = d'_{J\cap M}\cdot d'_{J/M},\ d'_M = d'_{J\cap M}\cdot d'_{M/J},$$

in which $d_{J\cap M}$ ($d'_{J\cap M}$) contains all common factors of d_J and d_M (d'_J and d'_M). Then we have from Lemma 1.1 that

(2.9a)
$$\frac{|d'_J d_M - d_J d'_M|}{d'_J d_M + d_J d'_M} = \frac{|d'_{J/M} d_{M/J} - d_{J/M} d'_{M/J}|}{d'_{J/M} d_{M/J} + d_{J/M} d'_{M/J}} \le \max\left\{\left|1 - \frac{d'_{M/J}}{d_{M/J}}\right|, \left|\frac{d'_{J/M}}{d_{J/M}} - 1\right|\right\}$$
$$\le ((1+\Delta)^{r-n(J\cap M)} - 1),$$

in which $n(J\cap M)$ is the number of common elements of J and M. Without loss of generality, let $d'_{J/M} d_{M/J} \ge d_{J/M} d'_{M/J}$; then, also from Lemma 1.1,

(2.9b)
$$\frac{|d'_J d_M - d_J d'_M|}{d'_J d_M + d_J d'_M} = \frac{d'_{J/M} d_{M/J} - d_{J/M} d'_{M/J}}{d'_{J/M} d_{M/J} + d_{J/M} d'_{M/J}} \le 1 - \frac{d_{J/M} d'_{M/J}}{d'_{J/M} d_{M/J}}$$
$$\le 1 - (1-\Delta)^{2(r-n(J\cap M))},$$

in which we have used the following inequalities:

$$\frac{d'_{M/J}}{d_{M/J}} = \Pi_{k\in M/J}\left(1 + \frac{\delta w_k}{w_k}\right) \ge (1-\Delta)^{(r-n(J\cap M))},$$
$$\frac{d_{J/M}}{d'_{J/M}} \ge \Pi_{k\in J/M,\ \delta w_k>0}\left(1 - \frac{\delta w_k}{w_k + \delta w_k}\right) \ge (1-\Delta)^{r-n(J\cap M)}.$$

Notice that $r - n(J\cap M) \le r$ and $r - n(J\cap M) \le m - r$. So substituting the inequalities of (2.9a) and (2.9b) into (2.8), we obtain the desired estimate in (2.5). □

Remark 2.1. It is easy to check that when $r - m$ or r is small, then the inequality in (2.5) is nearly optimal. Also note that $\eta = \|W^{-1}\delta W\|$ is the relative error bound in W.

Case 2. $W' = W \in \mathcal{D}$. In this case we have the following theorem.

THEOREM 2.2. *Suppose $A \in C_r^{m\times n}$ with $0 < r$. Then*

$$\sup_{W\in\mathcal{D}} \|(W^{\frac12}A')^\dagger W^{\frac12}(I - A(W^{\frac12}A)^\dagger W^{\frac12})\| < +\infty$$

for any $A' = A + \delta A \in C_r^{m\times n}$ with sufficiently small perturbation δA if and only if

(2.10) *any r rows of A are linearly independent.*

Moreover, if the condition in (2.10) *holds, then for any $A' = A + \delta A \in C_r^{m\times n}$ with $\|\delta A\| < \rho$, $J(A') = J(A)$, and we have the following estimates:*

(2.11)
$$\sup_{W\in\mathcal{D}} \|(W^{\frac12}A')^\dagger W^{\frac12}(I - A(W^{\frac12}A)^\dagger W^{\frac12})\| \le \frac{2}{\rho - \|\delta A\|}\left(\frac{\|\delta A\|}{\rho} + e_2\right),$$
$$e_2 = \min\left\{\left(1 + \frac{\|\delta A\|}{\rho}\right)^{2r} - 1,\ 1 - \left(1 - \frac{\|\delta A\|}{\rho}\right)^{4r}\right\}.$$

Proof. Necessity. If the condition in (2.10) does not hold, then there exist two vectors of r integers $J_0,\ J_1 \subset \{1,\dots,m\}$, such that $n(J_0\cap J_1) = r-1$, $\operatorname{rank}(A_{J_0}) = r-1$, and $\operatorname{rank}(A_{J_1}) = r$. Without loss of generality, we assume that $J_0 = \{1,\dots,r-1,r\}$

and $J_1 = \{1, \ldots, r-1, r+1\}$. Let a unitary decomposition of A be as in (2.3). Then from the assumption, $\operatorname{rank}(Q_{J_0}) = r-1$ and $\operatorname{rank}(Q_{J_1}) = r$. Let the SVD [11] of Q_{J_0} be

$$Q_{J_0} = (U_1, U_2)\Sigma(V_1, V_2)^H = U_1 \Sigma_1 V_1^H,$$

in which U and V are unitary matrices of order r, and U_1, V_1 (U_2, V_2) are, respectively, the first $r-1$ (remaining one) columns of U, V. For any $0 < \epsilon < \rho$, choose $Q_\epsilon = \epsilon U_2 V_2^H$ and let $A' = A + (I_{J_0})^T Q_\epsilon R$. Then $\operatorname{rank}(A') = r$, because $Q_{J_0} + Q_\epsilon$ is nonsingular and so $A'_{J_0} = (Q_{J_0} + Q_\epsilon)R$ has full row rank r. For a given $\delta > 0$, define $W(\delta) = W'(\delta) = \operatorname{diag}(I_{r-1}, \delta, \delta^2, \delta^3 I_{m-r-1}) \in \mathcal{D}$. Then with the notation in (2.2) and (2.2a) we have that

$$\begin{aligned} &J_0 \notin J(A),\ J_1 \in J(A),\ d_{J_1} = \delta^2,\ d_J = \delta^l \text{ with } l \geq 3 \text{ for } J_1 \neq J \in\ J(A), \\ &J_0 \in J(A'),\ d'_{J_0} = \delta,\ d'_J = \delta^l \text{ with } l \geq 2 \text{ for } J_0 \neq J \in\ J(A'). \end{aligned} \tag{2.12}$$

Then we have from (2.4) and (2.12) that

$$\begin{aligned} &\lim_{\delta \to 0} \| (W'(\delta)^{\frac{1}{2}} A')^\dagger W'(\delta)^{\frac{1}{2}} (I - A(W(\delta)^{\frac{1}{2}} A)^\dagger W(\delta)^{\frac{1}{2}}) \| \\ &\quad = \left\| \frac{a'_{J_0} a_{J_1} A_{J_0}^{'\dagger} (I_{J_0} - A_{J_0} A_{J_1}^\dagger I_{J_1})}{a'_{J_0} a_{J_1}} \right\| \\ &\quad = \| A_{J_0}^{'\dagger} (I_{J_0} - A_{J_0} A_{J_1}^\dagger I_{J_1}) \| = \| R^\dagger (Q_{J_0}^\dagger + Q_\epsilon^\dagger)(I_{J_0} - Q_{J_0} R A_{J_1}^\dagger I_{J_1}) \| \\ &\quad \geq \frac{\| R^\dagger V_2 \|}{\epsilon} - \| R^\dagger Q_{J_0}^\dagger (I_{J_0} - Q_{J_0} R A_{J_1}^\dagger I_{J_1}) \| \to +\infty \end{aligned}$$

when $\epsilon \to 0$. This proves the necessity part.

Sufficiency. Suppose that the condition in (2.10) holds. Then for each $A' = A + \delta A \in C_r^{m \times n}$ with $\|\delta A\| < \rho$ and each $J = \{i_1, \ldots, i_r\} \subset \{1, \ldots, m\}$, we have from the perturbation theory of singular values [11] that

$$\sigma_r(A'_J) \geq \sigma_r(A_J) - \|\delta A\| \geq \rho - \|\delta A\| > 0,$$

so $\operatorname{rank}(A'_J) = r$. Taking $W' = W$ in Lemma 2.2 so $d'_J = d_J$ for all $J = \{i_1, \ldots, i_r\} \subset \{1, \ldots, m\}$, we have that (for simplicity we use $\sum_J \equiv \sum_{J \in J(A)} = \sum_{J \in J(A')}$)

(2.13)

$$\begin{aligned} &\| (W^{\frac{1}{2}} A')^\dagger W^{\frac{1}{2}} (I - A(W^{\frac{1}{2}} A)^\dagger W^{\frac{1}{2}}) \| \\ &\quad = \left\| \frac{\sum_J \sum_M d_J d_M \left(a'_J a_M A_J^{'\dagger} (I_J - A_J A_M^\dagger I_M) + a_J a'_M A_M^{'\dagger} (I_M - A_M A_J^\dagger I_J) \right)}{\sum_K \sum_N d_K d_N (a'_K a_N + a_K a'_N)} \right\| \\ &\quad \leq \max_{J,M} \frac{\| a'_J a_M A_J^{'\dagger} (I_J - A_J A_M^\dagger I_M) + a_J a'_M A_M^{'\dagger} (I_M - A_M A_J^\dagger I_J) \|}{a'_J a_M + a_J a'_M} \\ &\quad \leq \max_{J,M} \left(\frac{|(a'_J a_M - a'_M a_J)|}{a'_J a_M + a_J a'_M} \| (A_J^{'\dagger} I_J - A_M^{'\dagger} I_M) \| + \| (A_J^{'\dagger} - A_M^{'\dagger} A_M A_J^\dagger) I_J \| \right). \end{aligned}$$

Furthermore, for fixed vectors of r integers $J,\ M \subseteq \{1, \ldots, m\}$, because

$$\frac{a'_M}{a_M} = \frac{\det(A'_M (A'_M)^H)}{\det(A_M (A_M)^H)} = \frac{\sigma_1(A'_M)^2 \cdots \sigma_r(A'_M)^2}{\sigma_1(A_M)^2 \cdots \sigma_r(A_M)^2},$$

so by applying Lemma 1.1 and using exactly the same strategy as proving (2.9a) and (2.9b), we obtain

$$\frac{|a'_J a_M - a_J a'_M|}{a'_J a_M + a_J a'_M} \le e_2. \tag{2.14}$$

When the condition in (2.10) holds, we have the following inequalities [27, Theorem 4.1]:

$$\rho - \|\delta A\| \le \rho' \le \rho + \|\delta A\|. \tag{2.15}$$

From (2.7), $A_M^{'\dagger} A'_M = A_J^{'\dagger} A'_J$ and $A_J A_J^\dagger = I_r$. So

$$\begin{aligned} \|(A_J^{'\dagger} - A_M^{'\dagger} A_M A_J^\dagger) I_J\| &= \left\| \left[A_M^{'\dagger}(A_M^{'} - A_M) A_J^\dagger - A_J^{'\dagger}(A_J^{'} - A_J) A_J^\dagger \right] I_J \right\| \\ &= \left\| \left[A_M^{'\dagger}(\delta A)_M - A_J^{'\dagger}(\delta A)_J \right] A_J^\dagger \right\| \le \frac{2\|\delta A\|}{\rho\rho'} \le \frac{2\|\delta A\|}{\rho(\rho - \|\delta A\|)}. \end{aligned} \tag{2.16}$$

We then complete the proof by substituting (2.14)–(2.16) into (2.13). ∎

Remark 2.2. Note that the condition in (2.10) guarantees that $J(A) = J(A')$ (which automatically holds when $A' = A$, as in the case of Theorem 2.1). The term e_2 is an upper bound of $\frac{|a'_J a_M - a_J a'_M|}{a'_J a_M + a_J a'_M}$. When r is large, this bound would be very rough where we have used the worst-case estimate. It is interesting if this bound can be improved.

Finally, we consider the case in which there are general perturbations for the matrices A and W.

Case 3. *General case.* In this case we have the following theorem.

THEOREM 2.3. *Suppose* $A \in C_r^{m\times n}$ *with* $0 < r$, $1 > \Delta > 0$ *is a constant. Then*

$$\sup_{W, W' \in \mathcal{D}, \eta \le \Delta} \|(W^{'\frac{1}{2}} A')^\dagger W^{'\frac{1}{2}} (I - A(W^{\frac{1}{2}} A)^\dagger W^{\frac{1}{2}})\| < +\infty$$

for any $A' = A + \delta A \in C_r^{m\times n}$ *with sufficiently small perturbations* δA *and* δW, *if and only if the condition in* (2.10) *holds. Moreover, if the condition in* (2.10) *holds, then for any* $A' = A + \delta A \in C_r^{m\times n}$ *with* $\|\delta A\| < \rho$, *we have the following estimate:*

$$\begin{aligned} &\sup_{W,\ W' \in \mathcal{D}, \eta \le \Delta} \|(W^{'\frac{1}{2}} A')^\dagger W^{'\frac{1}{2}} (I - A(W^{\frac{1}{2}} A)^\dagger W^{\frac{1}{2}})\| \le \frac{2}{\rho - \|\delta A\|} \left(\frac{\|\delta A\|}{\rho} + e_3 \right), \\ &e_3 = \min \left\{ (1+\Delta)^{r_m} \left(1 + \frac{\|\delta A\|}{\rho} \right)^{2r} - 1,\ 1 - (1-\Delta)^{2r_m} \left(1 - \frac{\|\delta A\|}{\rho} \right)^{4r} \right\}. \end{aligned} \tag{2.17}$$

Proof. The necessity part is proved in Theorem 2.1 because we can choose $W' = W \in \mathcal{D}$. To prove the sufficiency part, suppose that the condition in (2.10) holds. Then from (2.4) we get that

(2.18)

$$\begin{aligned}
&\|(W^{'\frac{1}{2}}A')^{\dagger}W^{'\frac{1}{2}}(I-A(W^{\frac{1}{2}}A)^{\dagger}W^{\frac{1}{2}})\| \\
&= \left\| \frac{\sum_J \sum_M \left(d'_J d_M a'_J a_M A_J^{'\dagger}(I_J - A_J A_M^{\dagger} I_M) + d_J d'_M a_J a'_M A_M^{'\dagger}(I_M - A_M A_J^{\dagger} I_J) \right)}{\sum_K \sum_N (d'_K d_N a'_K a_N + d_K d'_N a_K a'_N)} \right\| \\
&\le \max_{J,M} \frac{\|a'_J d'_J a_M d_M A_J^{'\dagger}(I_J - A_J A_M^{\dagger} I_M) + a'_M d'_M a_J d_J A_M^{'\dagger}(I_M - A_M A_J^{\dagger} I_J)\|}{a'_J d'_J a_M d_M + a'_M d'_M a_J d_J} \\
&\le \max_{J,M} \left(\frac{|a'_J d'_J a_M d_M - a'_M d'_M a_J d_J|}{a'_J d'_J a_M d_M + a'_M d'_M a_J d_J} \|A_J^{'\dagger} I_J - A_M^{'\dagger} I_M\| + \|(A_J^{'\dagger} - A_M^{'\dagger} A_M A_J^{\dagger}) I_J\| \right).
\end{aligned}$$

Also, by the same strategy used to prove (2.9a)–(2.9b), for given $J,\ M \in J(A) = J(A')$ we have that

$$\frac{|a'_J d'_J a_M d_M - a_J d_J a'_M d'_M|}{a'_J d'_J a_M d_M + a_J d_J a'_M d'_M} = \frac{|d'_{J/M} d_{M/J} a'_J a_M - d_{J/M} d'_{M/J} a_J a'_M|}{d'_{J/M} d_{M/J} a'_J a_M + d_{J/M} d'_{M/J} a_J a'_M} \le e_3. \tag{2.19}$$

We then complete the proof by substituting (2.16) and (2.19) into (2.18). ∎

Remark 2.3. From Theorems 2.1–2.3 we see that when $\|\delta A\| \to 0$ and $\Delta \to 0$,

$$\|(W^{'\frac{1}{2}}A')^{\dagger}W^{'\frac{1}{2}}(I-A(W^{\frac{1}{2}}A)^{\dagger}W^{\frac{1}{2}})\| \to 0$$

uniformly if and only if $\operatorname{rank}(A') = \operatorname{rank}(A) = r$ and any r rows of A are linearly independent. These conditions are also essential for the stability of weighted pseudo-inverses [26, 27], and in section 3 we will see that these conditions are also essential for upper perturbation bounds of solutions of the WLS and WLSE problems. Also notice that Theorems 2.1 and 2.2 are special cases of Theorem 2.3.

We now discuss the bound e_3. When r_m is small, $\Delta << 1$ and $\|\delta A\|/\rho << 1$,

$$(1+\Delta)^{r_m}\left(1+\frac{\|\delta A\|}{\rho}\right)^{2r} - 1 \le 1 - (1-\Delta)^{2r_m}\left(1-\frac{\|\delta A\|}{\rho}\right)^{4r}.$$

When r_m is large, Δ or $\|\delta A\|/\rho$ is relatively large, then

$$(1+\Delta)^{r_m}\left(1+\frac{\|\delta A\|}{\rho}\right)^{2r} - 1 \ge 1 - (1-\Delta)^{2r_m}\left(1-\frac{\|\delta A\|}{\rho}\right)^{4r},$$

and the error bound in (3.2) may be rough.

For the angle between weighted projections we have the following theorem.

THEOREM 2.4. *Suppose that A and $A' \in C_r^{m\times n}$ are such that the condition in* (2.10) *holds and that $\|\delta A\| < \rho$. Then for some constant $1 > \Delta > 0$ we have*

(2.20)

$$\sup_{W,\ W'\in\mathcal{D},\eta\le\Delta} \|A'(W^{'\frac{1}{2}}A')^{\dagger}W^{'\frac{1}{2}}(I-A(W^{\frac{1}{2}}A)^{\dagger}W^{\frac{1}{2}})\| \le \frac{2(\|A\|+\|\delta A\|)}{\rho - \|\delta A\|}\left(\frac{\|\delta A\|}{\rho} + e_3\right).$$

Proof. The estimate directly follows Theorem 2.3. ∎

3. Upper perturbation bounds of solutions of the WLS and WLSE problems when weight matrix ranges over $\mathcal{D}$. In this section, we will derive upper perturbation bounds of solutions of the WLS and WLSE problems when the weight matrices W, W' range over $\mathcal{D}$ by applying the results of previous section.

For $A(W)^\dagger$ (and so $A'(W')^\dagger$) defined in (1.4) or (1.7), we have the following decomposition by applying (1.10):

(3.1)
$$\begin{aligned} A'(W')^\dagger - A(W)^\dagger &= A'(W')^\dagger(I - A'A(W)^\dagger) - (I - A'(W')^\dagger A')A(W)^\dagger \\ &= -A'(W')^\dagger \delta A A(W)^\dagger + A'(W')^\dagger(I - AA(W)^\dagger) \\ &\quad - (I - A^{'\dagger}A')A^\dagger AA(W)^\dagger. \end{aligned}$$

We first consider the upper perturbation bounds of the WLS solutions. We have the following results.

THEOREM 3.1. *Suppose that A, $A' \in C_r^{m\times n}$, W, $W' = W + \delta W \in \mathcal{D}$ with $\|W^{-1}\delta W\| \le \Delta < 1$, b, $b' \in C^m$, such that the condition in* (2.10) *holds and that $\|\delta A\| < \rho$. Consider the following perturbed WLS problem:*

$$\min_{x\in C^n} \|W^{'\frac{1}{2}}(A'x - b')\|. \tag{1.2a}$$

Let x_{WLS} and x'_{WLS} be the minimum norm solutions of (1.2) *and* (1.2a)*, respectively. Then we have the following estimate:*

(3.2)
$$\begin{aligned} \|x'_{WLS} - x_{WLS}\| &\le \frac{1}{\rho - \|\delta A\|}\left\{\|\delta b\| + (1+\delta_{rn})\|\delta A\|\|x_{WLS}\| + 2\left(\frac{\|\delta A\|}{\rho} + e_3\right)\|r^{(0)}\|\right\} \\ &\le \frac{1}{\rho - \|\delta A\|}\left\{\|\delta b\| + (1+\delta_{rn})\|\delta A\|\|x^{(0)}\| + 2\left(\frac{\|\delta A\|}{\rho} + e_3\right)\|r^{(0)}\|\right\}, \end{aligned}$$

in which e_3 is defined in Theorem 2.3*, $\delta_{rn} = 0$ when $r = n$ and $\delta_{rn} = 1$ when $r < n$, $x^{(0)}$ and $r^{(0)}$ are, respectively, defined as*

$$\|x^{(0)}\| = \sup_{W\in\mathcal{D}} \|x_{WLS}\| \quad \text{and} \quad r^{(0)} = (I - AA^\dagger)b. \tag{3.3}$$

Furthermore, if $r < n$, then for any WLS solution x of (1.2)*,*

$$x = A(W)^\dagger b + (I - A^\dagger A)z \tag{3.4}$$

there exists a WLS solution x' of (1.2a)*, such that*

(3.5)
$$\begin{aligned} \|x' - x\| &\le \frac{1}{\rho - \|\delta A\|}\left\{\|\delta b\| + \|\delta A\|(\|x_{WLS}\| + \|(I - A^\dagger A)z\|) + 2\left(\frac{\|\delta A\|}{\rho} + e_3\right)\|r^{(0)}\|\right\} \\ &\le \frac{1}{\rho - \|\delta A\|}\left\{\|\delta b\| + \|\delta A\|(\|x^{(0)}\| + \|(I - A^\dagger A)z\|) + 2\left(\frac{\|\delta A\|}{\rho} + e_3\right)\|r^{(0)}\|\right\}. \end{aligned}$$

Proof. By using the relations in (1.4) and (3.1), for any $W,\ W' \in \mathcal{D}$ we have the following identity:

(3.6)
$$x'_{WLS} - x_{WLS} = A'(W')^\dagger b' - A(W)^\dagger b = A'(W')^\dagger \delta b + (A'(W')^\dagger - A(W)^\dagger)b$$
$$= A'(W')^\dagger(\delta b - \delta A x_{WLS}) + A'(W')^\dagger(I - AA(W)^\dagger)b - (I - A'^\dagger A')A^\dagger A x_{WLS}.$$

Notice that $(I - A'^\dagger A')A^\dagger A = 0$ when $r = n$, and $\|(I - A'^\dagger A')A^\dagger A\| \le \|\delta A\|\|A'^\dagger\| \le \|\delta A\|\rho'^{-1}$ when $r < n$. Then we obtain from (3.6) and (1.11)

$$\|x'_{WLS} - x_{WLS}\| \le \frac{1}{\rho'}\left(\|\delta b\| + (1+\delta_{rn})\|\delta A\|\|x_{WLS}\|\right) + \|A'(W')^\dagger(I - AA(W)^\dagger)\|\|r^{(0)}\|.$$

Then by applying Theorem 2.3 we obtain the desired estimate in (3.2).

When $r < n$, then for any solution x of (1.2) of the form (3.4) choose x' as

$$x' = A'(W')^\dagger b' + (I - A'^\dagger A')[x_{WLS} + (I - A^\dagger A)z]; \tag{3.4a}$$

then x' is a WLS solution of (1.2a) (see (1.9)) and from (3.1) we have the following identity:

$$x' - x = A'(W')^\dagger(\delta b - \delta A x_{WLS}) + A'(W')^\dagger(I - AA(W)^\dagger)r^{(0)} - A'^\dagger A'(I - A^\dagger A)z.$$

Then by applying Theorem 2.3 we obtain the desired estimate in (3.5). ∎

Remark 3.1. When the matrix A is of full column rank, Vavasis [22, Theorem 2.3] provides a perturbation analysis in the case $A' = A$. In fact Vavasis's estimate can be extended to a rank deficient case. Let $A' = A \in C_r^{m\times n}$. Then for any WLS solution x of (1.2) of the form in (3.4), let a WLS solution x' of (1.2a) be of the form in (3.4a). Because $\delta A = 0$ and $A'(W')^\dagger = A(W')^\dagger$, we have

$$x' - x = A(W')^\dagger \delta b + A(W')^\dagger(I - AA(W)^\dagger)r^{(0)} = A(W')^\dagger \delta b - A(W)^\dagger(I - AA(W')^\dagger)r^{(0)}.$$

Then by applying Theorem 2.1 we obtain the following estimate:

$$\|x' - x\| \le \frac{1}{\rho}\{\|\delta b\| + 2e_1\|r^{(0)}\|\},$$
$$e_1 = \min\{(1+\Delta)^{r_m} - 1,\ 1 - (1-\Delta)^{2r_m}\}. \tag{3.5a}$$

On the other hand, because $A^H W'(I - AA(W')^\dagger) = 0$, then

$$A(W)^\dagger(I - AA(W')^\dagger) = (A^H W A)^\dagger A^H W(I - AA(W')^\dagger)$$
$$= (A^H W A)^\dagger A^H W W^{-1}(W - W')(I - AA(W')^\dagger),$$

and so we obtain

$$\|x' - x\| \le \|A(W')^\dagger\|\|\delta b\| + \|A(W)^\dagger\|\|W^{-1}(W' - W)\|(1 + \|AA(W')^\dagger\|)\|b\|$$
$$\le \frac{1}{\rho}\left\{\|\delta b\| + \Delta\left(1 + \frac{1}{\bar\rho}\right)\|r^{(0)}\|\right\} \le \frac{1}{\rho}\left\{\|\delta b\| + \Delta\left(1 + \frac{1}{\bar\rho}\right)\|b\|\right\}, \tag{3.5b}$$

where $\frac{1}{\bar\rho} = \sup_{W\in\mathcal{D}} \|AA(W)^\dagger\| \ge 1$ [19, 16, 26]. Also in (3.6), by taking $A' = A$ we have,

$$x'_{WLS} - x_{WLS} = A(W')^\dagger \delta b + A(W')^\dagger(I - AA(W)^\dagger)r^{(0)}$$
$$= A(W')^\dagger \delta b - A(W)^\dagger(I - AA(W')^\dagger)r^{(0)},$$

from which we see that estimates in (3.5a) and (3.5b) also hold when we replace x by x_{WLS} and x' by x'_{WLS}. Notice that the bound (3.5b) is essentially the same as that of Vavasis in [22, Theorem 2.3].

Comparing (3.5a) with (3.5b), we see that when $\bar{\rho}$ is close to 1, r or $m-r$ is small, and then the error bounds (3.5a) and (3.5b) have the same order; when $\bar{\rho}$ is close to 1, both r and $m-r$ are large, then the bound in (3.5b) is better; when $\bar{\rho}$ is much large than 1, r or $m-r$ is small, then the bound in (3.5a) is better. In the following example, we provide the computational results.

EXAMPLE 3.1. *Let* $a \geq 1$, $\alpha = \sqrt{2+4a^2}$, *and* $0 < \Delta < 1$,

$$A = A' = \begin{pmatrix} 1 & 0 \\ a & a \\ 0 & 1 \end{pmatrix} = QR,\ Q = \begin{pmatrix} \alpha^{-1} & -\sqrt{2}/2 \\ 2a\alpha^{-1} & 0 \\ \alpha^{-1} & \sqrt{2}/2 \end{pmatrix},\ R = \frac{1}{2}\begin{pmatrix} \alpha & \alpha \\ \sqrt{2} & \sqrt{2} \end{pmatrix}$$

$$b = b' = \begin{pmatrix} 2 \\ 0.5 \\ 2 \end{pmatrix},\ W = \mathrm{diag}(1,1,0.2),\ W = \mathrm{diag}(1-\Delta, 1-\Delta, 0.2(1+\Delta)).$$

Then we have

$$\rho = \min_{J\in J(A)} \inf\nolimits_+(A_J) = \frac{a\sqrt{2}}{\sqrt{1+2a^2+\sqrt{1+4a^4}}},\ \bar{\rho} = \min_{J\in J(Q)} \inf\nolimits_+(Q_J) = \frac{1}{\sqrt{1+2a^2}}.$$

TABLE 3.1

$\Delta = 0.01$				
k	0	2	4	6
Rate a	0.243	0.198	0.198	0.198
Rate b	0.126	$2.73e-3$	$2.75e-5$	$2.75e-7$
$\Delta = 0.1$				
Rate a	0.255	0.210	0.210	0.210
Rate b	0.131	$2.91e-3$	$2.93e-5$	$2.93e-7$
$\Delta = 0.3$				
Rate a	0.280	0.246	0.246	0.246
Rate b	0.144	$3.39e-3$	$3.42e-5$	$3.42e-7$

In Table 3.1 we compare bounds (3.5a) and (3.5b) for this example. We take $a = 10^k$, $\Delta = 0.01,\ 0.1$, and 0.3, rate a is the ratio of the norms of true error and bound using (3.5a), and rate b is the ratio of the norms of true error and bound using (3.5b).

Because in this example, $r_m = 1$, and when $a \geq 1$ increases, $\bar{\rho}$ decreases and the bound in (3.5b) becomes worse, while the bound in (3.5a) is much better.

Remark 3.2. From Theorem 3.1 we see that if $\mathrm{rank}(A') = \mathrm{rank}(A) = r$ and any r rows of A are linearly independent; then when $\|\delta A\| \to 0$ and $\Delta \to 0$, $\|x_{WLS} - x'_{WLS}\| \to 0$ uniformly. As mentioned in Remark 2.2, when r is large, the error bound provided in Theorem 3.1 would be rough.

Now we provide one example to verify the condition in (2.10) and another one to verify the estimates in Theorem 3.1.

EXAMPLE 3.2. *Let* $\epsilon \neq 0$, $w_3 > 0$, *and*

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix},\ A' = \begin{pmatrix} 1 & 0 \\ 1 & \epsilon \\ 0 & 1 \end{pmatrix},\ b = b' = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix},\ W = W' = \mathrm{diag}(1,1,w_3).$$

Then $\operatorname{rank}(A) = \operatorname{rank}(A') = 2$, *and the condition* (2.10) *is violated. Now*

$$x_{WLS} = A(W)^\dagger b = \begin{pmatrix} 1/2 \\ 0 \end{pmatrix},\ x'_{WLS} = A'(W')^\dagger b' = \frac{1}{\epsilon^2 + 2w_3} \begin{pmatrix} \epsilon^2 + w_3 \\ -\epsilon \end{pmatrix},$$

and we have

$$\sup_{W,\ W' \in \mathcal{D}, \eta \le \Delta} \|x'_{WLS} - x_{WLS}\| \ge \lim_{w_3 \to 0} \frac{\sqrt{\epsilon^4 + 4\epsilon^2}}{2(\epsilon^2 + 2w_3)} \ge \frac{1}{|\epsilon|}.$$

When $\epsilon \to 0$, *the above quantity becomes unbounded. This example shows the importance of the condition* (2.10).

EXAMPLE 3.3. *Let* $0 < \Delta < 1$, $\epsilon > 0$,

$$A = \begin{pmatrix} 0.5 & 1 \\ 0.5\sqrt{3} & 0 \\ 0.5 & -1 \end{pmatrix},\ A' = A + \epsilon \begin{pmatrix} -1 & 1 \\ -1 & -1 \\ -1 & 1 \end{pmatrix},\ b = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix},$$
$$W = \operatorname{diag}(1, 1, 0.2),\ W' = \operatorname{diag}(1 - \Delta, 1, 0.2(1 + \Delta)).$$

Then we have

$$\rho = \min_{J \in J(A)} \inf\nolimits_+(A_J) = \frac{\sqrt{2}}{2}, \|\delta A\| = 2\epsilon.$$

TABLE 3.2
Ratios of $\|\delta x\|$/*bound* (3.2).

$\Delta \backslash \epsilon$	0.0	0.01	0.1	0.3
0.01	0.180	0.0977	0.0887	0.0359
0.1	0.184	0.110	0.0913	0.0360
0.3	0.198	0.130	0.0973	0.0363
0.6	0.245	0.171	0.110	0.0365

In Table 3.2 we verify the bound (3.2) for this example. We take $\Delta = 0.01,\ 0.1,$ 0.3, and 0.6, $\epsilon = 0.0,\ 0.01,\ 0.1,$ and 0.3. We list the ratios of the norms of true error and the bound derived in (3.2).

From Table 3.2 we see that when ϵ is small, the bound (3.2) is good; when ϵ becomes larger, the bound (3.2) becomes worse. Observe that in this example, ratios are relatively insensitive to the change of Δ.

We now study the upper perturbation bounds of the WLSE solutions. We first establish some relations between several suprema of weighted pseudoinverses.

LEMMA 3.1. *Suppose* $L \in C_p^{m_1 \times n}$, $K, \in C^{m_2 \times n}$, $A = \binom{L}{K} \in C_r^{m \times n}$, *and* $A(W) = \binom{L}{WKP(WKP)^\dagger K}$. *Then we have*

$$\operatorname{rank} \begin{pmatrix} L \\ K_J \end{pmatrix} = r \quad \textit{and } \operatorname{rank}(K_J P) = r - p \quad \textit{for any } J \in J(KP), \tag{3.7}$$

$$\begin{aligned} \rho_1 &\equiv \min_{J \in J(KP)} \inf\nolimits_+ \begin{pmatrix} L \\ K_J \end{pmatrix} = \left(\sup_{W \in \mathcal{D}_2} \|A(W)^\dagger\| \right)^{-1}, \\ \rho_2 &\equiv \min_{J \in J(KP)} \inf\nolimits_+ (K_J P) = \left(\sup_{W \in \mathcal{D}_2} \|(W^{\frac{1}{2}} KP)^\dagger W^{\frac{1}{2}}\| \right)^{-1}, \end{aligned} \tag{3.8}$$

in which D_2 is the set of positive diagonal matrices of order m_2, and

$$\rho \le \rho_1 \le \rho_2 \tag{3.9}$$

with $\rho \equiv \min_{K\in J(A)} \inf_+(A_K)$.

Proof. The rank relations in (3.7) are proved in [27]. The identity in (3.8) for ρ_1 is shown in [8, 27], and the identity in (3.8) for ρ_2 is shown in [26], as that in (2.11). That $\rho \le \rho_1$ is obvious because there are more restrictions on ρ_1 than ρ.

Now we prove that $\rho_1 \le \rho_2$. Let a unitary decomposition of L be

$$L = (G, 0)V^H = GV_1^H \text{ with } V^HV = I_n,\ V_1^HV_1 = I_p,\ V = (V_1, V_2),$$

and G has full column rank p. Then for any fixed $J \in J(KP)$ we have

$$\begin{pmatrix} L \\ K_J \end{pmatrix} = \begin{pmatrix} GV_1^H \\ K_J(V_1V_1^H + V_2V_2^H) \end{pmatrix} = \begin{pmatrix} G & 0 \\ K_JV_1 & K_JV_2 \end{pmatrix} V^H.$$

By using interlacing relation between the singular values of a matrix and its submatrix [11] we obtain

$$\begin{aligned}\sigma_{r-p}\begin{pmatrix} L \\ K_J \end{pmatrix} &= \sigma_{r-p}\begin{pmatrix} G & 0 \\ K_JV_1 & K_JV_2 \end{pmatrix} \ge \sigma_{r-p}\begin{pmatrix} 0 \\ K_JV_2 \end{pmatrix} \ge \sigma_r\begin{pmatrix} G & 0 \\ K_JV_1 & K_JV_2 \end{pmatrix} \\ &= \sigma_r\begin{pmatrix} L \\ K_J \end{pmatrix}.\end{aligned}$$

We then have

$$\rho_1 \le \sigma_r\begin{pmatrix} L \\ K_J \end{pmatrix} \le \sigma_{r-p}(K_JV_2) = \sigma_{r-p}(K_JP) = \sigma_{r-p}((KP)_J).$$

Because the above inequalities hold for any $J \in J(KP)$, we assert $\rho_1 \le \rho_2$. ∎

LEMMA 3.2 (see [27, Theorem 4.2]). *Under the conditions in Lemma* 3.1*, let $L' \in C_p^{m_1\times n}$, $K' \in C^{m_2\times n}$, and $A' = A + \delta A = \binom{L'}{K'} \in C_r^{m\times n}$ be, respectively, the perturbed versions of L, K, and A. Then $\rho_1' > 0$ for all sufficiently small perturbations δA if and only if*

$$\text{any } r-p \text{ rows of } KP \text{ are linearly independent.} \tag{3.10}$$

If the condition in (3.10) *holds, then for $\|\delta A\| < \rho_1$, we have the following estimate:*

$$\rho_1 - \|\delta A\| \le \rho_1' \le \rho_1 + \|\delta A\|. \tag{3.11}$$ ∎

Now we can establish the upper perturbation bounds of the WLSE problem.

THEOREM 3.2. *Suppose that $L,\ L' \in C_p^{m_1\times n}$, $K,\ K' \in C^{m_2\times n}$, $A = \binom{L}{K}$, $A' = A + \delta A = \binom{L'}{K'} \in C_r^{m\times n}$, $W,\ W' = W + \delta W \in D_2$ with $\|W^{-1}\delta W\| \le \Delta < 1$, $h,\ h' \in C^{m_1}$, $g,\ g' \in C^{m_2}$, $b = \binom{h}{g}$, and $b' = \binom{h'}{g'}$ such that $h \in R(L)$. Suppose that the condition in* (3.10) *holds, $\|\delta A\| < \rho_1$, and $\|\delta K\| + \|\delta L\|\|L^\dagger\|\|K\| < \rho_2$. Consider the following perturbed WLSE problem:*

$$\min_{x\in C^n} \|W'^{\frac{1}{2}}(K'x - g')\| \quad \text{subject to } \min_{x\in C^n} \|L'x - h'\|. \tag{1.5a}$$

Let x_{WLSE} and x'_{WLSE} be the minimum norm solutions of (1.5) *and* (1.5a)*, respectively. Then we have the following estimate:*

$$\|x'_{WLSE} - x_{WLSE}\| \le \frac{1}{\rho_1 - \|\delta A\|}\Bigg\{ (\|\delta b\| + (1+\delta_{rn})\|\delta A\|\|x_{WLSE}\|) + 2\left(\frac{\|\delta K\| + \|\delta L\|\|L^\dagger\|\|K\|}{\rho_2} + e_4\right)\|r_2^{(0)}\|\Bigg\},$$

$$e_4 = \min\Bigg\{(1+\Delta)^{r_{m2}}\left(1 + \frac{\|\delta K\| + \|\delta L\|\|L^\dagger\|\|K\|}{\rho_2}\right)^{2(r-p)} - 1,$$

$$1 - (1-\Delta)^{2r_{m2}}\left(1 - \frac{\|\delta K\| + \|\delta L\|\|L^\dagger\|\|K\|}{\rho_2}\right)^{4(r-p)}\Bigg\}, \tag{3.12}$$

where

$$r_{m2} = \min\{r-p, m_2 - r + p\} \text{ and } r_2^{(0)} = (I - KP(KP)^\dagger)(g - KL^\dagger h). \tag{3.13}$$

Furthermore, when $r < n$, for any WLSE solution x of (1.5) *that has the form*

$$x = A(W)^\dagger b + (I - A^\dagger A)z, \tag{3.14}$$

there exists a WLSE solution x' of (1.5a)*, such that*

$$\|x' - x\| \le \frac{1}{\rho_1 - \|\delta A\|}\Bigg\{ \big(\|\delta b\| + \|\delta A\|(\|x_{WLSE}\| + \|(I - A^\dagger A)z\|)\big) + 2\left(\frac{\|\delta K\| + \|\delta L\|\|L^\dagger\|\|K\|}{\rho_2} + e_4\right)\|r_2^{(0)}\|\Bigg\}. \tag{3.15}$$

Proof. By using the relations in (1.7) and (3.1), we have the following identity:

$$\begin{aligned} x'_{WLSE} - x_{WLSE} &= A'(W')^\dagger b' - A(W)^\dagger b = A'(W')^\dagger \delta b + (A'(W')^\dagger - A(W)^\dagger)b \\ &= A'(W')^\dagger(\delta b - \delta A x_{WLSE}) + A'(W')^\dagger(I - AA(W)^\dagger)b \\ &\quad - (I - A'^\dagger A')A^\dagger A x_{WLSE}. \end{aligned}$$

Taking norms in the above identity,

$$\|x'_{WLSE} - x_{WLSE}\| \le \frac{1}{\rho'_1}(\|\delta b\| + (1+\delta_{rn})\|\delta A\|\|x_{WLSE}\|) + \|A'(W')^\dagger(I - AA(W)^\dagger)b\|. \tag{3.16}$$

Now from the forms of $A(W)$ and $A(W)^\dagger$ in (1.7) and (1.8) we have

$$g - Kx_{WLSE} = (I - KP(W^{\frac{1}{2}}KP)^\dagger W^{\frac{1}{2}})(g - KL^\dagger h) = (I - KP(W^{\frac{1}{2}}KP)^\dagger W^{\frac{1}{2}})r_2^{(0)}$$

by applying (1.11). Taking into account that $h - Lx_{WLSE} = 0$ we have that

$$\begin{aligned} A'(W')^\dagger(I - AA(W)^\dagger)b &= A'(W')^\dagger(b - Ax_{WLSE}) = A'(W')^\dagger \begin{pmatrix} 0 \\ g - Kx_{WLSE} \end{pmatrix} \\ &= (W'^{\frac{1}{2}}K'P')^\dagger W'^{\frac{1}{2}}(I - KP(W^{\frac{1}{2}}KP)^\dagger W^{\frac{1}{2}})r_2^{(0)}. \end{aligned} \tag{3.17}$$

Using (3.9) and (3.11), we also have the following estimates:

$$\begin{aligned} &\rho_2' \geq \rho_1' \geq \rho_1 - \|\delta A\|, \\ &\|K'P' - KP\| \leq \|\delta K\| + \|P' - P\|\|K\| \leq \|\delta K\| + \|\delta L\|\|L^\dagger\|\|K\|. \end{aligned} \tag{3.18}$$

By applying Theorem 2.3 with KP replacing A and using the inequalities in Lemma 3.2, we obtain the desired estimate in (3.12).

When $r < n$, for any solution x of (1.5) in the form of (3.14), choose x' as

$$x' = A'(W')^\dagger b' + (I - A'^\dagger A')[x_{WLSE} + (I - A^\dagger A)z];$$

then x' is a WLSE solution of (1.5a), and from (3.1) we have the following identity:

$$x' - x = A'(W')^\dagger(\delta b - \delta A x_{WLSE}) + A'(W')^\dagger(I - AA(W)^\dagger)b - A'^\dagger A'(I - A^\dagger A)z.$$

Now applying (3.17) and (3.18) we obtain the desired estimate in (3.15). ∎

Remark 3.3. From Theorem 3.2 we see that if $\operatorname{rank}(L') = \operatorname{rank}(L) = p$, $\operatorname{rank}(A') = \operatorname{rank}(A) = r$, and any $r-p$ rows of KP are linearly independent; then when $\|\delta A\| \to 0$ and $\Delta \to 0$, $\|x_{WLSE} - x'_{WLSE}\| \to 0$ uniformly.

4. Upper perturbation bound of the weighted projections, WLS and WLSE solutions when W ranges over $\mathcal{P}$. In the case that $\mathcal{P}$ is a set of $m \times m$ real symmetric diagonally dominant positive semidefinite matrices, Forsgren [8] had constructed an interesting signature decomposition of $W = U\widehat{W}U^T \in \mathcal{P}$, where $U \equiv U(s)$ has the following form, with $l = \frac{m(m+1)}{2}$:

$$\begin{aligned} &U(s) \equiv (u(s)_1, \ldots, u(s)_l) \in R^{m\times l} \ \text{ with} \\ &u(s)_i = e_i \quad \text{for } i = 1, \ldots, m, \\ &u(s)_{m+t(i,j)} = e_i + s_{t(i,j)} e_j \quad \text{for } 1 \leq i < j \leq m, \end{aligned} \tag{4.1}$$

in which e_j is the jth column of the identity matrix I_m, and for $1 \leq i < j \leq m$,

$$t(i,j) = m(i-1) - \frac{i(i+1)}{2} + j \quad \text{and} \quad s_{t(i,j)} = \begin{cases} 1 & \text{if } w_{ij} \geq 0, \\ -1 & \text{if } w_{ij} < 0 \end{cases}$$

and

$$\begin{aligned} &\widehat{W} = \operatorname{diag}(\widehat{w}_1, \ldots, \widehat{w}_l) \text{ with } \widehat{w}_i = w_{ii} - \sum_{j=1,\ j\neq i}^{m} |w_{ij}| \text{ for } i = 1, \ldots m, \\ &\widehat{w}_{m+t(i,j)} = |w_{ij}| \text{ for } 1 \leq i < j \leq m. \end{aligned} \tag{4.2}$$

For a detailed description of the signature decomposition of W, we refer to [8].

In this section, we will derive upper perturbation bounds of weighted projections, WLS and WLSE solutions when weighted matrix W ranges over a set $\mathcal{P} \subset \mathcal{P}(A)$ which has the form

$$\begin{aligned} \mathcal{P} = \{W = U\widehat{W}U^T : \ &U \in \mathcal{U},\ \widehat{W} = \operatorname{diag}(\widehat{w}_1, \ldots, \widehat{w}_l) \geq 0 \\ &\text{such that } \operatorname{rank}(WA) = \operatorname{rank}(A)\} \end{aligned} \tag{4.3}$$

in which the given set $\mathcal{U} \subset R^{m\times l}$ satisfies

$$\operatorname{rank}(U^T A) = r \quad \text{for all } U \in \mathcal{U}, \quad \sup_{U\in\mathcal{U}} \max_{J\in J(U^T A)} \|(U_J^T A)^\dagger U_J^T\| < +\infty, \tag{4.4}$$

where $U_J^T \equiv (U^T)_J$. Then we have [8, 27]

(4.5)
$$\widehat{\rho} \equiv \inf_{U\in\mathcal{U}} \min_{J\in J(U^TA)} \inf\nolimits_+((U_J^T)^\dagger U_J^T A) = \frac{1}{\sup_{W\in\mathcal{P}} \|(W^{\frac12}A)^\dagger W^{\frac12}\|} = \frac{1}{\sup_{U\in\mathcal{U}} \max_{J\in J(\widehat{A})} \|(U_J^TA)^\dagger U^T\|}.$$

When discussing upper perturbation bounds, we will consider only the case $W' = W$. If $W' \neq W$, calculations become extremely complicated.

LEMMA 4.1. *Let $A,\ A' \in C_r^{m\times n}$, and $W = U\widehat{W}U^T \in \mathcal{P}$ be given. Then*

(4.6)
$$(W^{\frac12}A')^\dagger W^{\frac12}(I - A(W^{\frac12}A)^\dagger W^{\frac12}) = \frac{\sum_{J\in J(\widehat{A}')}\sum_{M\in J(\widehat{A})} \widehat{a}'_J\widehat{d}_J\widehat{a}_M\widehat{d}_M\widehat{A}_J^{\prime\dagger}(U_J^T - \widehat{A}_J\widehat{A}_M^\dagger U_M^T)}{\sum_{K\in J(\widehat{A}')}\sum_{N\in J(\widehat{A})} \widehat{a}'_K\widehat{d}_K\widehat{a}_N\widehat{d}_N},$$

in which

(4.7)
$$\widehat{A} = U^TA,\ \widehat{A}' = U^TA',\ \widehat{a}_J = \det(\widehat{A}_J(\widehat{A}_J)^H), \quad \widehat{a}'_J = \det(\widehat{A}'_J(\widehat{A}'_J)^H),\ \textit{and}\ \widehat{d}_J = \det(\widehat{W}_{JJ}).$$

Proof. It is very easy to show that

$$(W^{\frac12}A')^\dagger W^{\frac12}(I - A(W^{\frac12}A)^\dagger W^{\frac12}) = (\widehat{W}^{\frac12}\widehat{A}')^\dagger\widehat{W}^{\frac12}(I - \widehat{A}(\widehat{W}^{\frac12}\widehat{A})^\dagger\widehat{W}^{\frac12})U^T.$$

Applying Lemma 2.2 we obtain the desired result. ∎

THEOREM 4.1. *Suppose $A \in C_r^{m\times n}$ with $0 < r$. Then*

$$\sup_{W\in\mathcal{P}} \|(W^{\frac12}A')^\dagger W^{\frac12}(I - A(W^{\frac12}A)^\dagger W^{\frac12})\| < +\infty$$

for any $A' = A + \delta A \in C_r^{m\times n}$ with sufficiently small perturbation δA if and only if for any $U \in \mathcal{U}$ and any vector of sequence of increasing r integers $J = \{i_1, \ldots, i_r\} \subset \{1, \ldots, m\}$

(4.8)
$$\operatorname{rank}(U_J^T) = r \ \textit{implies}\ \operatorname{rank}(U_J^TA) = r.$$

If the condition in (4.8) *holds, then for any $A' = A + \delta A \in C_r^{m\times n}$ with $\|\delta A\| < \widehat{\rho}$ and $U \in \mathcal{U}$, $J(\widehat{A}') = J(\widehat{A})$ we have the following estimate:*

(4.9)
$$\sup_{W\in\mathcal{P}} \|(W^{\frac12}A')^\dagger W^{\frac12}(I - A(W^{\frac12}A)^\dagger W^{\frac12})\| \le \frac{2}{\widehat{\rho} - \|\delta A\|}\left(\frac{\|\delta A\|}{\widehat{\rho}} + e_5\right),$$
$$e_5 = \min\left\{\left(1 + \frac{\|\delta A\|}{\widehat{\rho}}\right)^{2r} - 1,\ 1 - \left(1 - \frac{\|\delta A\|}{\widehat{\rho}}\right)^{4r}\right\}.$$

Proof. Using Lemma 4.1 it is easy to see that the proof follows exactly the same steps as those in Theorem 2.2, except that we use the following inequalities:

$$\left(1-\frac{\|\delta A\|}{\widehat{\rho}}\right)^{2r} \le \frac{\widehat{a}'_M}{\widehat{a}_M} \le \left(1+\frac{\|\delta A\|}{\widehat{\rho}}\right)^{2r},\ \left(1-\frac{\|\delta A\|}{\widehat{\rho}}\right)^{2r} \le \frac{\widehat{a}_J}{\widehat{a}'_J}. \tag{4.10}$$

For a square matrix C, denote $\lambda_j(C)$ a nonzero eigenvalue of C for $j=1,\dots,\operatorname{rank}(C)$. Note that for $j=1,\dots,r$ we have that

$$\lambda_j(\widehat{A}_M\widehat{A}_M^H\cdot(U_M^T)^{\dagger T}(U_M^T)^\dagger)=\lambda_j((U_M^T)^\dagger\widehat{A}_M\widehat{A}_M^H\cdot(U_M^T)^{\dagger T})=\sigma_j((U_M^T)^\dagger\widehat{A}_M)^2,$$

and so from (4.5) we have that

$$\frac{\widehat{a}'_M}{\widehat{a}_M}=\prod\nolimits_{1\le k\le r}\left(\frac{\sigma_k((U_M^T)^\dagger\widehat{A}'_M)}{\sigma_k((U_M^T)^\dagger\widehat{A}_M)}\right)^2 \le \prod\nolimits_{1\le k\le r}\left(1+\frac{\|(U_M^T)^\dagger\delta\widehat{A}\|}{\sigma_k((U_M^T)^\dagger\widehat{A}_M)}\right)^2 \le \left(1+\frac{\|\delta A\|}{\widehat{\rho}}\right)^{2r},$$

$$\frac{\widehat{a}'_M}{\widehat{a}_M} \ge \prod\nolimits_{1\le k\le r,\ \sigma_k((U_M^T)^\dagger\widehat{A}'_M)<\sigma_k((U_M^T)^\dagger\widehat{A}_M)}\left(1-\frac{\|(U_M^T)^\dagger\delta\widehat{A}\|}{\sigma_k((U_M^T)^\dagger\widehat{A}_M)}\right)^2 \ge \left(1-\frac{\|\delta A\|}{\widehat{\rho}}\right)^{2r},$$

$$\frac{\widehat{a}_J}{\widehat{a}'_J} \ge \prod\nolimits_{1\le k\le r,\ \sigma_k((U_J^T)^\dagger\widehat{A}_J)<\sigma_k((U_J^T)^\dagger\widehat{A}'_J)}\left(1-\frac{\|(U_J^T)^\dagger\delta\widehat{A}\|}{\sigma_k((U_J^T)^\dagger\widehat{A}'_J)}\right)^2 \ge \left(1-\frac{\|\delta A\|}{\widehat{\rho}}\right)^{2r}. \quad \square$$

For the upper perturbation bound of the WLS problem when the weight matrix W ranges over $\mathcal{P}$, we have the following theorem.

THEOREM 4.2. *Suppose that $A,\ A'\in C_r^{m\times n}$, $b,\ b'\in C^m$, $W\in\mathcal{P}$, such that the condition in* (4.8) *holds and that $\|\delta A\|<\widehat{\rho}$. Consider the following perturbed WLS problem:*

$$\min_{x\in C^n}\|W^{\frac{1}{2}}(A'x-b')\|. \tag{1.2b}$$

Let x_{WLS} and x'_{WLS} be the minimum norm solutions of (1.2) *and* (1.2b)*, respectively. Then we have the following estimate:*

$$\|x'_{WLS}-x_{WLS}\| \le \frac{1}{\widehat{\rho}-\|\delta A\|}\left\{\Big(\|\delta b\|+(1+\delta_{rn})\|\delta A\|\|\widehat{x}^{(0)}\|\Big)+2\left(\frac{\|\delta A\|}{\widehat{\rho}}+e_5\right)\|r^{(0)}\|\right\}, \tag{4.11}$$

in which $\widehat{x}^{(0)}$ and $r^{(0)}$ are, respectively, defined by

$$\|\widehat{x}^{(0)}\|=\sup_{W\in\mathcal{P}}\|x_{WLS}\| \text{ and } r^{(0)}=(I-AA^\dagger)b. \tag{4.12}$$

Furthermore, when $r<n$, for any WLS solution x of (1.2) *of the form* (3.4)*, there exists a WLS solution x' of* (1.2b)*, such that*

$$\begin{aligned}\sup_{W\in\mathcal{P}}\|x'-x\| \le \frac{1}{\widehat{\rho}-\|\delta A\|}\Bigg\{&\Big(\|\delta b\|+\|\delta A\|(\|\widehat{x}^{(0)}\|+\|(I-A^\dagger A)z\|)\Big)\\&+2\Big(\frac{\|\delta A\|}{\widehat{\rho}}+e_5\Big)\|r^{(0)}\|\Bigg\}.\end{aligned} \tag{4.13}$$

Proof. The proof is exactly the same as in Theorem 3.1, except that we estimate $\|A'(W)^\dagger(I - AA(W)^\dagger)\|$ using (4.9) instead of (2.11). ∎

To derive an upper perturbation bound of the WLSE problem when the weight matrix W ranges over $\mathcal{P}$, we need the following results.

LEMMA 4.2. *Suppose* $L \in C_p^{m_1\times n}$, $K, \in C^{m_2\times n}$, $A = \binom{L}{K} \in C_r^{m\times n}$, *and a set* $\mathcal{P}_2 \subset \mathcal{P}(KP)$ *has the following form:*

$$\mathcal{P}_2 = \{W = U\widehat{W}U^T : \ U \in \mathcal{U},\ \widehat{W} = \operatorname{diag}(\widehat{w}_1, \dots, \widehat{w}_l) \geq 0 \ \text{such that } \operatorname{rank}(WKP) = \operatorname{rank}(KP)\}, \tag{4.3a}$$

in which the given set $\mathcal{U} \subset R^{m_2\times l}$ *satisfies*

$$\operatorname{rank}(U^TKP) = r - p \quad \text{for all } U \in \mathcal{U}, \quad \sup_{U\in\mathcal{U}} \max_{J\in J(U^TKP)} \|(U_J^TKP)^\dagger U_J^T\| < +\infty. \tag{4.4a}$$

Then we have

$$\begin{aligned} \widehat{\rho}_1 &\equiv \inf_{U\in\mathcal{U}} \min_{J\in J(U^TKP)} \inf\nolimits_+ \begin{pmatrix} L \\ U_J^{\dagger T}U_J^TK \end{pmatrix} = \left(\sup_{W\in\mathcal{P}_2} \|A(W)^\dagger\|\right)^{-1}, \\ \widehat{\rho}_2 &\equiv \inf_{U\in\mathcal{U}} \min_{J\in J(U^TKP)} \inf\nolimits_+ (U_J^{\dagger T}U_J^TK_JP) = \left(\sup_{W\in\mathcal{P}_2} \|(W^{\frac12}KP)^\dagger W^{\frac12}\|\right)^{-1}, \end{aligned} \tag{4.14}$$

and

$$\widehat{\rho}_1 \leq \widehat{\rho}_2. \tag{4.15}$$

Proof. The equalities in (4.14) are proved in [8, 27] and the inequality in (4.15) can be proved in exactly the same way as the proof of (3.9). ∎

LEMMA 4.3 (see [27]). *Suppose that* $L,\ L' \in C_p^{m_1\times n}$, $K,\ K' \in C^{m_2\times n}$, $A = \binom{L}{K}$, $A' = \binom{L'}{K'} \in C_r^{m\times n}$. *Then* $\widehat{\rho}_1' > 0$ *for all sufficiently small perturbations* δA *if and only if for any* $U \in \mathcal{U}$ *and any vector of* $r - p$ *increasing integers* $J = \{i_1, \dots, i_{r-p}\} \subset \{1, \dots, m_2\}$,

$$\operatorname{rank}(U_J^T) = r - p \text{ implies } \operatorname{rank}(U_J^TKP) = r - p. \tag{4.16}$$

If the condition (4.16) *holds, then for* $\|\delta A\| < \widehat{\rho}_1$ *we have*

$$\widehat{\rho}_1 - \|\delta A\| \leq \widehat{\rho}_1' \leq \widehat{\rho}_1 + \|\delta A\|. \tag{4.17}$$

∎

Now we have the following theorem.

THEOREM 4.3. *Suppose that* $L,\ L' \in C_p^{m_1\times n}$, $K,\ K' \in C^{m_2\times n}$, $A = \binom{L}{K}$, $A' = A + \delta A = \binom{L'}{K'} \in C_r^{m\times n}$, $h,\ h' \in C^{m_1}$, $g,\ g' \in C^{m_2}$, $b = \binom{h}{g}$, $b' = \binom{h'}{g'}$, *and* $h \in R(L)$, *such that the condition in* (4.16) *holds,* $\|\delta A\| < \widehat{\rho}_1$, *and* $\|\delta K\| + \|\delta L\|\|L^\dagger\|\|K\| < \widehat{\rho}_2$. *Consider the following perturbed WLSE problem:*

$$\min_{x\in C^n} \|W^{\frac12}(K'x - g')\| \quad \text{subject to } \min_{x\in C^n} \|L'x - h'\|. \tag{1.5b}$$

Let x_{WLSE} *and* x'_{WLSE} *be the minimum norm solutions of* (1.5) *and* (1.5b),

respectively. Then we have the following estimate:

$$\|x'_{WLSE} - x_{WLSE}\| \le \frac{1}{\widehat{\rho}_1 - \|\delta A\|}\left\{\left(\|\delta b\| + (1+\delta_{rn})\|\delta A\|\|\widehat{x}_2^{(0)}\|\right)\right.$$
$$\left. +2\left(\frac{\|\delta K\|+\|\delta L\|\|L^\dagger\|\|K\|}{\widehat{\rho}_2} + e_6\right)\|r_2^{(0)}\|\right\},$$
$$e_6 = \min\left\{\left(1+\frac{\|\delta K\|+\|\delta L\|\|L^\dagger\|\|K\|}{\widehat{\rho}_2}\right)^{2(r-p)} - 1,\right.$$
$$\left. 1-\left(1-\frac{\|\delta K\|+\|\delta L\|\|L^\dagger\|\|K\|}{\widehat{\rho}_2}\right)^{4(r-p)}\right\}, \tag{4.18}$$

in which $\widehat{x}_2^{(0)}$ *and* $r_2^{(0)}$ *are, respectively, defined by*

$$\|\widehat{x}_2^{(0)}\| = \sup_{W\in\mathcal{P}_2}\|x_{WLSE}\| \text{ and } r_2^{(0)} = (I-KP(KP)^\dagger)(g-KL^\dagger h). \tag{4.19}$$

Furthermore, when $r < n$*, then for any WLSE solution* x *of* (1.5) *in the form of* (3.14) *there exists a WLSE solution* x' *of* (1.5b)*, such that*

$$\|x'-x\| \le \frac{1}{\widehat{\rho}_1 - \|\delta A\|}\left\{\left(\|\delta b\| + \|\delta A\|(\|\widehat{x}_2^{(0)}\| + \|(I-A^\dagger A)z\|)\right)\right.$$
$$\left. + 2\left(\frac{\|\delta K\| + \|\delta L\|\|L^\dagger\|\|K\|}{\widehat{\rho}_2} + (1+e_6)\right)\|r_2^{(0)}\|\right\}. \tag{4.20}$$

Proof. The proof is exactly the same as in Theorem 3.3, except that we estimate $\|A'(W)^\dagger(I-AA(W)^\dagger)\|$ using (4.9) instead of (2.11). ∎

5. Concluding remarks. We derived upper perturbation bounds of weighted projections associated with the WLS and WLSE problems when W ranged over the set $\mathcal{D}$. We then applied these bounds to deduce upper perturbation bounds of solutions of WLS and WLSE problems when W ranged over $\mathcal{D}$. It was shown that when $\|\delta A\| \to 0$ and $\Delta \to 0$, then $\|x_{WLS} - x'_{WLS}\| \to 0$ uniformly if and only if $\operatorname{rank}(A') = \operatorname{rank}(A) = r$ and any r rows of A are linearly independent for the WLS problem. When $\|\delta A\| \to 0$ and $\Delta \to 0$, $\|x_{WLSE} - x'_{WLSE}\| \to 0$ uniformly if and only if $\operatorname{rank}(L') = \operatorname{rank}(L) = p$, $\operatorname{rank}(A') = \operatorname{rank}(A) = r$, and any $r-p$ rows of KP are linearly independent for the WLSE problem. We also extended the estimates to the case that W ranges over a subset of real symmetric positive semidefinite matrices.

Acknowledgments. The first author is grateful to L. Eldén for drawing his attention to a famous paper of Stewart [19] and to A. Forsgren for sending research material to him. Useful comments were suggested by two anonymous referees.

REFERENCES

[1] J. L. Barlow, *Error analysis and implementation aspects of deferred correction for equality constrained least squares problems*, SIAM J. Numer. Anal., 25 (1988), pp. 1340–1358.

[2] A. Ben-Israel and T. N. E. Greville, *Generalized Inverses: Theory and Applications*, John Wiley, New York, 1974.

[3] A. Ben-Tal and M. Teboulle, *A geometric property of the least squares solution of linear equations*, Linear Algebra Appl., 139 (1990), pp. 165–170,

[4] A. R. De Pierro and M. Wei, *Some new properties of the constrained and weighted least squares problem*, Linear Algebra Appl., to appear.
[5] I. I. Dikin, *On the speed of an iterative process*, Upravlyaemye Sistemi, 12 (1974), pp. 54–60.
[6] L. Eldén, *Perturbation theory for the least squares problem with linear equality constraints*, SIAM J. Numer. Anal., 17 (1980), pp. 338–350.
[7] L. Eldén, *A weighted pseudoinverse, generalized singular values, and constrained least square problems*, BIT, 22 (1982), pp. 487–502.
[8] A. Forsgren, *On linear least-squares problems with diagonally dominant weight matrices*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 763–788.
[9] A. Forsgren, P. E. Gill, and J. R. Shinnerl, *Stability of symmetric ill-conditioned systems arising in interior methods for constrained optimization*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 187–211.
[10] P. E. Gill, W. Murray, and M. H. Wright, *Numerical Linear Algebra and Optimization*, Vol. 1, Addison–Wesley, Reading, MA, 1991.
[11] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.
[12] C. C. Gonzaga, *Path-following methods for linear programming*, SIAM Rev., 34 (1992), pp. 167–224.
[13] M. Gullikson and P.-Å. Wedin, *Modifying the QR decomposition to constrained and weighted linear least squares*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1298–1313.
[14] M. Hanke and M. Neumann, *The geometry of the set of scaled projections*, Linear Algebra Appl., 190 (1993), pp. 137–148.
[15] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.
[16] D. P. O'Leary, *On bounds for scaled projections and pseudoinverses*, Linear Algebra Appl., 132 (1990), pp. 115–117.
[17] J. Miao and A. Ben-Israel, *The geometry of basic, approximate, and minimum-norm solution of linear equations*, Linear Algebra Appl., 216 (1995), pp. 25–41.
[18] G. W. Stewart, *On the perturbation of pseudo-inverses, projections and linear squares problems*, SIAM Rev., 19 (1977), pp. 634–662.
[19] G. W. Stewart, *On scaled projections and pseudoinverses*, Linear Algebra Appl., 112 (1989), pp. 189–193.
[20] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.
[21] R. J. Vanderbei and J. C. Lagarias, *Dikin's convergence result for the affine-scaling algorithm*, Contemp. Math., 114 (1990), pp. 109–119.
[22] S. A. Vavasis, *Stable numerical algorithms for equilibrium system*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1108–1131.
[23] P.-Å. Wedin, *Perturbation theory for pseudoinverses*, BIT, 13 (1973), pp. 217–232.
[24] P.-Å. Wedin, *Perturbation Theory and Condition Numbers for Generalized and Constrained Linear Least Squares Problems*, Tech. Report UMINF-125.85, Institute of Information Processing, Univ. Umeä, Sweden, 1985.
[25] M. Wei, *Perturbation theory for the rank-deficient equality constrained least squares problem*, SIAM J. Numer. Anal., 29 (1992), pp. 1462–1481.
[26] M. Wei, *Upper bound and stability of scaled pseudoinverses*, Numer. Math., 72 (1995), pp. 285–293.
[27] M. Wei, *Equivalent formulae for the supremum and stability of weighted pseudoinverses*, Math. Comp., 66 (1997), pp. 1487–1508.
[28] M. H. Wright, *Interior methods for constrained optimization*, in Acta Numerica, A. Iserles, ed., Cambridge University Press, Cambridge, UK, 1992, pp. 341–407.
[29] S. J. Wright, *Stability of linear equations solvers in interior-point methods*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1287–1307.
[30] S. Wright, *Stability of Linear Algebra Computations in Interior-Point Methods for Linear Programming*, Tech. Report MCS-P446-0694, Mathematics and Computer Science division, Argonne National Laboratory, Argonne, IL, 1994.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 3, pp. 952–969
© 2000 Society for Industrial and Applied Mathematics

FIXING VARIABLES IN SEMIDEFINITE RELAXATIONS*

CHRISTOPH HELMBERG†

Abstract. The standard technique of reduced cost fixing from linear programming is not trivially extensible to semidefinite relaxations because the corresponding Lagrange multipliers are usually not available. We propose a general technique for computing reasonable Lagrange multipliers for constraints that are not part of the problem description. Its specialization to the semidefinite $\{-1,1\}$ relaxation of quadratic 0-1 programming yields an efficient routine for fixing variables. The routine offers the possibility of exploiting problem structure. We extend the traditional bijective map between $\{0,1\}$ and $\{-1,1\}$ formulations to the constraints so that the dual variables remain the same and structural properties are preserved. Consequently, the fixing routine can be applied efficiently to optimal solutions of the semidefinite $\{0,1\}$ relaxation of constrained quadratic 0-1 programming as well. We provide numerical results showing the efficacy of this approach.

Key words. semidefinite programming, semidefinite relaxations, quadratic 0-1 programming, reduced cost fixing

AMS subject classifications. Primary, 90C31; Secondary, 90C25, 65K10, 49K40

PII. S089547989631442X

1. Introduction.

The power of semidefinite relaxations of combinatorial problems was first recognized in the seventies [27]. At that time it was not considered likely that practical algorithms for computing the associated bounds would ever be available. Research was primarily theoretical [13]. A new rush of theoretical results in the early nineties [28, 8, 31] and the development of interior-point algorithms for semidefinite programming [21, 29, 1, 33, 19, 24, 2, 30] spurred interest in the field. Within a short time several results in approximation theory [11, 22, 10, 6, 9] were published, giving further evidence of the high quality of semidefinite programming bounds. Although a general framework for designing semidefinite relaxations of linear and quadratic 0-1 programming problems is available [28, 16, 20], few papers presenting computational experience have been published [18, 23, 36, 35] (all are based on interior-point codes). The bounds prove to be of good quality in practice, but implementations suffer from the high computational cost involved in solving the semidefinite programs.

The task of solving hard combinatorial problems to optimality leads naturally to the *branch-and-bound* (or branch-and-cut) approach. In the branch-and-bound approach a bound on the optimal solution is computed by solving a tractable relaxation of the combinatorial problem. If the bound is not tight enough to prove optimality of the best combinatorial solution found so far, then in the so-called branching step the problem is split into two or more subproblems by partitioning the space of combinatorial solutions, e.g., by setting a $\{0,1\}$ variable to 0 in one subproblem and to 1 in the other. The same scheme is then applied recursively to the subproblems. The nature of this enumerative approach is best visualized by (in graph theoretic terms) a rooted tree (the *branch-and-bound tree*) with the root node corresponding to the original combinatorial problem, the arcs corresponding to the decisions, and the nodes corresponding to the subproblems. Subproblems that have not yet been solved

*Received by the editors December 9, 1996; accepted for publication (in revised form) April 9, 1999; published electronically March 6, 2000. A preliminary version of this article was published in the proceedings of ESA'97 *Lecture Notes in Comput. Sci.* 1284, pp. 259–270.
http://www.siam.org/journals/simax/21-3/31442.html
†Konrad-Zuse-Zentrum für Informationstechnik Berlin, Takustraße 7, D–14195 Berlin, Germany (helmberg@zib.de, http://www.zib.de/helmberg).

are located in the leaves. The decisions taken along a path (or *branch*) starting at the root node successively reduce the set of feasible solutions of the subproblems on this path. If the bound on a subproblem indicates that no better combinatorial solution can be found in the subproblem's feasible set, then the node may be *fathomed*, i.e., the recursion can be stopped for this branch of the tree. The problem is solved when all end nodes of the branches have been fathomed. *Branch-and-cut* refers to branch-and-bound approaches where the bounding procedure for each subproblem is a cutting plane algorithm. Cutting plane algorithms improve some initial relaxation by successively adding inequalities that cut off the last nonintegral optimal solution but preserve all feasible integral solutions.

In this setting the efficiency of an expensive bound hinges on the trade-off between the number of branch-and-bound nodes and the computation time needed for each node. Indeed, in spite of its high cost, the semidefinite programming relaxation outperforms any other approach in the case of unconstrained 0-1 quadratic programming with dense cost matrices [18] (for a comparison to other approaches in the literature, see [18]). Yet for problems of more than a hundred 0-1 variables, the approach must be considered impractical. Even though the number of branch-and-bound nodes may seem reasonably small, the computation time needed to solve the semidefinite relaxation becomes prohibitive. The sharp increase in computation time with growing dimension calls for routines that fix the "easy" variables quickly, thereby reducing the dimension of remaining problems in the branch-and-bound tree (not necessarily the number of nodes). In [18] we did not know how to exploit the dual to fix variables. Even when the relaxation displayed its preferences for some variables at the root node, it was necessary to run through all dimensions until the overall bound was good enough.

In linear programming relaxations, fixing variables using reduced costs is a standard procedure. In order to illustrate this technique, consider a quadratic 0-1 programming problem of the form

$$\text{(QP)} \qquad \max y^T B y \quad \text{s.t. } y \in \{0,1\}^n$$

with integral cost matrix B and assume that a "good" integral solution $c^* = \bar{y}^T B \bar{y}$ with $\bar{y} \in \{0,1\}^n$ is known. Linear relaxations of (QP) are usually formulated in variables y_{ij}, which represent linearizations of the products $y_i \cdot y_j$. The relaxations typically include the box constraints $0 \le y_{ij} \le 1$. Let the optimal solution of the relaxation have dual objective value $b^T \hat{u}$ where b is the vector of primal right-hand sides and $\hat{u}$ is the vector of Lagrange multipliers. If the optimal solution $\hat{y}$ of the linear relaxation yields, say, $\hat{y}_{ij} = 1$, then we can quickly compute an upper bound for the problem with the additional constraint $y_{ij} = 0$ by exploiting the following observation. In the dual problem, the change of the right-hand side in the primal constraint $y_{ij} \le 1 = b_k$ to $y_{ij} \le 0 = \bar{b}_k$ affects only the cost coefficients; b changes to $\bar{b}$. Thus the old Lagrange multipliers $\hat{u}$ remain feasible for the new cost coefficients $\bar{b}$ and the new dual objective value is, in this particular case, $\bar{b}^T \hat{u} = b^T \hat{u} - \hat{u}_k$. If this new bound satisfies $b^T \hat{u} - \hat{u}_k < c^* + 1$, then this proves the following: If there are better integral solutions than c^* in the feasible set, then they satisfy $y_{ij} = 1$.

The standard semidefinite programming relaxation for quadratic 0-1 programming already implies some of the box constraints. Consequently, these are not included in the relaxation, and the corresponding Lagrange multipliers are unknown. Yet if the optimal solution of the semidefinite relaxation displays some $y_{ij} = 1$, then there must be some corresponding active constraint buried in the semidefiniteness constraint. It

is the goal of this paper to develop a practical method for extracting this information. We will present two algorithms. The first is based on a line search with respect to the maximal eigenvalue function. It is suitable for general inequalities and might also be useful if the dual matrix variable is a huge structured matrix, as in [17, 5]. The second is tailored to box constraints on offdiagonal variables and assumes the availability of the eigenvalue decomposition of the dual matrix variable. It is worth noting that the discussion is completely independent of the actual algorithm used to solve the semidefinite relaxation.

There are two standard models for quadratic 0-1 programming: one formulated in $\{0,1\}$ variables, the other formulated in $\{-1,1\}$ variables. They lead, in a canonical way, to semidefinite relaxations that are slightly different in appearance. In particular, the dual of the $\{-1,1\}$ relaxation allows for a very efficient routine for fixing variables. It is well known that both problems and their primal relaxations are equivalent [7, 16, 25]. The dual variables, however, will differ for varying representations of the same primal set. We will show how the transformation used in [7, 16, 25] can be extended to the constraints so that the dual variables are the same for both. If the $n+1$ fundamental constraints of the $\{0,1\}$ relaxation are modeled correctly, this enables us to use the fixing procedure of the $\{-1,1\}$ formulation for optimal solutions computed in the $\{0,1\}$ setting.

In section 2 we introduce the semidefinite relaxation of quadratic 0-1 programming in $\{-1,1\}$ variables that motivated the considerations to follow. Section 3 provides the theoretical framework for extracting information from the dual in the general setting of semidefinite programming. Section 4 explains the practical difficulties in implementing the theoretic approach and presents an efficient alternative within the $\{-1,1\}$ setting. In section 5, the equivalence transformation between $\{0,1\}$ and $\{-1,1\}$ formulations is extended to the constraints so that dual variables and structural properties are preserved. Section 6 presents numerical results underlining the efficacy of the fixing routine. We conclude the paper in section 7.

NOTATION.

$\mathbb{R}$, $\mathbb{R}^n$	real numbers, real column vector of dimension n
$M_{m,n}$, M_n	$m \times n$, $n \times n$ real matrices
S_n	$n \times n$ symmetric real matrices
$A \succeq 0$, $A \succ 0$	A is (symmetric) positive semidefinite, positive definite
I, I_n	identity of appropriate size or of size n
e_i	ith column of I
e	vector of all ones of appropriate dimension
$\lambda_i(A)$	ith eigenvalue of $A \in M_n$, usually $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$
$\lambda_{\min}(A)$, $\lambda_{\max}(A)$	minimal and maximal eigenvalue of A
Λ_A	diagonal matrix with $(\Lambda_A)_{ii} = \lambda_i(A)$
$\operatorname{tr}(A)$	trace of $A \in M_n$, $\operatorname{tr}(A) = \sum_{i=1}^n a_{ii} = \sum_{i=1}^n \lambda_i(A)$
$\langle A, B\rangle$	inner product in $M_{m,n}$, $\langle A, B\rangle = \operatorname{tr}(B^T A)$
$\operatorname{rank}(A)$	rank of A
$\operatorname{diag}(A)$	$\operatorname{diag}(A) = [a_{11}, \ldots, a_{nn}]^T$
$\operatorname{Diag}(v)$	diagonal matrix with diagonal v
$\mathcal{A}X$	$= [\langle A_1, X\rangle, \ldots, \langle A_m, X\rangle]^T$ for given $A_i \in S_n$
$\mathcal{A}^T y$	$= \sum_{i=1}^m y_i A_i$

Unless explicitly stated otherwise, all matrices are symmetric and all vectors are column vectors.

2. Quadratic 0-1 programming in {-1,1} variables. Two canonical formulations of quadratic 0-1 programming appear in the literature: one in terms of $\{0,1\}$ variables and one in terms of $\{-1,1\}$ variables. For our purposes, the semidefinite relaxation of the $\{-1,1\}$ formulation (which is better known as the semidefinite relaxation of max-cut) is more convenient. We will return to the $\{0,1\}$ formulation and the equivalence of both in section 5.

The natural interpretation for a vector $x \in \{-1,1\}^n$ is that of a partition vector. Indices having the same sign belong to the same set. Formulated with respect to the product $x_i x_j$, i and j belong to the same set if $x_i x_j$ is equal to 1, and i and j belong to opposite sets if $x_i x_j$ is -1. The combinatorial problem to be investigated reads

$$\text{(MC)} \qquad \max x^T C x \quad \text{s.t. } x \in \{-1,1\}^n,$$

where $C \in S_n$ is the cost matrix. The standard semidefinite relaxation is derived by observing that $x^T C x = \langle C, xx^T \rangle$. For all $\{-1,1\}^n$ vectors, xx^T is a positive semidefinite matrix with all diagonal elements equal to 1. We relax xx^T to $X \succeq 0$ and $\operatorname{diag}(X) = e$ and obtain the following primal-dual pair of semidefinite programs:

$$\begin{array}{llll} & \max & \langle C, X \rangle & \\ \text{(PMC)} & \text{s.t.} & \operatorname{diag}(X) = e, & \\ & & X \succeq 0, & \end{array} \qquad \begin{array}{llll} & \min & e^T u & \\ \text{(DMC)} & \text{s.t.} & C + Z - \operatorname{Diag}(u) = 0, & \\ & & Z \succeq 0. & \end{array}$$

The relaxation can be strengthened by adding a few of the so-called triangle inequalities:

$$\begin{aligned} x_{ij} + x_{ik} + x_{jk} &\geq -1, \\ x_{ij} - x_{ik} - x_{jk} &\geq -1, \\ -x_{ij} + x_{ik} - x_{jk} &\geq -1, \\ -x_{ij} - x_{ik} + x_{jk} &\geq -1 \end{aligned}$$

for $i < j < k$ from $\{1, \ldots, n\}$. For later reference, we point out that, by exploiting the ones on the diagonal of X, these inequalities can be written in the form

$$v^T X v \geq 1, \tag{2.1}$$

with $v \in \mathbb{R}^n$ satisfying

$$v_h = \begin{cases} 1 \text{ or } -1 & \text{for } h \in \{i,j,k\}, \\ 0 & \text{otherwise.} \end{cases}$$

We briefly outline the branch-and-cut approach of [18] for solving (MC). The bound is computed by improving the semidefinite relaxation (PMC) with several cutting planes of type (2.1). From the optimal primal solution X^* of the relaxation, a good integral solution $\hat{x} \in \{-1,1\}^n$ with value $c^* = \hat{x}^T C \hat{x}$ is generated by rounding rows and applying simple local search techniques. If the bound $\langle C, X^* \rangle = e^T u^*$ is close enough to c^* (close meaning $e^T u^* < c^* + 1$ in the case that $x^T C x$ is integral for all $x \in \{-1,1\}^n$), then the feasible set cannot contain an integral solution with better value, and we are done (in the root node: stop; in the recursive step: fathom the node in the recursive step and proceed with an open problem in the branch-and-bound tree). Otherwise, select two indices i and j to generate two subproblems, one with i

and j in the same set, one with i and j in opposite sets. Both subproblems can be expressed as quadratic $\{-1,1\}^{n-1}$ problems. Proceed recursively.

If the bound is not good enough to fathom the node, but $x_{ij}=1$ ($x_{ij}=-1$) for some $i\neq j$ in the optimal solution of the relaxation, we can expect that indeed i and j belong together (apart) in all optimal solutions of the combinatorial subproblem. If we force the opposite, a reduction in the bound is to be expected. Can we prove that this reduction will be large enough to fathom this case without actually recomputing the bound?

In a linear cutting plane algorithm for the $\{-1,1\}$ model, the constraints $-1\leq x_{ij}\leq 1$ are typically included in the initial relaxation. If the optimal solution of the linear relaxation exhibits $|x_{ij}|=1$, then the dual variable of the corresponding active constraint yields a lower bound on the change of the objective value that would result from forcing x_{ij} to the opposite sign. This bound may be sufficient to prove that the current value of x_{ij} is correct for all optimal solutions of the combinatorial subproblem.

In the semidefinite relaxation (PMC) the constraints $-1\leq x_{ij}\leq 1$ are already implied by the diagonal constraints and the semidefiniteness of X. Therefore, they are not included in the semidefinite relaxation, and the corresponding dual variables are not available.

However, we can associate with each active constraint $x_{ij}\geq -1$ or $x_{ij}\leq 1$ an active constraint $v^TXv\geq 0$ from the set of constraints, ensuring the positive semidefiniteness of X as follows. Let $|x_{ij}|=1$ for some $i\neq j$ in the optimal solution (X^*,u^*,Z^*) of the current relaxation. Then the vector $v\in\mathbb{R}^n$ with

$$v_k=\begin{cases}1 & k=i,\\ -\operatorname{sgn}(x_{ij}) & k=j,\\ 0 & \text{otherwise}\end{cases}\tag{2.2}$$

is in the null space of X^*. Although the Lagrange multiplier corresponding to the constraint $\langle vv^T,X\rangle\geq 0$ is not directly available, it might be possible to construct one from information contained in the Lagrange multiplier to the semidefiniteness constraint on X, this multiplier being the dual slack matrix Z^*. We will do so in the general setting of semidefinite programming.

3. The theoretical framework. Consider a standard primal-dual pair of semidefinite programs:

$$\text{(P)}\quad \begin{array}{ll}\min & \langle C,X\rangle\\ \text{s.t.} & \mathcal{A}X=b,\\ & X\succeq 0,\end{array}\qquad\qquad \text{(D)}\quad \begin{array}{ll}\max & \langle b,u\rangle\\ \text{s.t.} & \mathcal{A}^Tu+Z=C,\\ & Z\succeq 0.\end{array}$$

$\mathcal{A}:S_n\to\mathbb{R}^m$ is a linear operator of the form

$$\mathcal{A}X=\begin{bmatrix}\langle A_1,X\rangle\\ \vdots\\ \langle A_m,X\rangle\end{bmatrix},$$

with $A_i\in S_n$, $i=1,\dots,m$. Its adjoint operator $\mathcal{A}^T:\mathbb{R}^m\to S_n$, satisfying $\langle\mathcal{A}X,u\rangle=\langle X,\mathcal{A}^Tu\rangle$ for all $X\in S_n$ and $u\in\mathbb{R}^m$, is

$$\mathcal{A}^Tu=\sum_{i=1}^m u_iA_i.$$

We examine possibilities to extract duality information for equality or inequality constraints that are not explicitly given in the problem description. Assume that optimal solutions X^* of (P) and (u^*, Z^*) of (D) are given and denote the optimal objective value by $p^* = \langle C, X^* \rangle = \langle b, u^* \rangle$. (In the last equation we tacitly assume some constraint qualification to hold.)

We are interested in the following question: How much does the optimal value of (P) increase if an additional constraint $\langle A_0, X \rangle = b_0$ is added to the problem? We would like to bound this quantity without actually computing the optimal solution of the new problem.

Let u_0 denote the new dual variable associated with the new constraint. The corresponding primal-dual pair reads

$$(\mathrm{P}_0) \quad \begin{array}{ll} \min & \langle C, X \rangle \\ \text{s.t.} & \langle A_0, X \rangle = b_0, \\ & \mathcal{A}X = b, \\ & X \succeq 0, \end{array} \qquad (\mathrm{D}_0) \quad \begin{array}{ll} \max & b_0 u_0 + \langle b, u \rangle \\ \text{s.t.} & u_0 A_0 + \mathcal{A}^T u + Z = C, \\ & Z \succeq 0. \end{array}$$

Computing the optimal solution is as hard as solving the original problem. However, we do already know a "good" dual feasible solution for (D_0), namely, $(u_0 = 0, u^*, Z^*)$. To improve this solution with reasonable effort, we restrict ourselves to a line search along an ascent direction $(\Delta u_0, \Delta u, \Delta Z)$ with

$$\begin{aligned} b_0 \Delta u_0 + \langle b, \Delta u \rangle &> 0, \\ \Delta u_0 A_0 + \mathcal{A}^T \Delta u + \Delta Z &= 0, \\ Z^* + t \Delta Z &\succeq 0 \qquad \text{for some } t \geq 0. \end{aligned}$$

Determining the best search direction is again as difficult as the problem itself. The choice of a good direction will depend on our understanding of the problem at hand.

Having fixed an ascent direction $(\Delta u_0, \Delta u, \Delta Z)$, it remains to compute the maximal step size t such that $Z^* + t\Delta Z$ is still positive semidefinite, because the objective function is linear. With

$$\begin{aligned} S &= C - \mathcal{A}^T u^* = Z^* \succeq 0, \\ B &= \Delta u_0 A_0 + \mathcal{A}^T \Delta u, \end{aligned}$$

the problem reduces to

$$(\mathrm{LS}) \quad \max t \quad \text{s.t. } S - tB \succeq 0.$$

There is no need to include the constraint $t \geq 0$, since $S \succeq 0$ implies that the optimal solution t^* is nonnegative. Problems of this form appear as matrix pencils in the literature (see, e.g., [12, Chapters 7.7 and 8.7], and the references therein). Indeed, the optimal t can be computed explicitly. To keep the paper self contained, we include the main steps.

Let $P \Lambda_S P^T = S$ denote an eigenvalue decomposition of S with P, an orthonormal matrix, and Λ_S, a diagonal matrix having the eigenvalues $\lambda_1(S) \geq \cdots \geq \lambda_n(S)$ on its diagonal, in this order. Then $S - tB \succeq 0$ is equivalent to

$$\Lambda_S - t P^T B P \succeq 0.$$

If the rank of S is k, then $\lambda_i(S) = 0$ for $i = k+1, \ldots, n$. Multiplying the equation above by $D = \mathrm{Diag}(\lambda_1(S)^{-\frac{1}{2}}, \ldots, \lambda_k(S)^{-\frac{1}{2}}, 1, \ldots, 1)$ from the left and right, we obtain

$$\begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix} - tDP^TBPD \succeq 0.$$

Assuming that $t > 0$ exists, we divide by t and impose the same block structure on DP^TBPD,

$$\begin{bmatrix} \frac{1}{t} I_k & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} B_{11} & B_{12} \\ B_{12}^T & B_{22} \end{bmatrix} \succeq 0,$$

with $B_{11} \in M_k$, $B_{22} \in M_{n-k}$, $B_{12} \in M_{k,n-k}$ such that

$$\begin{bmatrix} B_{11} & B_{12} \\ B_{12}^T & B_{22} \end{bmatrix} = DP^TBPD.$$

In the case where B_{12} and B_{22} are both zero, $\frac{1}{t} \geq \lambda_{\max}(B_{11})$ is the best choice (for $S \succ 0$ this specializes to $\frac{1}{t} = \lambda_{\max}(S^{-1}B)$; see [24]). Note that for $\lambda_{\max}(B_{11}) \leq 0$, the problem is unbounded. If $-B_{22}$ is nonzero and not positive semidefinite, then $t = 0$ is the optimal solution. If $-B_{22}$ is positive semidefinite with rank h, we can apply a similar sequence of steps to obtain a condition

$$\begin{bmatrix} \frac{1}{t} I_k - B_{11} & \bar{B}_{12} & \bar{B}_{13} \\ \bar{B}_{12}^T & I_h & 0 \\ \bar{B}_{13}^T & 0 & 0 \end{bmatrix} \succeq 0.$$

If $\bar{B}_{13}$ is nonzero, then again t must be zero. Otherwise, we can apply the Schur complement theorem to obtain the condition

$$\frac{1}{t} I_k - B_{11} \succeq \bar{B}_{12}\bar{B}_{12}^T.$$

This yields $\frac{1}{t} \geq \lambda_{\max}(B_{11} + \bar{B}_{12}\bar{B}_{12}^T)$.

We specialize this general procedure to a case of particular importance in semidefinite programming. For the purpose of explanation, assume that X^* and (u^*, Z^*) are a strictly complementary pair of optimal solutions, i.e., $\mathrm{rank}(X^*) + \mathrm{rank}(Z^*) = n$. (These do not necessarily exist; see, e.g., [2].) Furthermore, let A_0 be a dyadic product vv^T for some $v \in \mathbb{R}^n$ with $\langle vv^T, X^* \rangle = 0$, i.e., v is in the null space of X^*. The matrix vv^T may be interpreted as one of the active constraints ensuring the positive definiteness of X. The right-hand side b_0 of the new constraint must be greater than zero; otherwise there is no feasible primal solution for the new problem. As ascent direction we choose $\Delta u_0 = 1$ and $\Delta u = 0$. This yields the following line search problem:

$$\max t \quad \text{s.t. } Z^* - tvv^T \succeq 0.$$

Because X^* and Z^* are strictly complementary solutions and v is in the null space of X^*, we conclude that v lies in the span of the eigenvectors associated with nonzero eigenvalues of Z^*. Assume that $\mathrm{rank}(Z^*) = k$ and let $P\Lambda_{Z^*}P = Z^*$ denote the eigenvalue decomposition of Z^* with $P \in M_{n,k}$, $P^TP = I_k$, and the spectrum of nonzero eigenvalues $\Lambda_{Z^*} \in S_k$. Then the maximal t is given by

$$t^* = \frac{1}{v^TP\Lambda_{Z^*}^{-1}P^Tv}. \tag{3.1}$$

If, in particular, v happens to be an eigenvector of Z^*, then t^* is the corresponding eigenvalue of Z^*.

This procedure can be extended to the case where X^* and Z^* are not strictly complementary. For any vector v in the null space of X^*, but not in the span of the nonzero eigenvectors of Z^*, the optimal t is zero.

4. A practical algorithm. With respect to the semidefinite relaxation (PMC), formula (3.1) suggests a convenient procedure for constructing Lagrange multipliers for constraints of the form (2.2). Denote by (X^*, Z^*, u^*) an optimal primal-dual solution to (PMC). For some $i \neq j$ with $|x^*_{ij}| = 1$ in this solution, let v be the corresponding vector (2.2). Assuming that the eigenvalue decomposition of Z^* into $P\Lambda_{Z^*}P^T$ is available ($k = \operatorname{rank}(Z^*)$, $P \in M_{n,k}$, $P^TP = I_k$), it is easy to check whether v is in the span of the eigenvectors P. If it is not, then $t^* = 0$; otherwise $t^* = -1/(v^TP\Lambda^{-1}_{Z^*}P^Tv)$ is the best Lagrange multiplier for the constraint $v^TXv = 0$ obtainable from $Z^* = \operatorname{Diag}(u^*) - C$. (Since (PMC) is a maximization problem, the optimal t is now nonpositive.) Forcing the opposite assignment, $x_{ij} = -\operatorname{sgn}(x^*_{ij})$ can be modeled by changing the right-hand side of the (currently active) constraint $v^TXv = 0$ to $v^TXv = 4$ in the current relaxation. The optimal value of the relaxation corresponding to this opposite assignment is therefore at most $e^Tu^* + 4t^*$.

In theory this yields a very efficient algorithm for checking several pairs $\{i, j\}$. The eigenvalue decomposition has to be computed only once for all pairs, and the evaluation for a single pair requires roughly $O(nk)$ arithmetic operations. In implementing this approach, we encounter several difficulties. Indeed, we cannot expect any real-world algorithm to deliver a true optimal solution (X^*, u^*, Z^*) of (PMC) for arbitrary cost matrices. For a computed solution $(\hat{X}, \hat{u}, \hat{Z})$, both $\hat{X}$ and $\hat{Z}$ will be (rather ill-conditioned) full-rank matrices. Even in the case where the gap $\langle \hat{X}, \hat{Z}\rangle$ between the primal and dual solutions is almost zero, it is difficult to decide which of the eigenvalues of $\hat{X}$ and $\hat{Z}$ will eventually converge to zero. The space spanned by the eigenvectors corresponding to the "nonzero" eigenvalues of $\hat{X}$ and $\hat{Z}$ may still differ substantially from the true eigenspaces of X^* and Z^*. Eigenvalue decompositions are difficult to compute, because the eigenvalues tend to cluster at 0. The vectors v of (2.2) will be contained neither in the null space of $\hat{X}$ nor in the space spanned by the "nonzero" eigenvectors of $\hat{Z}$, because no $|x_{ij}|$ will be exactly one. As a result, the line search will allow only a very short step, and the approach fails.

However, in the case of (PMC) there is an obvious way to get around these difficulties. It is worth noting that the framework can be applied when there are additional primal constraints, but as these have no influence on the considerations to follow, we ignore them here.

Within the branch-and-bound scenario, let $(\hat{X}, \hat{u}, \hat{Z})$ be the solution computed for the relaxation of the current branch-and-bound node yielding the upper bound $e^T\hat{u}$. Let c^* denote the lower bound on the optimal solution of (MC). How much does the upper bound improve if we add a branching constraint $\langle A_0, \hat{X}\rangle = b_0$ (w.l.o.g. assume $b_0 > 0$) to the current relaxation? We denote the Lagrange multiplier for the new constraint by u_0. We would like to compute an upper bound, ideally smaller than c^*, for the problem

$$\begin{aligned} \min \quad & b_0u_0 + e^Tu \\ \text{s.t.} \quad & Z = u_0A_0 + \operatorname{Diag}(u) - C, \\ & Z \succeq 0. \end{aligned}$$

Consider the situation of setting u_0 to some (negative) value required for achieving

$b_0u_0 + e^T\hat{u} < c^*$. If $\hat{Z} + u_0A_0$ is still positive semidefinite, then we are done . If not, we add $-\lambda_{\min}(\hat{Z} + u_0A_0)e$ to $\hat{u}$, giving

$$u = \hat{u} - \lambda_{\min}(\hat{Z} + u_0A_0)e.$$

This worsens the original bound of $e^T\hat{u}$ by $-n\lambda_{\min}$, but the new Z is feasible. Thus we seek a u_0 such that

$$4u_0 + e^T\hat{u} - n\lambda_{\min}(\hat{Z} + u_0vv^T) < c^*.$$

Summing up, we specialize the semidefinite program above to

$$\min_{u_0\in\mathbb{R}} b_0u_0 + e^T\hat{u} - n\lambda_{\min}(u_0A_0 + \hat{Z}). \tag{4.1}$$

We first outline a general method for arriving at good estimates for u_0 and present a specialized algorithm for constraints of the form $x_{ij} = -\operatorname{sgn}(\hat{x}_{ij})$ afterwards.

The minimal eigenvalue is a concave function, so (4.1) is a convex optimization problem. The objective function is differentiable if and only if the minimal eigenvalue has multiplicity 1. In this case the gradient is determined by

$$\nabla_{u_0}\big(b_0u_0 - n\lambda_{\min}(u_0A_0 + \hat{Z})\big) = b_0 - n\, q(u_0)^TA_0q(u_0),$$

with $q(u_0)$ denoting the (normalized) eigenvector to the minimal eigenvalue of $u_0A_0 + \hat{Z}$. As explained above, it can be expected that $\hat{Z}$ has eigenvalue zero with high multiplicity. Therefore, the function is not differentiable for $u_0 = 0$, and it is hard to guess the initial descent. It seems appropriate to choose the starting value u_0 with respect to the gap $c^* - e^T\hat{u}$, e.g., $u_0 = (c^* - e^T\hat{u}) \cdot 1.2/b_0$. For reasonably large $|u_0|$ the minimal eigenvalue will be well separated, and we can use the gradient to decide whether it is worth increasing $|u_0|$ even further. If the value of the gradient implies that it might be possible to beat c^*, we do another step slightly overestimating the remaining gap. We repeat this procedure at most three times.

The computation of the gradient requires the computation of the eigenvector of the minimal eigenvalue of $u_0A_0 + \hat{Z}$. Extremal eigenvalues and eigenvectors are best determined via iterative methods such as the Lanczos method, which can exploit problem structure (see, e.g., [12]). In particular these methods are very fast if a good starting vector is known. For the first computation, we suggest choosing an eigenvector associated with the maximal eigenvalue of A_0; for all further iterations, the last eigenvector computed is the natural choice. We expect that this approach is efficiently applicable even in the case when $\hat{Z}$ is a huge structured matrix, as in [17, 5].

In the following we assume that the constraint is of the form $x_{ij} = 1$ (the argument for $x_{ij} = -1$ is analogous) and that the eigenvalue decomposition $\hat{Z} = P\Lambda P^T$ with $P^TP = I_n$ can be computed efficiently, i.e., the dimension of $\hat{Z}$ is not above, say, 1000. In practice the representation $\langle E_{ij}, X\rangle = 2$ (the matrix E_{ij} has a 1 in positions ij and ji and is 0 otherwise) of the constraint $x_{ij} = 1$ has proved superior to the representation $\langle vv^T, X\rangle = 4$ with v as in (2.2). This is probably due to the difference in the spectrum of the representation matrices.

The fact that E_{ij} is a rank-2 matrix can be exploited to avoid repeated eigenvalue computations in the line search[1]. Let $E_{ij} = vv^T - ww^T$ for appropriate $v, w \in \mathbb{R}^n$

[1]We are grateful to Kurt Anstreicher for pointing this out to us.

and let $\hat{Z} = P\Lambda P^T$ with $P^TP = I_n$ be an eigenvalue decomposition of $\hat{Z}$. Then (4.1) specializes to

$$\min_{u_0 \le 0, \lambda \ge 0} 2u_0 + n\lambda \quad \text{s.t. } u_0 P^T(vv^T - ww^T)P + \lambda I + \Lambda \succeq 0.$$

For a given $\lambda > 0$, we may rewrite the semidefinite constraint as

$$\left(\Lambda + \lambda I - u_0 \bar{w}\bar{w}^T\right) + u_0 \bar{v}\bar{v}^T \succeq 0,$$

where $\bar{v} = P^Tv$ and $\bar{w} = P^Tw$. Since, for $u_0 \le 0$, the matrix $-u_0\bar{w}\bar{w}^T$ is positive semidefinite and hence $\Lambda + \lambda I - u_0\bar{w}\bar{w}^T \succ 0$, we may employ (3.1) to obtain

$$u_0 \ge -\frac{1}{\bar{v}^T(\Lambda + \lambda I - u_0\bar{w}\bar{w}^T)^{-1}\bar{v}}.$$

Using the Sherman–Morrison–Woodbury formula for the inverse leads to

$$\bar{v}(\Lambda + \lambda I)^{-1}\bar{v} + u_0 \frac{[\bar{v}(\Lambda + \lambda I)^{-1}\bar{w}]^2}{1 - u_0\bar{w}^T(\Lambda + \lambda I)^{-1}\bar{w}} \le -\frac{1}{u_0},$$

which is equivalent to

$$\left(\sum_{i=1}^n \frac{\bar{v}_i\bar{w}_i}{\lambda_i(\hat{Z}) + \lambda}\right)^2 \ge \left(\sum_{i=1}^n \frac{\bar{v}_i^2}{\lambda_i(\hat{Z}) + \lambda} + \frac{1}{u_0}\right)\left(\sum_{i=1}^n \frac{\bar{w}_i^2}{\lambda_i(\hat{Z}) + \lambda} - \frac{1}{u_0}\right).$$

From this quadratic relation the best u_0 can be computed explicitly for any given $\lambda > 0$ in $O(n)$ arithmetic operations. This considerably speeds up the line search, which is now formulated in λ. The factorization of $\hat{Z}$ has to be computed only once for all x_{ij} that are considered for fixing.

In section 6 we will present some experimental results indicating the suitability of this approach. Note that the algorithms trivially extend to other semidefinite relaxations, e.g., for coloring [22] or max k-cut [10], allowing the possibility of shifting eigenvalues directly.

5. Quadratic 0-1 programming. It is well known that quadratic 0-1 programming in n variables is equivalent to quadratic $\{-1, 1\}$ programming in $n+1$ variables. The basic transformation appears already in [14]. De Simone [7] provided a formal prove that the transformation is indeed a bijective map between the cut polytope and the boolean quadric polytope. The equivalence extends, via the same transformation, to the canonical semidefinite relaxations [16, 25].

In general the $\{0, 1\}$ formulation is considered more intuitive and usually chosen for modeling combinatorial problems. In fact, most articles dealing with constrained quadratic 0-1 programming work within this setting [28, 3, 16, 20]. From a theoretical standpoint, the equivalence of the primal problems implies that the previous considerations can be applied (indirectly) to the $\{0, 1\}$ setting. From a practical standpoint, two other aspects are important: Problem transformations tend to destroy structure inherent in the natural problem formulation; therefore, transformations should be avoided or designed so as to preserve as much structure as possible. On the other hand, dual variables usually have a natural interpretation in the original formulation. It is difficult to translate this interpretation into a transformed model and typically it is even more difficult to construct dual variables for the original problem from the dual variables of the transformed problem.

In expressing the standard transformation between $\{0,1\}$ and $\{-1,1\}$ as a "scaling" of the matrix variables, we show that this transformation actually achieves both: problem structure is largely preserved, and the dual variables are the same. Previous representations concentrated on the bijection between variables; constraints were transformed indirectly via substitution and by moving all constants to the right-hand side. In such a procedure, structural properties, e.g., the low rank structure of a coefficient matrix, are lost and the "simplification" of the constants causes changes in the respective dual variables. In contrast, the scaling representation automatically provides the correct transformation of the constraints, so that dual variables remain unchanged and low rank structure is preserved. Since the transformation also preserves sparsity to a large extent, switching back and forth between both formulations will entail no serious loss in efficiency or information.

Quadratic 0-1 programming in $\{0,1\}$ variables seeks the optimal solution of

$$\text{(QP)} \qquad \max y^T B y \quad \text{s.t. } y \in \{0,1\}^n.$$

The canonical semidefinite relaxation for quadratic 0-1 programming is derived by adding an additional component 1 (with index 0) to the vector y and by looking at

$$\begin{bmatrix} 1 \\ y \end{bmatrix} \begin{bmatrix} 1 & y^T \end{bmatrix} = \begin{bmatrix} 1 & y^T \\ y & yy^T \end{bmatrix}.$$

This matrix is positive semidefinite, and its diagonal is equal to the first column and the first row for all $y \in \{0,1\}^n$. An intuitive way to write the semidefinite relaxation is

$$\text{(PQ)} \qquad \begin{array}{ll} \max & \langle B, Y \rangle \\ \text{s.t.} & \bar{Y} = \begin{bmatrix} 1 & \operatorname{diag}(Y)^T \\ \operatorname{diag}(Y) & Y \end{bmatrix} \succeq 0. \end{array}$$

There are several ways to model linear constraints so that the diagonal property of $\bar{Y}$ is preserved. We will construct a representation ensuring that the dual variables are the same as those of the equivalent problem (PMC) in $n+1$ variables.

To this end we present some well-known facts about transformations of the type $W = QXQ^T$ for nonsingular $Q \in M_n$ in the general setting of the primal-dual pair (P) and (D). These transformations belong to the automorphism group of the semidefinite cone (the set of all bijective linear maps leaving the semidefinite cone invariant) and appear several times in the interior-point literature in connection with scaling issues (see, e.g., [30, 32]). Clearly, $W = QXQ^T$ is positive semidefinite if and only if X is. How do we have to change the constraints of (P) such that we get the same semidefinite program in terms of W? Since $X = Q^{-1}WQ^{-T}$ and, for arbitrary $A \in S_n$,

$$\langle A, X \rangle = \left\langle A, Q^{-1}WQ^{-T} \right\rangle = \left\langle Q^{-T}AQ^{-1}, W \right\rangle,$$

the correct transformation of the matrix A is $Q^{-T}AQ^{-1}$. Note that this is the adjoint of the inverse transformation of QXQ^T. Using

$$\bar{C} = Q^{-T}CQ^{-1}, \qquad \bar{A}_i = Q^{-T}A_iQ^{-1}, \quad i = 1, \ldots, m,$$

and the linear operators $\bar{\mathcal{A}}$ and $\bar{\mathcal{A}}^T$ formed by the $\bar{A}_i$, we obtain the transformed primal-dual pair

$$\text{(P}_Q\text{)} \qquad \begin{array}{ll} \min & \left\langle \bar{C}, W \right\rangle \\ \text{s.t.} & \bar{\mathcal{A}}W = b, \\ & W \succeq 0, \end{array} \qquad\qquad \text{(D}_Q\text{)} \qquad \begin{array}{ll} \max & \langle b, \bar{u} \rangle \\ \text{s.t.} & \bar{\mathcal{A}}^T\bar{u} + \bar{Z} = \bar{C}, \\ & \bar{Z} \succeq 0. \end{array}$$

PROPOSITION 5.1. *X is a feasible solution of* (P) *if and only if the associated $W = QXQ^T$ is a feasible solution of* (P_Q). *Furthermore, X and W satisfy $\langle C, X\rangle = \langle \bar{C}, W\rangle$. (u, Z) is a feasible solution of* (D) *if and only if the associated $(\bar{u}, \bar{Z}) = (u, Q^{-T}ZQ^{-1})$ is a feasible solution of* (D_Q). *Trivially, $\langle b, u\rangle = \langle b, \bar{u}\rangle$.*

Proof. The statement follows directly from the construction above. ∎

In particular, this implies that, given an optimal primal-dual solution for one of the problems, we can construct an optimal primal-dual solution for the other.

We apply this approach to the transformation between $\{-1, 1\}$ and $\{0, 1\}$ representation of 0-1 quadratic programming .

PROPOSITION 5.2. *Let $Q \in M_{n+1}$ be the matrix*

$$Q = \begin{bmatrix} 1 & 0 \\ \frac{1}{2}e & \frac{1}{2}I_n \end{bmatrix};$$

then $\varphi : S_{n+1} \to S_{n+1}, X \mapsto \bar{Y} = QXQ^T$ bijectively maps feasible solutions of (PMC) *(for $n+1$ variables) to feasible solutions of* (PQ).

Proof. Q is nonsingular; therefore X is positive definite if and only if $\varphi(X)$ is. The properties concerning the diagonals are verified by direct computation. ∎

This is a slight simplification with respect to earlier proofs of this fact [16, 25]. The main advantage of this approach is that by Proposition 5.1 we know how to formulate the constraints such that we can go back and forth between both models without changing the dual variables.

To make this even simpler, we provide a table of the most important transformations. In order to introduce the necessary notation, observe that

$$Q^{-1} = \begin{bmatrix} 1 & 0 \\ -e & 2I_n \end{bmatrix}.$$

As the first row and column play a special role in (PQ), we give all transformations for partitioned matrices

$$\bar{A} = \begin{bmatrix} a_0 & a^T \\ a & A \end{bmatrix},$$

with $a_0 \in R$, $a \in \mathbb{R}^n$, and $A \in S_n$. The correct transformation of the coefficient matrices is achieved by the adjoint operator to φ^{-1} (with φ as in Proposition 5.2):

$$(\varphi^{-1})^* : S_{n+1} \to S_{n+1}, \bar{A} \mapsto \bar{B} = Q^{-T}\bar{A}Q^{-1}.$$

For implementational purposes, constraint matrices of the form $\bar{A} = \bar{v}\bar{v}^T$ or $\bar{A} = \bar{v}\bar{v'}^T + \bar{v'}\bar{v}^T$ are of special importance [28, 16, 20], being conveniently transformed by the linear bijective map

$$\sigma : \mathbb{R}^{n+1} \mapsto \mathbb{R}^{n+1}, \quad \bar{v} \mapsto \bar{w} = Q^{-T}\bar{v}.$$

Obviously,

$$\left(\varphi^{-1}\right)^*(\bar{v}\bar{v}^T) = \sigma(\bar{v})\sigma(\bar{v})^T \quad \text{and} \quad \left(\varphi^{-1}\right)^*\left(\bar{v}\bar{v'}^T + \bar{v'}\bar{v}^T\right) = \sigma(\bar{v})\sigma(\bar{v'})^T + \sigma(\bar{v'})\sigma(\bar{v})^T.$$

The explicit formulas are given in Table 5.1. These transformations preserve most of the structure (sparsity and low rank representations), which is of high practical importance.

TABLE 5.1
Transformations between the $\{-1,1\}$ and the $\{0,1\}$ model.

(MC) $\to$ (QP)	(QP) $\to$ (MC)
$\varphi(\bar{X})=Q\bar{X}Q^T=\begin{bmatrix} y_0 & y^T \\ y & Y \end{bmatrix}$	$\varphi^{-1}(\bar{Y})=Q^{-1}\bar{Y}Q^{-T}=\begin{bmatrix} x_0 & x^T \\ x & X \end{bmatrix}$
$y_0=x_0$	$x_0=y_0$
$y=\frac{1}{2}(x+x_0e)$	$x=2y-y_0e$
$Y=\frac{1}{4}(X+xe^T+ex^T+x_0ee^T)$	$X=4Y-2y_0e^T-2ey_0^T+y_0ee^T$
$(\varphi^{-1})^*(\bar{A})=Q^{-T}\bar{A}Q^{-1}=\begin{bmatrix} b_0 & b^T \\ b & B \end{bmatrix}$	$\varphi^*(\bar{B})=Q^T\bar{B}Q=\begin{bmatrix} a_0 & a^T \\ a & A \end{bmatrix}$
$b_0=a_0-2e^Ta+e^TAe$	$a_0=b_0+e^Tb+\frac{1}{4}e^TBe$
$b=2(a-Ae)$	$a=\frac{1}{2}b+\frac{1}{4}Be$
$B=4A$	$A=\frac{1}{4}B$
$\sigma(\bar{v})=Q^{-T}\bar{v}=\begin{bmatrix} w_0 \\ w \end{bmatrix}$	$\sigma^{-1}(\bar{w})=Q^T\bar{w}=\begin{bmatrix} v_0 \\ v \end{bmatrix}$
$w_0=v_0-e^Tv$	$v_0=w_0+\frac{1}{2}e^Tw$
$w=2v$	$v=\frac{1}{2}w$
$Q=\begin{bmatrix} 1 & 0 \\ \frac{1}{2}e & \frac{1}{2}I_n \end{bmatrix}$	$Q^{-1}=\begin{bmatrix} 1 & 0 \\ -e & 2I_n \end{bmatrix}$

To translate (PMC) to (PQ), we observe that $\operatorname{diag}(\bar{X})=\bar{e}$ can be modeled by

$$\bar{e}_i^T\bar{X}\bar{e}_i=1, \quad i=0,\dots,n.$$

Using σ, these translate to

$$\begin{aligned} \bar{e}_0 \quad &\to \quad \bar{w}^{(0)}=\bar{e}_0, \\ \bar{e}_i \quad &\to \quad \bar{w}^{(i)} \text{ with } \bar{w}^{(i)}_j=\begin{cases} -1 & j=0, \\ 2 & j=i, \\ 0 & \text{otherwise}, \end{cases} \qquad i=1,\dots,n. \end{aligned}$$

Collecting these $n+1$ constraints in an operator $\mathcal{A}$ with $\bar{A}_i=\bar{w}^{(i)}\bar{w}^{(i)T}$ $(i=0,\dots,n)$, we obtain a formulation of (PQ) having the same dual variables as (PMC):

$$\begin{array}{llll} & \max & \langle \bar{B},\bar{Y}\rangle & \\ (\mathrm{PQ}') & \text{s.t.} & \mathcal{A}\bar{Y}=e, & \\ & & \bar{Y}\succeq 0, & \end{array} \qquad \begin{array}{llll} & \min & e^Tu & \\ (\mathrm{DQ}') & \text{s.t.} & \bar{B}+S-\mathcal{A}^Tu=0, & \\ & & S\succeq 0. & \end{array}$$

Returning to the fixing procedure for (PMC), we mention that in the $\{0,1\}$ model it is not obvious how to guarantee the positive semidefiniteness of S in (DQ′) by a similar approach. However, for given optimal solutions of the $\{0,1\}$ model, we can switch to the $\{-1,1\}$ setting without changing the dual variables and compute appropriate Lagrange multipliers. These are also correct multipliers for the $\{0,1\}$ model.

We now interpret the fixing of x_{ij} to either $+1$ or -1 in the $\{0,1\}$ setting. Using σ again on v from (2.2), we obtain for $\bar{w} = \sigma(\bar{v})$

$$\begin{aligned} i = 0,\ 1 \le j \le n,\ x_{ij} = 1 \quad &\rightarrow \quad w_0 = 2,\ w_j = -2, \\ i = 0,\ 1 \le j \le n,\ x_{ij} = -1 \quad &\rightarrow \quad w_0 = 0,\ w_j = 2, \\ 1 \le i < j \le n,\ x_{ij} = 1 \quad &\rightarrow \quad w_0 = 0,\ w_i = 2,\ w_j = -2, \\ 1 \le i < j \le n,\ x_{ij} = -1 \quad &\rightarrow \quad w_0 = -2,\ w_i = 2,\ w_j = 2. \end{aligned}$$

Reinterpreted in $\{0,1\}$ variables y_i and y_j, the equations $\bar{w}^T \bar{Y} \bar{w} = 0$ correspond to

$$\begin{aligned} y_j &= 1, \\ y_j &= 0, \\ (y_i - y_j)^2 &= 0, \\ (y_i - y_j)^2 &= 1. \end{aligned}$$

The third equation states that both y_i and y_j must be 0 or both must be 1. The fourth states that exactly one of both must be 1. Using the analogous procedure as for (PMC), we can try to verify the validity of such an equation for the optimal solution of a particular problem by constructing the corresponding Lagrange multipliers with respect to an optimal solution $(\bar{Y}^*, u^*, S^*)$ of the relaxation (PQ′).

For completeness we include the interpretation of the box constraints $0 \le y_{ij} \le 1$ arising naturally in linear relaxations. Observe that they do not appear in the list above. The constraint $y_{ij} \le 1$ is guaranteed by the feasibility of $\bar{Y}$. The natural interpretation for $y_{ij} = 1$ is that both y_i and y_j must be 1. In the $\{-1,1\}$ setting this corresponds to requiring that indices 0, i, and j belong to the same set.

The constraint $y_{ij} \ge 0$ is not implied by the feasibility of $\bar{Y}$. In fact, this constraint is well known to correspond to a triangle inequality (in (2.1) take $v_0 = 1$, $v_i = 1$, and $v_j = 1$). The interpretation of $y_{ij} = 0$ is that at most one of y_i and y_j may attain the value 1. In the $\{-1,1\}$ setting, not all three, 0, i, and j, may belong to the same set.

6. Implementation. We have implemented both algorithms of section 4 for fixing variables within our branch-and-cut code for solving (MC) as described in [18]. In contrast to the implementation in [18], we improve the semidefinite relaxation (PMC) with triangle inequalities only[2]. Eigenvalues and eigenvectors are computed by the EISPACK routines `tred2` and `imtql2` (translated from Fortran to C with `f2c`). The fixing procedures are applied whenever a variable x_{ij} of the current optimal solution satisfies $|x_{ij}| > .98$. This leads to no additional cost for problems in which no variables satisfy this bound. When variables of this size appeared, usually some of them could be fixed. We tested the code on the same classes of problems as in [18]: $G_{.5}$, $G_{-1/0/1}$, G_p, $G_{\pm p}$, Q_{100}, and $Q_{100,.2}$.

$G_{.5}$ consists of unweighted graphs with edge probability 1/2, and $G_{-1/0/1}$ consists of weighted (complete) graphs with edge weights chosen uniformly from $\{-1,0,1\}$. G_p is a class of almost planar graphs with edge weights 1, and in $G_{\pm p}$ the edge weights of the same graphs are chosen randomly from $\{-1,1\}$. Q_{100} and $Q_{100,.2}$ were used in [34, 4]. Formulating Q_{100} with respect to (QP), the lower triangle of B is set to zero, the upper triangle (including the diagonal) is chosen uniformly from $\{-100,\dots,100\}$. The diagonal takes the role of the linear term. $Q_{100,.2}$ represents instances with a density of 20%.

[2]Further differences to the results published in [18] are due to some slight changes in the interior-point code and a faster machine.

TABLE 6.1
Average branch-and-bound results.

		no fixing		with fixing		gen. fixing		no fixing		with fixing		gen. fixing	
n	nr	h:mm:ss	nodes	h:mm:ss	nodes	h:mm:ss	nodes	h:mm:ss	nodes	h:mm:ss	nodes	h:mm:ss	nodes
		$G_{0.5}$						$G_{-1/0/1}$					
40	10	12	8	11	7	11	7	13	9	13	9	13	9
50	10	46	16	46	16	47	16	50	19	50	19	49	19
60	5	4:26	57	4:24	57	4:29	57	2:45	37	2:44	36	2:44	36
70	3	14:17	121	14:06	118	14:24	118	7:44	64	7:40	63	7:40	63
80	2	30:08	145	29:53	145	29:59	145	1:15	203	41:11	203	41:14	203
90	1	2:39:05	583	2:37:07	583	2:39:11	583	11:57:00	2363	11:52:28	2360	11:56:01	2360
		G_p						$G_{\pm p}$					
40	10	7	4	7	4	7	4	1	1	1	1	1	1
50	10	48	14	48	14	48	14	8	2	9	2	9	2
60	5	3:13	32	3:14	31	3:15	31	49	9	49	9	49	9
70	3	7:12	41	7:05	39	7:07	39	4:32	25	5:17	29	5:19	29
80	2	43:06	158	42:57	156	43:11	156	4:57	17	4:57	18	5:02	18
90	1	30:00	69	30:48	73	30:54	73	22:49	63	23:17	63	23:18	63
		Q_{100}						$Q_{100,.2}$					
41	10	23	23	12	10	16	10	1	1	1	1	1	1
51	10	1:12	39	34	14	42	14	6	1	5	1	7	1
61	5	2:47	51	1:11	16	1:26	16	56	15	40	9	47	9
71	3	5:57	71	3:42	35	4:38	35	7:20	71	3:41	32	4:19	32
81	2	11:23	86	9:24	45	10:50	45	7:32	33	3:22	11	5:28	11
91	1	18:35	79	1:21	3	6:04	3	18:17	73	2:56	7	4:33	7
101	1	31:39	137	28:47	104	32:34	104	31:31	113	4:29	7	8:30	7

The results are displayed in Table 6.1. Column n gives the dimension of the problem in the $\{-1,1\}$ setting (the increment 1 appearing in the rows corresponding to Q_{100} and $Q_{100,.2}$ is due to the transformation) and nr refers to the number of instances solved. For each problem class, we give results for the branch-and-cut code without fixing variables (column *no fixing*), with the specialized fixing routine exploiting the rank two property (column *with fixing*), and with the general fixing routine based on repeated eigenvalue computations (column *gen. fixing*). In each case the average computation time[3] and number of branch-and-cut nodes is listed. We did not recompute the examples of size $n = 100$ of the classes $G_{.5}$, $G_{-1/1/1}$, G_p, and $G_{\pm p}$ in view of their excessive computation times recorded in [18].

It was observed in [18] that in practice $G_{.5}$ and $G_{-1/0/1}$ are substantially more difficult to solve than G_p and $G_{\pm p}$, and these are again more difficult than Q_{100} and $Q_{100,.2}$. Indeed, for the classes $G_{.5}$, $G_{-1/0/1}$, G_p, and $G_{\pm p}$ the fixing routines were hardly ever called, because no variables satisfied $|x_{ij}| > .98$. Accordingly, the additional cost of the routine is negligible. However, for the "easy" classes of problems Q_{100} and $Q_{100,.2}$, the specialized fixing routine was very successful and led to large savings in most cases. The general fixing routine was almost equally successful in fixing variables, but the cost of the repeated eigenvalue computations is clearly visible.

Summing up, the experimental results show that the fixing procedure is an important addition to branch-and-cut algorithms. It might be possible to narrow the number of candidates by analyzing the relation of the respective vectors v to the spectrum and the eigenvectors of Z^*. If C is a huge structured matrix as in [17, 5], eigenvalue decompositions of Z^* are out of scope. In this case one may have to employ the general fixing routine in combination with Lanczos methods for eigenvalue computations.

7. Conclusions. We propose to compute Lagrange multipliers for constraints that are not included in the problem description by means of a line search. The optimal step size can be computed explicitly for any given direction. An open problem in practical implementations is the (fast) determination of a good search direction. Applied to constraints of the form $v^T X v \geq 0$, this approach suggests the interpretation of the dual slack matrix Z as a variable subsuming all Lagrange multipliers corresponding to the active constraints $v^T X v \geq 0$ that ensure the positive semidefiniteness of X.

In the special case of the $\{-1,1\}$ semidefinite relaxation of 0-1 quadratic programming, the diagonal variables can be used to guarantee dual feasibility. This leads to an efficient and comparatively robust procedure for fixing variables that offers the possibility of exploiting structure. In practice, first implementations show considerable savings in computation time whenever candidates for fixing appear.

We extended the standard equivalence transformation between $\{0,1\}$ and $\{-1,1\}$ representations of the semidefinite relaxation for 0-1 quadratic programming to the constraints in such a way that the dual variables and the structural properties of the constraints are preserved. This transformation allows us to apply the fixing procedure of the $\{-1,1\}$ formulation to optimal solutions of the corresponding $\{0,1\}$ relaxation as well.

Although the most important ingredients for a general constrained quadratic 0-1 programming solver seem to be available by now, the solution of real-world problems is still out of reach. The main obstacle is the high computational cost involved in

[3] All times were computed on a Sun Ultra 10 with a 299 MHz SUNW,UltraSPARC-IIi CPU.

solving the semidefinite relaxations. First steps towards algorithms that are able to exploit problem structure in semidefinite programs can be found in [17, 5]. Like the general fixing procedure, the spectral bundle method of [17] relies on efficient routines for computing repeatedly the maximal eigenvalue of tightly related huge structured matrices. These matrices are generically ill-conditioned in the sense that many eigenvalues cluster around the maximal eigenvalue. Current state-of-the-art implementations like ARPACK [26] perform poorly on these matrices. Due to the wide range of applications of semidefinite programming, better algorithms are in high demand, and we hope that this article will encourage specialists to look at these difficult problems.

Acknowledgments. I thank Kurt Anstreicher for drawing my attention to the possibility of exploiting the low rank structure of the constraint matrix and Stefan E. Karisch, Franz Rendl, and an anonymous referee for their constructive criticism with respect to the presentation.

REFERENCES

[1] F. Alizadeh, *Interior point methods in semidefinite programming with applications to combinatorial optimization*, SIAM J. Optim., 5 (1995), pp. 13–51.

[2] F. Alizadeh, J.-P. A. Haeberly, and M. L. Overton, *Complementarity and nondegeneracy in semidefinite programming*, Math. Programming, 77 (1997), pp. 111–128.

[3] E. Balas, S. Ceria, and G. Cornuejols, *A lift-and-project cutting plane algorithm for mixed* 0/1 *programs*, Math. Programming, 58 (1993), pp. 295–324.

[4] F. Barahona, M. Jünger, and G. Reinelt, *Experiments in quadratic* 0–1 *programming*, Math. Programming, 44 (1989), pp. 127–137.

[5] S. Benson, Y. Ye, and X. Zhang, *Solving large-scale sparse semidefinite programs for combinatorial optimization*, SIAM J. Optim., 10 (2000), pp. 443–461.

[6] B. Chor and M. Sudan, *A geometric approach to betweenness*, in Proceedings ESA'95, Lecture Notes in Comput. Sci. 979, Springer, New York, 1995, pp. 227–237.

[7] C. De Simone, *The cut polytope and the boolean quadric polytope*, Discrete Math., 79 (1989), pp. 71–75.

[8] C. Delorme and S. Poljak, *Laplacian eigenvalues and the maximum cut problem*, Math. Programming, 62 (1993), pp. 557–574.

[9] U. Feige and M. X. Goemans, *Approximating the value of two prover proof systems, with applications to MAX* 2*SAT and MAX DICUT*, in Proceedings 3rd Israel Symposium on Theory of Computing and Systems, Tel Aviv, Israel, 1995, pp. 182–189.

[10] A. Frieze and M. Jerrum, *Improved approximation algorithms for MAX k-CUT and MAX BISECTION*, in Integer Programming and Combinatorial Optimization, E. Balas and J. Clausen, eds., Lecture Notes in Comput. Sci. 920, Springer, New York, 1995, pp. 1–13.

[11] M. X. Goemans and D. P. Williamson, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming*, J. ACM, 42 (1995), pp. 1115–1145.

[12] G. H. Golub and C. F. van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.

[13] M. Grötschel, L. Lovász, and A. Schrijver, *Geometric Algorithms and Combinatorial Optimization*, 2nd ed., Algorithms and Combinatorics 2, Springer, New York, 1988.

[14] P. L. Hammer, *Some network flow problems solved with pseudo-boolean programming*, Oper. Res., 13 (1965), pp. 388–399.

[15] C. Helmberg, *Fixing variables in semidefinite relaxations*, in Algorithms—ESA'97, R. Burkard and G. Woeginger, eds., Lecture Notes in Comput. Sci. 1284, Springer, New York, 1997, pp. 259–270.

[16] C. Helmberg, S. Poljak, F. Rendl, and H. Wolkowicz, *Combining semidefinite and polyhedral relaxations for integer programs*, in Integer Programming and Combinatorial Optimization, E. Balas and J. Clausen, eds., Lecture Notes in Comput. Sci. 920, Springer, New York, 1995, pp. 124–134.

[17] C. Helmberg and F. Rendl, *A spectral bundle method for semidefinite programming*, SIAM J. Optim., 10 (2000), pp. 673–698.
[18] C. Helmberg and F. Rendl, *Solving quadratic* $(0,1)$*-problems by semidefinite programs and cutting planes*, Math. Programming, 82 (1998), pp. 291–315.
[19] C. Helmberg, F. Rendl, R. J. Vanderbei, and H. Wolkowicz, *An interior-point method for semidefinite programming*, SIAM J. Optim., 6 (1996), pp. 342–361.
[20] C. Helmberg, F. Rendl, and R. Weismantel, *Quadratic knapsack relaxations using cutting planes and semidefinite programming*, in Integer Programming and Combinatorial Optimization, W. H. Cunningham, S. T. McCormick, and M. Queyranne, eds., Lecture Notes in Comput. Sci. 1084, Springer, New York, 1996, pp. 175–189.
[21] F. Jarre, *An interior-point method for minimizing the maximum eigenvalue of a linear combination of matrices*, SIAM J. Control Optim., 31 (1993), pp. 1360–1377.
[22] D. Karger, R. Motwani, and M. Sudan, *Approximate graph coloring by semidefinite programming*, J. ACM, 45 (1998), pp. 246–265
[23] S. E. Karisch and F. Rendl, *Semidefinite programming and graph equipartition*, in Topics in Semidefinite and Interior-Point Methods, P.M. Pardalos and H. Wolkowicz, eds., Field Inst. Commun. 18, AMS, Providence, RI, 1998, pp. 77-95.
[24] M. Kojima, S. Shindoh, and S. Hara, *Interior-Point Methods for the Monotone Linear Complementarity Problem in Symmetric Matrices*, SIAM J. Optim., 7 (1997), pp. 86–125
[25] M. Laurent, S. Poljak, and F. Rendl, *Connections between semidefinite relaxations of the max-cut and stable set problems*, Math. Programming, 77 (1997), pp. 225–246.
[26] R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK USERS GUIDE: Solution of Large Scale Eigenvalue Problems by Implicitly Restarted Arnoldi Methods.*, Department of Computation and Applied Mathematics, Rice University, Houston, TX 1997, manuscript.
[27] L. Lovász, *On the Shannon capacity of a graph*, IEEE Trans. Inform. Theory, 25 (1979), pp. 1–7.
[28] L. Lovász and A. Schrijver, *Cones of matrices and set-functions and* 0-1 *optimization*, SIAM J. Optim., 1 (1991), pp. 166–190.
[29] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM Studies in Applied Mathematics 13, SIAM, Philadelphia, PA, 1994.
[30] Y. Nesterov and M. J. Todd, *Self-scaled barriers and interior-Point methods for convex programming*, Math. Oper. Res., 22 (1997), pp. 1–42.
[31] S. Poljak and F. Rendl, *Nonpolyhedral relaxations of graph-bisection problems*, SIAM J. Optim., 5 (1995), pp. 467–487.
[32] L. Tunçel, *Primal-dual symmetry and scale invariance of interior-point algorithms for convex optimization*, Math. Oper. Res., 23 (1998), pp. 708–718.
[33] L. Vandenberghe and S. Boyd, *A primal-dual potential reduction method for problems involving matrix inequalities*, Math. Programming, Ser. B, 69 (1995), pp. 205–236.
[34] A. Williams, *Quadratic* 0-1 *Programming Using the Roof Dual with Computational Results*, RUTCOR Research Report 8-85, Rutgers University, Piscataway, NJ, 1985.
[35] H. Wolkowicz and Q. Zhao, *Semidefinite Programming Relaxations for the Graph Partitioning Problem*, J. Combin. Optim., 2 (1998), pp. 71–109.
[36] Q. Zhao, S. E. Karisch, F. Rendl, and H. Wolkowicz, *Semidefinite Programming Relaxations for the Quadratic Assignment Problem*, CORR Report 95/27, University of Waterloo, Ontario, Canada, 1996.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 3, pp. 970–977
© 2000 Society for Industrial and Applied Mathematics

ON THE CONDITION OF SYMMETRIC QUASI-DEFINITE MATRICES*

ALAN GEORGE† AND KH. IKRAMOV‡

Abstract. A promising approach to solving sparse systems of equations involves creating an augmented system that is symmetric quasi-definite. The attraction of this approach stems from being able to use powerful algorithms and efficient data structures that have been developed for sparse positive definite systems. In this article, bounds are derived for the condition number of the Schur complements that emerge when symmetric Gaussian elimination is applied to these quasi-definite matrices. An application of these bounds to solving regularized linear least squares problems is discussed. For the corresponding linear systems, the bounds may be used to justify a reordering based solely on sparsity preservation considerations.

Key words. condition numbers, symmetric quasi-definite matrices, the Cholesky factorization

AMS subject classification. 65F10

PII. S0895479898333247

1. Introduction and motivation.

Suppose $\boldsymbol{Ax} = \boldsymbol{b}$ is a large sparse nonsingular system to be solved via Gaussian elimination. If $\boldsymbol{A}$ has no special properties (positive definite, diagonally dominant), rows and/or column interchanges are performed to promote sparsity and also to ensure numerical stability. Implementations of Gaussian elimination for such problems are complicated and incur substantial overhead. Similar difficulties arise in solving large sparse least squares problems, where $\boldsymbol{A}$ is $m \times n$ with $m \geq n$.

When $\boldsymbol{A}$ is symmetric and positive definite, very important advantages accrue. Symmetric Gaussian elimination (Cholesky factorization) applied to such matrices does not require interchanges to maintain numerical stability. Since $\boldsymbol{PAP}^T$ is also symmetric and positive definite for any permutation matrix $\boldsymbol{P}$, one can choose to reorder $\boldsymbol{A}$ symmetrically without regard to numerical stability and before the numerical factorization begins. The advantages of this with regard to implementation of the Cholesky method are well known. In particular, very reliable algorithms exist for finding a good $\boldsymbol{P}$, and highly efficient software exists for solving the reordered sparse positive definite systems.

An ideal circumstance would be to find a method that would handle general systems $\boldsymbol{Ax} = \boldsymbol{b}$ while also having the advantages enjoyed by symmetric positive definite systems. Recent work (e.g., Saunders [3, 4], Vanderbei [5]) has the potential to achieve this goal by using the *augmented system*

$$\boldsymbol{K}_\delta \begin{pmatrix} \boldsymbol{s} \\ \boldsymbol{x} \end{pmatrix} = \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{0} \end{pmatrix}, \tag{1}$$

*Received by the editors January 14, 1998; accepted for publication (in revised form) by P. Gill August 16, 1999; published electronically March 6, 2000. This work was supported in part by Natural Sciences and Engineering Research Council of Canada grant OGP 000811.

http://www.siam.org/journals/simax/21-3/33324.html

†Department of Computer Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada (jageorge@sparse.uwaterloo.ca).

‡Faculty of Computational Mathematics and Cybernetics, Moscow State University, Moscow 119899, Russia (ikramov@cmc.msk.su).

where

$$\boldsymbol{K}_\delta = \begin{pmatrix} \delta \boldsymbol{I}_m & \boldsymbol{A} \\ \boldsymbol{A}^T & -\delta \boldsymbol{I}_n \end{pmatrix}, \quad \delta > 0. \tag{2}$$

A solution to (1) provides a solution to the regularized least squares problem $\min_{\boldsymbol{x}} \left(\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 + \|\delta \boldsymbol{x}\|^2 \right)$. If $\boldsymbol{A}$ is square and nonsingular or the system is rectangular but compatible, the factorization of $\boldsymbol{K}_\delta$ can be treated as an approximate factorization of $\boldsymbol{K}_0$, the matrix obtained from $\boldsymbol{K}_\delta$ by setting $\delta = 0$, and iterative refinement can be used on the system

$$\boldsymbol{K}_0 \begin{pmatrix} \boldsymbol{s} \\ \boldsymbol{x} \end{pmatrix} = \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{0} \end{pmatrix} \tag{3}$$

to remove the effect of δ. Thus, if (1) can be solved accurately and efficiently, one has a versatile method for dealing with square and rectangular sparse systems of equations.

The matrix $\boldsymbol{K}_\delta$ is *symmetric quasi-definite* [1, 5]. Generally, a symmetric matrix $\boldsymbol{K}$ with real entries is called quasi-definite, or sqd, if there exists a permutation matrix Π such that $\Pi \boldsymbol{K} \Pi^T$ has the form

$$\begin{pmatrix} \boldsymbol{H} & \boldsymbol{A} \\ \boldsymbol{A}^T & -\boldsymbol{G} \end{pmatrix}, \tag{4}$$

where $\boldsymbol{H}$ is $m \times m$, $\boldsymbol{G}$ is $n \times n$, and both are positive definite. In what follows it is assumed that $\boldsymbol{K}$ itself (rather than some permutation of it) has the form (4).

The key fact is that, for an sqd matrix $\boldsymbol{K}$, any symmetric reordering $\Pi \boldsymbol{K} \Pi^T$ has a "Cholesky factorization" $\boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$, where $\boldsymbol{D}$ is diagonal with both positive and negative elements, *for any choice of permutation matrix* $\boldsymbol{P}$ [5]. It follows that the powerful algorithms and software that have been developed for sparse positive-definite systems can often be used with little or no modification to solve sqd systems. It is evident that when the matrix involved is $\boldsymbol{K}_\delta$, δ must not be too small if this approach is to be successful. In the general case, the norms of $\boldsymbol{H}$ and $\boldsymbol{G}$ must not be too small compared to that of $\boldsymbol{A}$.

The subject of this paper is to investigate the condition number of matrices that are derived from sqd matrices during their factorization, since this information can be used to obtain stability results about the solution to the corresponding system of equations [2]. Following Golub and Van Loan [6], Gill, Saunders, and Shinnerl [1] considered the stability of solving sqd systems. Their work provides a point of departure for the results in this paper. Their Theorem 5.1, which is restated below, is expressed in terms of

$$\omega(\boldsymbol{K}) \equiv \frac{\max\{\|\boldsymbol{A}^T \boldsymbol{H}^{-1} \boldsymbol{A}\|_2, \|\boldsymbol{A}\boldsymbol{G}^{-1}\boldsymbol{A}^T\|_2\}}{\|\boldsymbol{K}\|_2}. \tag{5}$$

Since only the 2-norm is used in this paper, the subscript 2 is omitted in what follows.

THEOREM 1.1 (see [1, Theorem 5.1]). *Let* $\hat{\boldsymbol{d}}$ *be the computed solution of the system* $\boldsymbol{K}\boldsymbol{d} = \boldsymbol{r}$ *with an sqd matrix* $\boldsymbol{K}$. *Then*

$$\frac{\|\boldsymbol{d} - \hat{\boldsymbol{d}}\|}{\|\boldsymbol{d}\|} \leq u\gamma n_K c_K \phi(\boldsymbol{K}), \tag{6}$$

where u *is the unit round-off,* γ *is some scalar of moderate size,* n_K *is the dimension of* $\boldsymbol{K}$, c_K *depends linearly on* n_K, *and*

$$\phi(\boldsymbol{K}) \equiv (1 + \omega(\boldsymbol{K}))cond(\boldsymbol{K}), \tag{7}$$

where

$$cond(\boldsymbol{K}) = \|\boldsymbol{K}\|\|\boldsymbol{K}^{-1}\|.$$

After some preliminary results are stated in section 2 (where Lemma 2.5 and its corollary may be of independent interest), section 3 contains a development showing that $\phi(\boldsymbol{K})$ is a uniform bound on the condition numbers of the Schur complements in the sqd matrix $\boldsymbol{K}$. In section 4, these results are specialized to the matrix $\boldsymbol{K}_\delta$, and their significance is elaborated upon in the final section 5.

Note that in the context of Gaussian elimination for finite-element systems, the relevant Schur complements are often referred to as active submatrices of the elimination.

In the remainder of this article, $\boldsymbol{A}$ is assumed to be an $m \times n$ matrix with $m \geq n$ and singular values

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq \sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_n = 0.$$

Given a matrix $\boldsymbol{M}$, the notation $\sigma_{max}(\boldsymbol{M})$ and $\sigma_{min}(\boldsymbol{M})$ refer, respectively, to the largest and smallest singular values of $\boldsymbol{M}$. Similarly, $\lambda_{max}(\boldsymbol{M})$ and $\lambda_{min}(\boldsymbol{M})$ refer, respectively, to the largest and smallest eigenvalues of $\boldsymbol{M}$.

The notation and definition below are taken from [7]. For index sets $\alpha \subseteq \{1, \ldots, m\}$ and $\beta \subseteq \{1, \ldots, n\}$, we denote the submatrix in the rows of $\boldsymbol{A}$ indexed by α and the columns indexed by β as $\boldsymbol{A}(\alpha, \beta)$. If $m = n$ and $\alpha = \beta$, then $\boldsymbol{A}(\alpha, \alpha)$ is a principal submatrix of $\boldsymbol{A}$ and is abbreviated $\boldsymbol{A}(\alpha)$. Suppose that $\boldsymbol{A}(\alpha)$ is nonsingular. Denote the complement of $\alpha \subseteq \{1, \ldots, n\}$ as α'. Then the matrix

$$\boldsymbol{A}(\alpha') - \boldsymbol{A}(\alpha', \alpha)\boldsymbol{A}(\alpha)^{-1}\boldsymbol{A}(\alpha, \alpha')$$

is called the Schur complement of $\boldsymbol{A}(\alpha)$ in $\boldsymbol{A}$ and denoted by $\boldsymbol{A}/\boldsymbol{A}(\alpha)$.

2. Preliminary results. Some preliminary observations and simple results are required for the development. They are presented as lemmas below.

LEMMA 2.1. *If $\boldsymbol{M}$ is a symmetric (positive or negative) definite matrix, then for any principal submatrix $\widehat{\boldsymbol{M}}$ of $\boldsymbol{M}$,*

$$\sigma_{max}(\widehat{\boldsymbol{M}}) \leq \sigma_{max}(\boldsymbol{M}),$$

$$\sigma_{min}(\widehat{\boldsymbol{M}}) \geq \sigma_{min}(\boldsymbol{M}),$$

and

$$cond(\widehat{\boldsymbol{M}}) \leq cond(\boldsymbol{M}).$$

LEMMA 2.2. *For any nonsingular matrix $\boldsymbol{M}$,*

$$cond(\boldsymbol{M}^{-1}) = cond(\boldsymbol{M}).$$

LEMMA 2.3. *Let $\boldsymbol{C}$ be a nonsingular block matrix*

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{pmatrix},$$

where $\boldsymbol{C}_{11}$ *is nonsingular. Denote* $\boldsymbol{D} = \boldsymbol{C}^{-1}$*, similarly partitioned, by*

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{D}_{11} & \boldsymbol{D}_{12} \\ \boldsymbol{D}_{21} & \boldsymbol{D}_{22} \end{pmatrix}.$$

Then

$$\boldsymbol{D}_{22} = (\boldsymbol{C}/\boldsymbol{C}_{11})^{-1}.$$

LEMMA 2.4 (see Vanderbei [5]). *If* $\boldsymbol{C}$ *is an sqd matrix, then* $\boldsymbol{D} = \boldsymbol{C}^{-1}$ *is an sqd matrix as well.*

LEMMA 2.5. *Let*

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{D}_{11} & \boldsymbol{D}_{12} \\ \boldsymbol{D}_{21} & \boldsymbol{D}_{22} \end{pmatrix}$$

be an sqd matrix with $p \times p$ *block* $\boldsymbol{D}_{11}$ *being positive definite and* $q \times q$ *block* $\boldsymbol{D}_{22}$ *being negative definite. If* $N = p + q$ *and*

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$$

are the eigenvalues of $\boldsymbol{D}$*, then*

$$\lambda_{min}(\boldsymbol{D}_{11}) \leq \lambda_p(\boldsymbol{D}) \tag{8}$$

and

$$\lambda_{max}(\boldsymbol{D}_{22}) \geq \lambda_{p+1}(\boldsymbol{D}). \tag{9}$$

This follows directly from the Courant–Fischer theorem.

COROLLARY 2.6.

$$\sigma_{min}(\boldsymbol{D}) \geq \min\{\sigma_{min}(\boldsymbol{D}_{11}), \sigma_{min}(\boldsymbol{D}_{22})\}.$$

Indeed,

$$\begin{aligned} \sigma_{min}(\boldsymbol{D}) &= \min\{\lambda_p(\boldsymbol{D}), |\lambda_{p+1}(\boldsymbol{D})|\} \\ &\geq \min\{\sigma_{min}(\boldsymbol{D}_{11}) = \lambda_{min}(\boldsymbol{D}_{11}), \sigma_{min}(\boldsymbol{D}_{22}) = |\lambda_{max}(\boldsymbol{D}_{22})|\}. \end{aligned} \tag{10}$$

The comparison of this corollary with Lemma 2.1 shows distinctly the difference between definite and quasi-definite matrices.

3. General sqd matrices. The following assertion is the main result of this paper.

THEOREM 3.1. *Let* $\boldsymbol{K}$ *be an sqd matrix, and let* $\boldsymbol{S}$ *be any principal submatrix of* $\boldsymbol{K}$*. Then*

$$cond(\boldsymbol{K}/\boldsymbol{S}) \leq \phi(\boldsymbol{K}). \tag{11}$$

Proof. Partition $\boldsymbol{K}^{-1}$ consistent with $\boldsymbol{K}$,

$$\boldsymbol{K}^{-1} = \begin{pmatrix} \boldsymbol{E} & \boldsymbol{B} \\ \boldsymbol{B}^T & -\boldsymbol{F} \end{pmatrix}. \tag{12}$$

Consider first the case $\boldsymbol{S} = \boldsymbol{H}$. Then $\boldsymbol{K}/\boldsymbol{H} = -(\boldsymbol{G} + \boldsymbol{A}^T\boldsymbol{H}^{-1}\boldsymbol{A})$, and

$$\begin{aligned} cond(\boldsymbol{K}/\boldsymbol{H}) &= \frac{\|\boldsymbol{G} + \boldsymbol{A}^T\boldsymbol{H}^{-1}\boldsymbol{A}\|}{\sigma_{min}(\boldsymbol{K}/\boldsymbol{H})} = \|\boldsymbol{G} + \boldsymbol{A}^T\boldsymbol{H}^{-1}\boldsymbol{A}\|\|\boldsymbol{F}\| \\ &\le \|\boldsymbol{G} + \boldsymbol{A}^T\boldsymbol{H}^{-1}\boldsymbol{A}\|\|\boldsymbol{K}^{-1}\| \le (\|\boldsymbol{G}\| + \|\boldsymbol{A}^T\boldsymbol{H}^{-1}\boldsymbol{A}\|)\|\boldsymbol{K}^{-1}\| \\ &\le (\|\boldsymbol{K}\| + \|\boldsymbol{A}^T\boldsymbol{H}^{-1}\boldsymbol{A}\|)\|\boldsymbol{K}^{-1}\| = \left(1 + \frac{\|\boldsymbol{A}^T\boldsymbol{H}^{-1}\boldsymbol{A}\|}{\|\boldsymbol{K}\|}\right) cond(\boldsymbol{K}) \\ &\le (1 + \omega(\boldsymbol{K}))cond(\boldsymbol{K}) = \phi(\boldsymbol{K}). \end{aligned} \tag{13}$$

In a similar way, the inequality

$$cond(\boldsymbol{K}/(-\boldsymbol{G})) \le \phi(\boldsymbol{K}) \tag{14}$$

can be proved.

Now let $\boldsymbol{S}$ be an *arbitrary* principal submatrix of $\boldsymbol{K}$. If $\boldsymbol{S}$ contains $\boldsymbol{H}$, then $(\boldsymbol{K}/\boldsymbol{S})^{-1}$ is a principal submatrix of $(\boldsymbol{K}/\boldsymbol{H})^{-1}$. By Lemma 2.1,

$$cond(\boldsymbol{K}/\boldsymbol{S}) = cond((\boldsymbol{K}/\boldsymbol{S})^{-1}) \le cond((\boldsymbol{K}/\boldsymbol{H})^{-1}) = cond(\boldsymbol{K}/\boldsymbol{H}) \le \phi(\boldsymbol{K}). \tag{15}$$

If $\boldsymbol{S}$ contains $-\boldsymbol{G}$, then again

$$cond(\boldsymbol{K}/\boldsymbol{S}) \le cond(\boldsymbol{K}/(-\boldsymbol{G})) \le \phi(\boldsymbol{K}). \tag{16}$$

Thus, one may assume that $\boldsymbol{S}$ contains neither $\boldsymbol{H}$ nor $-\boldsymbol{G}$. Then the matrix $\boldsymbol{D} = (\boldsymbol{K}/\boldsymbol{S})^{-1}$ is an sqd matrix. Let

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{D}_{11} & \boldsymbol{D}_{12} \\ \boldsymbol{D}_{21} & \boldsymbol{D}_{22} \end{pmatrix}$$

with $p \times p$ block $\boldsymbol{D}_{11}$ being positive definite and $q \times q$ block $\boldsymbol{D}_{22}$ being negative definite. According to Corollary 2.6 above,

$$\sigma_{min}(\boldsymbol{D}) \ge \min\{\sigma_{min}(\boldsymbol{D}_{11}), \sigma_{min}(\boldsymbol{D}_{22})\}.$$

Observe that $\boldsymbol{D}_{11}$ (resp., $\boldsymbol{D}_{22}$) is a principal submatrix of $\boldsymbol{E}$ (resp., $\boldsymbol{F}$). Hence, by Lemma 2.1,

$$\min\{\sigma_{min}(\boldsymbol{D}_{11}), \sigma_{min}(\boldsymbol{D}_{22})\} \ge \min\{\sigma_{min}(\boldsymbol{E}), \sigma_{min}(\boldsymbol{F})\}.$$

Combining these two inequalities, one has

$$\sigma_{min}(\boldsymbol{D}) \ge \min\{\sigma_{min}(\boldsymbol{E}), \sigma_{min}(\boldsymbol{F})\}.$$

Using Lemma 2.2 implies

$$\begin{aligned} cond(\boldsymbol{K}/\boldsymbol{S}) = cond(\boldsymbol{D}) &= \|\boldsymbol{D}\|(\sigma_{min}(\boldsymbol{D}))^{-1} \\ &\le \|\boldsymbol{K}^{-1}\|(\min\{\sigma_{min}(\boldsymbol{E}), \sigma_{min}(\boldsymbol{F})\})^{-1} \\ &\le \|\boldsymbol{K}^{-1}\| \max\{\|\boldsymbol{G} + \boldsymbol{A}^T\boldsymbol{H}^{-1}\boldsymbol{A}\|, \|\boldsymbol{H} + \boldsymbol{A}\boldsymbol{G}^{-1}\boldsymbol{A}^T\|\} \le \phi(\boldsymbol{K}). \end{aligned}$$

This completes the proof of the theorem.

4. The matrix K_δ. The following formulas can easily be verified [3]:

$$cond(\boldsymbol{K}_\delta) = \left(\frac{\delta^2+\sigma_1^2}{\delta^2+\sigma_n^2}\right)^{\frac{1}{2}} \qquad \text{if } r = n = m; \tag{17}$$

$$cond(\boldsymbol{K}_\delta) = \frac{(\delta^2+\sigma_1^2)^{\frac{1}{2}}}{\delta} \quad \text{otherwise.} \tag{18}$$

In what follows, it is assumed that $m > n = r$. Hence, (18) will be used as the expression for $cond(\boldsymbol{K}_\delta)$.

Now assume that, in Theorem 3.1 with $\boldsymbol{K} = \boldsymbol{K}_\delta$, one lets $\boldsymbol{S} = \delta \boldsymbol{I}_m$. Then

$$\boldsymbol{S} = \boldsymbol{K}_\delta/(\delta \boldsymbol{I}_m) = -\left(\delta \boldsymbol{I}_n + \frac{1}{\delta}\boldsymbol{A}^T\boldsymbol{A}\right) = -\frac{1}{\delta}(\delta^2 \boldsymbol{I}_n + \boldsymbol{A}^T\boldsymbol{A}) \tag{19}$$

and

$$cond(\boldsymbol{S}) = \frac{\delta^2+\sigma_1^2}{\delta^2+\sigma_n^2}. \tag{20}$$

On the other hand, for $\boldsymbol{S} = -\delta \boldsymbol{I}_n$,

$$\boldsymbol{S} = \boldsymbol{K}_\delta/(-\delta \boldsymbol{I}_n) = \delta \boldsymbol{I}_m + \frac{1}{\delta}\boldsymbol{A}\boldsymbol{A}^T = \frac{1}{\delta}(\delta^2 \boldsymbol{I}_m + \boldsymbol{A}\boldsymbol{A}^T) \tag{21}$$

and

$$cond(\boldsymbol{S}) = \frac{\delta^2+\sigma_1^2}{\delta^2}. \tag{22}$$

The latter is larger than the former and is equal to the square of the condition number of $\boldsymbol{K}_\delta$. However, one can readily verify that even this larger number is less than $\phi(\boldsymbol{K}_\delta)$. Note also that the maximal singular value is

$$(\delta^2+\sigma_1^2)/\delta$$

for both matrices $\boldsymbol{S}$.

Let $\boldsymbol{S}$ be an arbitrary principal submatrix of $\boldsymbol{K}_\delta$. If $\boldsymbol{S}$ contains $\delta \boldsymbol{I}_m$ or $-\delta \boldsymbol{I}_n$, then the argument in Theorem 3.1 applies, and

$$cond(\boldsymbol{K}_\delta/\boldsymbol{S}) \le \max\{cond(\boldsymbol{K}_\delta/\delta \boldsymbol{I}_m), cond(\boldsymbol{K}_\delta/(-\delta \boldsymbol{I}_n))\} = cond^2(\boldsymbol{K}_\delta).$$

Suppose $\boldsymbol{S}$ contains neither $\delta \boldsymbol{I}_m$ nor $-\delta \boldsymbol{I}_n$. Then, repeating the calculations in the final part of the proof of Theorem 3.1, one gets

$$cond(\boldsymbol{K}_\delta/\boldsymbol{S}) \le \|\boldsymbol{K}_\delta^{-1}\|(\delta^2+\sigma_1^2)/\delta = (\delta^2+\sigma_1^2)/\delta^2 = cond^2(\boldsymbol{K}_\delta).$$

Thus, $cond^2(\boldsymbol{K}_\delta)$ gives the uniform bound for the condition numbers of all the Schur complements in the matrix $\boldsymbol{K}_\delta$; this bound is slightly better than

$$\phi(\boldsymbol{K}_\delta) = \left(1 + \frac{1}{\delta}\frac{\sigma_1^2}{(\delta^2+\sigma_1^2)^{1/2}}\right) cond(\boldsymbol{K}_\delta) > cond^2(\boldsymbol{K}_\delta),$$

and the second of the matrices $\boldsymbol{S}$ is conditioned worse than any other Schur complement in $\boldsymbol{K}_\delta$.

5. Final comments. Theorem 1.1 is based on the classical Wilkinson backward round-off error analysis of Gaussian elimination. For some other nonorthogonal methods such as the Gauss–Jordan elimination or the bordering scheme, the local approach is more appropriate for analyzing the round-off behavior. If $\boldsymbol{K}\boldsymbol{x} = \boldsymbol{r}$ is the linear system to be solved, $\boldsymbol{K}_i$ is the matrix obtained after i steps of any algorithm above, and N is the overall number of steps in the algorithm, then the quantity

$$\sum_{i=1}^{N} cond(\boldsymbol{K}_i)$$

is more relevant for this type of analysis than the condition number of the original matrix $\boldsymbol{K}$. (More details on this approach can be found in [2] and the references therein.) From this point of view, the results of the two previous sections may be summarized as follows.

1. Let $\boldsymbol{K}$ be a general sqd matrix. If symmetric Gaussian elimination is applied to $\boldsymbol{K}$, then, whatever the order of elimination chosen, $cond(\boldsymbol{K}_i)$, the condition number of any active submatrix does not exceed $\phi(\boldsymbol{K})$.
2. Symmetric Gaussian elimination is applied to $\boldsymbol{K}_\delta$ in *natural order* if all pivots are chosen from one of the diagonal blocks before any are chosen from the other diagonal block. The result in section 4 shows that a natural order of elimination exhibits the worst case in terms of the conditioning of the active submatrices (and amounts to solving with $(\boldsymbol{A}^T\boldsymbol{A} + \delta^2\boldsymbol{I})$ or $(\boldsymbol{A}\boldsymbol{A}^T + \delta^2\boldsymbol{I})$); any other order of elimination cannot be any worse in terms of the condition of the active submatrices generated during elimination.
3. For $\boldsymbol{K}_\delta$, we have improved slightly on the bound $\phi(\boldsymbol{K}_\delta)$ of [1], thereby improving the error bound (6).

With respect to the first remark above, it is natural to question the relationship between $\sum_{i=1}^{N} cond(\boldsymbol{K}_i)$ and $\phi(\boldsymbol{K})$. We do not claim that our bounds improve on existing ones. Rather, our arguments are directed to showing that one can pivot the systems with matrices of the $\boldsymbol{K}_\delta$ type with regard solely to sparsity considerations, and the numerical quality of the solution obtained should not be worse than if a natural order of elimination is pursued. However, when comparing our bound with $\phi(\boldsymbol{K})$, one should remember that to obtain a bound for the error in the solution, one should multiply $\phi(\boldsymbol{K})$ by the bound for the norm of the equivalent perturbation accumulated during the whole elimination. In the local approach, the condition number of the current matrix is multiplied by the bound for the norm of the local equivalent perturbation only. Eventually, both types of analysis yield about the same results.

Acknowledgment. We thank both referees of our paper for many helpful remarks.

REFERENCES

[1] P. E. Gill, M. A. Saunders, and J. R. Shinnerl, *On the stability of Cholesky factorization for symmetric quasidefinite systems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 35–46.

[2] Kh. D. Ikramov, *Conditionality of the intermediate matrices of the Gauss, Jordan, and optimal elimination methods*, USSR Comput. Math. Math. Phys., 18 (1978), pp. 1–16.

[3] M. A. Saunders, *Solution of sparse rectangular systems using lsqr and craig*, BIT, 25 (1995), pp. 588–604.

[4] M. A. Saunders, *Cholesky based methods for sparse least squares: The benefits of regularizations*, in Proceedings of the AMS-IMS-SIAM Conference on Linear and Nonlinear CG-Related Methods, University of Washington, Seattle, July 1996.

[5] R. J. Vanderbei, *Symmetric quasidefinite matrices*, SIAM J. Optim., 5 (1995), pp. 100–113.
[6] G. H. Golub and C. F. Van Loan, *Unsymmetric positive definite linear systems*, Linear Algebra Appl., 28 (1979), pp. 85–98.
[7] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 3, pp. 978–988
© 2000 Society for Industrial and Applied Mathematics

DECOMPOSITION THEOREMS FOR CONDITIONAL SIGN-SOLVABILITY AND SIGN-SOLVABILITY OF GENERAL SYSTEMS*

SHAWN COKUS† AND VICTOR KLEE†

Abstract. For each $m \times n$ real matrix A, let Null(A) be the right nullspace of A and let $\mathcal{Q}(A)$ be the set of $m \times n$ real matrices with the same sign pattern as A. If it is true for all $\tilde{A} \in \mathcal{Q}(A)$ that (i) Null($\tilde{A}$) = $\{\mathbf{0}\}$, then A is an *L-matrix*, and if it is true for all $\tilde{A} \in \mathcal{Q}(A)$ that (ii) Null($\tilde{A}$) is a line penetrating the open positive orthant of $\mathbb{R}^n$, then A is an *S-matrix*. When (i) or (ii) holds for each $\tilde{A} \in \mathcal{Q}(A)$ and (ii) applies to at least one, then A is a CS-*matrix*. It is known that the study of sign-solvability of linear systems may be decomposed into the study of L- and S-matrices, while the study of conditional sign-solvability may be decomposed into that of L- and CS-matrices. When sign-solvability and conditional sign-solvability are regarded as properties of systems of sign cones in $\mathbb{R}^m$ associated with the relevant matrices' columns, they have natural generalizations to systems of arbitrary subsets of $\mathbb{R}^m$. This paper contains extended decomposition results for arbitrary systems in the case of sign-solvability, and in the case of conditional sign-solvability, for arbitrary systems of subsets in which each subset is relatively open in its linear span.

Key words. sign-solvable, conditionally sign-solvable, decomposition, sign cone, relatively open, L-matrix, S-matrix, CS-matrix, L-system, S-system, CS-system

AMS subject classifications. 15A06, 52A20, 52A37

PII. S0895479898336604

1. Introduction.

A linear system is *sign-solvable* if both the solvability of the system and the sign pattern of its solutions are implied by the sign pattern of its augmented matrix. A system is *conditionally sign-solvable* if the sign pattern of its augmented matrix is consistent with solvability and determines the sign pattern of any solutions, but does not necessarily imply solvability. The study of these and other properties in qualitative matrix theory can be traced to comments of Samuelson [9] concerning the qualitative analysis of economic models; a paper of Kasteleyn [4] on dimer statistics and phase transitions is also relevant. Seminal papers in the study of sign-solvability are those of Bassett, Maybee, and Quirk [1] and Maybee [8]. A comprehensive study of sign-solvability and related notions appears in the book of Brualdi and Shader [3].

Results of Manber [7] and Klee, Ladner, and Manber [5] reduce the study of sign-solvability to the study of two special classes of matrices termed *L-matrices* and *S-matrices*, and results of Brualdi, Chavey, and Shader [2] reduce the study of conditional sign-solvability to that of L-matrices and CS-*matrices*. Here we consider decompositions of these kinds for the "systems" setting introduced by Klee, Von Hohenbalken, and Lewis [6], in which the lists of sign cones in $\mathbb{R}^m$ that appear implicitly in the original theory are replaced by lists of more or less arbitrary subsets of $\mathbb{R}^m$. Terminology, notation, and standing assumptions are set forth in section 2, after which sections 3 and 4 develop our decomposition of conditional sign-solvability and ordinary sign-solvability, respectively. Roughly, the study of sign-solvability for arbitrary

*Received by the editors March 31, 1998; accepted for publication (in revised form) by R. Brualdi March 16, 1999; published electronically March 6, 2000. The research of both authors was supported in part by the National Science Foundation.

http://www.siam.org/journals/simax/21-3/33660.html

†Department of Mathematics, University of Washington, Box 354350, Seattle, WA 98195-4350 (cokus@math.washington.edu, klee@math.washington.edu).

systems of sets is reduced here to the study of L- and S-systems (see Theorem 4.8), while the study of conditional sign-solvability of systems of subsets in which each subset is relatively open in its linear span is reduced to the study of L- and CS-systems of such subsets (see Corollary 3.18). Thus the "sign cone assumptions" of earlier decomposition theorems are greatly weakened.

2. Terminology, notation, and standing assumptions. The base field for our vector spaces is $\mathbb{R}$, the reals. We use column vectors, $\mathbf{0}$ for the origin, $\top$ for transposition, and Span for linear span with $\mathrm{Span}(\emptyset) = \{\mathbf{0}\}$ by convention. $\mathbb{R}_+$ denotes the set of (strictly) positive reals and $\mathbb{R}_-$ the (strictly) negative reals. We use $\setminus$ for set difference, and when necessary to avoid confusion, $\wedge$ for logical "and". The precedence of $\wedge$ is higher than that of the implications and biconditional so that, for instance, "$P \iff Q \wedge R$" means "$P \iff (Q$ and $R)$" as opposed to "$(P \Leftrightarrow Q)$ and R." For integers a and c, define $a..c := \{\,\text{integers } b \mid a \le b \text{ and } b \le c\,\}$.

Let m and n be positive integers. If A is an $m \times n$ matrix and $B \subseteq \mathbb{R}^m$, the notation $A^{-1}B := \{\, x \in \mathbb{R}^n \mid Ax \in B \,\}$ is helpful. To illustrate, the nullspace $\mathrm{Null}(A)$ of A is $A^{-1}\{\mathbf{0}\}$. An $m \times n$ *system* is a list $\mathcal{A} := (A_1, \ldots, A_n)$ of nonempty subsets of $\mathbb{R}^m$. For each such $\mathcal{A}$ we define

$$\mathcal{M}(\mathcal{A}) := \{\, m \times n \text{ matrices } [\, a_1 \cdots a_n \,] \mid a_1 \in A_1, \ \ldots\ , a_n \in A_n \,\}.$$

For an $m \times n$ matrix A, the conditions of A having linearly independent columns, $\mathrm{Null}(A) = \{\mathbf{0}\}$, and the n columns of A forming with $\mathbf{0}$ the vertices of an n-dimensional simplex in $\mathbb{R}^m$ are equivalent. If every matrix in $\mathcal{M}(\mathcal{A})$ satisfies these conditions, then $\mathcal{A}$ is an *L-system*. The condition of there existing $x \in \mathbb{R}^n_+$ with $\mathrm{Null}(A) = \mathbb{R}x$ is equivalent to that of the n columns of A being the vertices of an $(n-1)$-dimensional simplex in $\mathbb{R}^m$ with $\mathbf{0}$ in its relative interior. If every member of $\mathcal{M}(\mathcal{A})$ satisfies this condition, then $\mathcal{A}$ is an *S-system*. If the set of members of $\mathcal{M}(\mathcal{A})$ that satisfy this condition is nonempty and $\mathrm{Null}(A) = \{\mathbf{0}\}$ for all other members (if any) of $\mathcal{M}(\mathcal{A})$, then $\mathcal{A}$ is a CS-*system*. By convention, the empty list $(\,)$ is an L-system.

As the term is used here, a *cone* is a set that is closed under scalar multiplication by elements of $\mathbb{R}_+$. The *sign cones of* $\mathbb{R}^n$ are the members of

$$\mathcal{X}_n := \{\, X_1 \times \cdots \times X_n \subseteq \mathbb{R}^n \mid X_1, \ \ldots\ , X_n \in \{\mathbb{R}_+, \{0\}, \mathbb{R}_-\} \,\}.$$

They partition $\mathbb{R}^n$ into 3^n nonempty convex cones, exactly one of which contains $\mathbf{0}$ and every one of which has the property of being relatively open in its linear span.

An *augmented $m \times n$ system* is merely an $m \times (n+1)$ system $(A_1, \ldots, A_n; B)$ in which we think of the last set B as playing a special role. Such an augmented system is *sign-solvable* if the solution set $A^{-1}\{b\}$ to $Ax = b$ is always nonempty and is always included in a fixed sign cone $X \in \mathcal{X}_n$ as A varies over $\mathcal{M}(A_1, \ldots, A_n)$ and b varies over B. If we relax this somewhat and permit occurrences of $A^{-1}\{b\} = \emptyset$, as long as not every pair (A, b) is of this new type, then the augmented system is *conditionally sign-solvable*. In both cases, X is clearly unique when it exists. The "augmented $m \times 0$ system" $(\,; B)$ is conditionally sign-solvable, by convention, if and only if $\mathbf{0} \in B$.

Our notion of sign-solvability includes the usual one (as given in [3], [5], [7], [8], etc.) as a special case, namely, that in which B and all the A_i are sign cones. Similarly, our definition of conditional sign-solvability generalizes the concept introduced by Brualdi, Chavey, and Shader in [2].

Observe that both the ordinary and conditional sign-solvability of an augmented system are unaffected by permutation of the A_i, except for solution vectors being

altered by having the same permutation applied to their coordinates. The same is true for the replacement of A_i by $-A_i$ for any $i \in 1..n$, but here the effect is to negate the ith coordinate of solution vectors. Hence, by suitable permutation and negation, any ordinarily or conditionally sign-solvable system may be brought into a form whose solution vectors have only positive components followed by zero components. This motivates the following definition: We say an augmented system $(A_1, \dots, A_n; B)$ is *in h-form* for $h \in 0..n$ if there exist $A \in \mathcal{M}(A_1, \dots, A_n)$ and $x := [\, x_1 \cdots x_n \,]^\top \in \mathbb{R}^n$ and $b \in B$ so that $Ax = b$ and $x_1, \dots, x_h > 0$ and $x_{h+1} = \cdots = x_n = 0$.

It is efficient to establish a common environment in which to state and prove all our propositions, lemmas, corollaries, and theorems. We presently introduce notation and make standing assumptions for this purpose. For typographical considerations we will make frequent use of shorthands such as

$$\begin{array}{lll} \xi_{i..j} \in \mathbb{R} & \text{for} & \xi_i \in \mathbb{R}, \dots, \xi_j \in \mathbb{R}, \\ a_{i..j} \in A_{i..j} & \text{for} & a_i \in A_i, \dots, a_j \in A_j, \qquad \text{and} \\ \xi_{i..j} \cdot a_{i..j} & \text{for} & \xi_i a_i + \cdots + \xi_j a_j. \end{array}$$

Let $(A_1, \dots, A_n; B)$ be an augmented $m \times n$ system in h-form for some $h \in 0..n$. By h-form, choose $\tilde{\xi}_{1..n} \in \mathbb{R}$, $\tilde{a}_{1..n} \in A_{1..n}$, and $\tilde{b} \in B$ with $\tilde{\xi}_{1..n} \cdot \tilde{a}_{1..n} = \tilde{b}$, $\tilde{\xi}_{1..h} > 0$, and $\tilde{\xi}_{(h+1)..n} = 0$. Put $V := \operatorname{Span}(A_1 \cup \cdots \cup A_h)$, let W be a subspace supplementary to V in $\mathbb{R}^m$, and write proj_W for the unique linear map from $\mathbb{R}^m$ to W that is the identity on W and has kernel V. Put $T := \mathbb{R}A_1 + \cdots + \mathbb{R}A_h \subseteq V$ and $U := \mathbb{R}A_{h+1} + \cdots + \mathbb{R}A_n$, and define

$$\xi^\flat := \begin{cases} \mathbb{R} & \text{if } \xi \geq 0 \\ \mathbb{R}_- & \text{if } \xi < 0 \end{cases} \qquad \text{and} \qquad \xi^* := \begin{cases} \mathbb{R} & \text{if } \xi = 0 \\ \mathbb{R} \setminus \{0\} & \text{if } \xi \neq 0 \end{cases}$$

for every $\xi \in \mathbb{R}$. Then $(\xi' \in \xi^\flat) \wedge (\xi' \geq 0) \Longrightarrow \xi \geq 0$ and $(\xi' \in \xi^*) \wedge (\xi' = 0) \Longrightarrow \xi = 0$.

We close this section with what we must agree is a long list of abbreviations. Some rather technical conditions are included, but every one of the sixteen appears to be essential for full elucidation of the relationship of ordinary and conditional sign-solvability to L-, S-, and CS-systems. For orientation, the reader may want at this time to spend a few minutes examining the statements of Theorem 3.15, Corollary 3.18, and Theorem 4.8.

$\mathbf{CSS_n} :\Longleftrightarrow (A_1, \dots, A_n; B)$ is conditionally sign-solvable
$\mathbf{CSS_h} :\Longleftrightarrow (A_1, \dots, A_h; B)$ is conditionally sign-solvable
$\mathbf{CS_h} :\Longleftrightarrow (A_1, \dots, A_h, -B)$ is a CS-system
$\mathbf{SS_n} :\Longleftrightarrow (A_1, \dots, A_n; B)$ is sign-solvable
$\mathbf{S_h} :\Longleftrightarrow (A_1, \dots, A_h, -B)$ is an S-system
$\mathbf{L_n} :\Longleftrightarrow (A_1, \dots, A_n)$ is an L-system
$\mathbf{L_h} :\Longleftrightarrow (A_1, \dots, A_h)$ is an L-system
$\mathbf{L_{h+1}} :\Longleftrightarrow (A_{h+1}, \dots, A_n)$ is an L-system
$\mathbf{L_{h+1}^W} :\Longleftrightarrow (\operatorname{proj}_W A_{h+1}, \dots, \operatorname{proj}_W A_n)$ is an L-system
$\mathbf{L_{NS}} :\Longleftrightarrow (A_1, \dots, A_n, B \setminus V)$ is an L-system, where $B \setminus V$ is omitted if empty
$\mathbf{NS} :\Longleftrightarrow (T + U) \cap B \subseteq V$
$\mathbf{XL} :\Longleftrightarrow (T - B) \cap U = \{\mathbf{0}\}$

$$\mathbf{UC} :\Longleftrightarrow \forall a_{1..h} \in A_{1..h},\ \ B \subseteq \mathbb{R}a_1 + \cdots + \mathbb{R}a_h$$

$$\mathbf{M_n} :\Longleftrightarrow \forall \xi_{1..n} \in \mathbb{R},\ \ \mathbf{0} \in \xi_{1..n} \cdot A_{1..n} \Longrightarrow \mathbf{0} \in \xi^{\flat}_{1..n} \cdot A_{1..n} - B$$

$$\mathbf{M_h} :\Longleftrightarrow \forall \xi_{1..h} \in \mathbb{R},\ \ \mathbf{0} \in \xi_{1..h} \cdot A_{1..h} \Longrightarrow \mathbf{0} \in \xi^{\flat}_{1..h} \cdot A_{1..h} + U - B$$

$$\mathbf{M^W_{h+1}} :\Longleftrightarrow \forall \xi_{(h+1)..n} \in \mathbb{R},\ \ \mathbf{0} \in \mathrm{proj}_W \textstyle\sum_{i=h+1}^{n} \xi_i A_i \Longrightarrow \mathbf{0} \in T + \sum_{i=h+1}^{n} \xi^*_i A_i - B$$

The first nine conditions are all natural generalizations of standard conditions used for characterizing ordinary and conditional sign-solvability. This is perhaps least evident in the case of L^W_{h+1}, but the common operation of selecting a subset of coordinates that appears in earlier characterizations is a form of projection, and in cases where V and W are cartesian products of various copies of $\mathbb{R}$ and $\{0\}$, the connection is clear. As results are developed in sections 3 and 4, the need for the latter seven conditions will become apparent.

Many of the conditions can be expressed in alternate forms that may aid the reader's understanding. Several such forms are given below, but it is surprisingly troublesome to express many of the conditions in alternate forms precisely. The original definitions were chosen in large part for their brevity and precision.

$\mathrm{L}^W_{h+1} \Longleftrightarrow$ if $A \in \mathcal{M}(A_1, \ldots, A_n)$ and $x \in \mathbb{R}^n$ are such that $Ax \in V$ and $x_{1..h} = 0$, then $x = \mathbf{0}$

$\mathrm{L_{NS}} \Longleftrightarrow$ $(A_1, \ldots, A_n)$ is an L-system, and for any $A \in \mathcal{M}(A_1, \ldots, A_n)$ and $b \in B$, if $Ax = b$ has a solution for $x \in \mathbb{R}^n$, then $b \in V$

$\mathrm{NS} \Longleftrightarrow$ for any $A \in \mathcal{M}(A_1, \ldots, A_n)$ and $b \in B$, if $Ax = b$ has a solution for $x \in \mathbb{R}^n$, then $b \in V$

$\mathrm{XL} \Longleftrightarrow$ for any $A \in \mathcal{M}(A_1, \ldots, A_n)$ and $x \in \mathbb{R}^n$ and $b \in B$, if $Ax = b$, then $Ax' = b$, where $x'_{1..h} := x_{1..h}$ and $x'_{(h+1)..n} := 0$

$\mathrm{UC} \Longleftrightarrow$ B is included in the column space of every member of $\mathcal{M}(A_1, \ldots, A_n)$

$\mathrm{M}_n \Longleftrightarrow$ if $A \in \mathcal{M}(A_1, \ldots, A_n)$ and $x \in \mathbb{R}^n$ are such that $Ax = \mathbf{0}$, then there is $A' \in \mathcal{M}(A_1, \ldots, A_n)$ and $x' \in \mathbb{R}^n$ so that $A'x' \in B$ and $x'_i < 0$ whenever $x_i < 0$ for $i \in 1..n$

$\mathrm{M}_h \Longleftrightarrow$ if $A \in \mathcal{M}(A_1, \ldots, A_n)$ and $x \in \mathbb{R}^n$ with $x_{(h+1)..n} = 0$ are such that $Ax = \mathbf{0}$, then there is $A' \in \mathcal{M}(A_1, \ldots, A_n)$ and $x' \in \mathbb{R}^n$ so that $A'x' \in B$ and $x'_i < 0$ whenever $x_i < 0$ for $i \in 1..h$

$\mathrm{M}^W_{h+1} \Longleftrightarrow$ if $A \in \mathcal{M}(A_1, \ldots, A_n)$ and $x \in \mathbb{R}^n$ with $x_{1..h} = 0$ are such that $Ax \in V$, then there is $A' \in \mathcal{M}(A_1, \ldots, A_n)$ and $x' \in \mathbb{R}^n$ so that $A'x' \in B$ and $x'_i \neq 0$ whenever $x_i \neq 0$ for $i \in (h+1)..n$

3. Conditional sign-solvability. This section develops our characterization of conditional sign-solvability. As it describes the interplay of so many conditions, it may be difficult to follow initially. We suggest that on first reading one should ignore all the proofs; the somewhat unusual format of this paper has been arranged to facilitate this approach. In doing this, one will find a continuous narrative that describes the interplay and contains examples that further illustrate the role of certain conditions.

Our view is that there are essentially two ingredients of conditional sign-solvability of systems in h-form:

(1) a relationship between $A_{1..h}$ and B, responsible for the positive components of solution vectors, and

(2) the "noninterference" of $A_{(h+1)..n}$ with (1), so as not to introduce new nonzero components in such vectors.

While perhaps never stated as such, these ingredients have been present in earlier characterizations. Matrix-based approaches typically conclude with the finding that

a certain matrix must have an L-type property while another must be of S- or CS-type. The S- and CS-type conditions generally assert the existence of solutions to $Ax = b$ along the lines of (1), while the L-type conditions forbid other kinds of solutions, as in (2). Graph-theoretic approaches have a similar structure, one statement giving a requirement on the signedness of directed cycles in a certain signed directed graph and another specifying a restriction on "chains," or dipaths, in such a graph; the former correspond with (2) and the latter with (1). However, for our purposes we need to refine (2) into the following two parts:

(2a) the need for an "internal consistency" within $A_{(h+1)..n}$, and

(2b) the presence of a "cross consistency" involving $A_{1..h}$ and B as well.

A precise formulation of the general view is then given by the following.

PROPOSITION 3.1. $\mathrm{CSS}_n \iff \mathrm{CSS}_h \wedge \mathrm{L}_{h+1} \wedge \mathrm{XL}$.

Proof. $\Longrightarrow$**:** Assume CSS_n. *To show* CSS_h: If $h = 0$, then $\mathbf{0} = \tilde{\xi}_{1..n} \cdot \tilde{a}_{1..n} = \tilde{b} \in B$ and we are done; take $h > 0$ without loss. Recall $[\,\tilde{a}_1 \cdots \tilde{a}_h\,]\,[\,\tilde{\xi}_1 \cdots \tilde{\xi}_h\,]^\top = \tilde{b}$. Suppose $\xi_{1..h} \in \mathbb{R}$ and $a_{1..h} \in A_{1..h}$ and $b \in B$ are such that $\xi_{1..h} \cdot a_{1..h} = b$. It suffices to show $\xi_{1..h} > 0$, but this follows from

$$[\,a_1 \cdots a_h\ \tilde{a}_{h+1} \cdots \tilde{a}_n\,]\,[\,\xi_1 \cdots \xi_h\ 0 \cdots 0\,]^\top = b,$$

as we have CSS_n and h-form.

To show L_{h+1}: Suppose $\xi_{(h+1)..n} \in \mathbb{R}$ and $a_{(h+1)..n} \in A_{(h+1)..n}$ satisfy $\xi_{(h+1)..n} \cdot a_{(h+1)..n} = \mathbf{0}$. Since $[\,\tilde{a}_1 \cdots \tilde{a}_h\ a_{h+1} \cdots a_n\,]\,[\,\tilde{\xi}_1 \cdots \tilde{\xi}_h\ \xi_{h+1} \cdots \xi_n\,]^\top = \tilde{b}$, we have $\xi_{(h+1)..n} = 0$ by CSS_n and h-form.

To show XL: We first show "$\subseteq \{\mathbf{0}\}$." Let $\xi_{1..n} \in \mathbb{R}$ and $a_{1..n} \in A_{1..n}$ and $b \in B$ be such that $\xi_{1..h} \cdot a_{1..h} - b = \xi_{(h+1)..n} \cdot a_{(h+1)..n}$. It is enough to show $\xi_{(h+1)..n} = 0$, but we have this as we have $[\,a_1 \cdots a_n\,]\,[\,\xi_1 \cdots \xi_h\ -\xi_{h+1} \cdots -\xi_n\,]^\top = b$ and CSS_n and h-form. To see "$\supseteq \{\mathbf{0}\}$," observe $\tilde{\xi}_{1..h} \cdot \tilde{a}_{1..h} - \tilde{b} = 0\,\tilde{a}_{h+1} + \cdots + 0\,\tilde{a}_n = \mathbf{0}$.

$\Longleftarrow$**:** Assume CSS_h, L_{h+1}, and XL. Recall $[\,\tilde{a}_1 \cdots \tilde{a}_n\,]\,[\,\tilde{\xi}_1 \cdots \tilde{\xi}_n\,]^\top = \tilde{b}$. Let $\xi_{1..n} \in \mathbb{R}$ and $a_{1..n} \in A_{1..n}$ and $b \in B$ satisfy $\xi_{1..h} \cdot a_{1..h} + \xi_{(h+1)..n} \cdot a_{(h+1)..n} = b$. It suffices to show $\xi_{1..h} > 0$ and $\xi_{(h+1)..n} = 0$. We have $-\xi_{(h+1)..n} \cdot a_{(h+1)..n} = \mathbf{0}$ by XL and so $\xi_{(h+1)..n} = 0$ by L_{h+1}. Thus, $\xi_{1..h} \cdot a_{1..h} = b$. By CSS_h and h-form we then also have $\xi_{1..h} > 0$. ∎

We now proceed to reduce CSS_h, L_{h+1}, and XL—the expressions in Proposition 3.1 of ingredients (1), (2a), and (2b), respectively—to conditions more like those that appear in earlier results such as Theorem 2.6 of [2]. We begin with CSS_h, but as is shown by the augmented system $(\,\{(1,0)\},\ \{(1,0),\ (0,1)\};\ \{(1,1)\}\,)$ in 2-form, for which CSS_h (and even CSS_n) holds but CS_h does not, we cannot hope to directly replace CSS_h by CS_h. To pinpoint the difficulty we have the following.

PROPOSITION 3.2. $\mathrm{CS}_h \iff \mathrm{CSS}_h \wedge \mathrm{L}_h$.

Proof. Assume $h > 0$ since the case $h = 0$ is clear. $\Longrightarrow$**:** Assume CS_h. *To show* CSS_h: Recall $[\,\tilde{a}_1 \cdots \tilde{a}_h\,]\,[\,\tilde{\xi}_1 \cdots \tilde{\xi}_h\,]^\top = \tilde{b}$. Let $\xi_{1..h} \in \mathbb{R}$ and $a_{1..h} \in A_{1..h}$ and $b \in B$ be such that $\xi_{1..h} \cdot a_{1..h} = b$. We are done if $\xi_{1..h} > 0$, but this is immediate from $\mathbf{0} \neq [\,\xi_1 \cdots \xi_h\ 1\,]^\top \in \mathrm{Null}[\,a_1 \cdots a_h\ -b\,]$ and CS_h.

To show L_h: Suppose $\xi_{1..h} \in \mathbb{R}$ and $a_{1..h} \in A_{1..h}$ are such that $\xi_{1..h} \cdot a_{1..h} = \mathbf{0}$. Since $[\,\xi_1 \cdots \xi_h\ 0\,]^\top \in \mathrm{Null}[\,a_1 \cdots a_h\ -\tilde{b}\,]$, we have $\xi_{1..h} = 0$ by CS_h.

$\Longleftarrow$**:** Assume CSS_h and L_h. Note $\mathbf{0} \neq [\,\tilde{\xi}_1 \cdots \tilde{\xi}_h\ 1\,]^\top \in \mathrm{Null}[\,\tilde{a}_1 \cdots \tilde{a}_h\ -\tilde{b}\,]$. Let $a_{1..h} \in A_{1..h}$ and $b \in B$. If $\mathrm{Null}[\,a_1 \cdots a_h\ -b\,] = \{\mathbf{0}\}$, we are done;

otherwise, choose $\xi'_{1..h}$, $c' \in \mathbb{R}$ not all 0 with $\xi'_{1..h} \cdot a_{1..h} - c'b = \mathbf{0}$. By L_h, $c' \neq 0$ so we may divide through by c' and assume $c' = 1$ without loss. Taking b to the other side and using CSS_h and h-form, we find $\xi'_{1..h} > 0$. Since we know $\mathrm{Null}[\, a_1 \cdots a_h \; -b\,] \supseteq \mathbb{R}[\, \xi'_1 \cdots \xi'_h \; 1\,]^\top$, it now suffices to show the reverse inclusion. Let $\xi_{1..h}, c \in \mathbb{R}$ satisfy $\xi_{1..h} \cdot a_{1..h} - c\, b = \mathbf{0}$. If $c = 0$, we are done, as then $\xi_{1..h} = 0$ as well by L_h. Otherwise, $(\xi'_1 - \xi_1/c)a_1 + \cdots + (\xi'_h - \xi_h/c)a_h = \mathbf{0}$. By L_h, each coefficient is 0. Hence, $[\, \xi_1 \cdots \xi_h \; c\,]^\top = c\,[\, \xi'_1 \cdots \xi'_h \; 1\,]^\top$. ∎

We thus require a condition able to supply L_h, for which M_h is well-suited.

PROPOSITION 3.3. $\mathrm{CSS}_n \wedge \mathrm{M}_h \Longrightarrow \mathrm{L}_h$.

Proof. Assume CSS_n and M_h. Suppose $\xi_{1..h} \in \mathbb{R}$ and $a_{1..h} \in A_{1..h}$ are such that $\xi_{1..h} \cdot a_{1..h} = \mathbf{0}$. It is enough to show $\xi_{1..h} \geq 0$, as then repetition of our argument with $-\xi_{1..h}$ will give $\xi_{1..h} \leq 0$ as well and we will have $\xi_{1..h} = 0$. By M_h, find $\xi'_{1..h} \in \xi^\flat_{1..h}$ and $\xi'_{(h+1)..n} \in \mathbb{R}$ and $a'_{1..n} \in A_{1..n}$ and $b' \in B$ such that $\xi'_{1..n} \cdot a'_{1..n} = b'$. By CSS_n and h-form, $\xi'_{1..h} > 0$ so that $\xi_{1..h} \geq 0$ as needed. ∎

The converse to Proposition 3.3 is false in general, as is shown by the augmented system $(\, \{(1,0)\}, \{(0,1)\}; \{(1,0), (0,1)\}\,)$ in 1-form. However, we do have the partial converse given by the following.

PROPOSITION 3.4. $\mathrm{L}_h \Longrightarrow \mathrm{M}_h$.

Proof. Assume L_h and let $\xi_{1..h} \in \mathbb{R}$ satisfy $\mathbf{0} \in \xi_{1..h} \cdot A_{1..h}$. By L_h, $\xi_{1..h} = 0$. Hence,

$$\mathbf{0} = \tilde{\xi}_{1..n} \cdot \tilde{a}_{1..n} - \tilde{b} \in \mathbb{R}A_1 + \cdots + \mathbb{R}A_n - B = \xi^\flat_{1..h} \cdot A_{1..h} + U - B. \qquad ∎$$

Since CS_h is our ultimate expression of ingredient (1), we are ready to concentrate on L_{h+1} and XL. But, in anticipation of the use of forthcoming Proposition 3.14 in the proof of Theorem 3.15, we first take this opportunity to show that when M_h is strengthened to M_n, we have the advantage of L_n over L_h.

PROPOSITION 3.5. $\mathrm{CSS}_n \wedge \mathrm{M}_n \Longrightarrow \mathrm{L}_n$.

Proof. Assume CSS_n and M_n. Let $\xi_{1..n} \in \mathbb{R}$, $a_{1..n} \in A_{1..n}$ satisfy $\xi_{1..n} \cdot a_{1..n} = \mathbf{0}$. It is sufficient to show $\xi_{1..n} \geq 0$, for then repeating the argument with $-\xi_{1..n}$ will give $\xi_{1..n} \leq 0$ as well and we will have $\xi_{1..n} = 0$. By M_n, choose $\xi'_{1..n} \in \xi^\flat_{1..n}$ and $a'_{1..n} \in A_{1..n}$ and $b' \in B$ so that $\xi'_{1..n} \cdot a'_{1..n} = b'$. Using CSS_n and h-form, we find $\xi'_{1..h} > 0$ and $\xi'_{(h+1)..n} = 0$ and thus we conclude $\xi_{1..n} \geq 0$ as desired. ∎

We say "strengthened" because of the following proposition, corollary, and also because M_h is not enough to guarantee L_n. This last fact is evident from the augmented system $(\, \{(1,0), (0,1)\}, \{(0,1)\}; \{(1,0)\}\,)$ in 1-form for which CSS_n and M_h hold but L_n does not.

PROPOSITION 3.6. $\mathrm{M}_n \Longrightarrow \mathrm{M}_h$.

Proof. Assume M_n and suppose $\xi_{1..h} \in \mathbb{R}$ are such that $\mathbf{0} \in \xi_{1..h} \cdot A_{1..h}$. If we take $\xi_{(h+1)..n} := 0$, then we may apply M_n to $\mathbf{0} \in \xi_{1..n} \cdot A_{1..n}$ and conclude $\mathbf{0} \in \xi^\flat_{1..n} \cdot A_{1..n} - B = \xi^\flat_{1..h} \cdot A_{1..h} + U - B$. ∎

LEMMA 3.7. *Let $\mathcal{A}$ be a finite list of nonempty subsets of $\mathbb{R}^m$ and let $\mathcal{A}'$ be $\mathcal{A}$ with any number of terms deleted. If $\mathcal{A}$ is an L-system, then $\mathcal{A}'$ is also an L-system.*

Proof. Suppose $\mathcal{A}$ is an L-system and let $A' \in \mathcal{M}(\mathcal{A}')$. Choose a matrix $A \in \mathcal{M}(\mathcal{A})$ whose list of columns is a superlist of the columns of A'. Since the columns of A are linearly independent by assumption and a sublist of a

linearly independent list is linearly independent, the columns of A' are linearly independent as needed. □

COROLLARY 3.8. $\mathrm{L}_n \Longrightarrow \mathrm{L}_h \wedge \mathrm{L}_{h+1}$.

Proof. Apply Lemma 3.7 twice. □

The converse to Proposition 3.5 is false (by the same counterexample as given for Proposition 3.3), but we do have as a partial converse the following.

PROPOSITION 3.9. $\mathrm{L}_n \Longrightarrow \mathrm{M}_n$.

Proof. Assume L_n. Suppose $\xi_{1..n} \in \mathbb{R}$ satisfy $\mathbf{0} \in \xi_{1..n} \cdot A_{1..n}$. By L_n, $\xi_{1..n} = 0$ so that $\mathbf{0} = \tilde{\xi}_{1..n} \cdot \tilde{a}_{1..n} - \tilde{b} \in \mathbb{R}A_1 + \cdots + \mathbb{R}A_n - B = \xi^\flat_{1..n} \cdot A_{1..n} - B$. □

We now consider L^W_{h+1} as an agent of ingredient (2). The augmented system ($\{(1,0,0,0), (0,1,0,0), (0,0,1,0)\}$, $\{(0,1,0,1)\}$, $\{(0,0,1,1)\}$; $\{(1,0,0,0)\}$) in 1-form, with $V = \mathbb{R}^3 \times \{0\}$ and $W = \{0\}^3 \times \mathbb{R}$, satisfies CSS_n, M_h, and M_n but not L^W_{h+1}, and so we are led to introduce M^W_{h+1} in the following.

PROPOSITION 3.10. $\mathrm{L}_{h+1} \wedge \mathrm{XL} \wedge \mathrm{M}^W_{h+1} \Longrightarrow \mathrm{L}^W_{h+1}$.

Proof. Assume L_{h+1}, XL, and M^W_{h+1}. Let $\xi_{(h+1)..n} \in \mathbb{R}$ and $a_{(h+1)..n} \in A_{(h+1)..n}$ be such that $\xi_{(h+1)..n} \cdot \mathrm{proj}_W\, a_{(h+1)..n} = \mathbf{0}$. By M^W_{h+1}, choose $t' \in T$ and $\xi'_{(h+1)..n} \in \xi^*_{(h+1)..n}$ and $a'_{(h+1)..n} \in A_{(h+1)..n}$ and $b' \in B$ with $t' - b' = -\xi'_{(h+1)..n} \cdot a'_{(h+1)..n}$. Then $\xi'_{(h+1)..n} \cdot a'_{(h+1)..n} = \mathbf{0}$ by XL and so $\xi'_{(h+1)..n} = 0$ by L_{h+1}. Hence, $\xi_{(h+1)..n} = 0$. □

As it stands, for Proposition 3.10 to be useful we need to be able to recover at least L_{h+1} and XL from L^W_{h+1}. We do have the following.

PROPOSITION 3.11. $\mathrm{L}^W_{h+1} \Longrightarrow \mathrm{L}_{h+1} \wedge \mathrm{M}^W_{h+1}$.

Proof. Assume L^W_{h+1}. *To show* L_{h+1}: Let $\xi_{(h+1)..n} \in \mathbb{R}$ and $a_{(h+1)..n} \in A_{(h+1)..n}$ satisfy $\xi_{(h+1)..n} \cdot a_{(h+1)..n} = \mathbf{0}$. Then $\xi_{(h+1)..n} \cdot \mathrm{proj}_W\, a_{(h+1)..n} = \mathbf{0}$ and so $\xi_{(h+1)..n} = 0$ by L^W_{h+1}.

To show M^W_{h+1}: Let $\xi_{(h+1)..n} \in \mathbb{R}$ satisfy $\mathbf{0} \in \mathrm{proj}_W(\xi_{(h+1)..n} \cdot A_{(h+1)..n})$. Then by L^W_{h+1} we have $\xi_{(h+1)..n} = 0$, whereby $\mathbf{0} = \tilde{\xi}_{1..n} \cdot \tilde{a}_{1..n} - \tilde{b} \in \mathbb{R}A_1 + \cdots + \mathbb{R}A_n = T + \xi^*_{(h+1)..n} \cdot A_{(h+1)..n} - B$. □

However, the augmented system ($\{(1,0)\}$, $\{(0,1)\}$; $\{(1,0),\ (1,1)\}$) in 1-form with $V = \mathbb{R} \times \{0\}$ and $W = \{0\} \times \mathbb{R}$ satisfies L^W_{h+1} but not XL. We thus conclude L^W_{h+1} is lacking as an expression of (2b) and look for a small condition able to retain what XL has but L^W_{h+1} may lose. Thus, we discover NS.

PROPOSITION 3.12. $\mathrm{XL} \Longrightarrow \mathrm{NS}$.

Proof. Assume XL and let $t \in T$ and $u \in U$ and $b \in B$ be such that $t + u = b$. Since $t - b = -u \in U$, we have $u = \mathbf{0}$ by XL. Hence, $b = t \in V$. □

PROPOSITION 3.13. $\mathrm{L}^W_{h+1} \wedge \mathrm{NS} \Longrightarrow \mathrm{XL}$.

Proof. Assume L^W_{h+1} and NS. Let $t \in T \subseteq V$ and $b \in B$, along with $\xi_{(h+1)..n} \in \mathbb{R}$ and $a_{(h+1)..n} \in A_{(h+1)..n}$, be given and put $u := \xi_{(h+1)..n} \cdot a_{(h+1)..n} \in U$. Suppose $t - b = u$. Since $t + (-u) = b$ and $-u \in U$, we have $b \in V$ by NS. Hence, $u = t - b \in V$ and so $\mathrm{proj}_W\, u = \mathbf{0}$. Thus, $\xi_{(h+1)..n} = 0$ by L^W_{h+1} and $u = \mathbf{0}$. □

Finally, results that are expressed as far as possible in terms of specific systems being CS- or *L*- are preferred to those that are not, and condition NS leads to the latter. Provided L_n is available, we have a remedy by way of the following.

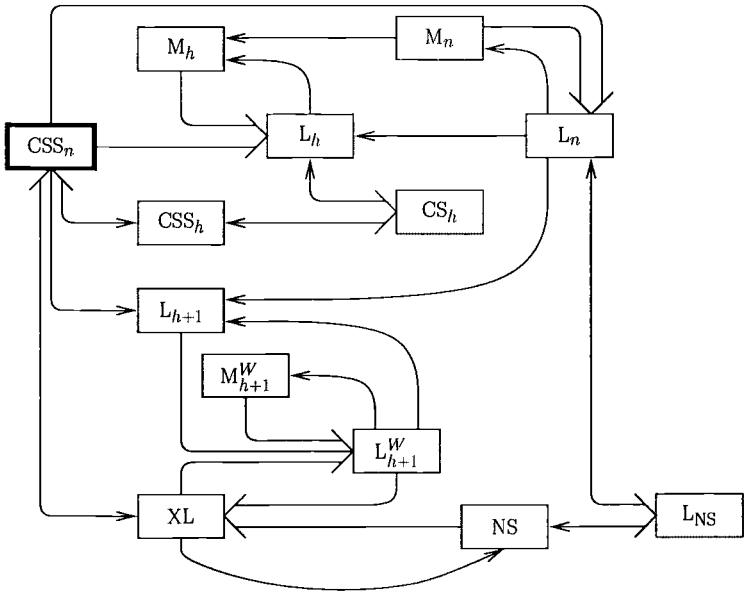


FIG. 3.1. *Summary of Propositions* 3.1–3.6, *Corollary* 3.8, *and Propositions* 3.9–3.14.

PROPOSITION 3.14. $\mathrm{L_{NS}} \iff \mathrm{NS} \wedge \mathrm{L}_n$.

Proof. $\Longrightarrow$**:** Assume $\mathrm{L_{NS}}$. *To show* NS: Let $\xi_{1..n} \in \mathbb{R}$ and $a_{1..n} \in A_{1..n}$ and $b \in B$ be such that $\xi_{1..h} \cdot a_{1..h} + \xi_{(h+1)..n} \cdot a_{(h+1)..n} = b$. Taking b to the other side, we conclude $b \in V$ as otherwise $\mathrm{L_{NS}}$ is contradicted.

To show L_n: Apply Lemma 3.7.

$\Longleftarrow$**:** Assume NS and L_n. If $B \setminus V$ is empty, then $\mathrm{L_{NS}}$ is L_n and we are done, so assume $B \setminus V$ is nonempty and let $\xi_{1..n}$, $c \in \mathbb{R}$ and $a_{1..n} \in A_{1..n}$ and $b \in B \setminus V$ satisfy $\xi_{1..n} \cdot a_{1..n} + c\, b = \mathbf{0}$. If c were nonzero, then dividing through by $-c$ and taking b to the other side would produce a contradiction with NS, so $c = 0$. By L_n, we thus have $\xi_{1..n} = 0$ as well. ∎

Collecting our results, we obtain our central theorem.

THEOREM 3.15 (Characterization of conditional sign-solvability 1).

$$\begin{array}{ccc}
\boxed{\mathrm{CSS}_n \Leftrightarrow \mathrm{CSS}_h \wedge \mathrm{L}_{h+1} \wedge \mathrm{XL}} & \Longleftarrow & \boxed{\mathrm{CSS}_n \wedge \mathrm{M}^W_{h+1} \Leftrightarrow \mathrm{CSS}_h \wedge \mathrm{L}^W_{h+1} \wedge \mathrm{NS}} \\
\Uparrow & & \Uparrow \\
\boxed{\mathrm{CSS}_n \wedge \mathrm{M}_h \Leftrightarrow \mathrm{CS}_h \wedge \mathrm{L}_{h+1} \wedge \mathrm{XL}} & & \boxed{\mathrm{CSS}_n \wedge \mathrm{M}_h \wedge \mathrm{M}^W_{h+1} \Leftrightarrow \mathrm{CS}_h \wedge \mathrm{L}^W_{h+1} \wedge \mathrm{NS}} \\
\Uparrow & & \Updownarrow \\
\boxed{\mathrm{CSS}_n \wedge \mathrm{M}_n \Leftrightarrow \mathrm{CS}_h \wedge \mathrm{L}_n \wedge \mathrm{XL}} & \Longleftarrow & \boxed{\mathrm{CSS}_n \wedge \mathrm{M}_n \wedge \mathrm{M}^W_{h+1} \Leftrightarrow \mathrm{CS}_h \wedge \mathrm{L}^W_{h+1} \wedge \mathrm{L_{NS}}}\ .
\end{array}$$

Proof. Refer to Figure 3.1, a diagrammatic summary of Propositions 3.1–3.6, Corollary 3.8, and Propositions 3.9–3.14, to see the boxed equivalences. The implications between boxes are variously trivial or from Proposition 3.6, except for $\mathrm{CSS}_n \wedge \mathrm{M}_h \wedge \mathrm{M}^W_{h+1} \Longrightarrow \mathrm{CSS}_n \wedge \mathrm{M}_n \wedge \mathrm{M}^W_{h+1}$, which we now prove. Assume the left-hand side. By Proposition 3.9, it is sufficient to show L_n, so let $\xi_{1..n} \in \mathbb{R}$ and $a_{1..n} \in A_{1..n}$ be such that $\xi_{1..n} \cdot a_{1..n} = \mathbf{0}$. We apply

L^W_{h+1} (by Propositions 3.1 and 3.10) after taking proj_W of both sides to get $\xi_{(h+1)..n} = 0$. Thus $\xi_{1..h} \cdot a_{1..h} = \mathbf{0}$, and applying L_h (by Proposition 3.3), we get $\xi_{1..h} = 0$ as well. ∎

Other characterization statements besides those in Theorem 3.15 can be identified from Figure 3.1, but the ones given above are the most interesting. The conditions M_h, M_n, and M^W_{h+1} are rather technical, though, and so we close this section by considering stronger—but much more applications-oriented—alternatives. In doing this we will arrive at Corollary 3.18, which is ostensibly more useful than Theorem 3.15.

LEMMA 3.16. *Suppose that each of $A_{1..h}$ is relatively open in its linear span (with respect to the usual topology on $\mathbb{R}^m$), and let $\xi_{1..h} \in \mathbb{R}\backslash\{0\}$ and $a_{1..h} \in A_{1..h}$ and $v \in V$ be given. Then for all sufficiently small $\epsilon > 0$ there holds $\xi_{1..h} \cdot a_{1..h} + \epsilon v \in \xi_{1..h} \cdot A_{1..h}$.*

Proof. Since $v \in V = \mathrm{Span}(A_1) + \cdots + \mathrm{Span}(A_h)$, we may write $v = v_1 + \cdots + v_h$ for suitable $v_1 \in \mathrm{Span}(A_1), \ldots, v_h \in \mathrm{Span}(A_h)$. Thus,

$$\xi_{1..h} \cdot a_{1..h} + \epsilon v = \xi_1 \left(a_1 + \frac{\epsilon}{\xi_1} v_1 \right) + \cdots + \xi_h \left(a_h + \frac{\epsilon}{\xi_h} v_h \right)$$

for all $\epsilon > 0$. The result follows as for each $i \in 1..h$ we have $a_i + (\epsilon/\xi_i)v_i \in A_i$ for all sufficiently small $\epsilon > 0$ due to A_i being relatively open in its linear span. ∎

PROPOSITION 3.17. *If each of $A_{1..h}$ is relatively open in its linear span, then the conditions M_h, M_n, and M^W_{h+1} are all satisfied.*

Proof. To show M_h and M_n: By Proposition 3.6, it suffices to show M_n. Let $\xi_{1..n} \in \mathbb{R}$ and $a_{1..n} \in A_{1..n}$ satisfy $\xi_{1..n} \cdot a_{1..n} = \mathbf{0}$, and let $\epsilon > 0$. For $i \in 1..n$, put

$$\xi'_i := \begin{cases} \tilde{\xi}_i & \text{if } \xi_i = 0 \\ \xi_i/\epsilon & \text{if } \xi_i \neq 0 \end{cases} \qquad \text{and} \qquad a'_i := \begin{cases} \tilde{a}_i & \text{if } \xi_i = 0 \\ a_i + (\epsilon\, \tilde{\xi}_i/\xi_i)\tilde{a}_i & \text{if } \xi_i \neq 0. \end{cases}$$

We have $\xi'_i \in \xi^\flat_i$ and, since A_i is relatively open in its linear span, we also have $a'_i \in A_i$ for all sufficiently small $\epsilon > 0$. Hence,

$$\mathbf{0} = (\xi_{1..n}/\epsilon) \cdot a_{1..n} + \tilde{\xi}_{1..n} \cdot \tilde{a}_{1..n} - \tilde{b} = \xi'_{1..n} \cdot a'_{1..n} - \tilde{b} \in \xi^\flat_{1..n} \cdot A_{1..n} - B$$

for a suitably small $\epsilon > 0$.

To show M^W_{h+1}: Let $\xi_{(h+1)..n} \in \mathbb{R}$ and $a_{(h+1)..n} \in A_{(h+1)..n}$ be such that $\xi_{(h+1)..n} \cdot \mathrm{proj}_W a_{(h+1)..n} = \mathbf{0}$. Then $-\xi_{(h+1)..n} \cdot a_{(h+1)..n} \in V$, so by Lemma 3.16 we may choose $a'_{1..h} \in A_{1..h}$ and $\epsilon > 0$ with

$$\tilde{\xi}_{1..h} \cdot \tilde{a}_{1..h} + \epsilon\, (-\xi_{(h+1)..n} \cdot a_{(h+1)..n}) = \tilde{\xi}_{1..h} \cdot a'_{1..h}.$$

We therefore have

$$\mathbf{0} = \tilde{\xi}_{1..h} \cdot a'_{1..h} + (\epsilon\, \xi_{(h+1)..n}) \cdot a_{(h+1)..n} - \tilde{b} \in T + \xi^*_{(h+1)..n} \cdot A_{(h+1)..n} - B. \quad ∎$$

COROLLARY 3.18 (Characterization of conditional sign-solvability 2). *If each of $A_{1..h}$ is relatively open in its linear span, then the following are equivalent:* (i) CSS_n, (ii) $\mathrm{CS}_h \wedge \mathrm{L}_{h+1} \wedge \mathrm{XL}$, (iii) $\mathrm{CS}_h \wedge \mathrm{L}_n \wedge \mathrm{XL}$, (iv) $\mathrm{CS}_h \wedge \mathrm{L}^W_{h+1} \wedge \mathrm{NS}$, (v) $\mathrm{CS}_h \wedge \mathrm{L}^W_{h+1} \wedge \mathrm{L}_{\mathrm{NS}}$.

Proof. Apply Theorem 3.15 and Proposition 3.17. ∎

Thus, given a system of subsets with each subset being relatively open in its linear span and the system being in h-form, the question of the system's conditional sign-solvability may be reduced to the question of whether certain closely related systems are CS- and L-systems. Also, as relatively open convex cones and open subsets of $\mathbb{R}^m$ are relatively open in their linear spans, Corollary 3.18 is a substantial generalization of Theorem 2.6 of [2]. There is for sign cones, however, a simplification we have not yet discussed and so to demonstrate more explicitly how Corollary 3.18 is a generalization, we include the following.

PROPOSITION 3.19. *If each of $A_{1..h}$ is a sign cone, then* NS *automatically holds.*

Proof. Using $\tilde{\xi}_{1..h} \cdot (\tilde{a}_{1..h})_i = \tilde{b}_i$ for $i \in 1..m$ at the second inclusion, we have

$$B \subseteq \prod_{i=1}^{m} \begin{Bmatrix} \mathbb{R} & \text{if } \tilde{b}_i \neq 0 \\ \{0\} & \text{if } \tilde{b}_i = 0 \end{Bmatrix}$$
$$\subseteq \prod_{i=1}^{m} \begin{Bmatrix} \mathbb{R} & \text{if } (\tilde{a}_j)_i \neq 0 \text{ for some } j \in 1..h \\ \{0\} & \text{if } (\tilde{a}_j)_i = 0 \text{ for every } j \in 1..h \end{Bmatrix} = V. \qquad \square$$

COROLLARY 3.20. *If each of $A_{1..h}$ is a sign cone, then* $\mathrm{CSS}_n \Longleftrightarrow \mathrm{CS}_h \wedge \mathrm{L}^W_{h+1}$.

Proof. Apply Corollary 3.18 and Proposition 3.19. $\square$

Note that we are now in the aforementioned situation of being able to take V and W as cartesian products of various copies of $\mathbb{R}$ and $\{0\}$, so that the projections in L^W_{h+1} are the more familiar selection of subsets of coordinates.

4. Sign-solvability. Due to the relationship between ordinary sign-solvability and conditional sign-solvability, in this section we are able to derive a general characterization of sign-solvability from the work of section 3. The connection between the two notions is expressed by the following pair of propositions, whose proofs are obvious. Note that the same condition, namely UC, mediates in both.

PROPOSITION 4.1. $\mathrm{SS}_n \Longleftrightarrow \mathrm{CSS}_n \wedge \mathrm{UC}$.

PROPOSITION 4.2. $\mathrm{S}_h \Longleftrightarrow \mathrm{CS}_h \wedge \mathrm{UC}$.

Our characterization of sign-solvability is simpler than section 3 because SS_n and S_h, respectively, are considerably more restrictive conditions than CSS_n and CS_h. Specifically, we have the following propositions and lemma.

PROPOSITION 4.3. $\mathrm{SS}_n \Longrightarrow \mathrm{L}_n$.

Proof. Assume SS_n and let $\xi_{1..n} \in \mathbb{R}$ and $a_{1..n} \in A_{1..n}$ be such that $\xi_{1..n} \cdot a_{1..n} = \mathbf{0}$. By SS_n, choose $\xi'_{1..n} \in \mathbb{R}$ with $\xi'_{1..n} \cdot a_{1..n} = \tilde{b}$. Observe $(\xi'_1 + c\,\xi_1)a_1 + \cdots + (\xi'_n + c\,\xi_n)a_n = \tilde{b}$ for all $c \in \mathbb{R}$. Hence, if any of $\xi_{1..n}$ are nonzero, then we obtain a contradiction with SS_n and h-form by taking c sufficiently large in absolute value and of an appropriate sign. Thus, $\xi_{1..n} = 0$. $\square$

LEMMA 4.4. *Let $a'_{1..h} \in A_{1..h}$ and put $V' := \mathbb{R}a'_1 + \cdots + \mathbb{R}a'_h$. If S_h, then $V = V'$.*

Proof. Since $V' \subseteq V$ it is sufficient to show $V \subseteq V'$, but since V' is a subspace of $\mathbb{R}^m$ we actually need only show for $j \in 1..h$ that $a_j \in V'$ for arbitrary $a_j \in A_j$. Using S_h twice, we choose $\xi'_{1..h}$, $c' > 0$ and $\xi_{1..h}$, $c > 0$ with

$$\xi'_{1..h} \cdot a'_{1..h} - c'\tilde{b} = \mathbf{0} \qquad \text{and} \qquad \sum_{i \in 1..h}^{i \neq j} \xi_i a'_i + \xi_j a_j - c\,\tilde{b} = \mathbf{0}$$

so that

$$a_j = \frac{c}{c'\xi_j}\left(c'\tilde{b}\right) + \frac{1}{\xi_j}\left(\xi_j a_j - c\,\tilde{b}\right) = \sum_{i\in 1..h}^{i\neq j}\left(\frac{c\,\xi_i'}{c'\xi_j} - \frac{\xi_i}{\xi_j}\right)a_i' + \frac{c\,\xi_j'}{c'\xi_j}a_j' \in V'. \quad \square$$

PROPOSITION 4.5. $\mathrm{S}_h \Longrightarrow \mathrm{M}_{h+1}^W$.

Proof. Assume S_h and suppose $\xi_{(h+1)..n} \in \mathbb{R}$ and $a_{(h+1)..n} \in A_{(h+1)..n}$ are such that $\mathrm{proj}_W(\xi_{(h+1)..n} \cdot a_{(h+1)..n}) = \mathbf{0}$, i.e., $v := \xi_{(h+1)..n} \cdot a_{(h+1)..n} \in V$. By Lemma 4.4, choose $\xi_{1..h} \in \mathbb{R}$ such that $\xi_{1..h} \cdot \tilde{a}_{1..h} = -v \in V$. We then have

$$\mathbf{0} = (\xi_1 + \tilde{\xi}_1)\tilde{a}_1 + \cdots + (\xi_h + \tilde{\xi}_h)\tilde{a}_h + v - \tilde{b} \in T + \xi^*_{(h+1)..n} \cdot A_{(h+1)..n} - B. \quad \square$$

The following two propositions simplify matters further.

PROPOSITION 4.6. $\mathrm{UC} \Longrightarrow \mathrm{NS}$.

Proof. By UC, $B \subseteq \mathbb{R}\tilde{a}_1 + \cdots + \mathbb{R}\tilde{a}_h \subseteq V$. $\square$

PROPOSITION 4.7. $\mathrm{L}_n \wedge \mathrm{UC} \Longrightarrow \mathrm{XL}$.

Proof. Assume L_n and UC. Let $\xi_{1..h} \in \mathbb{R}$ and $a_{1..h} \in A_{1..h}$ and $b \in B$ and $u \in U$ satisfy $\xi_{1..h} \cdot a_{1..h} - b = u$. By UC, choose $\xi'_{1..h} \in \mathbb{R}$ so that $\xi'_{1..h} \cdot a_{1..h} = b$. Then $(\xi_1 - \xi_1')a_1 + \cdots + (\xi_h - \xi_h')a_h - u = \mathbf{0}$ and so $u = \mathbf{0}$ by L_n. $\square$

Thus, we arrive at our characterization of sign-solvability.

THEOREM 4.8 (Characterization of sign-solvability). $\mathrm{SS}_n \Longleftrightarrow \mathrm{S}_h \wedge \mathrm{L}_{h+1}^W \Longleftrightarrow \mathrm{S}_h \wedge \mathrm{L}_n$.

Proof. Refer to Figure 3.1 and Propositions 4.1–4.3 and 4.5–4.7. $\square$

Hence, once an arbitrary system is placed into h-form, the question of its sign-solvability may be reduced to the question of whether certain closely related systems are S- and L-systems. The simplicity of the final result in this case is most appealing.

REFERENCES

[1] L. BASSETT, J. MAYBEE, AND J. QUIRK, *Qualitative economics and the scope of the correspondence principle*, Econometrica, 36 (1968), pp. 544–563.

[2] R. A. BRUALDI, K. L. CHAVEY, AND B. L. SHADER, *Conditional sign-solvability*, Math. Comput. Modelling, 17 (1993), pp. 141–148.

[3] R. A. BRUALDI AND B. L. SHADER, *Matrices of Sign-Solvable Linear Systems*, Cambridge Tracts in Mathematics 116, Cambridge University Press, Cambridge, UK, 1995.

[4] P. W. KASTELEYN, *Dimer statistics and phase transitions*, J. Math. Phys., 4 (1963), pp. 287–293.

[5] V. KLEE, R. LADNER, AND R. MANBER, *Sign-solvability revisited*, Linear Algebra Appl., 59 (1984), pp. 131–157.

[6] V. KLEE, B. VON HOHENBALKEN, AND T. LEWIS, *On the recognition of S-systems*, Linear Algebra Appl., 192 (1993), pp. 187–204.

[7] R. MANBER, *Graph-theoretical approach to qualitative solvability of linear systems*, Linear Algebra Appl., 48 (1982), pp. 457–470.

[8] J. S. MAYBEE, *Sign solvability*, in Computer-Assisted Analysis and Model Simplification, H. J. Greenberg and J. S. Maybee, eds., Academic Press, New York, London, 1981, pp. 201–257.

[9] P. A. SAMUELSON, *Foundations of Economic Analysis*, Harvard Economic Studies 80, Harvard University Press, Cambridge, MA, 1947.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 3, pp. 989–1003
© 2000 Society for Industrial and Applied Mathematics

NEW METHODS FOR ESTIMATING THE DISTANCE TO UNCONTROLLABILITY*

MING GU†

Abstract. Controllability is a fundamental concept in control theory. Given a linear control system, we present new algorithms for estimating its distance to uncontrollability, i.e., the norm of the normwise smallest perturbation that makes the given system uncontrollable. Many algorithms have been previously proposed to estimate this distance. Our new algorithms are the first that correctly estimate this distance at a cost polynomial in a dimension of the given system. We report results from some numerical experiments that demonstrate the reliability and effectiveness of these new algorithms.

Key words. controllability, numerical stability, complexity, QR factorization

AMS subject classifications. 65F99, 93B05

PII. S0895479897328856

1. Introduction.

One of the most fundamental concepts in control theory is that of controllability. A matrix pair $(A, B) \in \mathbf{C}^{n\times n} \times \mathbf{C}^{n\times m}$ is controllable (see Kailath [20, pp. 85–90]) if the state function $x = x(t)$ in the linear control system

$$\dot{x} = A\,x + B\,u \tag{1.1}$$

can be directed from any given state to a desired state in finite time by an input $u = u(t)$. Uncontrollability could signal fundamental trouble with the control model or the underlying physical system itself (Byers [11]).

A large number of algebraic and dynamic characterizations of controllability have been given (Laub [21], for example), but each and every one of these has difficulties when implemented in finite precision (Patel, Laub, and Van Dooren [27, p. 15]). For instance, it is well known that (A, B) is controllable if and only if

$$\operatorname{rank}([A - \lambda I, B]) = n \quad \text{for all } \lambda \in \mathcal{C}, \tag{1.2}$$

where $\mathcal{C}$ is the set of complex numbers. However, it is not clear how one can numerically verify whether a system is controllable through (1.2). More critically, (1.2) does not provide any means to detect systems that are nearly uncontrollable, systems that could be equally troublesome. From these considerations it became apparent (see Laub [21] and Paige [26]) that a more meaningful measure is the distance to uncontrollability, the norm distance of the pair (A, B) from the set of all uncontrollable pairs:

$$\rho(A, B) = \{\|[\Delta A, \Delta B]\|_F : (A + \Delta A, B + \Delta B) \quad \text{uncontrollable.}\} \tag{1.3}$$

It was later shown by Eising [15, 16] that

$$\rho(A, B) = \min_{\lambda\in\mathcal{C}} \sigma_n([A - \lambda I, B]), \tag{1.4}$$

*Received by the editors November 21, 1997; accepted for publication (in revised form) by P. Van Dooren February 15, 1999; published electronically March 8, 2000.
http://www.siam.org/journals/simax/21-3/32885.html

†Department of Mathematics, University of California, Los Angeles, CA 90095-1555. This research was supported in part by NSF Career Award CCR-9702866 and by the Applied Mathematical Sciences Subprogram of the Office of Energy, Research, U.S. Department of Energy, under contract DE-AC03-76SF00098.

where $\sigma_n(G)$ denotes the nth singular value of $G \in \mathbf{C}^{n\times(n+m)}$. Demmel [12] relates $\rho(A,B)$ to the sensitivity of the pole-assignment problem.

Many algorithms have been designed to compute $\rho(A,B)$. However, the function to be minimized in (1.4) is not convex and may have as many as n or more local minima. It is not clear just how many local minima there are for any given problem (Byers [11]). Methods that search for a local minimum tend to be efficient but have no guarantee of finding $\rho(A,B)$ with any accuracy since $\rho(A,B)$ is the global minimum (Boley [4, 6], Boley and Golub [5], Boley and Lu [7], Byers [11], Elsner and He [17], Miminis [24], and Wicks and DeCarlo [31]). Methods that search for the global minimum (Byers [11], Gao and Neumann [18], and He [19]) sometimes do have this guarantee but require computing time that is inverse proportional to $\rho^2(A,B)$, prohibitively expensive for nearly uncontrollable systems, which are the kind of systems for which computing $\rho(A,B)$ is important. While the backward stable algorithms of Beelen and Van Dooren [2, 3], Van Dooren [30], and Demmel and Kågström [13, 14] are efficient and very useful for detecting uncontrollability, they often fail to detect near uncontrollability.

In this paper, we propose new methods to correctly estimate $\rho(A,B)$ to within a factor of 2. They are based on the following bisection method.

ALGORITHM 1.1. BISECTION METHOD.

Set $\delta := \sigma_{\min}([A,B])/2$.

while $\delta \geq \rho(A,B)$

$\delta := \delta/2$.

endwhile

The bisection idea was used to compute the distance of a stable matrix to the unstable matrices (Byers [10]). It was then used to compute the $\mathbf{L}_\infty$ norm of a transfer matrix (Boyd, Balakrishnan, and Kabamba [9]); a quadratically convergent version of this later method was developed by Boyd and Balakrishnan [8].

There were past attempts to use Algorithm 1.1 to estimate $\rho(A,B)$ as well [11, 18], but they have resulted in potentially prohibitively expensive algorithms. The critical difference between our new approach and earlier attempts lies in how to numerically verify whether $\delta \geq \rho(A,B)$. Our new approach is based on a novel verifying scheme (see section 3.2). Paralleling the development of Boyd and Balakrishnan [8], we have also developed a generally quadratically convergent version of Algorithm 1.1. With very little modification, our new methods can be used to detect the uncontrollable modes for any given tolerance. The knowledge of such modes is essential if one wishes to remove them from the system.

Complexity-wise, these new algorithms differ from previous algorithms in that they are the first algorithms that correctly estimate the distance at a cost polynomial in the matrix size. In fact, they require $O(n^6)$ floating pointing operations. The main cost of these new algorithms is the computation of some eigenvalues of certain sparse generalized eigenvalue problems of size $O(n^2)$.

In section 2 we review methods of Byers and Gao and Neumann to minimize the function in (1.4) when λ is restricted to a straight line on the complex plane. In section 3 we present our new methods to minimize the function in (1.4) over the entire complex plane. In section 4 we present some numerical results. In section 5 we draw conclusions and discuss open questions.

2. Minimization methods over a straight line. Define

$$g(\tau) \equiv \sigma_n\left(\left[A - \left(\lambda_0 + e^{i\theta}\tau\right) I, B\right]\right), \tag{2.1}$$

where τ is a real variable. To motivate our new methods for minimizing the function in (1.4) over the entire complex plane, we present Algorithm 2.1 to estimate the global minimum of $g(\tau)$ to within a factor of 2 for a given complex number λ_0 and a real angle θ. This algorithm is a variation of the bisection schemes of [11, 18], which can actually compute the global minimum. Let τ_* be a global minimizer.

ALGORITHM 2.1. BISECTION METHOD OVER A STRAIGHT LINE.

Set $\delta := g(0)$.

while $\delta \geq g(\tau_*)$

$\delta := \delta/2$.

endwhile

We will also discuss a quadratically convergent version of Algorithm 2.1 in section 2.2.

2.1. The bisection method over a straight line. Let $g(\tau_*) = \min_\tau g(\tau)$. Algorithm 2.1 is missing a scheme to numerically verify whether $\delta \geq g(\tau_*)$ for a given $\delta > 0$. We discuss such a scheme below. Different versions of it were developed in Byers [11] and Gao and Neumann [18], and were based on earlier work of Byers [10].

We assume that $\delta \geq g(\tau_*)$. Since $g(\tau)$ is a continuous function with

$$\lim_{\tau \to +\infty} g(\tau + \tau_*) = \infty \quad \text{and} \quad \lim_{\tau \to -\infty} g(\tau + \tau_*) = \infty \,,$$

it follows that there exists at least two solutions[1] to the equation $g(\tau) = \delta$. By the definition of singular values, this implies that there exist nonzero vectors $\binom{x}{y}$ and z such that

$$\left(A - \left(\lambda_0 + e^{i\theta}\tau\right) I \quad B\right) \begin{pmatrix} x \\ y \end{pmatrix} = \delta z \quad \text{and} \quad \begin{pmatrix} A^* - \left(\bar{\lambda}_0 + e^{-i\theta}\tau\right) I \\ B^* \end{pmatrix} z = \delta \begin{pmatrix} x \\ y \end{pmatrix}.$$

These equations can be rewritten as

$$\begin{pmatrix} -\delta I & A - \left(\lambda_0 + e^{i\theta}\tau\right) I & B \\ A^* - \left(\bar{\lambda}_0 + e^{-i\theta}\tau\right) I & -\delta I & 0 \\ B^* & 0 & -\delta I \end{pmatrix} \begin{pmatrix} z \\ x \\ y \end{pmatrix} = 0 \,. \tag{2.2}$$

To simplify (2.2), we QR-factorize

$$\begin{pmatrix} B \\ -\delta I \end{pmatrix} = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \begin{pmatrix} R \\ 0 \end{pmatrix} \tag{2.3}$$

and define

$$\begin{pmatrix} z_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} Q_{11}^* & Q_{21}^* \\ Q_{12}^* & Q_{22}^* \end{pmatrix} \begin{pmatrix} z \\ y \end{pmatrix}.$$

These relations and (2.2) imply that $R^* z_1 = 0$. Since $\delta > 0$, R must be nonsingular. It follows that $z_1 = 0$. Hence (2.2) is reduced to

$$\begin{pmatrix} A - \left(\lambda_0 + e^{i\theta}\tau\right) I & BQ_{22} - \delta Q_{12} \\ -\delta I & \left(A^* - \left(\bar{\lambda}_0 + e^{-i\theta}\tau\right)\right) Q_{12} \end{pmatrix} \begin{pmatrix} x \\ y_1 \end{pmatrix} = 0 \,,$$

[1] If $\delta = g(\tau_*)$, then equation $g(\tau) = \delta$ has a double root at $\tau = \tau_*$.

which can be further rewritten as

$$\begin{pmatrix} A-\lambda_0 I & BQ_{22}-\delta Q_{12} \\ -\delta I & \left(A^*-\bar{\lambda}_0 I\right) Q_{12} \end{pmatrix} \begin{pmatrix} x \\ y_1 \end{pmatrix} = \tau \begin{pmatrix} e^{i\theta} I & 0 \\ 0 & e^{-i\theta} Q_{12} \end{pmatrix} \begin{pmatrix} x \\ y_1 \end{pmatrix}. \tag{2.4}$$

Since Q_{12} is part of the Q factor in the QR factorization (2.3), it follows that

$$Q_{12}^* B - \delta Q_{22}^* = 0 \quad \text{and} \quad Q_{12}^* Q_{12} + Q_{22}^* Q_{22} = I,$$

which implies

$$Q_{12}^* \left(\delta^2 I + BB^*\right) Q_{12} = \delta^2 I.$$

Hence Q_{12} is nonsingular and the pencil in (2.4) is regular. It is now easy to show that condition $\delta \geq g(\tau_*)$ holds if and only if the matrix pencil in (2.4) has a real eigenvalue τ.

To verify whether $\delta \geq g(\tau_*)$ in Algorithm 2.1, we compute the eigenvalues of the pencil in (2.4). If this pencil has real eigenvalues, then $\delta \geq g(\tau_*)$; otherwise, $\delta < g(\tau_*)$. Since Algorithm 2.1 guarantees that $2\delta \geq g(\tau_*)$ from the previous bisection step, the value of δ after Algorithm 2.1 exits from the **while** loop must satisfy

$$\frac{g(\tau_*)}{2} \leq \delta < g(\tau_*)\,. \tag{2.5}$$

We note that (2.2) was not reduced in [11], making it more time-consuming to verify whether $\delta \geq g(\tau_*)$. It was reduced to a regular eigenvalue problem by solving for y in [18], but the reduction appears to be less numerically reliable than our reduction to (2.4).

2.2. A quadratically convergent variation. Boyd and Balakrishnan [8] note that in the context of computing the $\mathbf{L}_\infty$-norm of a transfer matrix, the function to be minimized is, in general, approximately quadratic near the maximum, and they use this fact to design a quadratically convergent variation of a bisection method for computing the $\mathbf{L}_\infty$-norm.

Their idea applies equally well in minimizing (2.1). If $\delta - g(\tau_*)$ is small enough and if τ_1 and τ_2 are two roots of $g(\tau) = \delta$ closest to τ_*, then arguments similar to those of [8] show that $(\tau_1 + \tau_2)/2$ is in general a much better approximation to τ_*. We summarize this algorithm below.

ALGORITHM 2.2. QUADRATICALLY CONVERGENT VARIATION OF ALGORITHM 2.1.

Set $\delta := g(0)$.

while $\delta \geq g(\tau_*)$

Choose two real eigenvalues τ_1 and τ_2 of the pencil (2.4).

$\delta := \min\left\{\delta, g\left(\dfrac{\tau_1+\tau_2}{2}\right)\right\}/2\,.$

endwhile

With arguments similar to those in [8], it is easy to show that if τ_1 and τ_2 are chosen correctly, then

$$g\left(\frac{\tau_1+\tau_2}{2}\right) - g(\tau_*) = O\left(\left(\delta - g(\tau_*)\right)^2\right). \tag{2.6}$$

This estimate holds even if $g(\tau)$ is not approximately quadratic near τ_*. We caution that, strictly speaking, Algorithm 2.2 is not even asymptotically quadratically convergent, since it terminates as soon as it has found a δ that satisfies (2.5). Nevertheless, relation (2.6) does indicate rapid convergence of Algorithm 2.2 when $g(\tau_*)$ is tiny.

3. Minimization methods over the complex plane. Now we discuss methods to minimize the function in (1.4) over the entire complex plane. Let

$$f(\alpha,\beta) \equiv \sigma_n\left([A-(\alpha+\beta i)\,I,\,B]\right) \quad \text{and} \quad \rho(A,B) = f\left(\alpha_*,\beta_*\right) = \min_{\alpha,\beta} f(\alpha,\beta). \tag{3.1}$$

One such method is Algorithm 1.1 discussed in section 1. As in Algorithm 2.1, we need to develop a scheme to verify whether $\delta > \rho(A,B)$ in order to complete Algorithm 1.1. To do so, we first prove a fundamental theorem in section 3.1, then provide such a scheme in section 3.2, and finally we develop a generally quadratically convergent version of Algorithm 1.1 in section 3.3.

3.1. A fundamental theorem. Our scheme to verify whether $\delta > \rho(A,B)$ is based on Theorem 3.1 below.

THEOREM 3.1. *Assume that $\delta > \rho(A,B)$. Then there are at least two pairs of real numbers α and β such that*

$$\sigma\left([A-(\alpha+\beta i)\,I,\,B]\right) = \delta \quad \textit{and} \quad \sigma\left([A-(\alpha+\eta+\beta i)\,I,\,B]\right) = \delta, \tag{3.2}$$

where $0 < \eta \le 2\left(\delta - \rho(A,B)\right)$ and $\sigma(G)$ denotes a singular value of G.

Proof. From standard perturbation theory, we have

$$|(\alpha-\alpha_*)+(\beta-\beta_*)i| - \|[A-(\alpha_*+\beta_* i)\,I,\,B]\|_2 \le f(\alpha,\beta) \le |(\alpha-\alpha_*)+(\beta-\beta_*)i| + \rho(A,B). \tag{3.3}$$

Hence $f(\alpha,\beta)$ goes to infinity if $|\alpha+\beta i|$ does. It is well known that $f(\alpha,\beta)$ is a continuous function of α and β. Consequently, the fact that $\delta > \rho(A,B)$ immediately implies that there exists a pair of numbers (α_1,β_1) such that[2] $f(\alpha_1,\beta_1) = \delta$.

From the definition of singular values, α and β satisfy

$$\sigma\left([A-(\alpha+\beta i)\,I,\,B]\right) = \delta$$

if and only if they satisfy the algebraic equation

$$\det\left((A-(\alpha+\beta i)\,I)\cdot(A^*-(\alpha-\beta i)\,I) + B\,B^* - \delta^2 I\right) = 0.$$

It follows from $f(\alpha_1,\beta_1) = \delta$ that this algebraic equation has at least one solution and it follows from (3.3) that all its solutions are finite. Consequently, these solutions form a finite number of closed (continuous) algebraic curves on the α-β plane.

Now we claim that the point (α_*,β_*) must be in the interior of one of these closed curves. In fact, if this is not the case, then there exists a continuous curve $\lambda(\tau) = (\lambda_1(\tau),\lambda_2(\tau))$ on the α-β plane that does not intersect with any of these algebraic curves but connects (α_*,β_*) and infinity:

$$\lambda(0) = (\alpha_*,\beta_*) \quad \text{and} \quad \lim_{\tau\to\infty} |\lambda(\tau)| = \infty\ .$$

In other words,

$$f(\lambda_1(0),\lambda_2(0)) = \rho(A,B) \quad \text{and} \quad \lim_{\tau\to\infty} f\left(\lambda_1(\tau),\lambda_2(\tau)\right) = \infty\ .$$

It follows from the continuity argument that there exists a τ_1 such that $f\left(\lambda_1(\tau_1),\right.$ $\left.\lambda_2(\tau_1)\right) = \delta$. But this contradicts the assumption that the curve $\lambda(\tau)$ does not

[2] This fact is shown in Byers [11] and Gao and Neumann [18].

intersect with any of the algebraic curves. Consequently, the point (α_*, β_*) must be in the interior of one of these closed curves. Among all closed curves that have (α_*, β_*) in their interior, let $\mathcal{G}$ denote the one that covers the smallest area.

It follows from the same continuity argument that there exist two points

$$\mathcal{P}_1 = (\alpha_* - \eta_1, \beta_*) \quad \text{and} \quad \mathcal{P}_2 = (\alpha_* + \eta_2, \beta_*)$$

on $\mathcal{G}$ with $\eta_1 > 0$ and $\eta_2 > 0$. In other words,

$$\sigma\left([A - (\alpha_* - \eta_1, \beta_* i)\, I, B]\right) = \sigma\left([A - (\alpha_* + \eta_2, \beta_* i)\, I, B]\right) = \delta. \tag{3.4}$$

For simplicity, we assume that $\mathcal{P}_1$ and $\mathcal{P}_2$ are chosen so that η_1 and η_2 are the smallest positive numbers.

Since the point (α_*, β_*) is in the interior of $\mathcal{G}$ and also lies strictly inside the line segment between $\mathcal{P}_1$ and $\mathcal{P}_2$, it follows that any point that lies strictly inside this line segment is in the interior of curve $\mathcal{G}$. Combining (3.4) with relation (3.3), we get

$$\eta_1 \geq \delta - \rho(A, B) \quad \text{and} \quad \eta_2 \geq \delta - \rho(A, B). \tag{3.5}$$

Now we shift all the points on $\mathcal{G}$ horizontally by the same amount $-\eta$ to get a closed curve

$$\widehat{\mathcal{G}} = \{(\alpha - \eta, \beta) \,|(\alpha, \beta) \text{ is a point on } \mathcal{G}\}.$$

Since $\mathcal{P}_2$ is a point on $\mathcal{G}$, $\widehat{\mathcal{P}}_2 = (\alpha_* + \eta_2 - \eta, \beta_*)$ is a point on $\widehat{\mathcal{G}}$. Assume that $0 < \eta < 2\,(\delta - \rho(A, B))$. Then relation (3.5) implies that $\widehat{\mathcal{P}}_2$ is a point that lies strictly inside the line segment between $\mathcal{P}_1$ and $\mathcal{P}_2$. Hence $\widehat{\mathcal{P}}_2$ is in the interior of curve $\mathcal{G}$. Let $\mathcal{P}_3 = (\alpha_3, \beta_3)$ be the leftmost point on $\mathcal{G}$. Then $\widehat{\mathcal{P}}_3 = (\alpha_3 - \eta, \beta_3)$ is the leftmost point on $\widehat{\mathcal{G}}$. Since $\eta > 0$, we have that $\widehat{\mathcal{P}}_3$ is in the exterior of $\mathcal{G}$.

In other words, we have found a point $\widehat{\mathcal{P}}_2$ that is on $\widehat{\mathcal{G}}$ and in the interior of $\mathcal{G}$, and another point $\widehat{\mathcal{P}}_3$ on $\widehat{\mathcal{G}}$ that is in the exterior of $\mathcal{G}$. Since $\mathcal{G}$ and $\widehat{\mathcal{G}}$ are continuous closed curves, we conclude that these two curves intersect.

Let (α_4, β_4) be any intersecting point. It follows that both (α_4, β_4) and $(\alpha_4 + \eta, \beta_4)$ must be points on $\mathcal{G}$. Hence α_4 and β_4 are a solution to (3.2), and therefore equations (3.2) have at least one solution.

In the following argument we assume that (α_4, β_4) is the only intersecting point of $\mathcal{G}$ and $\widehat{\mathcal{G}}$. Let $\widehat{\mathcal{G}}_1$ denote the set of points on $\widehat{\mathcal{G}}$ that are either on $\mathcal{G}$ or in the interior of $\mathcal{G}$ and let $\mathcal{G}_1$ denote the corresponding set of points on $\mathcal{G}$. If $\widehat{\mathcal{G}}_1$ is not a closed curve itself, then $\widehat{\mathcal{G}}_1$ must be an open curve with one end point on $\mathcal{G}$ and the other in the interior of $\mathcal{G}$. It follows that $\widehat{\mathcal{G}}_1$, and hence $\mathcal{G}_1$, must have a positive arclength. Hence the portion of $\mathcal{G}$ without $\mathcal{G}_1$ is a closed curve. But this contradicts the way $\mathcal{G}$ is constructed. This contradiction implies that $\widehat{\mathcal{G}}_1$, and hence $\mathcal{G}_1$, must be closed curves themselves. Let $\widehat{\mathcal{G}}_2$ denote the set of points on $\widehat{\mathcal{G}}$ that are either on $\mathcal{G}$ or in the exterior of $\mathcal{G}$ and let $\mathcal{G}_2$ denote the corresponding set of points on $\mathcal{G}$. A similar argument shows that $\mathcal{G}_2$ must be a closed curve as well.

By construction, $\widehat{\mathcal{G}}_1$ and $\widehat{\mathcal{G}}_2$, and hence $\mathcal{G}_1$ and $\mathcal{G}_2$, do not share any common region with a positive area. Hence (α_*, β_*) can only be in the interior of one of these closed curves; this implies that $\mathcal{G}$ is not a closed curve that has (α_*, β_*) in its interior and covers the smallest area, a contradiction to the way $\mathcal{G}$ was constructed. This contradiction is the result of the assumption that $\mathcal{G}$ and $\widehat{\mathcal{G}}$ intersect only once. Hence $\mathcal{G}$ and $\widehat{\mathcal{G}}$ must intersect at least twice, so (3.2) must have at least two real solutions. By a continuity argument, equations (3.2) have two, possibly identical, real solutions, even if $\eta = 2\,(\delta - \rho(A, B))$. □

3.2. A new verifying scheme. Let $\delta > 0$ and $\eta > 0$. We now consider how to numerically verify whether (3.2) have a real solution. By the definition of singular values, (3.2) implies that there exist nonzero vectors $\binom{x}{y}$, z, $\binom{\widehat{x}}{\widehat{y}}$, and $\widehat{z}$ such that

$$[A-(\alpha+\beta i)\,I \quad B]\begin{pmatrix} x \\ y \end{pmatrix} = \delta z\,, \quad \begin{pmatrix} A^*-(\alpha-\beta i)\,I \\ B^* \end{pmatrix} z = \delta \begin{pmatrix} x \\ y \end{pmatrix},$$
$$[A-(\alpha+\eta+\beta i)\,I \quad B]\begin{pmatrix} \widehat{x} \\ \widehat{y} \end{pmatrix} = \delta \widehat{z}\,, \quad \begin{pmatrix} A^*-(\alpha+\eta-\beta i)\,I \\ B^* \end{pmatrix} \widehat{z} = \delta \begin{pmatrix} \widehat{x} \\ \widehat{y} \end{pmatrix}.$$

These equations can be rewritten as

$$\begin{pmatrix} -\delta I & A-\alpha I & B \\ A^*-\alpha I & -\delta I & 0 \\ B^* & 0 & -\delta I \end{pmatrix}\begin{pmatrix} z \\ x \\ y \end{pmatrix} = \beta i \begin{pmatrix} 0 & I & 0 \\ -I & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} z \\ x \\ y \end{pmatrix} \tag{3.6}$$

and

$$\begin{pmatrix} -\delta I & A-(\alpha+\eta)I & B \\ A^*-(\alpha+\eta)I & -\delta I & 0 \\ B^* & 0 & -\delta I \end{pmatrix}\begin{pmatrix} \widehat{z} \\ \widehat{x} \\ \widehat{y} \end{pmatrix} = \beta i \begin{pmatrix} 0 & I & 0 \\ -I & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} \widehat{z} \\ \widehat{x} \\ \widehat{y} \end{pmatrix}. \tag{3.7}$$

In the QR factorization (2.3), define

$$\begin{pmatrix} z_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} Q_{11}^* & Q_{21}^* \\ Q_{12}^* & Q_{22}^* \end{pmatrix}\begin{pmatrix} z \\ y \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \widehat{z}_1 \\ \widehat{y}_1 \end{pmatrix} = \begin{pmatrix} Q_{11}^* & Q_{21}^* \\ Q_{12}^* & Q_{22}^* \end{pmatrix}\begin{pmatrix} \widehat{z} \\ \widehat{y} \end{pmatrix}.$$

These relations and (3.6) and (3.7) imply that $R^* z_1 = 0$ and $R^* \widehat{z}_1 = 0$. Since R is nonsingular, it follows that $z_1 = 0$ and $\widehat{z}_1 = 0$. Hence (3.6) and (3.7) are reduced to

$$\begin{pmatrix} A-\alpha I & BQ_{22}-\delta Q_{12} \\ -\delta I & (A^*-\alpha I)\,Q_{12} \end{pmatrix}\begin{pmatrix} x \\ y_1 \end{pmatrix} = \beta i \begin{pmatrix} I & 0 \\ 0 & -Q_{12} \end{pmatrix}\begin{pmatrix} x \\ y_1 \end{pmatrix} \tag{3.8}$$

and

$$\begin{pmatrix} A-(\alpha+\eta)I & BQ_{22}-\delta Q_{12} \\ -\delta I & (A^*-(\alpha+\eta)I)\,Q_{12} \end{pmatrix}\begin{pmatrix} \widehat{x} \\ \widehat{y}_1 \end{pmatrix} = \beta i \begin{pmatrix} I & 0 \\ 0 & -Q_{12} \end{pmatrix}\begin{pmatrix} \widehat{x} \\ \widehat{y}_1 \end{pmatrix}. \tag{3.9}$$

As shown in section 2.1, Q_{12} is always nonsingular for $\delta > 0$. Hence the matrix on the right-hand sides of both (3.8) and (3.9) is nonsingular. In order for the two pencils defined in (3.8) and (3.9) to share a common pure imaginary eigenvalue βi, the following matrix equation for $X \in \mathbf{R}^{2n\times 2n}$

$$\begin{pmatrix} A-\alpha I & BQ_{22}-\delta Q_{12} \\ -\delta I & (A^*-\alpha I)\,Q_{12} \end{pmatrix} X \begin{pmatrix} I & 0 \\ 0 & -Q_{12} \end{pmatrix}^*$$
$$-\begin{pmatrix} I & 0 \\ 0 & -Q_{12} \end{pmatrix} X \begin{pmatrix} A-(\alpha+\eta)I & BQ_{22}-\delta Q_{12} \\ -\delta I & (A^*-(\alpha+\eta)I)\,Q_{12} \end{pmatrix}^* = 0 \tag{3.10}$$

must have a nonzero solution. Partition $X = \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix}$. This matrix equation becomes

$$\mathcal{H}u = 0 \quad \text{and} \quad \mathcal{A}u = 2\alpha \mathcal{B}u, \tag{3.11}$$

$$\text{with} \quad \mathcal{B} = \begin{pmatrix} 0 & 0 & -Q_{12} \otimes I & 0 \\ 0 & 0 & 0 & I \otimes Q_{12} \end{pmatrix}, \quad \hat{B} = BQ_{22} - \delta Q_{12} \, ,$$

$$u = \begin{pmatrix} \mathbf{vec}\,(X_{11}) \\ \mathbf{vec}\,(X_{22}) \\ \mathbf{vec}\,(X_{12}) \\ \mathbf{vec}\,(X_{21}) \end{pmatrix};$$

$$\mathcal{A} = \begin{pmatrix} \delta I & -Q_{12} \otimes \hat{B} & Q_{12} \otimes A - ((A^* - \eta)Q_{12}) \otimes I \\ -\delta I & \hat{B} \otimes Q_{12} & 0 \\ & & 0 \\ & & (A - \eta) \otimes Q_{12} - I \otimes (A^* Q_{12}) \end{pmatrix};$$

$$\mathcal{H} = \begin{pmatrix} I \otimes A - (A - \eta I) \otimes I & & & \\ 0 & & & \\ & 0 & -\hat{B} \otimes I & I \otimes \hat{B} \\ & ((A^* - \eta)Q_{12}) \otimes Q_{12} - Q_{12} \otimes (A^* Q_{12}) & \delta Q_{12} \otimes I & -\delta I \otimes Q_{12} \end{pmatrix}.$$

In these equations, $\otimes$ is the Kronecker product and $\mathbf{vec}(G)$ is a vector formed by stacking the column vectors of G.

To reduce (3.11) to a standard generalized eigenvalue problem, let

$$\mathcal{H} = (\mathcal{R} \quad 0) \begin{pmatrix} \mathcal{Q}_{11} & \mathcal{Q}_{12} \\ \mathcal{Q}_{21} & \mathcal{Q}_{22} \end{pmatrix} \tag{3.12}$$

be the RQ factorization of $\mathcal{H}$, where $\mathcal{Q}_{ij} \in \mathbf{C}^{2n^2 \times 2n^2}$, and define

$$\begin{pmatrix} w \\ v \end{pmatrix} = \begin{pmatrix} \mathcal{Q}_{11} & \mathcal{Q}_{12} \\ \mathcal{Q}_{21} & \mathcal{Q}_{22} \end{pmatrix} u \, .$$

Then the first equation in (3.11) reduces to $\mathcal{R}w = 0$. By setting $w = 0$, the second equation in (3.11) becomes

$$\mathcal{A}_1 v = 2\alpha \mathcal{B}_1 v \, , \tag{3.13}$$

where

$$\mathcal{A}_1 = \mathcal{A} \begin{pmatrix} \mathcal{Q}_{21}^* \\ \mathcal{Q}_{22}^* \end{pmatrix} \quad \text{and} \quad \mathcal{B}_1 = \begin{pmatrix} -Q_{12} \otimes I & 0 \\ 0 & I \otimes Q_{12} \end{pmatrix} \mathcal{Q}_{22}^* \, .$$

Equation (3.13) is now a $2n^2$-by-$2n^2$ generalized eigenvalue problem. Hence we have reduced the problem of finding a nonzero solution to (3.10) to the generalized eigenvalue problem (3.13).

To summarize, we have shown that in order for (3.2) to have at least one real solution (α, β), both matrix pencils in (3.8) and (3.9) must share a common pure imaginary eigenvalue βi. This requires that the matrix equation (3.10) must have a nonzero solution which, in turn, is equivalent to requiring that the generalized eigenvalue problem (3.13) have a real eigenvalue α.

In order to verify whether $\delta > \rho(A, B)$ in any bisection step of Algorithm 1.1, we set $\eta = \delta$ in (3.2) and check whether the generalized eigenvalue problem (3.13) has any real eigenvalues α. If it does, we then check for each real α whether the two matrix pencils in (3.8) and (3.9) share a common pure imaginary eigenvalue βi. If they do for at least one α, then we have found a pair of α and β such that

$$\sigma\left([A - (\alpha + \beta i)\, I, B]\right) = \delta \quad \text{and hence} \quad \delta \geq \rho(A, B).$$

On the other hand, if (3.13) does not have a real eigenvalue, or if the matrix pencils in (3.8) and (3.9) do not share a common pure imaginary eigenvalue for any real eigenvalue of (3.13), then we conclude by Theorem 3.1 that

$$\delta = \eta > 2(\delta - \rho(A, B)).$$

On the other hand, Algorithm 1.1 guarantees that $2\delta \geq \rho(A, B)$ from the previous bisection step. Thus the value of δ after Algorithm 1.1 exits from the **while** loop must satisfy

$$\frac{\rho(A, B)}{2} \leq \delta \leq 2\, \rho(A, B)\ . \tag{3.14}$$

3.3. A generally quadratically convergent variation. Algorithm 1.1 converges linearly. Following Boyd and Balakrishnan [8] (see section 2.2), we develop a generally quadratically convergent version of Algorithm 1.1 in this section. In the following development, we assume that $f(\alpha, \beta)$ is analytic in both α and β in a small neighborhood of (α_*, β_*) so that $f(\alpha, \beta)$ permits the expansions

$$f(\alpha + \eta, \beta) = f(\alpha, \beta) + \eta \frac{\partial}{\partial \alpha} f(\alpha + \frac{\eta}{2}, \beta) + O(\eta^3), \tag{3.15}$$

$$\begin{aligned} f(\alpha, \beta) = {} & f(\alpha_*, \beta_*) + \gamma\,(\alpha - \alpha_*)^2 + 2\nu\,(\alpha - \alpha_*)\,(\beta - \beta_*) + \xi\,(\beta - \beta_*)^2 \\ & + O\left((|\alpha - \alpha_*| + |\beta - \beta_*|)^3\right), \end{aligned} \tag{3.16}$$

where γ, ν, and ξ are the relevant second-order partial derivatives at (α_*, β_*). We further assume that the matrix $\Gamma \equiv \left(\begin{smallmatrix} \gamma & \nu \\ \nu & \xi \end{smallmatrix}\right)$ is positive definite. These expansions with a positive definite Γ imply that (α_*, β_*) is at least a local minimum.

Now assume that α and β are such that $f(\alpha + \eta, \beta) = f(\alpha, \beta)$. It follows from (3.15) that

$$\frac{\partial}{\partial \alpha} f\left(\alpha + \frac{\eta}{2}, \beta\right) + O(\eta^2) = 0\ .$$

Expanding the partial derivative at (α_*, β_*) to get

$$\gamma\left(\alpha + \frac{\eta}{2} - \alpha_*\right) + \nu\,(\beta - \beta_*) = O\left(\eta^2 + (\alpha - \alpha_*)^2 + (\beta - \beta_*)^2\right), \tag{3.17}$$

where we have used the fact that $\frac{\partial}{\partial \alpha} f(\alpha_*, \beta_*) = 0$.

Let $\mu = \delta - f\,(\alpha_*, \beta_*) = \delta - \rho(A, B)$; and let (α_1, β_1) and (α_2, β_2) be the two solutions to (3.2) that are near (α_*, β_*) (see Theorem 3.1). It follows that α_i and β_i satisfy both (3.17) and (3.16) for $i = 1$ and 2. Now let $\zeta_{1,i}$ and $\zeta_{2,i}$ denote the error terms in (3.17) and (3.16) for (α_i, β_i). Consequently,

$$\begin{aligned} \gamma\left(\alpha_i + \frac{\eta}{2} - \alpha_*\right) + \nu\,(\beta_i - \beta_*) &= \zeta_{1,i}\ , \\ \gamma\,(\alpha_i - \alpha_*)^2 + 2\nu\,(\alpha_i - \alpha_*)\,(\beta_i - \beta_*) + \xi\,(\beta_i - \beta_*)^2 &= \mu + \zeta_{2,i}\ . \end{aligned}$$

It follows from the first equation that

$$(\alpha_i - \alpha_*) = -\frac{\eta}{2} - \frac{\nu}{\gamma}\,(\beta_i - \beta_*) + \frac{\zeta_{1,i}}{\gamma}\ . \tag{3.18}$$

Plugging this into the second equation and simplifying lead to

$$(\beta_i - \beta_*)^2 = \frac{\mu - \gamma\eta^2/4}{\xi - \dfrac{\nu^2}{\gamma}} + \zeta_{3,i} \ ,$$

where

$$\begin{aligned}\zeta_{3,i} &= \frac{\zeta_{2,i} + 2\zeta_{1,i}\left(\dfrac{\eta}{2} + \dfrac{\nu}{\gamma}\left(\beta_i - \beta_*\right)\right) - \dfrac{\zeta_{1,i}^2}{\gamma} - 2\zeta_{1,i}\left(\beta_i - \beta_*\right)\dfrac{\nu}{\gamma}}{\xi - \dfrac{\nu^2}{\gamma}} \\ &= O\left(\left(\eta + |\alpha_i - \alpha_*| + |\beta_i - \beta_*|\right)^3\right) .\end{aligned}$$

On the other hand, since Γ is assumed to be positive definite, (3.16) implies that $|\alpha_i - \alpha_*|^2 + |\beta_i - \beta_*|^2 = \mathcal{O}(\mu)$. Furthermore, the choice of $\eta = \delta$ in Algorithm 1.1 ensures that $\eta = \mathcal{O}(\mu)$ (see section 3.2). Hence,

$$\begin{aligned}(\beta_i - \beta_*) &= \pm\sqrt{\frac{\mu - \gamma\eta^2/4}{\xi - \dfrac{\nu^2}{\gamma}} + \zeta_{3,i}} = \pm\sqrt{\frac{\mu - \gamma\eta^2/4}{\xi - \dfrac{\nu^2}{\gamma}}} + O\left(\frac{\zeta_{3,i}}{\sqrt{\mu}}\right) \\ &= \pm\sqrt{\frac{\mu - \gamma\eta^2/4}{\xi - \dfrac{\nu^2}{\gamma}}} + O\left(\mu\right) .\end{aligned}$$

The result can be rewritten as[3]

$$(\beta_1 - \beta_*) = \sqrt{\frac{\mu - \gamma\eta^2/4}{\xi - \frac{\nu^2}{\gamma}}} + O(\mu) \quad \text{and} \quad (\beta_2 - \beta_*) = -\sqrt{\frac{\mu - \gamma\eta^2/4}{\xi - \frac{\nu^2}{\gamma}}} + O(\mu) \ .$$

Combining these equations,

$$\beta_* = \frac{\beta_1 + \beta_2}{2} + O(\mu) \ .$$

Plugging this relation into (3.18) and combining the equations for $i = 1, 2$, we obtain

$$\alpha_* = \frac{\alpha_1 + \alpha_2 + \eta}{2} + O(\mu).$$

Define

$$\alpha_{\mathbf{new}} = \frac{\alpha_1 + \alpha_2 + \eta}{2} \quad \text{and} \quad \beta_{\mathbf{new}} = \frac{\beta_1 + \beta_2}{2} \ .$$

It follows that

$$\alpha_* = \alpha_{\mathbf{new}} + O(\mu) \quad \text{and} \quad \beta_* = \beta_{\mathbf{new}} + O(\mu)$$

[3]These relations hold as long as $\gamma\eta^2/4 \ll \mu$. It is likely that under certain conditions Theorem 3.1 holds for much larger values of η as well.

and that

$$f(\alpha_{\mathbf{new}}, \beta_{\mathbf{new}}) - f(\alpha_*, \beta_*) = \mathcal{O}\left(|\alpha_{\mathbf{new}} - \alpha_*|^2 + |\beta_{\mathbf{new}} - \beta_*|^2\right) = (\mathcal{O}(\mu))^2. \tag{3.19}$$

We note that this relation is very similar to (2.6). Now we modify Algorithm 1.1 to get the following.

ALGORITHM 3.1. QUADRATICALLY CONVERGENT VARIATION OF ALGORITHM 2.1.

Set $\delta := g(0)$.

while $\delta \geq \rho(A, B)$

Choose two real solutions (α_1, β_1) and (α_2, β_2) of (3.2).

$$\delta := \min\left\{\delta, f\left(\frac{\alpha_1 + \alpha_2 + \eta}{2}, \frac{\beta_1 + \beta_2}{2}\right)\right\}/2.$$

endwhile

In our implementation, we computed

$$f\left(\frac{\alpha_1 + \alpha_2 + \eta}{2}, \frac{\beta_1 + \beta_2}{2}\right)$$

among adjacent pairs of real solutions and chose the pair with smallest f value. We note that in both Algorithms 1.1 and 3.1, we can compute a better initial guess δ by using Algorithm 2.2 with some values of λ_0 and θ, such as $\lambda_0 = 0$ and $\theta = 0$.

Algorithm 3.1 was derived under the assumptions noted at the beginning of section 3.3, which need not hold for all linear control systems of the form (1.1). Hence estimate (3.19) may not hold for some linear control systems. However, it is clear that Algorithm 3.1 converges at least linearly to arrive at an estimate δ that satisfies (3.14). Similar to Algorithm 2.2, Algorithm 3.1 is not, strictly speaking, quadratically convergent since it terminates as soon as it has found a δ that satisfies (3.14). Nevertheless, Algorithm 3.1 does converge much more rapidly than Algorithm 1.1 when $\rho(A, B)$ is tiny. We discuss this point further in section 4.

3.4. Further considerations. Sometimes it may be more important to find the uncontrollable modes of (1.1) for a given tolerance ε. In this case, we solve (3.2) with $\delta = \eta = \varepsilon$. If there are no solutions to (3.2), then the system (1.1) is controllable; otherwise, each solution to (3.2) corresponds to an uncontrollable mode. Conversely, it is easy to see from the proof of Theorem 3.1 that any uncontrollable mode will result in at least two solutions to (3.2). Hence the set of all solutions to (3.2) provide approximations to the uncontrollable modes of (3.2). The formulas for $\alpha_{\mathbf{new}}$ and $\beta_{\mathbf{new}}$ provide more accurate approximations to these modes.

If $\rho(A, B)$ is very small, then $\delta = \eta$ will also become very small during the execution of Algorithms 1.1 and 3.1. In fact, for small enough η, the two different points $\alpha + \beta i$ and $\alpha + \eta + \beta i$ will look identical. Hence the solutions to (3.2) are potentially ill conditioned. See section 4 for more details.

Like many other algorithms in engineering computations, such as those for semidefinite programming [1, 25, 29], both Algorithms 1.1 and 3.1 are expensive for large problems, since both the reduction to and the solution of the pencil (3.13) require $O(n^6)$ floating point operations. However, the eigenvalue problem (3.11) is highly sparse as a $4n^2 \times 4n^2$ problem. It is likely that sparse matrix computation technologies, such as the implicitly restarted Arnoldi iteration [22, 23, 28], can be used to compute the real eigenvalues of (3.13) quickly. The effectiveness of this approach is currently under thorough investigation.

4. Numerical experiments. We have done some elementary numerical experiments with Algorithms 1.1 and 3.1. In this section we report some of the results obtained from these experiments. The experiments were done in `matlab` in double precision.

Matrices in Examples 2 through 5 were taken from Gao and Neumann [18]. These are systems with small $\rho(A, B)$. Global optimization methods (Byers [11], Gao and Neumann [18], and He [19]) could require prohibitively expensive computation time to correctly estimate $\rho(A, B)$ in these cases. On the other hand, both Algorithms 1.1 and 3.1 worked well on them, with Algorithm 3.1 converging much faster than Algorithm 1.1 as expected.

Example 1. In this example we took $A \in \mathbf{R}^{5\times 5}$ and $B \in \mathbf{R}^5$ to be random matrices. This is a matrix pair with fairly large $\rho(A, B)$. Both Algorithms 1.1 and 3.1 took two iterations to terminate. This example illustrates that, for linear systems (1.1) far away from the set of uncontrollable systems, both Algorithms 1.1 and 3.1 take very few iterations.

Example 2. In this example we took

$$A = \begin{pmatrix} 1 & 1 & 2 & 3 \\ -1 & 1 & 4 & 5 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & -2 & 1 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}.$$

This pair is uncontrollable since the smallest singular value of $[A - (1 \pm 2i)\, I \;\; B]$ is zero. Algorithms 1.1 and 3.1 took 42 and 5 iterations, respectively, to find $\rho(A, B) \leq 10^{-15}$.

Example 3. In this example we took

$$A = \begin{pmatrix} -0.32616458 & -0.09430266 & 0.05207847 & -.08481401 & 0.05829280 \\ 0.01158922 & -.39787419 & -.14901699 & -.01394125 & -.10626942 \\ 0.05623810 & -.03153954 & -.50160557 & -.05748511 & -.00552321 \\ 0.07474298 & .01205259 & -.00249719 & -.26871590 & ..23107147 \\ ..07829650 & -.19442017 & -.05631780 & .17761876 & -.45847313 \end{pmatrix}$$

and $B = (-.13705019, -.18309206, -.13735654, .80258791, ..53355446)^T$. Algorithm 1.1 took 20 iterations to return $\delta = 4.49639424 \times 10^{-7}$. For $\delta = 8.99278848 \times 10^{-7}$, it returned six distinct solutions to (3.2): $(\alpha, \beta) = \left(-6.5094391 \times 10^{-1}, \pm 5.02193288 \times 10^{-7}\right)$ and

$$\left(-2.60379928 \times 10^{-1}, \pm 5.20865745 \times 10^{-7}\right),$$

$$\left(-5.20755942 \times 10^{-1}, \pm 9.06572930 \times 10^{-7}\right).$$

On the other hand, Algorithm 3.1 took 4 iterations to return $\delta = 3.81345287 \times 10^{-7}$. For $\delta = 7.62690574 \times 10^{-7}$, Algorithm 3.1 returned one distinct solution, $(\alpha, \beta) = \left(-5.207554634 \times 10^{-1}, 0\right)$.

Example 4. In this example we took

$$A = \begin{pmatrix} -.22907968 & 0.08886286 & -.18085425 & -.03469234 & -.32819211 \\ ..11868229 & -.43816868 & -.27812914 & -.04200964 & -.07784618 \\ -.02507663 & .30736050 & -.24819024 & .21852948 & -.06260819 \\ ..16055050 & -.00818190 & -.19591208 & .08940924 & .22683641 \\ -.19138555 & .13088864 & -.22839105 & -.23175762 & ..12274100 \end{pmatrix}$$

and $B = (-.73491186, -.35694241, .04637973, .52703303, ..22930713)^T$. Algorithm 1.1 took 7 iterations to return $\delta = 3.64238211 \times 10^{-5}$. For $\delta = 7.28476422 \times 10^{-5}$, it returned four distinct solutions:

$$(\alpha, \beta) = \left(-9.48126863 \times 10^{-2}, \pm 3.91642858 \times 10^{-2}\right),$$
$$\left(-9.47712590 \times 10^{-2}, \pm 4.63687903 \times 10^{-2}\right).$$

On the other hand, Algorithm 3.1 took 3 iterations to return $\delta = 3.40238900 \times 10^{-5}$. For $\delta = 6.80477800 \times 10^{-5}$, Algorithm 3.1 returned two distinct solutions, $(\alpha, \beta) = \left(-9.67874713, \pm 4.031215939 \times 10^{-2}\right)$.

Example 5. In this example we took

$$A = \begin{pmatrix} -.27422658 & -.21968089 & -.21065336 & -.22134064 & 0.19235875 \\ -.07210867 & .18848014 & -.29068998 & .28936270 & 0.10007703 \\ -.03547166 & .17931676 & .14590007 & .00556579 & .38838791 \\ ..00029995 & .14755893 & -.25420697 & -.12193382 & -.14071387 \\ -.07780546 & -.29477373 & .01366200 & .32749991 & -.0131683 \end{pmatrix}$$

and $B = (.81475593, -.30523653, -.34286610, -.05815542, ..34937688)^T$. Algorithm 1.1 took 4 iterations to return $\delta = 1.60740324 \times 10^{-7}$. For $\delta = 3.21480648 \times 10^{-7}$, it returned 14 distinct solutions to (3.2). On the other hand, Algorithm 3.1 took 2 iterations to return $\delta = 1.08714072 \times 10^{-7}$. And for $\delta = 2.17428144 \times 10^{-7}$, it returned one distinct solution $(\alpha, \beta) = \left(8.37424478 \times 10^{-3}, 0\right)$.

Example 6. In this example we took the matrix pair (A, B) from Example 2, and set

$$A := Q\ A\ Q^T \quad \text{and} \quad B := Q\ B,$$

where Q is a random orthogonal matrix. This new matrix pair is still uncontrollable. But Algorithm 1.1 took 28 iterations to return $\delta = 4.06303199 \times 10^{-9}$, while Algorithm 3.1 took 6 iterations to return $\delta = 2.33160388 \times 10^{-10}$. This example illustrates that both algorithms can have numerical difficulties in correctly estimating $\rho(A, B)$ if it is very tiny.

5. Conclusions and extensions. In this paper we have presented the first algorithms that require a cost polynomial in the matrix size to correctly estimate the controllability distance $\rho(A, B)$ for a given linear control system. And we have demonstrated their effectiveness and reliability through some numerical experiments.

The biggest open question is how to further reduce the cost. At the core of these algorithms is the computation of all real eigenvalues of a sparse $4n^2 \times 4n^2$ eigenvalue problem. Currently, we find these eigenvalues by treating the eigenvalue problem as a dense one, resulting in algorithms that are too expensive for large problems. In the future, we plan to exploit the possibility of finding these real eigenvalues via sparse matrix computation technologies, such as the implicitly restarted Arnoldi iteration [22, 23, 28], to significantly reduce the computation cost.

Another open question is how to better understand the effects of finite precision arithmetic on the estimated distance $\rho(A, B)$. As we observed in section 4, if $\rho(A, B)$ is very tiny, then the distance estimated by the new algorithms in finite precision could be much larger than the exact distance.

Finally, the perturbation $[\Delta A, \Delta B]$ in (1.3) can be complex even if both A and B are real. It is known (Byers [11]) that the normwise smallest real perturbation can

be much larger than $\rho(A, B)$. Whether our new algorithms shed new light on the computation of the normwise smallest real perturbation remains to be seen.

REFERENCES

[1] F. Alizadeh, *Interior-point methods in semidefinite programming with applications to combinatorial optimization*, SIAM J. Optim., 5 (1995), pp. 13–51.

[2] T. Beelen and P. Van Dooren, *An improved algorithm for the computation of Kronecker's canonical form of a singular pencil*, Linear Algebra Appl., 105 (1988), pp. 9–65.

[3] T. Beelen, P. Van Dooren, and M. Verhaegen, *A class of staircase algorithms for generalized state space systems*, in Proceedings of the American Control Conference, Seattle, WA, 1986, pp. 425–426.

[4] D. Boley, *Computing the Controllability/Observability Decomposition of a Linear Time-Invariant Dynamic System: A Numerical Approach*, Ph.D. thesis, Stanford University, Stanford, CA, 1981.

[5] D. Boley and G. Golub, *The Lanczos-Arnoldi algorithm and controllability*, Systems Control Lett., 4 (1984), pp. 317–324.

[6] D. L. Boley, *Computing rank-deficiency of rectangular matrix pencils*, Systems Control Lett., 9 (1987), pp. 207–214.

[7] D. L. Boley and W.-S. Lu, *Measuring how far a controllable system is from uncontrollable one*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 249–251.

[8] S. Boyd and V. Balakrishnan, *A regularity result for the singular values of a transfer matrix and a quadratically convergent algorithm for computing its* $\mathbf{L}_\infty$*-norm*, Systems Control Lett., 15 (1990), pp. 1–7.

[9] S. Boyd, V. Balakrishnan, and P. Kabamba, *A bisection method for computing the* $\mathbf{H}_\infty$ *norm of a transfer matrix and related problems*, Math. Control Signals Systems, 2 (1989), pp. 207–219.

[10] R. Byers, *A bisection method for measuring the distance of a stable matrix to the unstable matrices*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 875–881.

[11] R. Byers, *Detecting nearly uncontrollable pairs*, in Numerical Methods Proceedings of the International Symposium MTNS-89, vol. III, M. A. Kaashoek, J. H. van Schuppen, and A. C. M. Ran, eds., Springer-Verlag, New York, 1990, pp. 447–457.

[12] J. Demmel, *On condition numbers and the distance to the nearest ill-posed problem*, Numer. Math., 51 (1987), pp. 251–289.

[13] J. Demmel and B. Kågström, *Accurate solutions of ill-posed problems in control theory*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 126–145.

[14] J. Demmel and B. Kågström, *The generalized Schur decomposition of an arbitrary pencil* $A - \lambda B$*: Robust software with error bounds and applications. Parts* I *and* II, ACM Trans. Math. Software, 19 (1993), pp. 160–201.

[15] R. Eising, *The Distance Between a System and the Set of Uncontrollable Systems*, Memo COSOR 82-19, Eindhoven Univ. Technol., Eindhoven, The Netherlands, 1982.

[16] R. Eising, *Between controllable and uncontrollable*, Systems Control Lett., 4 (1984), pp. 263–264.

[17] L. Elsner and C. He, *An algorithm for computing the distance to uncontrollability*, Systems Control Lett., 17 (1991), pp. 453–464.

[18] M. Gao and M. Neumann, *A global minimum search algorithm for estimating the distance to uncontrollability*, Linear Algebra Appl., 188-189 (1993), pp. 305–350.

[19] C. He, *Estimating the distance to uncontrollability: A fast method and a slow one*, Systems Control Lett., 26 (1995), pp. 275–281.

[20] T. Kailath, *Linear Systems*, Prentice–Hall, Englewood Cliffs, NJ, 1980.

[21] A. J. Laub, *Survey of computational methods in control theory*, in Electric Power Problems: The Mathematical Challenge, A. M. Erisman, K. W. Neves, and M. H. Dwarakanath, eds., 1980, SIAM, Philadelphia, PA, pp. 231–260.

[22] R. Lehoucq, *Analysis and Implementation of an Implicitly Restarted Arnoldi Iteration*, Ph.D. thesis, Rice University, Houston, TX, 1995.

[23] R. B. Lehoucq and D. C. Sorensen, *Deflation techniques for an implicitly restarted Arnoldi iteration*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 789–821.

[24] G. Miminis, *Numerical Algorithms for Controllability and Eigenvalue Location*, Masters thesis, School of Computer Science, McGill University, Montreal, Quebec, Canada, 1981.

[25] R. D. C. Monteiro, *Primal-Dual Path Following Algorithms for Semidefinite Programming*, manuscript, School of Industrial and Systems Engineering, Georgia Institute of Technology,

Atlanta, GA, 1995.

[26] C. C. Paige, *Properties of numerical algorithms related to computing controllability*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 130–138.

[27] R. V. Patel, A. J. Laub, and P. M. Van Dooren, *Numerical Linear Algebra Techniques for Systems and Control*, IEEE Press, New York, 1994.

[28] D. Sorensen, *Implicit application of polynomial filters in a k-step Arnoldi method*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 357–385.

[29] M. J. Todd, K. C. Toh, and R. H. Tütüncü, *On the Nesterov-Todd Direction in Semidefinite Programming*, Technical report 1154, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, 1996.

[30] P. Van Dooren, *The computation of Kronecker's canonical form of a singular pencil*, Linear Algebra Appl., 27 (1979), pp. 103–141.

[31] M. Wicks and R. A. DeCarlo, *Computing the distance to an uncontrollable system*, IEEE Trans. Automat. Control, AC-36 (1991), pp. 39–49.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 3, pp. 1004–1025
© 2000 Society for Industrial and Applied Mathematics

STAIRCASE FAILURES EXPLAINED BY ORTHOGONAL VERSAL FORMS*

ALAN EDELMAN† AND YANYUAN MA‡

Abstract. Treating matrices as points in n^2-dimensional space, we apply geometry to study and explain algorithms for the numerical determination of the Jordan structure of a matrix. Traditional notions such as sensitivity of subspaces are replaced with angles between tangent spaces of manifolds in n^2-dimensional space. We show that the subspace sensitivity is associated with a small angle between complementary subspaces of a tangent space on a manifold in n^2-dimensional space. We further show that staircase algorithm failure is related to a small angle between what we call staircase invariant space and this tangent space. The matrix notions in n^2-dimensional space are generalized to pencils in $2mn$-dimensional space. We apply our theory to special examples studied by Boley, Demmel, and Kågström.

Key words. staircase algorithm, Jordan structure, Kronecker structure, versal deformation, SVD

AMS subject classification. 65F99

PII. S089547989833574X

1. Introduction.

The problem of accurately computing Jordan–Kronecker canonical structures of matrices and pencils has captured the attention of many specialists in numerical linear algebra. Standard algorithms for this process are denoted "staircase algorithms" because of the shape of the resulting matrices [22, p. 370]. Little is understood concerning how and why they fail, and in this paper, we study the geometry of matrices in n^2-dimensional space and pencils in $2mn$-dimensional space to explain these failures. This follows a geometrical program to complement and perhaps replace traditional numerical concepts associated with matrix subspaces that are usually viewed in n-dimensional space.

This paper targets readers who are already familiar with the staircase algorithm. We refer them to [22, p. 370] and [10] for excellent background material and list other literature in section 1.1 for those who wish to have a comprehensive understanding of the algorithm. On the mathematical side, it is also helpful if the reader has some knowledge of Arnold's theory of versal forms, though a dedicated reader should be able to read this paper without such knowledge, perhaps skipping section 3.2.

The most important contributions of this paper may be summarized as follows:

- A geometrical explanation of staircase algorithm failures is given.
- Three significant subspaces are identified that decompose matrix or pencil space: $\mathcal{T}_b$, $\mathcal{R}$, $\mathcal{S}$. The most important of these spaces is $\mathcal{S}$, which we choose to call the "staircase invariant space."
- The idea that the staircase algorithm computes an Arnold normal form that is numerically more appropriate than Arnold's "matrices depending on parameters" is discussed.

*Received by the editors March 16, 1998; accepted for publication (in revised form) by P. Van Dooren May 3, 1999; published electronically March 8, 2000.
http://www.siam.org/journals/simax/21-3/33574.html

†Department of Mathematics, Massachusetts Institute of Technology, Room 2-380, Cambridge, MA 02139-4307 (edelman@math.mit.edu, http://www-math.mit.edu/~edelman). This author was supported by NSF grants 9501278-DMS and 9404326-CCR.

‡Department of Mathematics, Massachusetts Institute of Technology, Room 2-333, Cambridge, MA 02139-4307 (yanyuan@math.mit.edu, http://www-math.mit.edu/~yanyuan). This author was supported by NSF grant 9501278-DMS.

- A first order perturbation theory for the staircase algorithm is given.
- The theory is illustrated using an example by Boley [3].

The paper is organized as follows: In section 1.1 we briefly review the literature on staircase algorithms. In section 1.2 we introduce concepts that we call *pure, greedy,* and *directed* staircases to emphasize subtle distinctions on how the algorithm might be used. Section 1.3 contains some important messages that result from the theory to follow.

Section 2 presents two similar-looking matrices with very different staircase behavior. Section 3 studies the relevant n^2-dimensional geometry of matrix space, while section 4 applies this theory to the staircase algorithm. The main result may be found in Theorem 6.

Sections 5, 6, and 7 mimic sections 2, 3, and 4 for matrix pencils. Section 8 applies the theory toward special cases introduced by Boley [3] and Demmel and Kågström [12].

1.1. Jordan–Kronecker algorithm history. The first staircase algorithm was given by Kublanovskaya for Jordan structure in 1966 [32], where a normalized QR factorization is used for rank determination and nullspace separation. Ruhe [35] first introduced the use of the SVD into the algorithm in 1970. The SVD idea was further developed by Golub and Wilkinson [23, section 10]. Kågström and Ruhe [28, 29] wrote the first library-quality software for the complete Jordan normal form reduction, with the capability of returning after different steps in the reduction. Recently, Chaitin-Chatelin and Frayssé [6] developed a nonstaircase "qualitative" approach.

The staircase algorithm for the Kronecker structure of pencils was given by Van Dooren [13, 14, 15] and Kågström and Ruhe [30]. Kublanovskaya [33] fully analyzed the AB algorithm; however, earlier work on the AB algorithm goes back to the 1970s. Kågström [26, 27] gave an RGDSVD/RGQZD algorithm and this provided a base for later work on software. Error bounds for this algorithm were given by Demmel and Kågström [8, 9]. Beelen and Van Dooren [2] gave an improved algorithm which requires $O(m^2n)$ operations for $m \times n$ pencils. Boley [3] studied the sensitivity of the algebraic structure. Error bounds are given by Demmel and Kågström [10, 11].

Staircase algorithms are used both theoretically and practically. Elmroth and Kågström [19] used the staircase algorithm to test the set of 2×3 pencils; hence to analyze the algorithm Demmel and Edelman [7] used the algorithm to calculate the dimension of matrices and pencils with a given form. Van Dooren [14], Emami-Naeini and Van Dooren [20], Kautsky, Nichols, and Van Dooren [31], Boley and Van Dooren [5], and Wicks and DeCarlo [36] considered systems and control applications. Software for control theory was provided by Demmel and Kågström [12].

A number of papers used geometry to understand Jordan–Kronecker structure problems. Fairgrieve [21] regularized by taking the most degenerate matrix in a neighborhood; Edelman, Elmroth, and Kågström [17, 18] studied versality and stratifications; and Boley [4] concentrates on stratifications.

1.2. The staircase algorithms. Staircase algorithms for the Jordan–Kronecker form work by making sequences of rank decisions in combination with eigenvalue computations. We coin the terms *pure*, *greedy*, and *directed* staircases to emphasize a few variations on how the algorithm might be used. Pseudocode for the Jordan versions appears near the end of this subsection. In combination with these three choices, one can choose an option of *zeroing*. These choices are explained below.

The three variations for purposes of discussion are considered in exact arithmetic. The *pure* version is the pure mathematician's algorithm: It gives precisely the Jordan

structure of a given matrix. The *greedy* version (also useful for a pure mathematician!) attempts to find the most "interesting" Jordan structure near the given matrix. The *directed* staircase attempts to find a nearby matrix with a preconceived Jordan structure. Roughly speaking, the difference between pure, greedy, and directed is whether the Jordan structure is determined by the matrix, by a user-controlled neighborhood of the matrix, or directly by the user, respectively.

In the *pure* staircase algorithm, rank decisions are made using the singular value decomposition. An explicit distinction is made between zero singular values and nonzero singular values. This determines the exact Jordan form of the input matrix.

The *greedy* staircase algorithm attempts to find the most interesting Jordan structure near the given matrix. Here the word "interesting" (or "degenerate") is used in the same sense as it is with precious gems—the rarer, the more interesting. Algorithmically, as many singular values as possible are thresholded to zero with a user-defined threshold. The more singular values that are set to zero, the rarer in the sense of codimension (see [7, 17, 18]).

The *directed* staircase algorithm allows the user to decide in advance what Jordan structure is desired. The Jordan structure dictates which singular values are set to 0. Directed staircase is used in a few special circumstances. For example, it is used when separating the zero Jordan structure from the right singular structure (used in `GUPTRI` [10, 11]). Moreover, Elmroth and Kågström imposed structures by the staircase algorithm in their investigation of the set of 2×3 pencils [19]. Recently, Lippert and Edelman [34] use directed staircase to compute an initial guess for a Newton minimization approach to computing the nearest matrix with a given form in the Frobenius norm.

In the greedy and directed modes, if we explicitly *zero* the singular values, we end up computing a new matrix in staircase form that has the same Jordan structure as a matrix *near* the original one. If we do not explicitly zero the singular values, we end up computing a matrix that is orthogonally similar to the original one (in the absence of roundoff errors), which is *nearly* in staircase form. For example, in `GUPTRI` [11], the choice of whether to zero the singular values is made by the user with an input parameter named `zero` which may be true or false.

To summarize the many choices associated with a staircase algorithm, there are really five distinct algorithms worth considering: The pure algorithm stands on its own; otherwise, the two choices of combinatorial structure (greedy and directed) may be paired with the choice to zero or not. Thereby we have the five algorithms:

1. pure staircase,
2. greedy staircase with zeroing,
3. greedy staircase without zeroing,
4. directed staircase with zeroing, and
5. directed staircase without zeroing.

Notice that in the pure staircase, we do not specify zeroing or not zeroing, since both will give the same result vacuously.

Of course, algorithms run in finite precision. One further detail is that there is some freedom in the singular value calculations which leads to an ambiguity in the staircase form: In the case of unequal singular values, an order must be specified, and when singular values are equal, there is a choice of basis to be made. We will not specify any order for the SVD, except that all singular values considered to be zero appear first.

In the ith loop iteration, we use w_i to denote the number of singular values that are considered to be zero. For the directed algorithm, w_i are input; otherwise, w_i are

computed. In pseudocode, we have the following staircase algorithms for computing the Jordan form corresponding to eigenvalue λ.

INPUT:
1) matrix A
2) specify pure, greedy, or direct mode
3) specify zeroing or not zeroing

OUTPUT:
1) matrix A that may or may not be in staircase form
2) Q (optional)

$i = 0, \quad Q = I$
$A_{tmp} = A - \lambda I$
while A_{tmp} not full rank
  $i = i + 1$
  Let $n' = \sum_{j=1}^{i-1} w_j$ and $n_{tmp} = n - n' = dim(A_{tmp})$
  Use the SVD to compute an n_{tmp} by n_{tmp} unitary matrix V whose leading w_i columns span the nullspace or an approximation
    Choice I: Pure: Use the SVD algorithm to compute w_i and the exact nullspace
    Choice II: Greedy: Use the SVD algorithm and threshold the small singular values with a user specified tolerance, thereby defining w_i. The corresponding singular vectors become the first w_i vectors of V.
    Choice III: Directed: Use the SVD algorithm, the w_i are defined from the input Jordan structure. The w_i singular vectors are the first w_i columns of V.
  $A = \text{diag}(I_{n'}, V^*) \cdot A \cdot \text{diag}(I_{n'}, V), \quad Q = Q \cdot \text{diag}(I_{n'}, V)$
  Let A_{tmp} be the lower right $n_{tmp} - w_i$ by $n_{tmp} - w_i$ corner of A
  $A_{tmp} = A_{tmp} - \lambda I$
endwhile
If zeroing, return A in the form $\lambda I +$ a block strictly upper triangular matrix.

While the staircase algorithm often works very well, it has been known to fail. We can say that the greedy algorithm fails if it does not detect a matrix with the least generic form [7] possible within a given tolerance. We say that the directed algorithm fails if the staircase form it produces is very far (orders of magnitude, in terms of the usual Frobenious norm of matrix space) from the staircase form of the nearest matrix with the intended structure. In this paper, we mainly concentrate on the greedy staircase algorithm and its failure, but the theory is applicable to both approaches. We emphasize that we are intentionally vague about how far is "far" as this may be application dependent, but we will consider several orders of magnitude to constitute this notion.

1.3. Geometry of staircase and Arnold forms. Our geometrical approach is inspired by Arnold's theory of versality [1]. For readers already familiar with Arnold's theory, we point out that we have a new normal form that enjoys the same properties as Arnold's original form, but is more useful numerically. For numerical analysts, we point out that these ideas are important for understanding the staircase algorithm. Perhaps it is safe to say that numerical analysts have had an "Arnold normal form" for years, but we did not recognize it as such—the computer was doing it for us automatically.

TABLE 1.1

		Angles			Components	
A	Staircase fails	$\langle \mathcal{S}, \mathcal{T}_b \oplus \mathcal{R} \rangle$	$\langle \mathcal{T}_b, \mathcal{R} \rangle$	$\langle \mathcal{S}, \mathcal{R} \rangle$	$\mathcal{S}$	$\mathcal{R}$
No weak stair	no	large	large	$\pi/2$	small	small
Weak stair	no	large	small	$\pi/2$	small	large
Weak stair	yes	small	small	$\pi/2$	large	large

The strength of the normal form that we introduce in section 3 is that it provides a first order rounding theory of the staircase algorithm. We will show that instead of decomposing the perturbation space into the normal space and a tangent space at a matrix A, the algorithm chooses a so-called staircase invariant space to take the place of the normal space. When some directions in the staircase invariant space are very close to the tangent space, the algorithm can fail.

From the theory, we decompose the matrix space into three subspaces that we call $\mathcal{T}_b$, $\mathcal{R}$, and $\mathcal{S}$, the precise definitions of which are given in Definitions 1 and 3. Here, $\mathcal{T}_b$ and $\mathcal{R}$ are two subspaces of the tangent space and $\mathcal{S}$ is a certain complementary space of the tangent space in the matrix space. For the eager reader, we point out that angles between these spaces are related to the behavior of the staircase algorithm; note that $\mathcal{R}$ is always orthogonal to $\mathcal{S}$. (We use $\langle \cdot, \cdot \rangle$ to represent the angle between two spaces.) See Table 1.1.

Here, by a weak stair [16] we mean the near rank deficiency of any superdiagonal block of the strictly block upper triangular matrix A.

2. A staircase algorithm's failure to motivate the theory. Consider the two matrices

$$A_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & \delta \\ 0 & 0 & 0 \end{pmatrix} \text{ and } A_2 = \begin{pmatrix} 0 & \delta & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix},$$

where $\delta =$`1.5e-9` is approximately on the order of the square root of the double precision machine $\epsilon = 2^{-52}$, roughly, `2.2e-16`. Both of these matrices clearly have the Jordan structure $J_3(0)$, but the staircase algorithm on A_1 and A_2 can behave very differently.

To test this, we used the `GUPTRI` [11] algorithm. `GUPTRI`[1] requires an input matrix A and two tolerance parameters `EPSU` and `GAP`. We ran `GUPTRI` on $\tilde{A_1} \equiv A_1 + \epsilon \mathtt{E}$ and $\tilde{A_2} \equiv A_2 + \epsilon \mathtt{E}$, where

$$\mathtt{E} = \begin{pmatrix} .3 & .4 & .2 \\ .8 & .3 & .6 \\ .4 & .9 & .6 \end{pmatrix},$$

and $\epsilon =$ `2.2e-14` is roughly 100 times the double-precision machine ϵ. The singular values of each of the two matrices $\tilde{A_1}$ and $\tilde{A_2}$ are $\sigma_1 =$ `1.0000e00`, $\sigma_2 =$ `1.4901e-09`, and $\sigma_3 =$ `8.8816e-15`. We set `GAP` to be always ≥ 1 and let `EPSU` $= a/(\|\tilde{A_i}\| *$ `GAP`$)$,

[1] GUPTRI [10, 11] is a "greedy" algorithm with a sophisticated thresholding procedure based on two input parameters EPSU and GAP ≥ 1. We threshold σ_{k-1} if $\sigma_{k-1} < GAP \times \max(\sigma_k, EPSU \times \|A\|)$ (defining $\sigma_{n+1} \equiv 0$). The first argument of the maximum σ_k ensures a large gap between thresholded and nonthresholded singular values. The second argument ensures that σ_{k-1} is small. Readers who look at the GUPTRI software should note that singular values are ordered from smallest to largest, contrary to modern convention.

TABLE 2.1

a	Computed Jordan structure for $\tilde{A}_1$	Computed Jordan structure for $\tilde{A}_2$
$a \geq \sigma_2$	$J_2(0) \oplus J_1(0) + O(10^{-9})$	$J_2(0) \oplus J_1(0) + O(10^{-9})$
$\gamma \leq a < \sigma_2$	$J_3(0) + O(10^{-6})$ ☹	$J_3(0) + O(10^{-14})$
$a < \gamma$	$J_1(0) \oplus J_1(\alpha) \oplus J_1(\beta) + O(10^{-14})$	$J_1(0) \oplus J_1(\alpha) \oplus J_1(\beta) + O(10^{-14})$

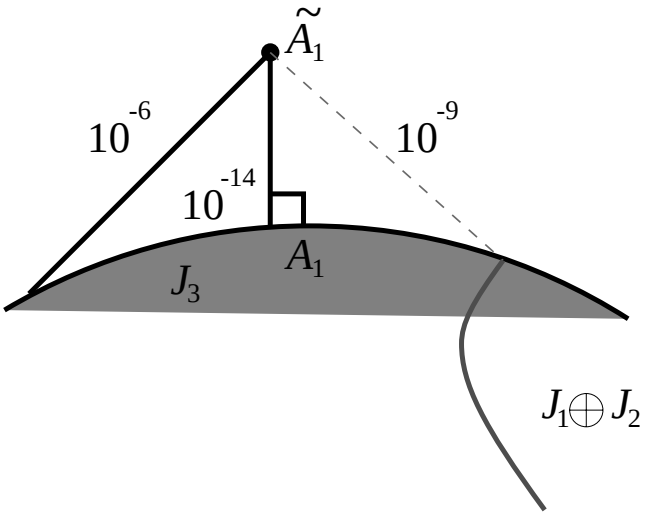


FIG. 2.1. *The staircase algorithm fails to find A_1 at distance* 2.2e-14 *from $\tilde{A}_1$ but does find a $J_3(0)$ or a $J_2(0) \oplus J_1(0)$ if given a much larger tolerance. (The latter is δ away from $\tilde{A}_1$.)*

where we vary the value of a. (The tolerance is effectively a.) Our observations are tabulated in Table 2.1.

Here, we use $J_k(\lambda)$ to represent a $k \times k$ Jordan block with eigenvalue λ. In the table, typically, $\alpha \neq \beta \neq 0$. Setting a small (smaller than $\gamma = 1.9985$`e-14` here, which is the smaller singular value in the second stage), the software returns two nonzero singular values in the first and second stages of the algorithm and one nonzero singular value in the third stage. Setting `EPSU` $\times$ `GAP` large (larger than σ_2 here), we zero two singular values in the first stage and one in the second stage, giving the structure $J_2(0) \oplus J_1(0)$ for both $\tilde{A}_1$ and $\tilde{A}_2$. (There is a matrix within $O(10^{-9})$ of A_1 and A_2 of the form $J_2(0) \oplus J_1(0)$.) The most interesting case is in between. For appropriate `EPSU` $\times$ `GAP` $\approx a$ (between γ and σ_2 here), we zero one singular value in each of the three stages, getting a $J_3(0)$ which is $O(10^{-14})$ away for A_2, while we can only get a $J_3(0)$ which is $O(10^{-6})$ away for A_1. In other words, the staircase algorithm fails for A_1 but not for A_2. As pictured in Figure 2.1, the A_1 example indicates that a matrix of the correct Jordan structure may be within the specified tolerance, but the staircase algorithm may fail to find it.

Consider the situation when A_1 and A_2 are transformed using a random orthogonal matrix Q. As a second experiment, we pick

$$Q \approx \begin{pmatrix} -.39878 & .20047 & -.89487 \\ -.84538 & -.45853 & .27400 \\ -.35540 & .86577 & .35233 \end{pmatrix}$$

TABLE 2.2

a	Computed Jordan structure for $\tilde{A}_1$	Computed Jordan structure for $\tilde{A}_2$
$a \geq \sigma_2$	$J_2(0) \oplus J_1(0) + O(10^{-5})$	$J_2(0) \oplus J_1(0) + O(10^{-6})$
$\gamma \leq a < \sigma_2$	$J_1(0) \oplus J_1(\alpha) \oplus J_1(\beta)$ ☹	$J_3(0) + O(10^{-6})$
$a < \gamma$	$J_1(0) \oplus J_1(\alpha) \oplus J_1(\beta) + O(10^{-14})$	$J_1(0) \oplus J_1(\alpha) \oplus J_1(\beta) + O(10^{-14})$

and take $\tilde{A}_1 = Q(A_1 + \epsilon E)Q^T$, $\tilde{A}_2 = Q(A_2 + \epsilon E)Q^T$. This will impose a perturbation of order ϵ. We ran `GUPTRI` on these two matrices; Table 2.2 shows the result.

In the table, $\gamma =$ `2.6980e-14`, all other values are the same as in the previous table.

In this case, `GUPTRI` is still able to detect a J_3 structure for $\tilde{A}_2$, although the one it finds is $O(10^{-6})$ away. But it fails to find any J_3 structure at all for $\tilde{A}_1$. The comparison of A_1 and A_2 in the two experiments indicates that the explanation is more subtle than the notion of a weak stair (a superdiagonal block that is almost column rank deficient) [16].

In this paper we present a geometrical theory that clearly predicts the difference between A_1 and A_2. The theory is based on how close certain directions that we will denote *staircase invariant directions* are to the tangent space of the manifold of matrices similar to the matrix with specified canonical form. It turns out that for A_1, these directions are nearly in the tangent space, but not for A_2. This is the crucial difference!

The tangent directions and the staircase invariant directions combine to form a "versal deformation" in the sense of Arnold [1], but one with more useful properties for our purposes.

3. Staircase invariant space and versal deformations.

3.1. The staircase invariant space and related subspaces. We consider block matrices as in Figure 3.1. Dividing a matrix A into blocks of row and column sizes $n_1, \ldots, n_k$, we obtain a *general block matrix*. A block matrix is *conforming to A* if it is also partitioned into blocks of size $n_1, \ldots, n_k$ in the same manner as A. If a general block matrix has nonzero entries only in the upper triangular blocks excluding the diagonal blocks, we call it a *block strictly upper triangular matrix*. If a general block matrix has nonzero entries only in the lower triangular blocks including the diagonal blocks, we call it a *block lower triangular matrix*. A matrix A is in *staircase form* if we can divide A into blocks of sizes $n_1 \geq n_2 \geq \cdots \geq n_k$ such that (s.t.) A is a strictly block upper triangular matrix and every superdiagonal block has full column rank. If a general block matrix has only nonzero entries on its diagonal blocks and each diagonal block is an orthogonal matrix, we call it a *block diagonal orthogonal matrix*. We call the matrix e^B a *block orthogonal matrix (conforming to A)* if B is a block antisymmetric matrix (conforming to A). (That is, B is antisymmetric with zero diagonal blocks. Here, we abuse the word "conforming" since e^B does not have a block structure.)

DEFINITION 1. *Suppose A is a matrix in staircase form. We call S a* staircase invariant matrix *of A if $S^TA = 0$ and S is block lower triangular. We call the space of matrices consisting of all such S the staircase invariant space of A, and denote it by $\mathcal{S}$.*

We remark that the columns of S will not be independent except possibly when $A = 0$; S can be the zero matrix as an extreme case. However the generic sparsity structure of S may be determined by the sizes of the blocks. For example, let A have

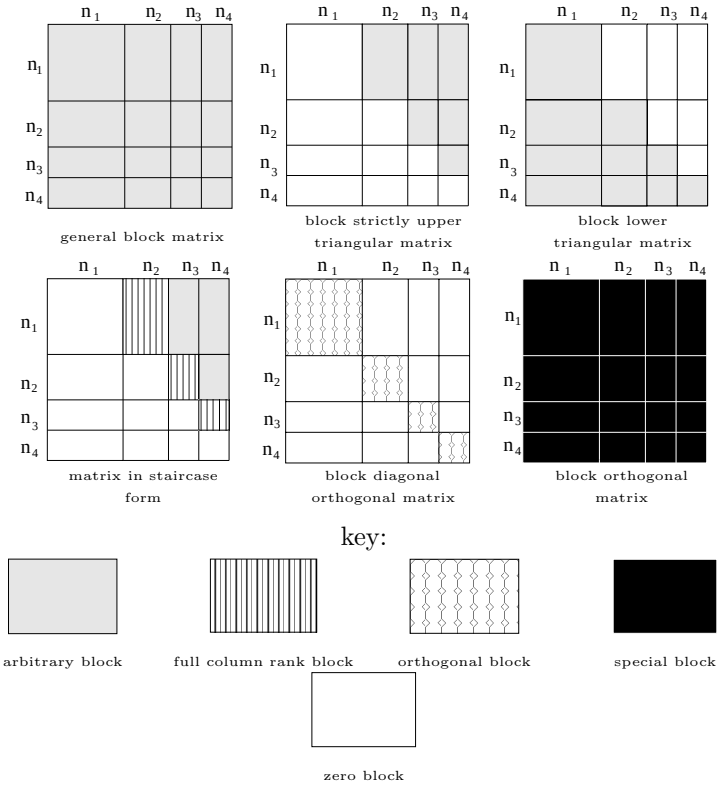


FIG. 3.1. *A schematic of the block matrices defined in the text.*

the staircase form

$$A = \begin{pmatrix} 0\,0\,0 & \times\,\times & \times\,\times & \times \\ 0\,0\,0 & \times\,\times & \times\,\times & \times \\ 0\,0\,0 & \times\,\times & \times\,\times & \times \\ & 0\;\;0 & \times\,\times & \times \\ & 0\;\;0 & \times\,\times & \times \\ & & 0\;\;0 & \times \\ & & 0\;\;0 & \times \\ & & & 0 \end{pmatrix}.$$

Then

$$S = \begin{pmatrix} \times\,\times\,\times & & & \\ \times\,\times\,\times & & & \\ \times\,\times\,\times & & & \\ \times\,\times\,\times & \circ\;\;\circ & & \\ \times\,\times\,\times & \circ\;\;\circ & & \\ \times\,\times\,\times & \times\,\times & \times\,\times & \\ \times\,\times\,\times & \times\,\times & \times\,\times & \\ \times\,\times\,\times & \times\,\times & \times\,\times & \times \end{pmatrix}$$

is a staircase invariant matrix of A if every column of S is a left eigenvector of A. Here, the $\circ$ notation indicates 0 entries in the block lower triangular part of S that are a consequence of the requirement that every column be a left eigenvector. This

may be formulated as a general rule: If we find more than one block of size $n_i \times n_i$, then only those blocks on the lowest block row appear in the sparsity structure of S. For example, the $\circ$'s do not appear because they are above another block of size 2. As a special case, if A is strictly upper triangular, then S is 0 above the bottom row as is shown below. Readers familiar with Arnold's normal form will notice that if A is a given single Jordan block in normal form, then S contains the versal directions.

$$A = \begin{pmatrix} \times & \times & \times & \times & \times & \times \\ & \times & \times & \times & \times & \times \\ & & \times & \times & \times & \times \\ & & & \times & \times & \times \\ & & & & \times & \times \\ & & & & & \times \end{pmatrix}, \quad S = \begin{pmatrix} & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ \times & \times & \times & \times & \times & \times \end{pmatrix}.$$

DEFINITION 2. *Suppose A is a matrix. We call $\mathcal{O}(A) \equiv \{XAX^{-1} : X$ is a nonsingular matrix$\}$ the* orbit *of a matrix A. We call $\mathcal{T} \equiv \{AX - XA : X$ is any matrix$\}$ the* tangent space *of $\mathcal{O}(A)$ at A.*

THEOREM 1. *Let A be an $n \times n$ matrix in staircase form; then the staircase invariant space $\mathcal{S}$ of A and the tangent space $\mathcal{T}$ form an oblique decomposition of $n \times n$ matrix space, i.e., $\mathbb{R}^{n^2} = \mathcal{S} \oplus \mathcal{T}$.*

Proof. Assume that $A_{i,j}$, the (i,j) block of A, is $n_i \times n_j$ for $i, j = 1, \ldots, k$ and, of course, $A_{i,j} = 0$ for all $i \le j$.

There are n_1^2 degrees of freedom in the first block column of S because there are n_1 columns and each column may be chosen from the n_1-dimensional space of left eigenvectors of A. Indeed there are n_i^2 degrees of freedom in the ith block, because each of the n_i columns may be chosen from the n_i-dimensional space of left eigenvectors of the matrix obtained from A by deleting the first $i-1$ block rows and columns. The total number of degrees of freedom is $\sum_{i=1}^{k} n_i^2$, which combined with $\dim(\mathcal{T}) = n^2 - \sum_{i=1}^{k} n_i^2$ [7], gives the dimension of the whole space n^2.

If $S \in \mathcal{S}$ is also in $\mathcal{T}$, then S has the form $AX - XA$ for some matrix X. Our first step will be to show that X must have block upper triangular form after which we will conclude that $AX - XA$ is strictly block upper triangular. Since S is block lower triangular, it will then follow that if it is also in $\mathcal{T}$, it must be 0.

Let i be the first block column of X which does not have block upper triangular structure. Clearly the ith block column of XA is 0 below the diagonal block, so that the ith block column of $S = AX - XA$ contains vectors in the column space of A. However, every column of S is a left eigenvector of A from the definition and therefore is orthogonal to the column space of A. (Notice that we do not require these column vectors of S to be independent; the one Jordan block case is a good example.) Thus the ith block column of S is 0, and from the full column rank conditions on the superdiagonal blocks of A, we conclude that X is 0 below the block diagonal. ∎

DEFINITION 3. *Suppose A is a matrix. We call $\mathcal{O}_b(A) \equiv \{Q^T A Q : Q = e^B, B$ is a block antisymmetric matrix conforming to $A\}$ the* block orthogonal orbit *of a matrix A. We call $\mathcal{T}_b \equiv \{AX - XA : X$ is a block antisymmetric matrix conforming to $A\}$ the* block tangent space *of the block orthogonal orbit $\mathcal{O}_b(A)$ at A. We call $\mathcal{R} \equiv \{$block strictly upper triangular matrix conforming to $A\}$ the* strictly upper block space *of A.*

Note that because of the complementary structure of the two matrices R and S, we can see that $\mathcal{S}$ is always orthogonal to $\mathcal{R}$.

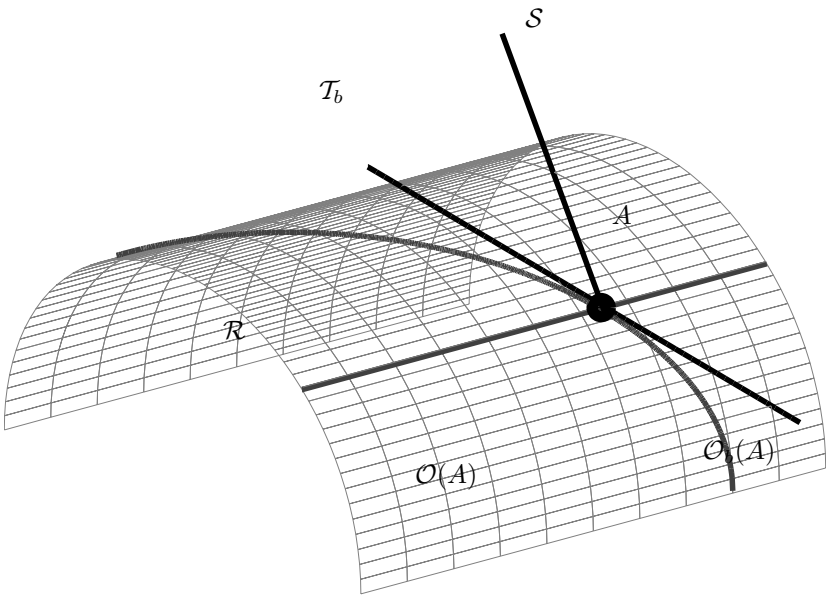


FIG. 3.2. *A diagram of the orbits and related spaces. The similarity orbit at A is indicated by a surface $\mathcal{O}(A)$; the block orthogonal orbit is indicated by a curve $\mathcal{O}_b(A)$ on the surface; the tangent space of $\mathcal{O}_b(A)$, $\mathcal{T}_b$, is indicated by a line, $\mathcal{R}$, which lies on $\mathcal{O}(A)$ is pictured as a line too; and the staircase invariant space $\mathcal{S}$ is represented by a line pointing away from the plane.*

THEOREM 2. *Let A be an $n \times n$ matrix in staircase form; then the tangent space $\mathcal{T}$ of the orbit $\mathcal{O}(A)$ can be split into the block tangent space $\mathcal{T}_b$ of the orbit $\mathcal{O}_b(A)$ and the strictly upper block space $\mathcal{R}$, i.e., $\mathcal{T} = \mathcal{T}_b \oplus \mathcal{R}$.*

Proof. We know that the tangent space $\mathcal{T}$ of the orbit at A has dimension $n^2 - \sum_{i=1}^{k} n_i^2$. If we decompose X into a block upper triangular matrix and a block antisymmetric matrix, we can decompose every $AX - XA$ into a block strictly upper triangular matrix and a matrix in $\mathcal{T}_b$. Since $\mathcal{T} = \mathcal{T}_b + \mathcal{R}$, each of $\mathcal{T}_b$ and $\mathcal{R}$ has dimension $\leq 1/2(n^2 - \sum_{i=1}^{k} n_i^2)$, they must both be exactly of dimension $1/2(n^2 - \sum_{i=1}^{k} n_i^2)$. Thus we know that they actually form a decomposition of $\mathcal{T}$, and the strictly upper block space $\mathcal{R}$ can also be represented as $\mathcal{R} \equiv \{AX - XA : X \text{ is block upper triangular matrix conforming to } A\}$. ∎

COROLLARY 1. $\mathbb{R}^{n^2} = \mathcal{T}_b \oplus \mathcal{R} \oplus \mathcal{S}$. *See Figure* 3.2.

In Definition 3, we really do not need the whole set $\{e^B : B \text{ is block antisymmetric}\} \equiv \{e^B\}$, we merely need a small neighborhood around $B = 0$. Readers may well wish to skip ahead to section 4, but for those interested in mathematical technicalities we review a few simple concepts. Suppose that we have partitioned $n = n_1 + \cdots + n_k$. An orthogonal decomposition of n-dimensional space into k mutually orthogonal subspaces of dimensions $n_1, n_2, \ldots, n_k$ is a point on the *flag manifold*. (When $k = 2$, this is the *Grassmann manifold*.) Equivalently, a point on the flag manifold is specified by a filtration, i.e., a nested sequence of subspaces V_i of dimension $n_1 + \cdots + n_i$ $(i = 1, \ldots, k)$:

$$0 \subset V_1 \subset \cdots \subset V_k = \mathbb{C}^n.$$

The corresponding decomposition can be written as

$$\mathbb{C}^n = V_k = V_1 \oplus V_2 \backslash V_1 \oplus \cdots \oplus V_k \backslash V_{k-1}.$$

This may be expressed concretely. If from a unitary matrix U, we define only V_i for $i = 1, \ldots, k$ as the span of the first $n_1 + n_2 + \cdots + n_i$ columns, then we have $V_1 \subset \cdots \subset V_k$, i.e., a point on the flag manifold. Of course, many unitary matrices U will correspond to the same flag manifold point. In an open neighborhood of $\{e^B\}$, near the point $e^0 = I$, the map between $\{e^B\}$ and an open subset of the flag manifold is a one-to-one homeomorphism. The former set is referred to as a local cross section [25, Lemma 4.1, p. 123] in Lie algebra. No two unitary matrices in a local cross section would have the same sequence of subspaces $V_i, i = 1, \ldots, k$.

3.2. Staircase as a versal deformation. Next, we are going to build up the theory of our versal form. Following Arnold [1], a *deformation* of a matrix A is a matrix $A(\lambda)$ with entries that are power series in the complex variables λ_i, where $\lambda = (\lambda_1, \ldots, \lambda_k)^T \in \mathbb{C}^k$, convergent in a neighborhood of $\lambda = 0$, with $A(0) = A$.

Good introductions to versal deformations may be found in [1, section 2.4] and [17]. The key property of a versal deformation is that it has enough parameters so that no matter how the matrix is perturbed, it may be made equivalent by analytic transformations to the versal deformation with some choice of parameters. The advantage of this concept for a numerical analyst is that we might make a rounding error in any direction and yet still think of this as a perturbation to a standard canonical form.

Let $N \subset M$ be a smooth submanifold of a manifold M. We consider a smooth mapping $A : \Lambda \to M$ of another manifold Λ into M and let λ be a point in Λ such that $A(\lambda) \in N$. The mapping A is called *transversal* to N at λ if the tangent space to M at $A(\lambda)$ is the direct sum

$$TM_{A(\lambda)} = A_* T\Lambda_\lambda \oplus TN_{A(\lambda)}.$$

Here, $TM_{A(\lambda)}$ is the tangent space of M at $A(\lambda)$, $TN_{A(\lambda)}$ is the tangent space of N at $A(\lambda)$, $T\Lambda_\lambda$ is the tangent space of Λ at λ, and A_* is the mapping from $T\Lambda_\lambda$ to $TM_{A(\lambda)}$ induced by A. (It is the Jacobian.)

THEOREM 3. *Suppose A is in staircase form. Fix $S_i \in \mathcal{S}$, $i = 1, \ldots, k$, s.t. span$\{S_i\} = \mathcal{S}$ and $k \geq \dim(\mathcal{S})$. It follows that*

$$A(\lambda) \equiv A + \sum_i \lambda_i S_i \tag{3.1}$$

is a versal deformation of every particular $A(\lambda)$ for λ small enough. $A(\lambda)$ is mini*versal at $\lambda = 0$ if $\{S_i\}$ is a basis of $\mathcal{S}$.*

Proof. Theorem 1 tells us the mapping $A(\lambda)$ is transversal to the orbit at A. From the equivalence of transversality and versality [1], we know that $A(\lambda)$ is a versal deformation of A. Since the dimension of the staircase invariant space $\mathcal{S}$ is the codimension of the orbit, $A(\lambda)$ given by (3.1) is a miniversal deformation if the S_i are a basis for $\mathcal{S}$ (i.e., $k = \dim(\mathcal{S})$). Moreover, $A(\lambda)$ is a versal deformation of every matrix in a neighborhood of A; in other words, the space $\mathcal{S}$ is transversal to the orbit of every $A(\lambda)$. Take a set of matrices X_i s.t. the $X_i A - A X_i$ form a basis of the tangent space $\mathcal{T}$ of the orbit at A. We know $\mathcal{T} \oplus \mathcal{S} = \mathbb{R}^{n^2}$ (here $\oplus$ implies $\mathcal{T} \cap \mathcal{S} = 0$), so there is a fixed minimum angle θ between $\mathcal{T}$ and $\mathcal{S}$. For small enough λ, we can guarantee that the $X_i A(\lambda) - A(\lambda) X_i$ are still linearly independent of each other and

span a subspace of the tangent space at $A(\lambda)$ that is at least, say, $\theta/2$ away from $\mathcal{S}$. This means that the tangent space at $A(\lambda)$ is transversal to $\mathcal{S}$. ∎

Arnold's theory concentrates on general similarity transformations. As we have seen above, the staircase invariant directions are a perfect versal deformation. This idea can be refined to consider similarity transformations that are block orthogonal. Everything is the same as above, except that we add the block strictly upper triangular matrices R to compensate for the restriction to block orthogonal matrices. We now spell this out in detail, as follows.

DEFINITION 4. *If the matrix $C(\lambda)$ is block orthogonal for every λ, then we refer to the deformation as a* block orthogonal deformation.

We say that two deformations $A(\lambda)$ and $B(\lambda)$ are block orthogonally equivalent *if there exists a block orthogonal deformation $C(\lambda)$ of the identity matrix such that $A(\lambda) = C(\lambda)B(\lambda)C(\lambda)^{-1}$.*

We say that a deformation $A(\lambda)$ is block orthogonally versal *if any other deformation $B(\mu)$ is block orthogonally equivalent to the deformation $A(\phi(\mu))$. Here, ϕ is a mapping analytic at 0 with $\phi(0) = 0$.*

THEOREM 4. *A deformation $A(\lambda)$ of A is block orthogonally versal iff the mapping $A(\lambda)$ is transversal to the block orthogonal orbit of A at $\lambda = 0$.*

Proof. The proof follows Arnold [1, sections 2.3 and 2.4] except that we use the block orthogonal version of the relevant notions, and we remember that the tangents to the block orthogonal group are the commutators of A with the block antisymmetric matrices. ∎

Since we know that $\mathcal{T}$ can be decomposed into $\mathcal{T}_b \oplus \mathcal{R}$, we get the following theorem.

THEOREM 5. *Suppose a matrix A is in staircase form. Fix $S_i \in \mathcal{S}, i = 1, \ldots, k$, s.t. span$\{S_i\} = \mathcal{S}$ and $k \geq \dim(\mathcal{S})$. Fix $R_j \in \mathcal{R}, j = 1, \ldots, l$, s.t. span$\{R_j\} = \mathcal{R}$ and $l \geq \dim(\mathcal{R})$. It follows that*

$$A(\lambda) \equiv A + \sum_i \lambda_i S_i + \sum_j \lambda_j R_j$$

is a block orthogonally versal deformation of every particular $A(\lambda)$ for λ small enough. $A(\lambda)$ is block orthogonally mini*versal at A if $\{S_i\}$, $\{R_j\}$ are bases of $\mathcal{S}$ and $\mathcal{R}$.*

It is not hard to see that the theory we set up for matrices with all eigenvalues 0 can be generalized to a matrix A with different eigenvalues. The staircase form is a block upper triangular matrix, each of its diagonal blocks of the form $\lambda_i I + A_i$, with A_i in staircase form defined at the beginning of this chapter, and superdiagonal blocks arbitrary matrices. Its staircase invariant space is spanned by the block diagonal matrices, each diagonal block being in the staircase invariant space of the corresponding diagonal block A_i. $\mathcal{R}$ space is spanned by the block strictly upper triangular matrices s.t. every diagonal block is in the $\mathcal{R}$ space of the corresponding A_i. $\mathcal{T}_b$ is defined exactly the same as in the one eigenvalue case. All our theorems are still valid. When we give the definitions or apply the theorems, we do not really use the values of the eigenvalues. All that is important is how many different eigenvalues A has. In other words, we are working with bundle instead of orbit.

These forms are normal forms that have the same property as the Arnold's normal form: They are continuous under perturbation. The reason that we introduce block orthogonal notation is that the staircase algorithm is a realization to first order of the block orthogonally versal deformation, as we will see in the next section.

4. Application to matrix staircase forms. We are ready to understand the staircase algorithm described in section 1.2. We concentrate on matrices with all eigenvalues 0, since otherwise the staircase algorithm will separate other structures and continue recursively.

We use the notation stair(A) to denote the output A of the staircase algorithm as described in section 1.2. Now suppose that we have a matrix A which is in staircase form. To zeroth order, any instance of the staircase algorithm replaces A with $\hat{A} = Q_0^T A Q_0$, where Q_0 is block diagonal orthogonal. Of course this does not change the staircase structure of A; the Q_0 represents the arbitrary rotations within the subspaces and can depend on how the software is written and the subtlety of roundoff errors when many singular values are 0. Next, suppose that we perturb A by ϵE. According to Corollary 1, we can decompose the perturbation matrix uniquely as $E = S + R + T_b$ with $S \in \mathcal{S}$, $R \in \mathcal{R}$, and $T_b \in \mathcal{T}_b$. Theorem 6 states that, in addition to some block diagonal matrix Q_0, the staircase algorithm will apply a block orthogonal similarity transformation $Q_1 = I + \epsilon X + o(\epsilon)$ to $A + \epsilon E$ to kill the perturbation in T_b.

THEOREM 6. *Suppose that A is a matrix in staircase form and E is any perturbation matrix. The staircase algorithm (without zeroing) on $A + \epsilon E$ will produce an orthogonal matrix Q (depending on ϵ) and the output matrix* $\text{stair}(A + \epsilon E) = Q^T(A + \epsilon E)Q = \hat{A} + \epsilon(\hat{S} + \hat{R}) + o(\epsilon)$*, where $\hat{A}$ has the same staircase structure as A, $\hat{S}$ is a staircase invariant matrix of $\hat{A}$, and $\hat{R}$ is a block strictly upper triangular matrix. If singular values are zeroed out, then the algorithm further kills $\hat{S}$ and outputs $\hat{A} + \epsilon\hat{R} + o(\epsilon)$.*

Proof. After the first stage of the staircase algorithm, the first block column is orthogonal to the other columns, and this property is preserved through the completion of the algorithm. Generally, after the ith iteration, the ith block column below (including) the diagonal block is orthogonal to all other columns to its right, and this property is preserved all through. So when the algorithm terminates, we will have a matrix whose columns below (including) the diagonal block are orthogonal to all the columns to the right; in other words, it is a matrix in staircase form plus a staircase invariant matrix.

We can always write the similarity transformation matrix as $Q = Q_0(I + \epsilon X + \text{o}(\epsilon))$, where Q_0 is a block diagonal orthogonal matrix and X is a block antisymmetric matrix that does not depend on ϵ because of the local cross section property that we mentioned at the beginning of section 3. Notice that Q_0 is not a constant matrix decided by A; it depends on ϵE to its first order. We should have written $(Q_0)_0 + \epsilon(Q_0)_1 + \text{o}(\epsilon)$ instead of Q_0 . However, we do not expand Q_0 since as long as it is a block diagonal orthogonal transformation, it does not change the staircase structure of the matrix. Hence, we get

$$\begin{aligned}
\text{stair}(A + \epsilon E) &= \text{stair}(A + \epsilon S + \epsilon R + \epsilon T_b) \\
&= (I + \epsilon X^T + o(\epsilon))Q_0^T(A + \epsilon S + \epsilon R + \epsilon T_b)Q_0(I + \epsilon X + o(\epsilon)) \\
&= (I + \epsilon X^T + o(\epsilon))(\hat{A} + \epsilon\hat{S} + \epsilon\hat{R} + \epsilon\hat{T}_b)(I + \epsilon X + o(\epsilon)) \\
&= \hat{A} + \epsilon(\hat{S} + \hat{R} + \hat{T}_b + \hat{A}X - X\hat{A}) + o(\epsilon) \\
&= \hat{A} + \epsilon(\hat{S} + \hat{R}) + o(\epsilon).
\end{aligned} \tag{4.1}$$

Here, $\hat{A}, \hat{S}, \hat{R}$, and $\hat{T}_b$ are, respectively, $Q_0^T A Q_0, Q_0^T S Q_0, Q_0^T R Q_0$, and $Q_0^T T_b Q_0$. It is easy to check that $\hat{S}, \hat{R}, \hat{T}_b$ is still in the $\mathcal{S}, \mathcal{R}, \mathcal{T}_b$ space of $\hat{A}$. X is a block antisymmetric matrix satisfying $\hat{T}_b = X\hat{A} - \hat{A}X$. We know that X is uniquely determined because the dimensions of $\hat{T}_b$ and the block antisymmetric matrix space are the same.

The reason that $\hat{T}_b = X\hat{A} - \hat{A}X$, and hence the last equality in (4.1), holds is because the algorithm forces the output form as described in the first paragraph of this proof: $\hat{A} + \epsilon\hat{R}$ is in staircase form and $\epsilon\hat{S}$ is a staircase invariant matrix. Since $(\mathcal{S} \oplus \mathcal{R}) \cap \mathcal{T}_b$ is the zero matrix, the T_b term must vanish. ∎

To understand more clearly what this observation tells us, let us check some simple situations. If the matrix A is only perturbed in the direction $\mathcal{S}$ or $\mathcal{R}$, then the similarity transformation will simply be a block diagonal orthogonal matrix Q_0. If we ignore this transformation, which does not change any structure, we can think of the output to be unchanged from the input. This is the reason we call $\mathcal{S}$ the staircase invariant space. The reason we did not include R in the staircase invariant space is that $A + \epsilon R$ is still within $O_b(A)$. If the matrix A is only perturbed along the block tangent direction $\mathcal{T}_b$, then the staircase algorithm will kill the perturbation and do a block diagonal orthogonal similarity transformation.

Although the staircase algorithm decides this Q_0 step by step all through the algorithm (due to SVD rank decisions), we can actually think of the Q_0 as decided at the first step. We can even ignore this Q_0 because the only reason it comes up is that the SVD we use follows a specific method of sorting singular values when they are different and choosing the basis of the singular vector space when the same singular values appear.

We know that every matrix A can be reduced to a staircase form under an orthogonal transformation. In other words, we can always think of any general matrix M as P^TAP, where A is in staircase form. Thus, in general, the staircase algorithm always introduces an orthogonal transformation and returns a matrix in staircase form and a first order perturbation in its staircase invariant direction, i.e., $\text{stair}(M + \epsilon E) = \text{stair}(P^TAP + \epsilon E) = \text{stair}(A + \epsilon PEP^T)$.

It is now obvious that if a staircase form matrix A has its $\mathcal{S}$ and $\mathcal{T}$ almost normal to each other, then the staircase algorithm will behave very well. On the other hand, if $\mathcal{S}$ is very close to $\mathcal{T}$, then it will fail. To emphasize this, we write it as a conclusion.

CONCLUSION 1. *The angle between the staircase invariant space $\mathcal{S}$ and the tangent space $\mathcal{T}$ decides the behavior of the staircase algorithm. The smaller the angle, the worse the algorithm behaves.*

In the one Jordan block case, we have an if-and-only-if condition for $\mathcal{S}$ to be near $\mathcal{T}$.

THEOREM 7. *Let A be an $n \times n$ matrix in staircase form and suppose that all of its block sizes are 1×1; then $\mathcal{S}(A)$ is close to $\mathcal{T}(A)$ iff the following two conditions hold:*

1. *(Row condition) there exists a nonzero row in A s.t. every entry on this row is $o(1)$.*
2. *(Chain condition) there exists a chain of length $n - k$ with the chain value $O(1)$, where k is the lowest row satisfying* (1).

Here we call $A_{i_1,i_2}, A_{i_2,i_3}, \ldots, A_{i_t,i_{t+1}}$ a chain of length t and the product $A_{i_1,i_2}A_{i_2,i_3}\cdots A_{i_t,i_{t+1}}$ the chain value.

Proof sketch. Notice that $\mathcal{S}$ being close to $\mathcal{T}$ is equivalent to $\mathcal{S}$ being almost perpendicular to $\mathcal{N}$, the normal space of A. In this case, $\mathcal{N}$ is spanned by $\{I, A^T, A^{T2}, \ldots, A^{T(n-1)}\}$ and $\mathcal{S}$ consists of matrices with nonzero entries only in the last row. Considering the angle between any two matrices from the two spaces, it is straightforward to show that for $\mathcal{S}$ to be almost perpendicular to $\mathcal{N}$ is equivalent to the following:

1. There exists a k s.t. the (n, k) entry of each of the matrices $I, A^T, \ldots, A^{T(n-1)}$ is $o(1)$ or 0.
2. If the entry is $o(1)$, then it must have some other O(1) entry in the same matrix. Assume k is the largest choice if there are different k's. By a combi-

natorial argument, we can show that these two conditions are equivalent to the row and chain conditions, respectively, in our theorem. ∎

Remark 1. Note that saying that there exists an O(1) entry in a matrix is equivalent to saying that there exists a singular value of the matrix of O(1). So, the chain condition is the same as saying that the singular values of A^{n-k} are not all $O(\epsilon)$ or smaller.

Generally, we do not have an if-and-only-if condition for $\mathcal{S}$ to be close to $\mathcal{T}$. We only have a necessary condition, that is, only if at least one of the superdiagonal blocks of the original unperturbed matrix has a singular value almost zero, i.e., it has a weak stair, will $\mathcal{S}$ be close to $\mathcal{T}$. Actually, it is not hard to show that the angle between $\mathcal{T}_b$ and $\mathcal{R}$ is at most in the same order as the smallest singular value of the weak stair. So, when the perturbation matrix E is decomposed into $R + S + T_b$, R and T_b are typically very large, but whether S is large depends on whether $\mathcal{S}$ is close to $\mathcal{T}$.

Notice that (4.1) is valid for sufficiently small ϵ. What range of ϵ is considered sufficiently small? Clearly, ϵ has to be smaller than the smallest singular value δ of the weak stairs. Moreover, the algorithm requires the perturbations along both $\mathcal{T}$ and $\mathcal{S}$ to be smaller than δ. Assuming the angle between $\mathcal{T}$ and $\mathcal{S}$ is θ, then generally, when θ is large, we would expect an ϵ smaller than δ to be sufficiently small. However, when θ is close to zero, for a random perturbation, we would expect an ϵ in the order of δ/θ to be sufficiently small. Here, again, we can see that the angle between $\mathcal{S}$ and $\mathcal{T}$ decides the range of effective ϵ. For small θ, when ϵ is not sufficiently small, we observed some discontinuity in the zeroth order term in (4.1) caused by the ordering of singular values during certain stages of the algorithm. Thus, instead of the identity matrix, we get a permutation matrix in the zeroth order term.

The theory explains why the staircase algorithm behaves so differently on the two matrices A_1 and A_2 in section 2. Using Theorem 7, we can see that A_1 is a staircase failure ($k = 2$), while A_2 is not ($k = 1$). By a direct calculation, we find that the tangent space and the staircase invariant space of A_1 is very close ($\sin(\langle \mathcal{S}, \mathcal{T} \rangle) = \delta/\sqrt{1+\delta^2}$), while this is not the situation for A_2 ($\sin(\langle \mathcal{S}, \mathcal{T} \rangle) = 1/\sqrt{3}$). When transforming to get $\tilde{A}_1$ and $\tilde{A}_2$ with Q, which is an approximate orthogonal matrix up to the order of square root of machine precision ϵ_m, another error in the order of $\sqrt{\epsilon_m}$ (10^{-7}) is introduced, it is comparable with δ in our experiment, so the staircase algorithm actually runs on a shifted version $A_1 + \delta E_1$ and $A_2 + \delta E_1$. That is why we see R as large as an $O(10^{-6})$ added to J_3 in the second table for $\tilde{A}_2$ (see Table 2.2). We might as well call A_2 a staircase failure in this situation, but A_1 suffers a much worse failure under the same situation, in that the staircase algorithm fails to detect a J_3 structure at all. This is because the tangent space and the staircase invariant space are so close that the S and T component are very large, hence (4.1) no longer applies.

5. A staircase algorithm failure to motivate the theory for pencils. The pencil analog to the staircase failure in section 2 is

$$(A_1, B_1) = \left(\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} \delta & 0 & 0 & 0 \\ 0 & \delta & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \right),$$

where $\delta =$ `1.5e-8`. This is a pencil with the structure $L_1 \oplus J_2(0)$. After we add a random perturbation of size `1e-14` to this pencil, `GUPTRI` fails to return back the original pencil no matter which `EPSU` we choose. Instead, it returns back a more

generic $L_2 \oplus J_1(0)$ pencil $O(\epsilon)$ away.

On the other hand, for another pencil with the same $L_1 \oplus J_2(0)$ structure,

$$(A_2, B_2) = \left(\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \delta & 0 \end{bmatrix} \right),$$

`GUPTRI` returns an $L_1 \oplus J_2(0)$ pencil $O(\epsilon)$ away.

At this point, readers may correctly expect that the reason behind this is again the angle between two certain spaces, as in the matrix case.

6. Matrix pencils. Parallel to the matrix case, we can set up a similar theory for the pencil case. For simplicity, we concentrate on the case when a pencil only has L-blocks and $J(0)$-blocks. Pencils containing L^T-blocks and nonzero (including ∞) eigenvalue blocks can always be reduced to the previous case by transposing and exchanging the two matrices of the pencil and/or shifting.

6.1. The staircase invariant space and related subspaces for pencils. A pencil (A, B) is in *staircase form* if we can divide both A and B into block rows of sizes $r_1, \ldots, r_k$ and block columns of sizes $s_1, \ldots, s_{k+1}$, s.t. A is strictly block upper triangular with every superdiagonal block having full column rank and B is block upper triangular with every diagonal block having full row rank and the rows orthogonal to each other. Here we allow s_{k+1} to be zero. A pencil is called *conforming to* (A, B) if it has the same block structure as (A, B). A square matrix is called *row (column) conforming to* (A, B) if it has diagonal block sizes the same as the row (column) sizes of (A, B).

DEFINITION 5. *Suppose* (A, B) *is a pencil in staircase form and* B_d *is the block diagonal part of* B. *We call* (S_A, S_B) *a* staircase invariant pencil of (A, B) *if* $S_A^T A = 0$, $S_B B_d^T = 0$, *and* (S_A, S_B) *has complementary structure to* (A, B). *We call the space consisting of all such* (S_A, S_B) *the* staircase invariant space of (A, B) *and denote it by* $\mathcal{S}$.

For example, let (A, B) have the staircase form

$$(A, B) = \left(\begin{bmatrix} 0\,0\,0 & \times\,\times & \times\,\times & \times \\ 0\,0\,0 & \times\,\times & \times\,\times & \times \\ & 0\;0 & \times\,\times & \times \\ & 0\;0 & \times\,\times & \times \\ & & 0\;0 & \times \\ & & 0\;0 & \times \\ & & & 0 \end{bmatrix}, \begin{bmatrix} \times\,\times\,\times & \times\,\times & \times\,\times & \times \\ \times\,\times\,\times & \times\,\times & \times\,\times & \times \\ & \times\;\times & \times\,\times & \times \\ & \times\;\times & \times\,\times & \times \\ & & \times\;\times & \times \\ & & \times\;\times & \times \\ & & & \times \end{bmatrix} \right);$$

then

$$(S_A, S_B) = \left(\begin{bmatrix} \circ\;\circ\;\circ & & & \\ \circ\;\circ\;\circ & & & \\ \circ\;\circ\;\circ & \circ\;\circ & & \\ \circ\;\circ\;\circ & \circ\;\circ & & \\ \times\,\times\,\times & \times\,\times & \times\,\times & \\ \times\,\times\,\times & \times\,\times & \times\,\times & \\ \times\,\times\,\times & \times\,\times & \times\,\times & \times \end{bmatrix}, \begin{bmatrix} & & & \\ \times\,\times\,\times & & & \\ \times\,\times\,\times & & & \\ \times\,\times\,\times & \circ\;\circ & & \\ \times\,\times\,\times & \circ\;\circ & & \\ \times\,\times\,\times & \circ\;\circ & \circ\;\circ & \end{bmatrix} \right)$$

is a staircase invariant pencil of (A, B) if every column of S_A is in the left null space of A and every row of S_B is in the right null space of B. Notice that the sparsity structure of S_A and S_B is at most complementary to that of A and B, respectively, but S_A and S_B are often less sparse because of the requirement on the nullspace. To

be precise, if we find more than one diagonal block with the same size, then among the blocks of this size, only the blocks on the lowest block row appear in the sparsity structure of S_A. If any of the diagonal blocks of B is a square block, then S_B has all zero entries throughout the corresponding block column.

As special cases, if A is a strictly upper triangular square matrix and B is an upper triangular square matrix with diagonal entries nonzero, then S_A has only nonzero entries in the bottom row and S_B is simply a zero matrix. If A is a strictly upper triangular $n \times (n+1)$ matrix and B is an upper triangular $n \times (n+1)$ matrix with diagonal entries nonzero, then (S_A, S_B) is the zero pencil.

DEFINITION 6. *Suppose* (A, B) *is a pencil. We call* $\mathcal{O}(A, B) \equiv \{X(A, B)Y : X, Y$ *are nonsingular square matrices* $\}$ *the* orbit *of a pencil* (A, B). *We call* $\mathcal{T} \equiv \{X(A, B) - (A, B)Y : X, Y$ *are any square matrices* $\}$ *the* tangent space *of* $\mathcal{O}(A, B)$ *at* (A, B).

THEOREM 8. *Let* (A, B) *be an* $m \times n$ *pencil in staircase form; then the staircase invariant space* $\mathcal{S}$ *of* (A, B) *and the tangent space* $\mathcal{T}$ *form an oblique decomposition of* $m \times n$ *pencil space, i.e.,* $\mathbb{R}^{2mn} = \mathcal{S} + \mathcal{T}$.

Proof. The proof of the theorem is similar to that of Theorem 1; first we prove the dimension of $\mathcal{S}(A, B)$ is the same as the codimension of $\mathcal{T}(A, B)$, then we prove $\mathcal{S} \cap \mathcal{T} = \{0\}$ by induction. The reader may fill in the details. ∎

DEFINITION 7. *Suppose* (A, B) *is a pencil. We call* $\mathcal{O}_b(A, B) \equiv \{P(A, B)Q : P = e^X, X$ *is a block antisymmetric matrix row conforming to* $(A, B), Q = e^Y, Y$ *is a block antisymmetric matrix column conforming to* $(A, B)\}$ *the* block orthogonal orbit *of a pencil* (A, B). *We call* $\mathcal{T}_b \equiv \{X(A, B) - (A, B)Y : X$ *is a block antisymmetric matrix row conforming to* (A, B), Y *is a block antisymmetric matrix column conforming to* $(A, B)\}$ *the* block tangent space *of the block orthogonal orbit* $\mathcal{O}_b(A, B)$ *at* (A, B). *We call* $\mathcal{R} \equiv \{U(A, B) - (A, B)V : U$ *is a block upper triangular matrix row conforming to* (A, B), V *is a block upper triangular matrix column conforming to* $(A, B)\}$ *the* block upper pencil space *of* (A, B).

THEOREM 9. *Let* (A, B) *be an* $m \times n$ *pencil in staircase form; then the tangent space* $\mathcal{T}$ *of the orbit* $\mathcal{O}(A, B)$ *can be split into the block tangent space* $\mathcal{T}_b$ *of the orbit* $\mathcal{O}_b(A, B)$ *and the block upper pencil space* $\mathcal{R}$, *i.e.,* $\mathcal{T} = \mathcal{T}_b \oplus \mathcal{R}$.

Proof. This can be proved by a very similar argument concerning the dimensions as for matrix, in which the dimension of $\mathcal{R}$ is $2\sum_{i<j} r_i s_j + \sum r_i s_i$, the dimension of $\mathcal{T}_b$ is $\sum_{i<j} r_i r_j + \sum_{i<j} s_i s_j$, the codimension of the orbit $\mathcal{O}(A, B)$ (or $\mathcal{T}$) is $\sum s_i r_i - \sum_{j>i} s_i s_j + 2\sum_{j>i} s_i r_j - \sum_{j>i} r_i r_j$ [7]. ∎

COROLLARY 2. $\mathbb{R}^{2mn} = \mathcal{T}_b \oplus \mathcal{R} \oplus \mathcal{S}$.

6.2. Staircase as a versal deformation for pencils. The theory of versal forms for pencils [17] is similar to that for matrices. A *deformation* of a pencil (A, B) is a pencil $(A, B)(\lambda)$ with entries power series in the real variables λ_i. We say that two deformations $(A, B)(\lambda)$ and $(C, D)(\lambda)$ are *equivalent* if there exist two deformations $P(\lambda)$ and $Q(\lambda)$ of identity matrices such that $(A, B)(\lambda) = P(\lambda)(C, D)(\lambda)Q(\lambda)$.

THEOREM 10. *Suppose* (A, B) *is in staircase form. Fix* $S_i \in \mathcal{S}$, $i = 1, \ldots, k$, *s.t. span*$\{S_i\} = \mathcal{S}$ *and* $k \geq \dim(\mathcal{S})$. *It follows that*

$$(A, B)(\lambda) \equiv (A, B) + \sum_i \lambda_i S_i \tag{6.1}$$

is a versal deformation of every particular $(A, B)(\lambda)$ *for* λ *small enough.* $(A, B)(\lambda)$ *is* mini*versal at* $\lambda = 0$ *if* $\{S_i\}$ *is a basis of* $\mathcal{S}$.

DEFINITION 8. *We say two deformations* $(A,B)(\lambda)$ *and* $(C,D)(\lambda)$ *are* block orthogonally equivalent *if there exist two block orthogonal deformations* $P(\lambda)$ *and* $Q(\lambda)$ *of the identity matrix such that* $(A,B)(\lambda) = P(\lambda)(C,D)(\lambda)Q(\lambda)$. *Here,* $P(\lambda)$ *and* $Q(\lambda)$ *are exponentials of matrices which are conforming to* (A,B) *in row and column, respectively.*

We say that a deformation $(A,B)(\lambda)$ *is* block orthogonally versal *if any other deformation* $(C,D)(\mu)$ *is block orthogonally equivalent to the deformation* $(A,B)(\phi(\mu))$. *Here,* ϕ *is a mapping holomorphic at 0 with* $\phi(0) = 0$.

THEOREM 11. *A deformation* $(A,B)(\lambda)$ *of* (A,B) *is block orthogonally versal iff the mapping* $(A,B)(\lambda)$ *is transversal to the block orthogonal orbit of* (A,B) *at* $\lambda = 0$.

This is the corresponding result to Theorem 4.

Since we know that $\mathcal{T}$ can be decomposed into $\mathcal{T}_b \oplus \mathcal{R}$, we get the following theorem.

THEOREM 12. *Suppose a pencil* (A,B) *is in staircase form. Fix* $S_i \in \mathcal{S}, i = 1,\dots,k$, *s.t. span*$\{S_i\} = \mathcal{S}$ *and* $k \geq \dim(\mathcal{S})$. *Fix* $R_j \in \mathcal{R}, j = 1,\dots,l$, *s.t. span*$\{R_j\} = \mathcal{R}$ *and* $l \geq \dim(\mathcal{R})$. *It follows that*

$$(A,B)(\lambda) \equiv (A,B) + \sum_i \lambda_i S_i + \sum_j \lambda_j R_j$$

is a block orthogonally versal deformation of every particular $(A,B)(\lambda)$ *for* λ *small enough.* $(A,B)(\lambda)$ *is block orthogonally* miniversal *at* (A,B) *if* $\{S_i\}$, $\{R_j\}$ *are bases of* $\mathcal{S}$ *and* $\mathcal{R}$.

Notice that as in the matrix case, we can also extend our definitions and theorems to the general form containing L^T-blocks and nonzero eigenvalue blocks, and again, we will not specify what eigenvalues they are and hence get into the bundle case. We want to point out only one particular example here. If (A,B) is in the staircase form of $L_n + J_1(\cdot)$, then A will be a strictly upper triangular matrix with nonzero entries on the super diagonal and B will be a triangular matrix with nonzero entries on the diagonal except the $(n+1,n+1)$ entry. S_A will be the zero matrix and S_B will be a matrix with the only nonzero entry on its $(n+1,n+1)$ entry.

7. Application to pencil staircase forms. We concentrate on $L \oplus J(0)$ structures only, since otherwise the staircase algorithm will separate all other structures and continue similarly after a shift and/or transpose on that part only. As in the matrix case, the staircase algorithm basically decomposes the perturbation pencil into three spaces $\mathcal{T}_b$, $\mathcal{R}$, and $\mathcal{S}$ and kills the perturbation in $\mathcal{T}_b$.

THEOREM 13. *Suppose that* (A,B) *is a pencil in staircase form and* E *is any perturbation pencil. The staircase algorithm (without zeroing) on* $(A,B) + \epsilon E$ *will produce two orthogonal matrices* P *and* Q *(depending on* ϵ*) and the output pencil* $\text{stair}((A,B) + \epsilon E) = P^T((A,B) + \epsilon E)Q = (\hat{A},\hat{B}) + \epsilon(\hat{S} + \hat{R}) + o(\epsilon)$, *where* $(\hat{A},\hat{B})$ *has the sane staircase structure as* (A,B), $\hat{S}$ *is a staircase invariant pencil of* $(\hat{A},\hat{B})$, *and* $\hat{R}$ *is in the block upper pencil space* $\mathcal{R}$. *If singular values are zeroed out, then the algorithm further kills* $\hat{S}$ *and output* $(\hat{A},\hat{B}) + \epsilon\hat{R} + o(\epsilon)$.

We use a formula to explain the statement more clearly:

$$\begin{aligned}
&(I + \epsilon X + o(\epsilon))P_1((A,B) + \epsilon S + \epsilon R + \epsilon T_b)Q_1(I - \epsilon Y + o(\epsilon)) \\
&\quad = (I + \epsilon X + o(\epsilon))((\hat{A},\hat{B}) + \epsilon\hat{S} + \epsilon\hat{R} + \epsilon\hat{T_b})(I - \epsilon Y + o(\epsilon)) \\
&\quad = (\hat{A},\hat{B}) + \epsilon(\hat{S} + \hat{R} + \hat{T_b} + X(\hat{A},\hat{B}) - (\hat{A},\hat{B})Y) + o(\epsilon) \\
&\quad = (\hat{A},\hat{B}) + \epsilon(\hat{S} + \hat{R}) + o(\epsilon).
\end{aligned} \tag{7.1}$$

Similarly, we can see that when a pencil has its $\mathcal{T}$ and $\mathcal{S}$ almost normal to each other, the staircase algorithm will behave well. On the other hand, if $\mathcal{S}$ is very close to $\mathcal{T}$, then it will behave badly. This is exactly the situation in the two pencil examples in section 5. Although the two pencils are both ill-conditioned, a direct calculation shows that the first pencil has its staircase invariant space very close to the tangent space (the angle $\langle \mathcal{S}, \mathcal{T} \rangle = \delta/\sqrt{\delta^2 + 2}$) while the second one does not (the angle $\langle \mathcal{S}, \mathcal{T} \rangle = 1/\sqrt{2 + \delta^2}$).

The if-and-only-if condition for $\mathcal{S}$ to be close to $\mathcal{T}$ is more difficult than in the matrix case. One necessary condition is that one super diagonal block of A almost is not full column rank or one diagonal block of B almost is not full row rank. This is usually referred to as *weak coupling*.

8. Examples: The geometry of the Boley pencil and others. Boley [3, Example 2, p. 639] presents an example of a 7×8 pencil (A, B) that is controllable (has generic Kronecker structure) yet it is known that an uncontrollable system (nongeneric Kronecker structure) is nearby at a distance `6e-4`. What makes the example interesting is that the staircase algorithm fails to find this nearby uncontrollable system while other methods succeed (for example, [24]). Our theory provides a geometrical understanding of why this famous example leads to staircase failure: The staircase invariant space is very close to the tangent space.

The pencil that we refer to is $(A, B(\epsilon))$, where

$$A = \begin{pmatrix} \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 \end{pmatrix} \quad \text{and} \quad B(\epsilon) = \begin{pmatrix} 1 & -1 & -1 & -1 & -1 & -1 & -1 & 7 \\ \cdot & 1 & -1 & -1 & -1 & -1 & -1 & 6 \\ \cdot & \cdot & 1 & -1 & -1 & -1 & -1 & 5 \\ \cdot & \cdot & \cdot & 1 & -1 & -1 & -1 & 4 \\ \cdot & \cdot & \cdot & \cdot & 1 & -1 & -1 & 3 \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 & -1 & 2 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \epsilon & 1 \end{pmatrix}.$$

(The dots refer to zeros, and in the original Boley example $\epsilon = 1$.)

When $\epsilon = 1$, the staircase algorithm predicts a distance of 1, and is therefore off by nearly four orders of magnitude. To understand the failure, our theory works best for smaller values of ϵ, but it is still clear that even for $\epsilon = 1$, there will continue to be difficulties.

It is useful to express the pencil $(A, B(\epsilon))$ as $P_0 + \epsilon E$, where $P_0 = (A, B(0))$ and S is zero except for a 1 in the (7,7) entry of its B part. P_0 is in the bundle of pencils whose Kronecker form is $L_6 + J_1(\cdot)$ and the perturbation E is exactly in the unique staircase invariant direction (hence the notation "S") as we pointed out at the end of section 6.

The relevant quantity is then the angle between the staircase invariant space and the pencil space. An easy calculation reveals that the angle is very small: $\theta_S = 0.0028$ radians. In order to get a feeling for what range of ϵ first order theory applies, we calculated the exact distance $d(\epsilon) \equiv d(P(\epsilon), \text{bundle})$ using the nonlinear eigenvalue template software [34]. To first order, $d(\epsilon) = \theta_S \cdot \epsilon$. Figure 8.1 plots the distances first for $\epsilon \in [0, 2]$ and then a close-up for $\epsilon = [0, 0.02]$.

Our observation based on this data suggests that first order theory is good to two decimal places for $\epsilon \leq 10^{-4}$ and one place for $\epsilon \leq 10^{-2}$. To understand the geometry of staircase algorithmic failure, one decimal place or even merely an order of magnitude is quite sufficient.

In summary, we see clearly that the staircase invariant direction is at a small

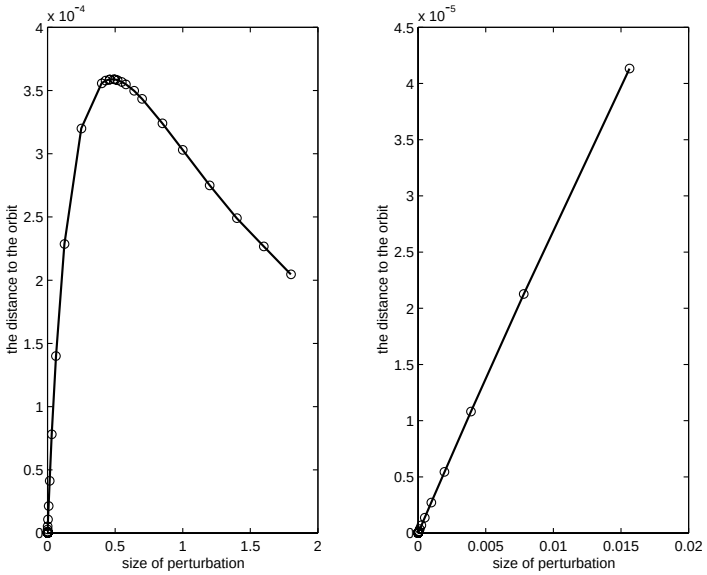


FIG. 8.1. *The picture to explain the change of the distance of the pencils $P_0 + \epsilon E$ to the bundle of $L_6 + J(\cdot)$ as ϵ changes. The second subplot is part of the first one at the points near $\epsilon = 0$.*

angle to the tangent space, and therefore the staircase algorithm will have difficulty finding the nearest pencil on the bundle or predicting the distance. This difficulty is quantified by the angle θ_S.

Since the Boley example is for $\epsilon = 1$, we computed the distance well past $\epsilon = 1$. The breakdown of first order theory is attributed to the curving of the bundle towards S. A three-dimensional schematic is portrayed in Figure 8.2.

The relevant picture for control theory is a planar intersection of the above picture. In control theory, we set the special requirement that the A matrix has the form $[0\ I]$. Pencils on the intersection of this hyperplane and the bundle are termed *uncontrollable*.

We analytically calculated the angle θ_c between S and the tangent space for the "uncontrollable surfaces." We found that $\theta_c = 0.0040$. Using the nonlinear eigenvalue template software [34], we numerically computed the true distance from $P_0 + \epsilon E$ to the uncontrollable surfaces and calculated the ratio of this distance to ϵ. We found that for $\epsilon < 8e-4$, the ratio agrees with $\theta_c = 0.0040$ very well.

We did a similar analysis on the three pencils C_1, C_2, C_3 given by Demmel and Kågström [12]. We found that the *sin* values of the angles between $\mathcal{S}$ and $\mathcal{T}$ are, respectively, `2.4325e-02, 3.4198e-02`, and `8.8139e-03` and the *sin* values between $\mathcal{T}_b$ and $\mathcal{R}$ are, respectively, `1.7957e-02`, `7.3751e-03`, and `3.3320e-06`. This explains why we saw the staircase algorithm behave progressively worse on them. Especially, it explains why, when a perturbation about 10^{-3} is added to these pencils, C_3 behaves dramatically worse than C_1 and C_2. The component in $\mathcal{S}$ is almost of the same order as the entries of the original pencil.

So we conclude that the reason the staircase algorithm does not work well on this example is because $P_0 = (A, B(0))$ is actually a staircase failure, in that its tangent space, is very close to its staircase invariant space, and also the perturbation is so large that even if we know the angle in advance we cannot estimate the distance well.

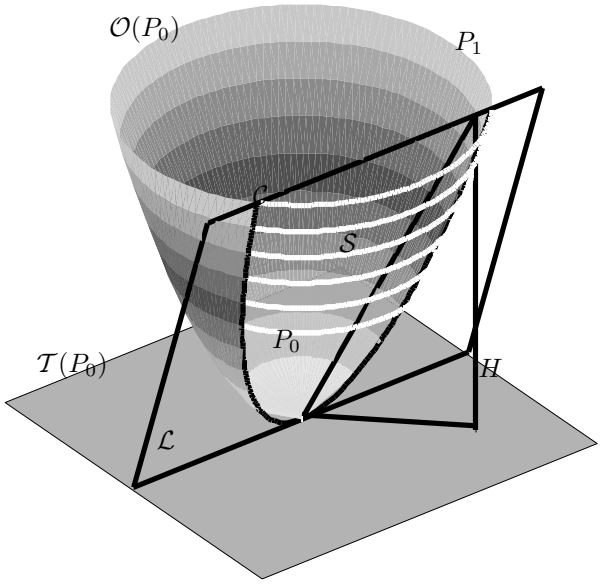


FIG. 8.2. *The staircase algorithm on the Boley example. The surface represents the orbit* $\mathcal{O}(P_0)$*. Its tangent space at the pencil* P_0*,* $\mathcal{T}(P_0)$*, is represented by the plane on the bottom.* P_1 *lies on the staircase invariant space* $\mathcal{S}$ *inside the "bowl." The hyperplane of uncontrollable pencils is represented by the plane cutting through the surface along the curve* $\mathcal{C}$*. It intersects* $\mathcal{T}(P_0)$ *along* $\mathcal{L}$*. The angle between* $\mathcal{L}$ *and* $\mathcal{S}$ *is* θ_c*. The angle between* $\mathcal{S}$ *and* $\mathcal{T}(P_0)$*,* θ_S*, is represented by the angle* $\angle HP_0P_1$*.*

Acknowledgments. The authors thank Bo Kågström and Erik Elmroth for their helpful discussion and their `conlab` software for easy interactive numerical testing. The staircase invariant directions were originally discovered for single Jordan blocks with Erik Elmroth while he was visiting MIT during the fall of 1996.

REFERENCES

[1] V. ARNOLD, *On matrices depending on parameters*, Russian Math. Surveys, 26 (1971), pp. 29–43.

[2] T. BEELEN AND P. V. DOOREN, *An improved algorithm for the computation of Kronecker's canonical form of a singular pencil*, Linear Algebra Appl., 105 (1988), pp. 9–65.

[3] D. BOLEY, *Estimating the sensitivity of the algebraic structure of pencils with simple eigenvalue estimates*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 632–643.

[4] D. BOLEY, *The algebraic structure of pencils and block Toeplitz matrices*, Linear Algebra Appl., 279 (1998), pp. 255–279.

[5] D. BOLEY AND P. V. DOOREN, *Placing zeroes and the Kronecker canonical form*, Circuits Systems Signal Process., 13 (1994), pp. 783–802.

[6] F. CHAITIN-CHATELIN AND V. FRAYSSÉ, *Lectures on Finite Precision Computations*, SIAM, Philadelphia, 1996.

[7] J. DEMMEL AND A. EDELMAN, *The dimension of matrices (matrix pencils) with given Jordan (Kronecker) canonical forms*, Linear Algebra Appl., 230 (1995), pp. 61–87.

[8] J. DEMMEL AND B. KÅGSTRÖM, *Stably computing the Kronecker structure and reducing subspace of singular pencils* $A - \lambda B$ *for uncertain data*, in Large Scale Eigenvalue Problems, J. Cullum and R. Willoughby, eds., North-Holland Math. Stud. 127, 1986, pp. 283–323.

[9] J. DEMMEL AND B. KÅGSTRÖM, *Computing stable eigendecompositions of matrix pencils*, Linear Algebra Appl., 88/89 (1987), pp. 139–186.

[10] J. DEMMEL AND B. KÅGSTRÖM, *The generalized Schur decomposition of an arbitrary pencil*

$A - \lambda B$: Robust software with error bounds and applications. I. *Theory and algorithms*, ACM Trans. Math. Software, 19 (1993), pp. 160–174.

[11] J. Demmel and B. Kågström, *The generalized Schur decomposition of an arbitrary pencil $A-\lambda B$: Robust software with error bounds and applications.* II. *Software and applications*, ACM Trans. Math. Software, 19 (1993), pp. 175–201.

[12] J. Demmel and B. Kågström, *Accurate solutions of ill-posed problems in control theory*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 126–145.

[13] P. V. Dooren, *The computation of Kronecker's canonical form of a singular pencil*, Linear Algebra Appl., 27 (1979), pp. 103–140.

[14] P. V. Dooren, *The generalized eigenstructure problem in linear system theory*, IEEE Trans. Automat. Control, 26 (1981), pp. 111–129.

[15] P. V. Dooren, *Reducing subspaces: Definitions, properties and algorithms*, in Matrix Pencils, B. Kågström and A. Ruhe, eds., Lecture Notes in Math. 973, Springer-Verlag, Berlin, 1983, pp. 58–73.

[16] P. V. Dooren, *private communication*, 1996.

[17] A. Edelman, E. Elmroth, and B. Kågström, *A geometric approach to perturbation theory of matrices and matrix pencils.* I. *Versal deformations*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 653–692.

[18] A. Edelman, E. Elmroth, and B. Kågström, *A geometric approach to perturbation theory of matrices and matrix pencils.* II. *A Stratification-enhanced staircase algorithm*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 667–699.

[19] E. Elmroth and B. Kågström, *The set of 2-by-3 matrix pencils—Kronecker structures and their transitions under perturbations*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 1–34.

[20] A. Emami-Naeini and P. V. Dooren, *Computation of zeros of linear multivariable systems*, Automatica, 18 (1982), pp. 415–430.

[21] T. F. Fairgrieve, *The Application of Singularity Theory to the Computation of Jordan Canonical Form*, Master's thesis, Univ. of Toronto, Toronto, ON, Canada, 1986.

[22] G. Golub and C. V. Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, London, 1996.

[23] G. H. Golub and J. H. Wilkinson, *Ill-conditioned eigensystems and the computation of the Jordan canonical form*, SIAM Rev., 18 (1976), pp. 578–619.

[24] M. Gu, *New methods for estimating the distance to uncontrollability*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 989–1003.

[25] S. Helgason, *Differential Geometry, Lie Groups, and Symmetric Spaces*, Academic Press, New York, San Francisco, London, 1978.

[26] B. Kågström, *The generalized singular value decomposition and the general $A - \lambda B$ problem*, BIT, 24 (1984), pp. 568–583.

[27] B. Kågström, *RGSVD—An algorithm for computing the Kronecker structure and reducing subspaces of singular $A-\lambda B$ pencils*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 185–211.

[28] B. Kågström and A. Ruhe, *ALGORITHM* 560: *JNF, an algorithm for numerical computation of the Jordan normal form of a complex matrix [F2]*, ACM Trans. Math. Software, 6 (1980), pp. 437–443.

[29] B. Kågström and A. Ruhe, *An algorithm for numerical computation of the Jordan normal form of a complex matrix*, ACM Trans. Math. Software, 6 (1980), pp. 398–419.

[30] B. Kågström and A. Ruhe, *Matrix Pencils*, Lecture Notes in Math. 973, Springer-Verlag, New York, 1982.

[31] J. Kautsky, N. K. Nichols, and P. V. Dooren, *Robust pole assignment in linear state feedback*, Institute J. Control, 41 (1985), pp. 1129–1155.

[32] V. Kublanovskaya, *On a method of solving the complete eigenvalue problem of a degenerate matrix*, USSR Comput. Math. Phys., 6 (1966), pp. 1–14.

[33] V. Kublanovskaya, *AB-algorithm and its modifications for the spectral problem of linear pencils of matrices*, Numer. Math., 43 (1984), pp. 329–342.

[34] R. Lippert and A. Edelman, *Nonlinear eigenvalue problems*, in Templates for Eigenvalue Problems, Z. Bai, ed., to appear.

[35] A. Ruhe, *An algorithm for numerical determination of the structure of a general matrix*, BIT, 10 (1970), pp. 196–216.

[36] M. Wicks and R. DeCarlo, *Computing the distance to an uncontrollable system*, IEEE Trans. Automat. Control, 36 (1991), pp. 39–49.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 3, pp. 1026–1050
© 2000 Society for Industrial and Applied Mathematics

NEW ACCURATE ALGORITHMS FOR SINGULAR VALUE DECOMPOSITION OF MATRIX TRIPLETS*

ZLATKO DRMAČ†

Abstract. This paper presents a new algorithm for accurate floating-point computation of the singular value decomposition (SVD) of the product $A = B^\tau SC$, where $B \in \mathbf{R}^{p\times m}$, $C \in \mathbf{R}^{q\times n}$, $S \in \mathbf{R}^{p\times q}$, and $p \le m$, $q \le n$. The new algorithm uses diagonal scalings, the LU factorization with complete pivoting, the QR factorization with column pivoting, and matrix multiplication to replace A by $A' = B'^\tau S'C'$, where A and A' have the same singular values and the matrix A' is computed explicitly. The singular values of A' are computed using the Jacobi SVD algorithm. It is shown that the accuracy of the new algorithm is determined by (i) the accuracy of the QR factorizations of B^τ and C^τ; (ii) the accuracy of the LU factorization with complete pivoting of S; and (iii) the accuracy of the computation of the SVD of a matrix A' with moderate $\min_{D=\mathrm{diag}} \kappa_2(A'D)$. Theoretical analysis and numerical evidence show that, in the case of $\mathrm{rank}(B) = \mathrm{rank}(C) = p$ and full rank S, the accuracy of the new algorithm is unaffected by replacing B, S, C with, respectively, D_1B, D_2SD_3, D_4C, where D_i, $i = 1, \ldots, 4$, are arbitrary diagonal matrices. As an application, the paper proposes new accurate algorithms for computing the (H, K)–SVD and (H^{-1}, K)–SVD of S.

Key words. generalized singular value decomposition, matrix triplets, relative accuracy, singular values

AMS subject classifications. 65F15, 65F25, 65G05

PII. S0895479897321209

1. Introduction.

In this paper, we study floating-point computation of the singular value decomposition (SVD) of the product

$$A = B^\tau SC, \qquad B \in \mathbf{R}^{p\times m},\ \ C \in \mathbf{R}^{q\times n},\ \ S \in \mathbf{R}^{p\times q},\ \ p \le m,\ \ q \le n. \tag{1}$$

Our goal is to develop an efficient stable algorithm for computing the SVD of the matrix A and to compute the singular values and the singular vectors with high relative accuracy in the following regular case:

$$\mathrm{rank}(B) = p,\ \ \mathrm{rank}(C) = q,\ \ \mathrm{rank}(S) = \rho \equiv \min\{p, q\}. \tag{2}$$

If relation (2) holds, we call the 3-tuple (B, S, C) *a regular matrix triplet*. In that case, the matrix A has $\min\{m, n\} - \rho$ well-determined zero singular values. To compute the remaining nonzero singular values with high relative accuracy means that the computed approximations $\tilde\sigma_1 \ge \cdots \ge \tilde\sigma_\rho$ of the exact singular values $\sigma_1 \ge \cdots \ge \sigma_\rho$ satisfy a uniform error bound

$$\max_{1\le i\le\rho} \frac{|\tilde\sigma_i - \sigma_i|}{\sigma_i} \le f(m, n, p)\kappa(B, S, C)\boldsymbol{\varepsilon} < 1, \tag{3}$$

*Received by the editors May 7, 1997; accepted for publication (in revised form) by A. Edelman March 31, 1998; published electronically March 8, 2000. This work was supported in part by Croatian Ministry of Science and Technology grant 037012. This work was done during the author's stay at the Department of Computer Science, Engineering Center ECOT 7–7, University of Colorado at Boulder, Boulder, CO, and was supported in part by National Science Foundation grants ACS–9357812 and ASC–9625912 and Department of Energy grant DE–FG03–94ER25215. Parts of this work were presented at the International Workshop on Accurate Eigensolving and Applications, Split, Croatia, 1996, and at the First Congress of the Croatian Mathematical Society, Zagreb, Croatia, 1996.
http://www.siam.org/journals/simax/21-3/32120.html

†Department of Mathematics, University of Zagreb, Bijenička 30, 10000 Zagreb, Croatia (drmac@math.hr, http://www.math.hr/~drmac/).

where $\kappa(B,S,C)$ is a certain condition number, $f(m,n,p)$ is a modestly growing function of the dimensions, and $\boldsymbol{\varepsilon}$ is the round-off unit. We use perturbation theory to identify $\kappa(B,S,C)$ and we describe a class of regular triplets (B,S,C) for which $\kappa(B,S,C)$ is moderate.

The SVD computation of the product of three matrices arises in a number of applications. It is an implicit way to compute the (H,K)-SVD, introduced by Van Loan [46], [47]. Given symmetric and positive definite matrices $H \in \mathbf{R}^{p\times p}$ and $K \in \mathbf{R}^{q\times q}$, then the (H,K)-SVD of a matrix $S \in \mathbf{R}^{p\times q}$ is the decomposition $Y^{-1}SZ = D$, where D is diagonal and $Y^\tau HY = I_p$, $Z^\tau KZ = I_q$. It is easy to show that the (H,K)-SVD of S can be computed using the SVD of the product $B^\tau SC^{-1}$, where $BB^\tau = H$, $C^\tau C = K$. Positive definiteness of the matrix H is stable in the presence of floating-point rounding errors if and only if H can be written as $H = \Delta_H H_s \Delta_H$, where Δ_H is diagonal, H_s has unit diagonal, and the spectral condition number $\kappa_2(H_s)$ is moderate (cf. [10]). In that case, any factorization $H = BB^\tau$ yields full row rank matrix B such that $B = \Delta_B B_r$, where Δ_B is diagonal, the rows of B_r have unit Euclidean length, and $\kappa_2(B_r)$ is moderate. Similarly, if $PKP^\tau = CC^\tau$ is the Cholesky factorization with pivoting of K, then the rows of C^{-1} can be scaled so that the scaled matrix is well-conditioned. Van Loan [47], Larimore and Luk [34], and Ewerbring and Luk [23] show that the (H,K)-SVD provides theoretical and computational framework for solution of linear algebra problems such as weighted least squares, canonical correlations, and optimal prediction. Hence, all these problems can be solved using the SVD of the matrix A from relation (1), where B and C are full row rank matrices.

Computing the SVD of $B^\tau SC$ cannot be simply reduced to the SVD computation of the explicitly computed matrix $A = B^\tau SC$. Even if we compute the matrix $\tilde{A}$ as exactly rounded exact product $A = B^\tau SC$, and if we compute the SVD of $\tilde{A}$ exactly, the result might have unacceptably large relative error. For instance, let $\epsilon \neq 0$ and let

$$A = B^\tau SC = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \epsilon \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 1+\epsilon & 1-\epsilon \\ 1-\epsilon & 1+\epsilon \end{bmatrix}. \tag{4}$$

If $|\epsilon|$ is small (or large) enough, the values of $1 \pm \epsilon$ are rounded to one (or to $\pm\epsilon$), and the matrix $\tilde{A}$ is exactly singular. Thus, even an exact SVD algorithm cannot recover the minimal (nonzero) singular value of A, $\sigma_{\min}(A) = \min\{2, 2|\epsilon|\}$. (Note that $\sigma_{\min}(A)$ has small relative variation as a function of the parameter ϵ.)

To avoid numerical difficulties illustrated in example (4), Ewerbring and Luk [23] and Bojanczyk, Ewerbring, Luk, and van Dooren [5] used generalizations of the Kogbetliantz algorithm [33], [40], [30]. These algorithms start by reducing the problem to the computation of the SVD of the product of triangular matrices (cf. [30], [23]). In the next phase, these algorithms iteratively use plane rotations that solve a sequence of 2×2 subproblems. More precisely, these algorithms first use orthogonal transformations to replace the input data B, S, C with a triplet $A_1^{(0)}$, $A_2^{(0)}$, $A_3^{(0)}$ of upper triangular matrices such that $B^\tau SC$ and $A_1^{(0)}A_2^{(0)}A_3^{(0)}$ have the same nonzero singular values. Then, starting with $A' = A_1^{(0)}A_2^{(0)}A_3^{(0)}$, the iterative process is defined by

$$\left(A_i^{(k+1)} = Q_i^{(k)} A_i^{(k)} \left(Q_{i+1}^{(k)}\right)^\tau,\ \ i = 1,2,3\right), \qquad k = 0,1,2,\ldots, \tag{5}$$

where $Q_i^{(k)}$, $i = 1,2,3,4$, $k \geq 0$, are suitable orthogonal plane transformations, and all matrices $A_i^{(k)}$ are upper triangular. In the limit, $\lim_{k\to\infty}(A_1^{(k)}A_2^{(k)}A_3^{(k)})$ is diagonal matrix with the singular values of $A_1^{(0)}A_2^{(0)}A_3^{(0)}$ along the diagonal. Numerical

experiments presented in [5] show that the algorithm (5) is more accurate than the computation of the SVD of the explicitly computed matrix A. Note, however, that the computation and the memory access in algorithm (5) are quite involved: in each step, the algorithm determines four rotations and applies them to all three matrices both rowwise and columnwise, using Level 1 BLAS [37], [36]. Furthermore, if the singular vectors are needed, the algorithm updates two additional square arrays (accumulated products of the matrices $Q_1^{(k)}$, $k \geq 0$, and $Q_4^{(k)}$, $k \geq 0$, respectively).

On the other hand, a straightforward algorithm that computes $A = B^\tau SC$ explicitly and then computes the SVD of A is rather simple. It can efficiently use optimized Level 3 BLAS for matrix products, and the SVD of A can be computed using fast algorithms such as the divide and conquer algorithm [29]. Unfortunately, the cost of this simplicity is a potentially large relative error in the computed singular values. This large relative error may be caused by explicit computation of the product (cf. example (4)) or by inaccuracy of the algorithm for SVD computation (cf. [13]).

In this paper, we show that it is possible to follow the simplicity of the straightforward algorithm and, with proper preconditioning of the input matrices, to preserve the numerical stability. In our new algorithm, we first replace the triplet (B, S, C) with an equivalent triplet $(B', I, C') \equiv (B', C')$, and we compute the SVD of $B'^\tau C'$ using an algorithm from [15]. In that algorithm, the pair (B', C') is replaced by an equivalent pair (B'', C''), and the SVD is computed by an application of the Jacobi SVD algorithm to the explicitly computed matrix $B''^\tau C''$. In the transformation of (B, S, C) to $B''^\tau C''$, we use diagonal scalings, the LU and QR factorizations with pivoting, and the standard matrix multiplication that involves triangular or trapezoidal matrices. We use a similar strategy to compute the (H, K)-SVD of a rectangular matrix S, where we use the Cholesky factorizations of H and M to reduce the problem to the computation of the SVD of matrix triplet. Modular structure of the new algorithms makes them suitable for LAPACK- [1] and ScaLAPACK-style [4] implementations on top of high performance libraries such as BLAS [14] and PBLAS [7].

We show that the accuracy of the new algorithm for the SVD of $B^\tau SC$ is determined by our ability to compute (i) accurate QR factorizations of B^τ and C^τ; (ii) accurate LU factorization with complete pivoting of S; and (iii) accurate SVD of a matrix A with moderate $\min_{D=\mathrm{diag}} \kappa_2(AD)$. Our theoretical analysis and numerical evidence show that the accuracy of the new algorithm is unaffected by replacing B, S, C with, respectively, D_1B, D_2SD_3, D_4C, where D_i, $i = 1, \ldots, 4$, are arbitrary diagonal matrices. A similar conclusion holds for the computation of the (H, K)-SVD of S and scalings D_1HD_1, D_2SD_3, D_4KD_4.

The paper is organized as follows. In section 2, we review the algorithm from [15] for accurate computation of the SVD of the product of two matrices. The new algorithm is described and analyzed in section 3. We give a backward error estimate for the general case (1) and we prove that in the case of the regular triplet (B, S, C) the new algorithm is capable of achieving high relative accuracy. In section 4, we use the technique developed in section 3 to develop new accurate algorithms for computation of the (H, K)-SVD of general rectangular matrix S. We also analyze a new algorithm based on the generalized SVD (GSVD; cf. [46], [47], [41]) of regular matrix pairs. In section 5, we show that floating-point implementations of the new algorithms run as predicted by the theory.

2. The SVD of the product $B^\tau C$. The SVD of the product of two matrices (product induced SVD; [24]) arises in a number of applications. An example is computation of the canonical correlations of normally scaled matrix pairs; see [28]. Further,

if $M = CC^\tau$ and $H = BB^\tau$ are the Cholesky factorizations of positive definite matrices M and H, then the eigenvalue problem $MHx = \lambda x$ can be reduced to the SVD of $B^\tau C$. The eigenvalue problem $MHx = \lambda x$ has applications in computation of the contragredient transformation in the design of reduced-order linear systems (see [35]) or in statistical computation such as principal relations (see [48]). In this paper, we use the SVD of the product of two matrices in process of computing the SVD of the matrix product in relation (1).

We compute the SVD of $B^\tau C$ using the following algorithm from [15].

ALGORITHM 2.1. PSVD(B, C).

Input $B \in \mathbf{R}^{p\times m}$, $C \in \mathbf{R}^{p\times n}$, $p \le \min\{m, n\}$.

Step 1 Compute $\Delta_B = \mathrm{diag}(\|B^\tau e_i\|_2)$ and $B_r = \Delta_B^\dagger B$, $C_1 = \Delta_B C$.

Step 2 Compute the QR factorization of C_1^τ with column pivoting,

$$C_1^\tau \Pi = Q \begin{bmatrix} R \\ \mathbf{O}_{n-\gamma,p} \end{bmatrix}, \quad R \in \mathbf{R}^{\gamma\times p}, \ \mathrm{rank}(R) = \gamma, \ Q \text{ orthogonal.}$$

Step 3 Compute the matrix $F = B_r^\tau \Pi R^\tau$, using the standard matrix multiply algorithm.

Step 4 Compute the QR factorization (optionally, with column pivoting) of F:

$$F\Pi_F = Q_F \begin{bmatrix} R_F \\ \mathbf{O} \end{bmatrix}.$$

Step 5 Apply the Jacobi SVD algorithm to R_F^τ to compute the SVD of R_F as $\Sigma = V^\tau R_F W$.

Output The SVD of $B^\tau C$ is $\begin{bmatrix} \Sigma \\ \mathbf{O} \end{bmatrix} = \begin{bmatrix} V^\tau & \\ & I \end{bmatrix} Q_F^\tau (B^\tau C)(Q(W \oplus I_{n-p}))$.

Algorithm 2.1 replaces the computation of the SVD of $B^\tau C$ with the computation of the SVD of the explicitly computed matrix F. Since $F = B^\tau CQ$, the first three steps of Algorithm 2.1 can be viewed as a way to find an orthogonal matrix Q such that the product $F = B^\tau(CQ)$ can be computed without loss of accuracy. The key idea is to combine diagonal scaling and the following two important properties of the QR factorization of a general full column rank matrix Y: (i) *If $Y = Y_c\Delta_Y$, where Δ_Y is a diagonal matrix and Y_c is a well-conditioned matrix with unit column norms, then floating-point QR factorization of Y is accurate independent of $\kappa_2(\Delta_Y)$* (*cf.* [17], [31]). (ii) *If $Y\Pi = QR_Y$ is the QR factorization with column pivoting* [26]*, then the matrix R_Y is of the form $R_Y = D(R_Y)_r$, where D is diagonal and $(R_Y)_r$ is well-conditioned independent of Y*. The resulting matrix F is of the form $F = F_c\Delta_F$, where Δ_F is diagonal and $\kappa_2(F_c)$ is moderate if $\kappa_2(B_r)$ is moderate. Accurate SVD of F is possible independent of $\kappa_2(\Delta_F)$ due to excellent stability properties of the Jacobi SVD algorithm (cf. [13], [18]).

In Step 1 of Algorithm 2.1 we use the generalized inverse $\Delta_B^\dagger$ for the case that some rows of B are zero. In that case, z (say) zero rows of B produce z zero columns in C_1^τ. The pivoting in the QR factorization of C_1^τ ensures that the zero columns in C_1^τ are permuted to the last z positions. As a result, the matrix F equals $F = [\hat{F}, \mathbf{O}]$, where $\hat{F}$ is $m \times (p - z)$, and the corresponding z zero singular values are deflated without error.

The QR factorization of C_1^τ is computed with column pivoting as in [26]. If we use the LAPACK procedure `SGEQPF()` (cf. [1]), the matrix Q is computed in factored form, as a product of Householder reflections. The Householder vectors that define these reflections are stored in the lower trapezoidal part of the array C_1^τ. In Step 5,

the Jacobi SVD algorithm computes the SVD of R_F as $W\Sigma = R_F^\tau V$, where V denotes the accumulated product of the Jacobi rotations and $W\Sigma$ is the limit matrix. Note, however, that in the case of trapezoidal R_F (that is, rank deficient F) we need the QR factorization of the limit matrix $R_F^\tau V$ in order to obtain $p \times p$ orthogonal matrix W. As the output from Algorithm 2.1 indicates, the Jacobi rotations are accumulated only if both the left and the right singular vectors of $B^\tau C$ are needed. Indeed, if only the left singular matrix is needed, we can apply Algorithm 2.1 to the matrix $C^\tau B$.

To analyze Algorithm 2.1 in floating-point arithmetic, we use the standard model

$$\boldsymbol{fl}(a \odot b) = (a \odot b)(1+\xi), \quad \boldsymbol{fl}(\sqrt{c}) = \sqrt{c}(1+\zeta), \quad |\xi|, |\zeta| \leq \boldsymbol{\varepsilon}, \tag{6}$$

where a, b, and c are floating-point numbers; $\odot$ denotes any of the four elementary operations $+$, $-$, $\cdot$, and $\div$; and $\boldsymbol{\varepsilon}$ is the round-off unit. From relation (6) it follows that the floating-point product Z of an $m \times n$ matrix X and an $n \times p$ matrix Y satisfies

$$Z = XY + E, \quad |E| \leq \boldsymbol{\varepsilon}_{MM}(n)|X| \cdot |Y|, \quad 0 \leq \boldsymbol{\varepsilon}_{MM}(n) \leq (1+\boldsymbol{\varepsilon})^{n+1} - 1, \tag{7}$$

and that the Euclidean length $\|x\|_2$ of a floating-point vector $x \in \mathbf{R}^n$ is computed as

$$\boldsymbol{fl}(\|x\|_2) = \|x\|_2(1+\epsilon), \quad |\epsilon| \leq \boldsymbol{\varepsilon}_{\ell_2}(n) \leq (1+\boldsymbol{\varepsilon})^{(n+2)/2} - 1. \tag{8}$$

Further, floating-point QR factorization of a $n \times p$ matrix Y can be represented as

$$Y + \delta Y = \tilde{Q} \begin{bmatrix} \tilde{R}_Y \\ \mathbf{O} \end{bmatrix}, \quad \|\delta Y e_i\|_2 \leq \boldsymbol{\varepsilon}_{QR}(n,p)\|Y e_i\|_2, \ \ 1 \leq i \leq p, \tag{9}$$

where $\tilde{Q}$ is certain orthogonal matrix, $\tilde{R}_Y$ is the computed triangular factor, and $\boldsymbol{\varepsilon}_{QR}(n,p)$ is bounded by round-off $\boldsymbol{\varepsilon}$ times a modest polynomial in n, p (cf. [17], [31]). Similarly, if the Jacobi SVD algorithm is applied to a $p \times p$ matrix G, then the computed matrix $\tilde{G}^{(\ell)}$ that satisfies the stopping criterion (all columns mutually orthogonal up to a given tolerance `tol`; cf. [13]) can be represented as

$$\tilde{G}^{(\ell)} = (G + \delta G)U, \ \ \|(\delta G)^\tau e_i\|_2 \leq \boldsymbol{\varepsilon}_J(p)\|G^\tau e_i\|_2, \ \ 1 \leq i \leq p,$$

where U is a certain orthogonal matrix and $\boldsymbol{\varepsilon}_J(p)$ is bounded by round-off $\boldsymbol{\varepsilon}$ times a modest polynomial in p. (For more details about the Jacobi SVD algorithm see [13], [17], [18].)

Accuracy and stability properties of Algorithm 2.1 are given in the following theorem from [15].

THEOREM 2.1. *Let $\tilde{R} \in \mathbf{R}^{\tilde{\gamma} \times p}$ be the computed triangular (generally, trapezoidal) factor in Step 2 of Algorithm 2.1, and let $\tilde{R}_{r,1} = \mathrm{diag}(\|\tilde{R}^\tau e_i\|_1)^{-1}\tilde{R}$. Let $\tilde{F}$ be the floating-point approximation of the matrix F in Step 3, and let $\tilde{R}_F$ be the computed upper triangular factor in floating-point QR factorization of $\tilde{F}$. Let the Jacobi SVD algorithm be applied on $G = \tilde{R}_F^\tau$, and let the columns of the output matrix $\tilde{G}^{(\ell)}$ be mutually orthogonal up to $\tau(p) \leq \mathtt{tol} + O(p\boldsymbol{\varepsilon})$. Let $\tilde{F}^{(\ell)} = (\tilde{G}^{(\ell)})^\tau$. Then there exist backward perturbations δB, δC such that the diagram in Figure 1 commutes. Furthermore, it holds for all i that*

$$\|(\delta B)^\tau e_i\|_2 \leq \eta_B \|B^\tau e_i\|_2, \quad \|(\delta C)^\tau e_i\|_2 \leq \eta_C \|C^\tau e_i\|_2, \tag{10}$$

with $\eta_C \leq \boldsymbol{\varepsilon}_{QR}(n,p)(1+\boldsymbol{\varepsilon}) + \boldsymbol{\varepsilon}$, and

$$\eta_B = \frac{1+\boldsymbol{\varepsilon}_{\ell_2}(m)}{1-\boldsymbol{\varepsilon}_{\ell_2}(m)}(1+\boldsymbol{\varepsilon})\left\{\boldsymbol{\varepsilon}_{MM}(p)\| \ |\tilde{R}^\dagger| \cdot |\tilde{R}| \ \|_\infty + \eta(m,p)(1+\boldsymbol{\varepsilon}_{MM}(p))\|(\tilde{R}_{r,1})^\dagger\|_\infty\right\} + \boldsymbol{\varepsilon},$$

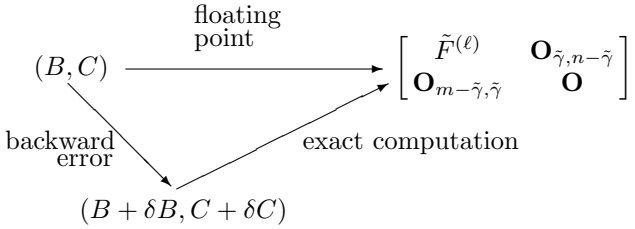


FIG. 1. *Commutative diagram of PSVD*(B, C).

where $\eta(m,p) = \varepsilon_{QR}(m,p) + \varepsilon_J(p) + \varepsilon_{QR}(m,p)\varepsilon_J(p)$. *Let* B *and* C *have full row ranks, let* $\sigma_1 \geq \cdots \geq \sigma_p > 0$ *be the singular values of* $B^\tau C$, *and let* $\tilde{\sigma}_1 \geq \cdots \geq \tilde{\sigma}_p$ *be the sorted floating-point approximations of the Euclidean column norms of* $\tilde{F}^{(\ell)}$. *Let* $B = \mathrm{diag}(\|B^\tau e_i\|_2)B_r$, $C = \mathrm{diag}(\|C^\tau e_i\|_2)C_r$. *If* $\sqrt{p}\eta_B\|B_r^\dagger\|_2 < 1$, $\sqrt{p}\eta_C\|C_r^\dagger\|_2 < 1$, *then*

$$\max_{1\leq i\leq p} \frac{|\tilde{\sigma}_i - \sigma_i|}{\sigma_i} \leq (1 + \sqrt{p}\eta_B\|B_r^\dagger\|_2)(1 + \sqrt{p}\eta_C\|C_r^\dagger\|_2)\frac{1 + \varepsilon_{\ell_2}(p)}{1 - p\tau(p)} - 1. \tag{11}$$

Note that Algorithm 2.1 has the same error bound in the family of matrices $\{B^\tau DC,\ D \in \mathcal{D}_p\}$, where $\mathcal{D}_p$ is the set of $p \times p$ diagonal nonsingular matrices. If B and C are not full row rank matrices, then the angle between the column spaces of B and C plays an important role in the stability of the singular values of $B^\tau C$. More precisely, if the column spaces of B and C are nearly mutually orthogonal, computation with high relative accuracy is not possible. For a proof of Theorem 2.1 and further discussion see [15], [28].

3. The SVD of the product $B^\tau SC$. In this section, we analyze the new algorithm for the SVD of $B^\tau SC$. In section 3.1, we give a detailed description of the algorithm. In section 3.2, we give a backward error analysis of a floating-point implementation of the new algorithm, and in section 3.3 we estimate the relative errors in the computed singular values and singular vectors. In section 3.4 we briefly discuss applications of the new algorithm to computation of the SVD of $B^\tau S^{-1}C$ (S square and nonsingular).

3.1. The algorithm. The main idea in the new algorithm is to reduce the triplet (B, S, C) to an equivalent triplet $(B', I, C') \equiv (B', C')$. We use diagonal scalings, the LU factorization with complete pivoting, and matrix multiplications. The new pair (B', C') is then given as input to Algorithm 2.1.

ALGORITHM 3.1. PSVD(B, S, C).

Input $B \in \mathbf{R}^{p\times m}$, $C \in \mathbf{R}^{q\times n}$, $S \in \mathbf{R}^{p\times q}$, $p \leq m$, $q \leq n$.

Step 1 Compute $\Delta_B = \mathrm{diag}(\|B^\tau e_i\|_2)$ and $\Delta_C = \mathrm{diag}(\|C^\tau e_i\|_2)$. Then compute $B_r = \Delta_B^\dagger B$, $C_r = \Delta_C^\dagger C$, $S_1 = \Delta_B S \Delta_C$.

Step 2 Compute the LU factorization with complete pivoting of S_1,

$$\Pi_1 S_1 \Pi_2 = LU,$$
$$L \in \mathbf{R}^{p\times\rho}, \quad U \in \mathbf{R}^{\rho\times q}, \quad \rho = \mathrm{rank}(L) = \mathrm{rank}(U), \quad L_{ii} = 1, \quad 1 \leq i \leq \rho.$$

Step 3 Compute $M = L^\tau \Pi_1 B_r$, $N = U\Pi_2^\tau C_r$, and apply Algorithm 2.1 to the pair (M, N).

Output Let Q, Q_F, V, and W be as in Algorithm 2.1. The SVD of $B^\tau SC$ is

$$\begin{bmatrix} \Sigma \oplus \mathbf{O} \\ \mathbf{O} \end{bmatrix} = \begin{bmatrix} V^\tau & \\ & I \end{bmatrix} Q_F^\tau (B^\tau SC)(Q(W \oplus I_{n-p})).$$

The key idea in Algorithm 3.1 (and in Algorithm 2.1) is to ensure that important condition numbers of certain scaled matrices do not increase too much after explicit computation of matrix products. For instance, due to pivoting in the LU factorization, the matrix U is a product of a diagonal matrix and a well-conditioned matrix, and, as a result, the matrix $N = U(\Pi_2^\tau C_r)$ is the product of a diagonal matrix and a well-conditioned matrix. A similar conclusion holds for the matrix M.

Algorithm 3.1 is simple and efficient. Its efficiency depends on the efficiency of the LU and the QR factorizations, matrix multiplication that involves triangular or trapezoidal matrix, and the efficiency of the ordinary SVD computation. This high-level structure of the algorithm also allows easy adaptation to modern computer architectures.

Our next observation is that Algorithm 3.1 computes the SVD of $B^\tau SC$ without any additional square array. To illustrate this feature, we give a brief description of the memory usage in Algorithm 3.1. On input, B and C are stored in transposed form, that is, we are given B^τ and C^τ as $m \times p$ and $n \times q$ arrays, respectively. The matrices L and U are stored in the standard way: L is stored in the lower trapezoidal part of S, while U occupies the upper trapezoidal part of S. The matrix M^τ is stored in the leading $m \times \min\{p, q\}$ submatrix of the initial array B^τ, and N^τ is stored in the $n \times \min\{p, q\}$ leading submatrix of C^τ. Both M^τ and N^τ are efficiently computed using the Level 3 BLAS procedures `STRMM()`, `SGEMM()` and without additional working space. In the algorithm PSVD(M, N) (Algorithm 2.1), most of the computation can be performed without additional square arrays. If we need the accumulated product of the Jacobi rotations, we can use the leading $\min\{p, q\} \times \min\{p, q\}$ submatrix of S.

3.2. Backward error analysis. Although Algorithm 3.1 reduces the triplet (B, S, C) to a single matrix by explicit computation of matrix products, we can show that, in certain well-conditioned cases, the computed singular values approximate the true values with high relative accuracy. Furthermore, we show that floating-point computation is equivalent to exact calculations with a triplet $(B+\delta B, S+\delta S, C+\delta C)$, where δB, δS, and δC are small backward errors.

THEOREM 3.1. *Let $\tilde{\Delta}_B$, $\tilde{\Delta}_C$, $\tilde{B}_r$, $\tilde{C}_r$, $\tilde{L}$, $\tilde{U}$, $\tilde{M}$, and $\tilde{N}$ be the computed values of Δ_B, Δ_C, B_r, C_r, L, U, M, and N, respectively. Furthermore, let $\tilde{L} = \tilde{L}_{c,1}\,\mathrm{diag}(\|\tilde{L}e_i\|_1)$ and $\tilde{U} = \mathrm{diag}(\|\tilde{U}^\tau e_i\|_1)\tilde{U}_{r,1}$ and let $\mathrm{rank}(\tilde{L}) = \mathrm{rank}(\tilde{U}) = \tilde{\rho}$. Let $\tilde{F}^{(\ell)}$ be the matrix computed as the output matrix from Algorithm 2.1 in Step 3 of Algorithm 3.1 (cf. Theorem 2.1), and let $\delta\tilde{M}$ and $\delta\tilde{N}$ be the backward errors, where for some small constants $\eta_{\tilde{M}}$ and $\eta_{\tilde{N}}$ (cf. Theorem 2.1),*

$$\|(\delta\tilde{M})^\tau e_i\|_2 \le \eta_{\tilde{M}}\|\tilde{M}^\tau e_i\|_2, \quad \|(\delta\tilde{N})^\tau e_i\|_2 \le \eta_{\tilde{N}}\|\tilde{N}^\tau e_i\|_2, \quad 1 \le i \le p. \tag{12}$$

There exist perturbations δB, δC, and δS such that the diagram in Figure 2 is commutative. More precisely, there exist orthogonal matrices W' and V' such that

$$\begin{bmatrix} \tilde{F}^{(\ell)} & \mathbf{O}_{\tilde{\rho},n-\tilde{\rho}} \\ \mathbf{O}_{m-\tilde{\rho},\tilde{\rho}} & \mathbf{O}_{m-\tilde{\rho},n-\tilde{\rho}} \end{bmatrix} = W'^\tau (B+\delta B)^\tau (S+\delta S)(C+\delta C)V'$$

and such that for all i, $\|(\delta B)^\tau e_i\|_2 \le \eta_B\|B^\tau e_i\|_2$, $\|(\delta C)^\tau e_i\|_2 \le \eta_C\|C^\tau e_i\|_2$, where

$$\eta_B = \varepsilon + (1+\varepsilon)\frac{\varepsilon_{MM}(p)\| \,|\tilde{L}| \cdot |\tilde{L}^\dagger| \,\|_1 + \eta_{\tilde{M}}(1+\varepsilon_{MM}(p))\|(\tilde{L}_{c,1})^\dagger\|_1}{1-\varepsilon_{\ell_2}(m)}, \tag{13}$$

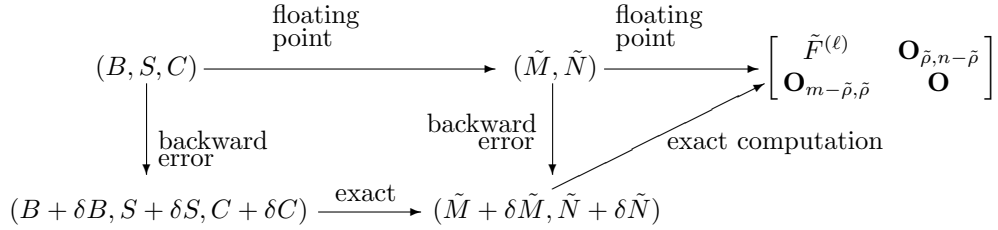


FIG. 2. *Commutative diagram of PSVD*(B,S,C).

$$\eta_C = \varepsilon + (1+\varepsilon)\frac{\varepsilon_{MM}(q)\| \, |\tilde{U}^\dagger| \cdot |\tilde{U}| \, \|_\infty + \eta_{\tilde{N}}(1+\varepsilon_{MM}(q))\|(\tilde{U}_{r,1})^\dagger\|_1}{1-\varepsilon_{\ell_2}(n)}. \tag{14}$$

Furthermore, δS is such that

$$|\delta S| \le 3\varepsilon|S| + \varepsilon_{LU}(q)\tilde{\Delta}_B^{-1}\tilde{\Pi}_1^\tau|\tilde{L}| \cdot |\tilde{U}|\tilde{\Pi}_2^\tau\tilde{\Delta}_C^{-1}. \tag{15}$$

Proof. From relation (7), it follows that

$$\tilde{M} = \tilde{L}^\tau\tilde{\Pi}_1\tilde{B}_r + \mathcal{E}_M, \quad |\mathcal{E}_M| \le \varepsilon_{MM}(p)|\tilde{L}|^\tau \cdot \tilde{\Pi}_1 \cdot |\tilde{B}_r|, \tag{16}$$

$$\tilde{N} = \tilde{U}\tilde{\Pi}_2^\tau\tilde{C}_r + \mathcal{E}_N, \quad |\mathcal{E}_N| \le \varepsilon_{MM}(q)|\tilde{U}| \cdot \tilde{\Pi}_2^\tau \cdot |\tilde{C}_r|. \tag{17}$$

On the other hand, from Theorem 2.1, it follows that there exist orthogonal matrices W', V' such that

$$\begin{bmatrix} \tilde{F}^{(\ell)} & \mathbf{O}_{\tilde{\rho},n-\tilde{\rho}} \\ \mathbf{O}_{m-\tilde{\rho},\tilde{\rho}} & \mathbf{O}_{m-\tilde{\rho},n-\tilde{\rho}} \end{bmatrix} = W'^\tau(\tilde{M}+\delta\tilde{M})^\tau(\tilde{N}+\delta\tilde{N})V' \tag{18}$$

and such that relation (12) holds. Further, we can write $\tilde{M}+\delta\tilde{M}$ as

$$\tilde{M}+\delta\tilde{M} = \tilde{L}^\tau\tilde{\Pi}_1(\tilde{B}_r+\delta\tilde{B}_r), \;\; \delta\tilde{B}_r = \tilde{\Pi}_1^\tau(\tilde{L}^\tau)^\dagger\mathcal{E}_M + \tilde{\Pi}_1^\tau(\tilde{L}^\tau)^\dagger\delta\tilde{M},$$

where $(\tilde{L}^\tau)^\dagger = (\tilde{L}^\dagger)^\tau = \tilde{L}(\tilde{L}^\tau\tilde{L})^{-1}$. Note that

$$|\tilde{\Pi}_1^\tau(\tilde{L}^\tau)^\dagger\mathcal{E}_M| \le \varepsilon_{MM}(p)\tilde{\Pi}_1^\tau \cdot |\tilde{L}^\dagger|^\tau \cdot |\tilde{L}|^\tau \cdot \tilde{\Pi}_1 \cdot |\tilde{B}_r|$$

and that for all i and i' such that $\tilde{\Pi}_1 e_i = e_{i'}$,

$$\|(\tilde{\Pi}_1^\tau(\tilde{L}^\dagger)^\tau\delta\tilde{M})^\tau e_i\|_2 \le \eta_{\tilde{M}}\sum_{k=1}^{\tilde{\rho}}|\tilde{L}^\dagger|_{ki'}\|\tilde{M}^\tau e_k\|_2 \tag{19}$$

$$\le \eta_{\tilde{M}}(1+\varepsilon_{MM}(p))\max_{1\le j\le p}\|\tilde{B}_r^\tau e_j\|_2\sum_{k=1}^{\tilde{\rho}}|\tilde{L}^\dagger|_{ki'}\|\tilde{L}e_k\|_1 \tag{20}$$

$$\le \eta_{\tilde{M}}(1+\varepsilon_{MM}(p))\max_{1\le j\le p}\|\tilde{B}_r^\tau e_j\|_2\|(\tilde{L}_{c,1})^\dagger\|_1. \tag{21}$$

Here we have used the fact that, for any diagonal nonsingular D, $(\tilde{L}D)^\dagger = D^{-1}\tilde{L}^\dagger$. Thus, we can estimate $\delta\tilde{B}_r$ by $\|(\delta\tilde{B}_r)^\tau e_i\|_2 \le \zeta_B$, $1 \le i \le p$, where

$$\zeta_B = (1+\varepsilon)\frac{\varepsilon_{MM}(p)\| \, |\tilde{L}| \cdot |\tilde{L}^\dagger| \, \|_1 + \eta_{\tilde{M}}(1+\varepsilon_{MM}(p))\|(\tilde{L}_{c,1})^\dagger\|_1}{1-\varepsilon_{\ell_2}(m)}. \tag{22}$$

Hence, $\tilde{M}+\delta\tilde{M}=\tilde{L}^{\tau}\tilde{\Pi}_1\tilde{\Delta}_B^{-1}(B+\delta B_e+\tilde{\Delta}_B\delta\tilde{B}_r)$, $|\delta B_e|\le\varepsilon|B|$, and the estimate for δB follows by defining $\delta B=\delta B_e+\tilde{\Delta}_B\delta\tilde{B}_r$ and $\eta_B=\varepsilon+\zeta_B$.

Similarly, we can write $\tilde{N}+\delta\tilde{N}$ as $\tilde{N}+\delta\tilde{N}=\tilde{U}\tilde{\Pi}_2^{\tau}\tilde{\Delta}_C^{-1}(C+\delta C_e+\tilde{\Delta}_C\delta\tilde{C}_r)$, where for all i, $\|(\delta\tilde{C}_r)^{\tau}e_i\|_2\le\zeta_C$ with

$$\zeta_C=(1+\varepsilon)\frac{\varepsilon_{MM}(q)\|\,|\tilde{U}^{\dagger}|\cdot|\tilde{U}|\,\|_{\infty}+\eta_{\tilde{N}}(1+\varepsilon_{MM}(q))\|(\tilde{U}_{r,1})^{\dagger}\|_1}{1-\varepsilon_{\ell_2}(n)}. \tag{23}$$

The estimate for δC follows by defining $\delta C=\delta C_e+\tilde{\Delta}_C\delta\tilde{C}_r$ and $\eta_C=\varepsilon+\zeta_C$. Thus, we have

$$\begin{bmatrix}\tilde{F}^{(\ell)} & \mathbf{O}_{\tilde{\rho},n-\tilde{\rho}}\\ \mathbf{O}_{m-\tilde{\rho},\tilde{\rho}} & \mathbf{O}_{m-\tilde{\rho},n-\tilde{\rho}}\end{bmatrix}=W'^{\tau}(B+\delta B)^{\tau}\tilde{\Delta}_B^{-1}\tilde{\Pi}_1^{\tau}\tilde{L}\tilde{U}\tilde{\Pi}_2^{\tau}\tilde{\Delta}_C^{-1}(C+\delta C)V'. \tag{24}$$

It remains to estimate the backward error in S. If $\tilde{S}_1$ is the approximation of S_1, then $\tilde{S}_1=\tilde{\Delta}_B(S+\delta S_e)\tilde{\Delta}_C$, where $|\delta S_e|\le 3\varepsilon|S|$. Furthermore, the computed triangular factors $\tilde{L}$ and $\tilde{U}$ satisfy the relation

$$\tilde{L}\tilde{U}=\tilde{\Pi}_1\tilde{S}_1\tilde{\Pi}_2+\delta(\tilde{\Pi}_1\tilde{S}_1\tilde{\Pi}_2),\quad |\delta(\tilde{\Pi}_1\tilde{S}_1\tilde{\Pi}_2)|\le\varepsilon_{LU}(q)|\tilde{L}|\cdot|\tilde{U}|, \tag{25}$$

where $\varepsilon_{LU}(q)\le q\varepsilon/(1-q\varepsilon)$. Hence, if we define $\delta S'=\tilde{\Delta}_B^{-1}\tilde{\Pi}_1^{\tau}\delta(\tilde{\Pi}_1\tilde{S}_1\tilde{\Pi}_2)\tilde{\Pi}_2^{\tau}\tilde{\Delta}_C^{-1}$, we can write

$$\tilde{L}\tilde{U}=\tilde{\Pi}_1\tilde{\Delta}_B(S+\delta S_e+\delta S')\tilde{\Delta}_C\tilde{\Pi}_2, \tag{26}$$

where $\delta S'$ satisfies $|\delta S'|\le\varepsilon_{LU}(q)\tilde{\Delta}_B^{-1}\tilde{\Pi}_1^{\tau}|\tilde{L}|\cdot|\tilde{U}|\tilde{\Pi}_2^{\tau}\tilde{\Delta}_C^{-1}$. Now, define $\delta S=\delta S_e+\delta S'$ and use relation (26) to replace $\tilde{L}\tilde{U}$ in (24) with $\tilde{\Pi}_1\tilde{\Delta}_B(S+\delta S)\tilde{\Delta}_C\tilde{\Pi}_2$. ∎

Remark 3.1. From relation (25), we see that the transition from S to $\tilde{L}\tilde{U}$ can be equivalently described using mixed error analysis: after small rounding errors ($|\delta S_e|\le 3\varepsilon|S|$) followed by exact scaling and pivoting ($\tilde{\Pi}_1\tilde{\Delta}_B(S+\delta S_e)\tilde{\Delta}_C\tilde{\Pi}_2$), we have backward perturbation followed by exact LU factorization.

Remark 3.2. Using the results from [43], [2], [19], [8], we conclude that in Step 2 of Algorithm 3.1 we can also use the QR factorization with complete pivoting of Powell and Reid to obtain similar accuracy in many cases. However, the LU approach is generally more accurate.

Remark 3.3. If the matrix C_1^{τ} in Algorithm 2.1 does not admit accurate QR factorization, i.e., if $\kappa_2(C_r)$ is large, or if the Powell–Reid complete pivoting does not improve the accuracy of the QR factorization, we can first compute the LU factorization with complete pivoting of $C_1=\Delta_B C$, $P_C^{(1)}C_1P_C^{(2)}=L_CU_C$, and then call PSVD$(P_C^{(1)}B_r,L_C,U_C(P_C^{(2)})^{\tau})$. Let $\tilde{P}_C^{(1)}(\tilde{C}_1+\delta\tilde{C}_1)P_C^{(2)}=(I+E_L)L_CU_C(I+E_U)$ be the computed LU factorization, where $\tilde{C}_1=\boldsymbol{fl}(\Delta_BC)$. Then on input to Algorithm 3.1 we seek the SVD of

$$\left(I+\tilde{B}_r^{\tau}(\tilde{P}_C^{(1)})^{\tau}E_L(\tilde{B}_r^{\tau})^{\dagger}\right)\tilde{B}_r^{\tau}\underbrace{(\tilde{P}_C^{(1)})^{\tau}L_CU_C(\tilde{P}_C^{(2)})^{\tau}}_{\tilde{C}_1}\left(I+\tilde{P}_C^{(2)}E_U(\tilde{P}_C^{(2)})^{\tau}\right).$$

Similarly, if $P_B^{(1)}BP_B^{(2)}=L_BU_B$ is the LU factorization with complete pivoting of B, then we can use PSVD$(U_B(P_B^{(2)})^{\tau},L_B^{\tau},P_B^{(1)}C)$. Combining the two LU factorizations yields an application of PSVD$(U_B(P_B^{(2)})^{\tau},L_B^{\tau}P_B^{(1)}(P_C^{(1)})^{\tau}L_C,U_C(P_C^{(2)})^{\tau})$. These procedures can be easily analyzed using the results and techniques from previous sections.

3.3. Relative error estimate for the singular values and vectors. In this section, we examine the sensitivity of the singular values of $A = B^\tau SC$ from relation (1), if B, S, and C are changed to $B+\delta B$, $S+\delta S$, and $C+\delta C$, respectively. Since we are interested in the relative accuracy of the singular values computed by Algorithm 3.1 in floating-point arithmetic, we deduce singular value perturbation estimates for perturbations δB, δS, and δC from the backward error analysis of Algorithm 3.1 (cf. Theorem 3.1).

We start with an estimate of the accuracy of the floating-point LU factorization in Step 2 of the algorithm. For the sake of simplicity, we assume that $q = p$ and $\mathrm{rank}(S) = p$.

PROPOSITION 3.1. *Let the notation of Theorem* 3.1 *hold, and let* $\tilde{L}$ *and* $\tilde{U}$ *be* $p \times p$ *nonsingular matrices. Furthermore, let*

$$S_1' \equiv \tilde{\Pi}_1 \tilde{\Delta}_B S \tilde{\Delta}_C \tilde{\Pi}_2 = L'U' \tag{27}$$

be the exact LU factorization of S_1', *and let* $\delta S_1' = \tilde{\Pi}_1 \tilde{\Delta}_B \delta S_e \tilde{\Delta}_C \tilde{\Pi}_2 + \delta(\tilde{\Pi}_1 \tilde{S}_1 \tilde{\Pi}_2)$. *Note that* $S_1' + \delta S_1' = \tilde{L}\tilde{U}$ *and that*

$$|\delta S_1'| \le 3\boldsymbol{\varepsilon}|S_1'| + \boldsymbol{\varepsilon}_{LU}(p)\,|\tilde{L}| \cdot |\tilde{U}|. \tag{28}$$

If the spectral radius of $|\tilde{L}^{-1}\delta S_1'\tilde{U}^{-1}|$ *is less than one, there exist a strictly lower triangular matrix* E_L *and an upper triangular matrix* E_U *such that* $S_1' + \delta S_1' = (I + E_L)S_1'(I + E_U)$ *and such that*

$$|E_L| \le \boldsymbol{\varepsilon}_{LU}(p)\,|\tilde{L}|\,\mathbf{tril}\left(|\tilde{L}^{-1}| \cdot |\tilde{L}| \cdot |\tilde{U}| \cdot |\tilde{U}^{-1}|\right)|L'^{-1}| + O(\boldsymbol{\varepsilon}^2), \tag{29}$$

$$|E_U| \le \boldsymbol{\varepsilon}_{LU}(p)\,|U'^{-1}|\,\overline{\mathbf{triu}}\left(|\tilde{L}^{-1}| \cdot |\tilde{L}| \cdot |\tilde{U}| \cdot |\tilde{U}^{-1}|\right)|\tilde{U}| + O(\boldsymbol{\varepsilon}^2). \tag{30}$$

Here $\overline{\mathbf{triu}}(\cdot)$ *and* $\mathbf{tril}(\cdot)$ *denote, respectively, the upper and the strictly lower triangular parts of a matrix.*

For the proof of Proposition 3.1 see [11], [20]. The next important fact is that the LU factorization of S_1' is accurate if S_1' can be written as $D_1 Z D_2$, where D_1, D_2 are diagonal matrices with nonincreasing diagonal elements and the LU factorization of Z is accurate. Hence, if S has accurate LU factorization, we can also expect similar accuracy in the LU factorization of S_1'. For the proofs and further details on the accuracy of the LU factorization and its application in SVD computation see [11], [20], [19].

Proposition 3.1 and Theorem 3.1 yield the following theorem.

THEOREM 3.2. *Let* (B, S, C) *be regular matrix triplet, let* δB, δS, *and* δC *be perturbations described in Theorem* 3.1, *and let* $\sigma_1 \ge \cdots \ge \sigma_p$ *and* $\tilde{\sigma}_1 \ge \cdots \ge \tilde{\sigma}_p$ *be the singular values of* (B, S, C) *and* $(B + \delta B, S + \delta S, C + \delta C)$, *respectively. Let* $B_r' = \tilde{\Delta}_B^{-1} B$, $C_r' = \tilde{\Delta}_C^{-1} C$ *and let*

$$\eta_1 = \|B^\dagger \delta B\|_2, \quad \eta_2 = \kappa_2(B_r')\|E_L\|_2 \le \frac{1 + \boldsymbol{\varepsilon}_{\ell_2}(m)}{1 - \boldsymbol{\varepsilon}_{\ell_2}(m)}\kappa_2(B_r)\|E_L\|_2, \tag{31}$$

$$\eta_3 = \|C^\dagger \delta C\|_2, \quad \eta_4 = \kappa_2(C_r')\|E_U\|_2 \le \frac{1 + \boldsymbol{\varepsilon}_{\ell_2}(n)}{1 - \boldsymbol{\varepsilon}_{\ell_2}(n)}\kappa_2(C_r)\|E_U\|_2. \tag{32}$$

If $\max_i |\eta_i| < 1$, *then*

$$\max_{1 \le i \le p} \frac{|\tilde{\sigma}_i - \sigma_i|}{\sigma_i} \le \prod_{i=1}^{4}(1 + \eta_i) - 1. \tag{33}$$

Furthermore, let $\eta_{12} = \eta_1 + \eta_2 + \eta_1\eta_2$, $\eta_{34} = \eta_3 + \eta_4 + \eta_3\eta_4$, $\eta = \max\{\eta_{12}, \eta_{34}\}$, *and* $\overline{\eta} = 2\eta + \eta^2$. *If*

$$\overline{\eta} < \mathrm{gap}(\sigma_i) \equiv \min\left\{\min_{j\neq i} \frac{|\sigma_i - \tilde{\sigma}_j|}{\sigma_i}, 2\right\},$$

then the acute angles between the corresponding perturbed ($\tilde{u}_i$, $\tilde{v}_i$) *and unperturbed* (u_i, v_i) *singular vectors satisfy*

$$\max\{\sin\angle(u_i, \tilde{u}_i), \sin\angle(v_i, \tilde{v}_i)\} \leq \sqrt{2}\left(\frac{1+\overline{\eta}}{1-\overline{\eta}} \cdot \frac{\overline{\eta}}{\mathrm{gap}(\sigma_i) - \overline{\eta}} + \eta\right). \tag{34}$$

Proof. Using the notation from the proof of Theorem 3.1, we can write the perturbed matrix $A + \delta A = (B + \delta B)^\tau (S + \delta S)(C + \delta C)$ as

$$A + \delta A = (\tilde{\Delta}_B^{-1} B + \tilde{\Delta}_B^{-1}\delta B)^\tau \tilde{\Pi}_1^\tau (S_1' + \delta S_1')\tilde{\Pi}_2^\tau (\tilde{\Delta}_C^{-1} C + \tilde{\Delta}_C^{-1}\delta C), \tag{35}$$

where $S_1' + \delta S_1' = \tilde{L}\tilde{U} = (I + E_L)S_1'(I + E_U)$ is as in Proposition 3.1. Using $B_r' = \tilde{\Delta}_B^{-1} B$, $C_r' = \tilde{\Delta}_C^{-1} C$, we easily obtain the relation

$$\begin{aligned} A + \delta A &= (I + B^\dagger \delta B)^\tau (B_r')^\tau (I + \tilde{\Pi}_1^\tau E_L \tilde{\Pi}_1)\tilde{\Delta}_B S \tilde{\Delta}_C (I + \tilde{\Pi}_2 E_U \tilde{\Pi}_2^\tau) C_r' (I + C^\dagger \delta C) \\ &= (I + B^\dagger \delta B)^\tau (I + (B_r')^\dagger \tilde{\Pi}_1^\tau E_L \tilde{\Pi}_1 B_r')^\tau B^\tau S C (I + (C_r')^\dagger \tilde{\Pi}_2 E_U \tilde{\Pi}_2^\tau C_r')(I + C^\dagger \delta C). \end{aligned}$$

Finally, note that for any nonzero $x \in \mathbf{R}^n$ and $y = (I + (C_r')^\dagger \tilde{\Pi}_2 E_U \tilde{\Pi}_2^\tau C_r')(I + C^\dagger \delta C)x$ it holds that

$$\frac{\|Ay\|_2}{\|y\|_2} \prod_{i=1}^{4}(1 - \eta_i) \leq \frac{\|(A + \delta A)x\|_2}{\|x\|_2} \leq \frac{\|Ay\|_2}{\|y\|_2} \prod_{i=1}^{4}(1 + \eta_i),$$

and relation (33) follows from the variational characterization of the singular values (see [25, Lemma 6.4 and Corollary 6.1], [32, section 3.3, Problem 12], [22], [38]). To prove (34), note that $A + \delta A = (I + E)A(I + F)$, where $\|E\|_2 \leq \eta_{12}$, $\|F\|_2 \leq \eta_{34}$, and apply a result from [39]. ∎

3.4. Application to SVD computation of $B^\tau S^{-1} C$. In this section, we show that a modification of Algorithm 3.1 can be used for accurate computation of the singular values of the product

$$A = B^\tau S^{-1} C, \quad B \in \mathbf{R}^{p\times m}, \quad C \in \mathbf{R}^{p\times n}, \quad S \in \mathbf{R}^{p\times p}, \ \det(S) \neq 0. \tag{36}$$

The singular values of the product $B^\tau S^{-1} C$ arise in the restricted singular value computation. The restricted singular values are used in rank estimation in the presence of structured perturbations; see [49]. For a general matrix triplet (S, C, B^τ) with compatible dimensions, the restricted singular values are defined by

$$\zeta_k(S, C, B^\tau) = \min_X\{\|X\|_2 \ : \ \mathrm{rank}(S + CXB^\tau) \leq k - 1\}, \ \ k = 1, 2, \ldots.$$

If B, S, C are as in relation (36), and if $\sigma_1 \geq \cdots \geq \sigma_p$ are the nonzero singular values of $B^\tau S^{-1} C$, then the (finite) restricted singular values of (S, C, B^τ) are $1/\sigma_p \geq \cdots \geq 1/\sigma_1$.

The following algorithm computes the SVD of $B^\tau S^{-1} C$ with accuracy similar to Algorithm 3.1.

ALGORITHM 3.2. PSVD(B, S^{-1}, C).

Input $B \in \mathbf{R}^{p\times m}$, $C \in \mathbf{R}^{p\times n}$, $S \in \mathbf{R}^{p\times p}$, $\mathrm{rank}(S) = p$.

Step 1 Compute $\Delta_B = \mathrm{diag}(\|B^\tau e_i\|_2)$ and $\Delta_C = \mathrm{diag}(\|C^\tau e_i\|_2)$. Then compute $B_r = \Delta_B^{-1}B$, $C_r = \Delta_C^{-1}C$, $S_1 = \Delta_C^{-1}S\Delta_B^{-1}$.

Step 2 Compute the LU factorization with complete pivoting of S_1,

$$\Pi_1 S_1 \Pi_2 = LU, \;\; L_{ii} = 1, \;\; 1 \le i \le p.$$

Step 3 Compute $M = U^{-\tau}\Pi_2 B_r$, $N = L^{-1}\Pi_1^\tau C_r$, and apply Algorithm 2.1 to the product $M^\tau N$.

Output Let Q, Q_F, V, and W be as in Algorithm 2.1. The SVD of $B^\tau S^{-1}C$ is

$$\begin{bmatrix} \Sigma \oplus \mathbf{O} \\ \mathbf{O} \end{bmatrix} = \begin{bmatrix} V^\tau & \\ & I \end{bmatrix} Q_F^\tau (B^\tau S^{-1}C)(Q(W \oplus I_{n-p})).$$

The error analysis of Algorithm 3.2 is similar to the analysis of Algorithm 3.1. The only difference is in the analysis of Step 3. We estimate the errors in the computation of M and N using the perturbation estimate for floating-point inversion of triangular matrices. As in Theorem 3.1, we use $\tilde{}$ to denote the computed quantities. Obviously, there is no major difference in the analysis of the LU factorization. Consider the backward errors in B and C. From [31, Theorem 8.5] it follows that the computed matrix $\tilde{M}$ satisfies $\tilde{U}^\tau\tilde{M} - \tilde{\Pi}_2\tilde{B}_r = \mathcal{E}'_M$, $|\mathcal{E}'_M| \le \varepsilon_T(p)|\tilde{U}|^\tau|\tilde{M}|$, that is, $\tilde{M} = \tilde{U}^{-\tau}\tilde{\Pi}_2(\tilde{B}_r + \tilde{\Pi}_2^\tau\mathcal{E}'_M)$. On the other hand, an easy calculation shows that

$$\left(I - \varepsilon_T(p)\tilde{\Pi}_2^\tau(|\tilde{U}^{-1}|\cdot|\tilde{U}|)^\tau\tilde{\Pi}_2\right)(\tilde{\Pi}_2^\tau|\mathcal{E}'_M|) \le \varepsilon_T(p)\tilde{\Pi}_2^\tau(|\tilde{U}^{-1}|\cdot|\tilde{U}|)^\tau\tilde{\Pi}_2|\tilde{B}_r|. \tag{37}$$

Since $\left(I - \varepsilon_T(p)\tilde{\Pi}_2^\tau(|\tilde{U}^{-1}|\cdot|\tilde{U}|)^\tau\tilde{\Pi}_2\right)$ is an M-matrix, it follows that

$$\tilde{\Pi}_2^\tau|\mathcal{E}'_M| \le \varepsilon_T(p)\left(I - \varepsilon_T(p)\tilde{\Pi}_2^\tau(|\tilde{U}^{-1}|\cdot|\tilde{U}|)^\tau\tilde{\Pi}_2\right)^{-1}\tilde{\Pi}_2^\tau(|\tilde{U}^{-1}|\cdot|\tilde{U}|)^\tau\tilde{\Pi}_2|\tilde{B}_r|. \tag{38}$$

Note that for all i it holds that

$$\|(\tilde{\Pi}_2^\tau|\mathcal{E}'_M|)^\tau e_i\|_2 \le \max_{1\le j\le p}\|\tilde{B}_r^\tau e_j\|_2\varepsilon_T(p)\chi(\tilde{U}), \quad \chi(\tilde{U}) = \frac{\|\,|\tilde{U}^{-1}|\cdot|\tilde{U}|\,\|_1}{1 - \varepsilon_T(p)\|\,|\tilde{U}^{-1}|\cdot|\tilde{U}|\,\|_1}.$$

If $\delta\tilde{M}$ is the backward error from the algorithm PSVD$(\tilde{M}, \tilde{N})$ (Algorithm 2.1), then for all i it holds that $\|(\delta\tilde{M})^\tau e_i\|_2 \le \eta_{\tilde{M}}\|\tilde{M}^\tau e_i\|_2$ (cf. Theorem 2.1) and we can write

$$\tilde{M} + \delta\tilde{M} = \tilde{U}^{-\tau}\tilde{\Pi}_2(\tilde{B}_r + \tilde{\Pi}_2^\tau\mathcal{E}'_M + \tilde{\Pi}_2^\tau\tilde{U}^\tau\delta\tilde{M}).$$

Furthermore, for all i and i' such that $\tilde{\Pi}_2 e_i = e_{i'}$ it holds that

$$\begin{aligned}\|(\tilde{\Pi}_2^\tau\tilde{U}^\tau\delta\tilde{M})^\tau e_i\|_2 &\le \eta_{\tilde{M}}\sum_{k=1}^{p}|\tilde{U}_{ki'}|\|\tilde{M}^\tau e_k\|_2 \\ &\le \eta_{\tilde{M}}(1 + \varepsilon_T(p)\chi(\tilde{U}))\max_{1\le j\le p}\|\tilde{B}_r^\tau e_j\|_2\|(\tilde{U}_{r,1})^{-1}\|_{\ell_1},\end{aligned}$$

where $\|\cdot\|_{\ell_1}$ is the ℓ_1 vector norm of a matrix. Hence, the backward error defined by $\delta\tilde{B}'_r \equiv \tilde{\Pi}_2^\tau\mathcal{E}'_M + \tilde{\Pi}_2^\tau\tilde{U}^\tau\delta\tilde{M}$ satisfies $\max_i\|(\delta\tilde{B}'_r)^\tau e_i\|_2 \le \zeta'_B$, where

$$\zeta'_B = (1+\varepsilon)\frac{\varepsilon_T(p)\chi(\tilde{U}) + \eta_{\tilde{M}}(1 + \varepsilon_T(p)\chi(\tilde{U}))\|(\tilde{U}_{r,1})^{-1}\|_{\ell_1}}{1 - \varepsilon_{\ell_2}(m)}.$$

The value of ζ'_B is comparable with the value of ζ_C in the proof of Theorem 3.1. A similar analysis applies to C. We omit the details for the sake of brevity.

4. The (H,K)-SVD of S. Let $H \in \mathbf{R}^{p\times p}$, $K \in \mathbf{R}^{q\times q}$ be symmetric and positive definite and let $\|y\|_H = \sqrt{y^\tau H y}$, $\|x\|_K = \sqrt{x^\tau K x}$ be the corresponding elliptic norms. Consider the weighted least squares problem

$$\|Sx - b\|_H \longrightarrow \min, \quad \|x\|_K \longrightarrow \min, \tag{39}$$

with the coefficient matrix $S \in \mathbf{R}^{p\times q}$, $p \geq q$. Let $H = R_H^\tau R_H$, $K = R_K^\tau R_K$ be the Cholesky factorizations of H and K and let $W^\tau R_H S R_K^{-1} V = \Sigma$ be the SVD of $A = R_H S R_K^{-1}$. It is known that the solution of the problem (39) is obtained using the solution of the simple problem

$$\|\Sigma z - c\|_2 \longrightarrow \min, \quad \|z\|_2 \longrightarrow \min$$

and the substitution $z = V^\tau R_K x$, $c = W^\tau R_H b$. The matrices $Y = R_H^{-1} W$ and $Z = R_K^{-1} V$ satisfy

$$Y^{-1} S Z = \Sigma, \quad Y^\tau H Y = I_p, \quad Z^\tau K Z = I_q. \tag{40}$$

Relation (40) defines the (H,K)-SVD of S. Note that for the solution of the least squares problem (39) we need Σ, Y^{-1}, and Z. The (H,K)-SVD of S also implicitly solves the eigenvalue problem for the pencil $S^\tau H S - \lambda K$. In that case, only the matrices Σ and Z are of interest.

Ewerbring and Luk [23] describe an algorithm that computes the decomposition (40). In the first phase (reduction to triangular form) this algorithm computes the Cholesky factorizations of H and K, $H = R_H^\tau R_H$, $K = R_K^\tau R_K$, the QR factorization of S, $S = Q_S \left[\begin{smallmatrix} R_S \\ \mathbf{O} \end{smallmatrix}\right]$, and the QR factorization of $R_H Q_S$, $R_H Q_S = Q'_H R'_H$. In the second phase, the algorithm computes the SVD of the product $\hat{R}'_H R_S R_K^{-1}$, where $\hat{R}'_H$ is the leading $q \times q$ submatrix of R'_H. To avoid loss of accuracy, the product $\hat{R}'_H R_S R_K^{-1}$ is not explicitly computed. Instead of that, the algorithm uses plane rotations to generate a sequence of triplets of triangular matrices, $(R_1^{(k)}, R_2^{(k)}, R_3^{(k)})$, $k \geq 0$, $((R_1^{(0)}, R_2^{(0)}, R_3^{(0)}) \equiv (R'_H, R_S, R_K))$. For sufficiently large k, the matrix $C_k = R_1^{(k)} R_2^{(k)} (R_3^{(k)})^{-1}$ is close to diagonal form and its diagonal entries approximate the singular values of $\hat{R}'_H R_S R_K^{-1}$.

4.1. On sensitivity to perturbations of H and K. Before we proceed with the analysis of numerical properties of algorithms for computation of the decomposition (40), let us briefly analyze some necessary conditions for accurate floating-point computation. Demmel [10] shows that the smallest ϵ such that there exists a perturbation δH, $|\delta H| \leq \epsilon |H|$, for which $H + \delta H$ is singular is between $\|H_s^{-1}\|_2^{-1}/p$ and $\|H_s^{-1}\|_2^{-1}$, where $H_s = \Delta_H^{-1} H \Delta_H^{-1}$, $\Delta_H = \mathrm{diag}(\sqrt{H_{ii}})$. Furthermore, it is shown in [10] that, in the case $\|H_s^{-1}\|_2 > 1/\boldsymbol{\varepsilon}$, there exist rounding errors $(|\delta H_{ij}| \leq \boldsymbol{\varepsilon} |H_{ij}|)$ such that $H + \delta H$ is not positive definite, and that if $\|H_s^{-1}\|_2 < 1/(p\boldsymbol{\varepsilon}_C(p))$ the Cholesky factorization is guaranteed to succeed in floating-point arithmetic. Here the factor $\boldsymbol{\varepsilon}_C(p)$ estimates the backward error in the floating-point Cholesky factorization: if the factorization completes without breakdown, the computed triangular factor is the exact factor of $H + \delta H$, where $\max_{i,j}(|\delta H_{ij}|/\sqrt{H_{ii}H_{jj}}) \leq \boldsymbol{\varepsilon}_C(p) \leq (p+5)\boldsymbol{\varepsilon}$. Hence, in this section we make a reasonable assumption that the matrices

$$H_s = \Delta_H^{-1} H \Delta_H^{-1} \ \text{ and } \ K_s = \Delta_K^{-1} K \Delta_K^{-1}, \tag{41}$$

where $\Delta_H = \mathrm{diag}(\sqrt{H_{ii}})$, $\Delta_K = \mathrm{diag}(\sqrt{K_{ii}})$, have inverses bounded in the spectral norm by a modest constant. In that case we say that H and K are well-conditioned. If

$\tilde{R}_H$ and $\tilde{R}_K$ are the floating-point Cholesky factors of H and K, then there exist upper triangular matrices Ω_H and Ω_K such that $\tilde{R}_H = (I+\Omega_H)R_H$, $\tilde{R}_K = (I+\Omega_K)R_K$ and such that $\|\Omega_H\|_2 \leq \sqrt{2}p\boldsymbol{\varepsilon}_C(p)\|H_s^{-1}\|_2$, $\|\Omega_K\|_2 \leq \sqrt{2}q\boldsymbol{\varepsilon}_C(q)\|K_s^{-1}\|_2$. Since $\tilde{R}_H S\tilde{R}_K^{-1} = (I+\Omega_K)R_H S R_K^{-1}(I+\Omega_K)^{-1}$, we see that using computed triangular factors of H and K introduces an $O(\|\Omega_H\|_2 + \|\Omega_K\|_2)$ relative error in the (H,K)-singular values of S. Hence, any algorithm that computes the singular values of $\tilde{R}_H S\tilde{R}_K^{-1}$ with relative accuracy of the order of $f(p,q)\boldsymbol{\varepsilon}(\|H_s^{-1}\|_2 + \|K_s^{-1}\|_2)$ in the practice is almost as good as exact computation, because even an exact computation generally cannot correct the initial uncertainty. (Here $f(p,q)$ is a modestly growing polynomial in p and q.)

4.2. Illustrative examples. To illustrate the accuracy of the floating-point Cholesky factorization, consider the matrix

$$H = \begin{bmatrix} \xi^2 & \xi \\ \xi & 2 \end{bmatrix} = \begin{bmatrix} \xi & 0 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} \xi & 1 \\ 0 & 1 \end{bmatrix}, \quad \xi \text{ arbitrary nonzero scalar.} \tag{42}$$

The Cholesky factorization of H is as accurate as the Cholesky factorization of the matrix $H_s = \Delta_H^{-1} H \Delta_H^{-1}$,

$$H_s = \begin{bmatrix} 1/\xi & 0 \\ 0 & 1/\sqrt{2} \end{bmatrix}\begin{bmatrix} \xi^2 & \xi \\ \xi & 2 \end{bmatrix}\begin{bmatrix} 1/\xi & 0 \\ 0 & 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 1 & 1/\sqrt{2} \\ 1/\sqrt{2} & 1 \end{bmatrix},$$

or the Cholesky factorization of

$$K = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}. \tag{43}$$

The QR factorization of the matrix S is accurate if the columns of S can be scaled so that the resulting matrix is well-conditioned (cf. relation (9)). For instance, if S is

$$S = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 3\zeta \\ -1 & \zeta \end{bmatrix} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 1 & \zeta \\ 0 & 2\zeta \end{bmatrix} \equiv Q_S R_S, \tag{44}$$

then the floating-point QR factorization of S is accurate for any ζ.

Consider now the following experiment: We use the above-described algorithm of Ewerbring and Luk to solve the problem (39) with H as in (42), with S as in (44), and with K as in (43). Let $\zeta = 1$ and let $|\xi| \leq \boldsymbol{\varepsilon}$ so that $\boldsymbol{fl}(1 \pm \xi) = 1$. For the sake of simplicity, we assume that the computation is exact with the exception of the computation of the matrix $R_H Q_S$. The perturbed value of the matrix $R_H Q_S$ is set to be an elementwise rounded exact product, that is,

$$R_H Q_S = \frac{1}{\sqrt{2}}\begin{bmatrix} -1+\xi & 1+\xi \\ -1 & 1 \end{bmatrix} \text{ is replaced with } \widetilde{R_H Q_S} = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}. \tag{45}$$

Obviously, $|R_H Q_S - \widetilde{R_H Q_S}| \leq \boldsymbol{\varepsilon}|R_H Q_S|$. However, $\widetilde{R_H Q_S}$ is exactly singular. If we want to explain this singularity in "backward mode" of the analysis, we must conclude that the backward error in R_H has made R_H (and, hence, H) exactly singular. Thus, in the overall backward error analysis, the actually solved problem is posed using semidefinite norm $\|\cdot\|_{H+\delta H}$.

A closer look at the matrix product $R_H Q_S$ shows that the infinitely ill-conditioned result (singular matrix) is caused by linear combinations between vectors with different lengths. This is precisely the same problem that occurs in the computation of $R_H S$:

$$R_H S = \frac{1}{\sqrt{2}}\begin{bmatrix} \xi - 1 & \zeta(1+3\xi) \\ -1 & \zeta \end{bmatrix} \approx \frac{1}{\sqrt{2}}\begin{bmatrix} -1 & \zeta \\ -1 & \zeta \end{bmatrix}.$$

Remark 4.1. Note that Q_S is an exactly orthogonal matrix. It is a very often overlooked (or ignored) fact that a multiplication by an orthogonal matrix can increase the condition number of a well-conditioned matrix problem.

Also note that the singular values of $A = R_H S R_K^{-1}$ are well determined by the data since

$$R_H S R_K^{-1} = \frac{1}{\sqrt{2}} \begin{bmatrix} -1+\xi & 4\xi \\ -1 & 0 \end{bmatrix}.$$

To avoid the loss of accuracy illustrated in relation (45), we may scale the columns of R_H to equilibrate their ℓ_2 norms. In other words, we write $R_H = (R_H)_c \Delta_H$ and consider the SVD of $(R_H)_c(\Delta_H S)R_K^{-1}$, where we compute the QR factorization $\Delta_H S = Q_S R_S$ and ensure that the multiplication $(R_H)_c Q_S$ is accurate. However, the matrix $\Delta_H S$,

$$\Delta_H S = \frac{1}{\sqrt{2}} \begin{bmatrix} \xi & 3\zeta\xi \\ -\sqrt{2} & \sqrt{2}\zeta \end{bmatrix} = \begin{bmatrix} -\xi/\sqrt{2} & 3\xi/\sqrt{2} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & \zeta \end{bmatrix},$$

has almost linearly dependent columns and the QR factorization is not guaranteed to be accurate. Generally, even if S is well-conditioned for the floating-point QR factorization, the matrix $\Delta_H S$ might be ill-conditioned with respect to the QR factorization. In addition, for a reliable floating-point Householder QR factorization it is necessary to use row pivoting. This is a well-known problem in the least squares computation community. On the other hand, in the factorization

$$\Delta_H S = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \xi/\sqrt{2} \end{bmatrix} \begin{bmatrix} -1 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \zeta \end{bmatrix} \tag{46}$$

we see how two-sided scaling combined with pivoting reveals the structure of a well-conditioned problem (cf. [11]).

4.3. A new algorithm based on the product-induced SVD. We propose the following algorithm for the computation of the (H, K)-SVD of a real matrix S.

ALGORITHM 4.1. (H, K)-SVD(S).

Input $H \in \mathbf{R}^{p\times p}$, $K \in \mathbf{R}^{q\times q}$, $S \in \mathbf{R}^{p\times q}$. H and K positive definite.

Step 1 Compute $\Delta_H = \mathrm{diag}(\sqrt{H_{ii}})$ and $\Delta_K = \mathrm{diag}(\sqrt{K_{ii}})$. Then compute $H_s = \Delta_H^{-1} H \Delta_H^{-1}$, $K_s = \Delta_K^{-1} K \Delta_K^{-1}$, $S_1 = \Delta_H S \Delta_K^{-1}$.

Step 2 Compute the LU factorization with complete pivoting of S_1,

$$\Pi_1 S_1 \Pi_2 = LU, \;\; L_{ii} = 1, \;\; 1 \le i \le \rho = \mathrm{rank}(L) = \mathrm{rank}(U).$$

Also, compute the Cholesky factorizations $H_s = (R_H)_c^\tau (R_H)_c$, $K_s = (R_K)_c^\tau (R_K)_c$.

Step 3 Compute $M^\tau = (R_H)_c \Pi_1^\tau L$, $N = U \Pi_2^\tau (R_K)_c^{-1}$.

Step 4 Apply Algorithm 2.1 to the product $M^\tau N$. (Note that $M^\tau N = R_H S R_K^{-1}$.)

Output Let $W^\tau (M^\tau N) V = \Sigma$ be the computed SVD of $M^\tau N$. Compute $Y = R_H^{-1} W$ and $Z = R_K^{-1} V$. The (H, K)-SVD of S is $Y^{-1} S Z = \Sigma$.

The analysis of Algorithm 4.1 is a straightforward application of the technique developed in the previous section. Let $\tilde{H}_c = H_c + \delta H_c$, $\tilde{K}_c = K_c + \delta K_c$, $\tilde{S}_1 = S_1 + \delta S_1$ be computed in Step 1. Then $|\delta H_c| \le \varepsilon_1 |H_c|$, $|\delta K_c| \le \varepsilon_1 |K_c|$, $|\delta S_1| \le \varepsilon_1 |S_1|$, where $0 \le \varepsilon_1 \le (1+\varepsilon)(1-\varepsilon)^{-3} - 1$. For simplicity, we explicitly set the diagonals of $\tilde{H}_c$ and $\tilde{K}_c$ to 1. Hence, the computed Cholesky factors $\widetilde{(R_H)_c}$ and $\widetilde{(R_K)_c}$ satisfy $\widetilde{(R_H)_c}^\tau \widetilde{(R_H)_c} = \tilde{H}_c + \delta\tilde{H}_c$, $\widetilde{(R_K)_c}^\tau \widetilde{(R_K)_c} = \tilde{K}_c + \delta\tilde{K}_c$, where $\max_{i,j} |\delta\tilde{H}_c|_{ij} \le \varepsilon_C(p)$, $\max_{i,j} |\delta\tilde{K}_c|_{i,j} \le$

$\varepsilon_C(q)$. We conclude that the (exact) SVD of the matrix $\widetilde{(R_H)_c}\tilde{S}_1\widetilde{(R_K)_c}^{-1}$ is equivalent to the (exact) $(H+\delta H, K+\delta K)$-SVD of $S+\delta S$, where

$$|\delta S| \le \varepsilon_1 |S|, \quad \max_{1\le i,j\le p} \frac{|\delta H_{ij}|}{\sqrt{H_{ii}H_{jj}}} \le \varepsilon_C(p) + \varepsilon_1, \quad \max_{1\le i,j\le q} \frac{|\delta K_{ij}|}{\sqrt{K_{ii}K_{jj}}} \le \varepsilon_C(q) + \varepsilon_1.$$

(This conclusion follows from undoing in exact arithmetic the scaling in Step 1. We first note that the SVD of $\widetilde{(R_H)_c}\tilde{S}_1\widetilde{(R_K)_c}^{-1}$ is equivalent to the $(\tilde{H}_c+\delta\tilde{H}_c, \tilde{K}_c+\delta\tilde{K}_c)$-SVD of $\tilde{S}_1$. Then, we undo the scaling.) Furthermore, there exist upper triangular matrices Γ_H and Γ_K such that (cf. [21])

$$\widetilde{(R_H)_c} = (I+\Gamma_H)(R_H)_c, \quad \widetilde{(R_K)_c} = (I+\Gamma_K)(R_K)_c, \tag{47}$$

and such that $\|\Gamma_H\|_2 \le \sqrt{2}p\varepsilon_C(p)\|\tilde{H}_c^{-1}\|_2$, $\|\Gamma_K\|_2 \le \sqrt{2}q\varepsilon_C(q)\|\tilde{K}_c^{-1}\|_2$. If we replace $\tilde{S}_1$ with $\tilde{\Pi}_1^\tau\tilde{L}\tilde{U}\tilde{\Pi}_2^\tau$, then we compensate the uncertainties in $\tilde{L}$ and $\tilde{U}$ by replacing $\tilde{S}_1$ with $\tilde{S}_1+\delta\tilde{S}_1$, where the structure of $\delta\tilde{S}_1$ is explained in section 3.

Consider now the computation of $\tilde{M} = M+\delta M$ and $\tilde{N} = N+\delta N$. Using relation (7), we conclude that there exists an error matrix $\mathcal{E}_M$ such that

$$\tilde{M}^\tau = \widetilde{(R_H)_c}\tilde{\Pi}_1^\tau\tilde{L} + \mathcal{E}_M, \quad |\mathcal{E}_M| \le \varepsilon_{MM}(p)|\widetilde{(R_H)_c}|\tilde{\Pi}_1^\tau|\tilde{L}|, \tag{48}$$

or, equivalently,

$$\tilde{M}^\tau = (I+\Gamma'_H)\widetilde{(R_H)_c}\tilde{\Pi}_1^\tau\tilde{L}, \quad \Gamma'_H = \mathcal{E}_M\tilde{L}^\dagger\tilde{\Pi}_1\widetilde{(R_H)_c}^{-1}. \tag{49}$$

Relation (49) has a two-part interpretation. In the forward mode of the analysis, it gives multiplicative error in the computed value of the product $\widetilde{(R_H)_c}\tilde{\Pi}_1^\tau\tilde{L}$. In the backward mode, it states that $\tilde{M}^\tau$ is the exact product of $(I+\Gamma'_H)\widetilde{(R_H)_c}$ and $\tilde{\Pi}_1^\tau\tilde{L}$. Since

$$\|\Gamma'_H\|_2 \le \varepsilon_{MM}(p)\| \, |\widetilde{(R_H)_c}| \, \|_2 \|\widetilde{(R_H)_c}^{-1}\|_2 \| \, |\tilde{L}| \cdot |\tilde{L}^\dagger| \, \|_2,$$

the columnwise relative uncertainty in $(I+\Gamma'_H)\widetilde{(R_H)_c}$ is comparable with the uncertainty already present in $\widetilde{(R_H)_c} = (I+\Gamma_H)(R_H)_c$. To represent the error in the computation of $\tilde{M}$ in terms of the backward error in H, we first rewrite relation (48) to $\tilde{M}^\tau = (\widetilde{(R_H)_c} + \mathcal{E}_M\tilde{L}^\dagger\tilde{\Pi}_1)\tilde{\Pi}_1^\tau\tilde{L}$ and then we note that

$$\max_{1\le i\le p} \|\mathcal{E}_M\tilde{L}^\dagger\tilde{\Pi}_1 e_i\|_2 \le \eta_H^{(1)} \equiv \varepsilon_{MM}(p)(1+\|\Gamma_H\|_2)\| \, |\tilde{L}| \cdot |\tilde{L}^\dagger| \, \|_1.$$

Hence, the matrix $\widetilde{(R_H)_c} + \mathcal{E}_M\tilde{L}^\dagger\tilde{\Pi}_1$ is an exact factor of $\tilde{H}_c + \delta\tilde{H}_c + \delta\tilde{H}'_c$, where

$$\max_{i,j} |(\delta\tilde{H}'_c)_{ij}| \le 2\eta_H^{(1)}(1+\varepsilon_C(p)) + (\eta_H^{(1)})^2(1+\varepsilon_C(p))^2. \tag{50}$$

(The fact that $\widetilde{(R_H)_c}+\mathcal{E}_M\tilde{L}^\dagger\tilde{\Pi}_1$ is not triangular does not matter here.) The analysis of the computation of $\tilde{N}$ is similar. From [31, Theorem 8.5] it follows that there exists an error matrix $\mathcal{E}_N$ such that $\tilde{N}\widetilde{(R_K)_c} - \tilde{U}\tilde{\Pi}_2^\tau = \mathcal{E}_N$, $|\mathcal{E}_N| \le \varepsilon_T(q)|\tilde{N}| \cdot |\widetilde{(R_K)_c}|$, or, equivalently,

$$\tilde{N} = \tilde{U}\tilde{\Pi}_2^\tau\widetilde{(R_K)_c}^{-1}(I+\Gamma'_K), \quad \Gamma'_K = \widetilde{(R_K)_c}\tilde{\Pi}_2\tilde{U}^{-1}\mathcal{E}_N\widetilde{(R_K)_c}^{-1}. \tag{51}$$

An easy calculation shows that (cf. (37) and (38))

$$(52)\ |\mathcal{E}_N| \leq \boldsymbol{\varepsilon}_T(q)|\tilde{U}|\tilde{\Pi}_2^\tau|\widetilde{(R_K)_c}^{-1}|\cdot|\widetilde{(R_K)_c}|\cdot\left(I-\boldsymbol{\varepsilon}_T(q)|\widetilde{(R_K)_c}^{-1}|\cdot|\widetilde{(R_K)_c}|\right)^{-1}.$$

Hence,

$$(53)\ \|\Gamma'_K\|_2 \leq \boldsymbol{\varepsilon}_T(q)\kappa_2(\widetilde{(R_K)_c})\| \ |\tilde{U}^{-1}|\cdot|\tilde{U}|\ \|_2\frac{\| \ |\widetilde{(R_K)_c}^{-1}|\cdot|\widetilde{(R_K)_c}|\ \|_2}{1-\boldsymbol{\varepsilon}_T(q)\| \ |\widetilde{(R_K)_c}^{-1}|\cdot|\widetilde{(R_K)_c}|\ \|_2}.$$

The bound in relation (53) remains valid if the Cholesky factorization of $\tilde{K}_s$ is computed with pivoting. In that case, the factor $\| \ |\widetilde{(R_K)_c}^{-1}|\cdot|\widetilde{(R_K)_c}|\ \|_2$ is bounded by a function of the dimension q and it can be much smaller (and it is never much larger) than $\kappa_2(\widetilde{(R_K)_c}) \approx \sqrt{\kappa_2(H_c)}$. Hence, the relative perturbations of the singular values in Step 3 are comparable with the perturbations caused by the initial Cholesky factorizations. For the overall error estimate in the first three steps, we need to analyze the accuracy of the LU factorization of $\tilde{S}_1$. We use the representation $\tilde{L} = (I+E_L)L$, $\tilde{U} = U(I+E_U)$, where $\tilde{\Pi}_1 S_1 \tilde{\Pi}_2 = LU$ is the exact LU factorization. (For estimates of E_L and E_U see [45], [20], [11]. For the sake of brevity, we omit the details.) Now, as in the proof of Theorem 3.2, we can write

$$\begin{aligned}\tilde{M}^\tau\tilde{N} &= (I+\Gamma'_H)\widetilde{(R_H)_c}\tilde{\Pi}_1^\tau\tilde{L}\tilde{U}\tilde{\Pi}_2^\tau\widetilde{(R_K)_c}^{-1}(I+\Gamma'_K)\\ &= (I+\Gamma'_H)(I+E'_L)(I+\Gamma_H)R_H S R_K^{-1}(I+\Gamma_K)^{-1}(I+E'_U)(I+\Gamma'_K),\end{aligned}$$

where $E'_L = \widetilde{(R_K)_c}\tilde{\Pi}_1^\tau E_L\tilde{\Pi}_1\widetilde{(R_K)_c}^{-1}$, $E'_U = \widetilde{(R_K)_c}\tilde{\Pi}_2 E_U\tilde{\Pi}_2^\tau\widetilde{(R_K)_c}^{-1}$. Hence, the accuracy of the first three steps is determined by the accuracy of the Cholesky factorizations of $\tilde{H}_s$ and $\tilde{K}_s$ and by the accuracy of the LU factorization with complete pivoting of $\tilde{S}_1$. (Note that $\tilde{S}_1$ and S have the same zero pattern and the same sign distribution. Hence, the LU of $\tilde{S}_1$ inherits the properties of the LU of S which are based on zero and sign structures of S; cf. [11].)

The accuracy of the algorithm PSVD$(\tilde{M},\tilde{N})$ (Step 4 of Algorithm 4.1) is determined by the condition numbers $\|\tilde{M}_r^\dagger\|_2$ and $\|\tilde{N}_r^\dagger\|_2$, where $\tilde{M}_r$ and $\tilde{N}_r$ are obtained from $\tilde{M}$ and $\tilde{N}$, respectively, by scaling their rows to have unit Euclidean norms. Simple estimates for $\|\tilde{M}_r^\dagger\|_2$ and $\|\tilde{N}_r^\dagger\|_2$ are derived as follows. Let $\tilde{U} = \mathrm{diag}(\tilde{U}_{ii})\tilde{U}_d$. Then

$$\begin{aligned}&\|\tilde{N}_r^\dagger\|_2 \leq \leq \sqrt{\tilde{\rho}}\kappa_2(\tilde{U}_d\tilde{\Pi}_2^\tau\widetilde{(R_K)_c}^{-1}(I+\Gamma'_K)) \leq \sqrt{\tilde{\rho}}\kappa_2(\tilde{U}_d)\kappa_2(\widetilde{(R_K)_c})\frac{1+\|\Gamma'_K\|_2}{1-\|\Gamma'_K\|_2},\\ &\|\tilde{M}_r^\dagger\|_2 \leq \sqrt{\tilde{\rho}}\kappa_2(\tilde{L})\kappa_2(\widetilde{(R_H)_c})\frac{1+\|\Gamma'_H\|_2}{1-\|\Gamma'_H\|_2}.\end{aligned}$$

This implies that Algorithm 2.1 computes the SVD of $\tilde{M}^\tau\tilde{N}$ with the relative error that is comparable with the uncertainty on input to PSVD$(\tilde{M},\tilde{N})$.

To analyze the overall backward error, one can use the technique from the proof of Theorem 3.1. We omit the details for the sake of brevity. The results of numerical experiments with Algorithm 4.1 are given in section 5.

4.4. Application to canonical correlation analysis and (H^{-1},K)-SVD. In the canonical correlation analysis, we are given two sets of random variables, $x =$

$(x_1, \ldots, x_p)^\tau$, $y = (y_1, \ldots, y_q)^\tau$, $p \geq q$, with joint distribution and the covariance matrix

$$C = \begin{bmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{bmatrix} = \mathbf{E}\left(\begin{bmatrix} x \\ y \end{bmatrix} [x^\tau, y^\tau] \right) \equiv \begin{bmatrix} H & S \\ S^\tau & K \end{bmatrix}. \tag{54}$$

The goal is to find certain mutually uncorrelated linear combinations of the first set of variables that maximize correlations with certain mutually uncorrelated linear combinations of the second set of variables. Application of canonical correlations include, for example, well log analysis where the two sets of interest might be (cf. [9]) {*gamma ray intensity, sonic transmissivity, electrical resistance of the rock*} and {*permeability, porosity, oil saturation, water saturation*}.

The canonical correlation problem is related to the SVD of $C_{xx}^{-1/2} C_{xy} C_{yy}^{-1/2} \equiv H^{-1/2} S K^{-1/2}$. It is shown in [48] that the canonical correlation problem can also be solved as a principal relations problem with the weighting matrix $M = C_{xx}^{-1} \oplus C_{yy}^{-1}$. Here we work in the framework of the (H^{-1}, K)-SVD of S, where H, K, and S are as in (54). The (H^{-1}, K)-SVD of S implicitly solves the eigenvalue problem for the pencil $S^\tau H^{-1} S - \lambda K$ and the weighted least squares problem (39) with $\|\cdot\|_H$ replaced with $\|\cdot\|_{H^{-1}}$. These problems arise in such applications as aircraft wing flutter analysis, system identification, and optimal prediction (cf. [34], [23]).

We propose the following algorithm for accurate computation of the (H^{-1}, K)-SVD of S.

ALGORITHM 4.2. CC$(x, y, C) \equiv (H^{-1}, K)$-SVD$(S)$ $(C = \left[\begin{smallmatrix} H & S \\ S^\tau & K \end{smallmatrix}\right])$.

Input $H \in \mathbf{R}^{p \times p}$, $K \in \mathbf{R}^{q \times q}$, $S \in \mathbf{R}^{p \times q}$. H and K positive definite.

Step 1 Compute $\Delta_H = \mathrm{diag}(\sqrt{H_{ii}})$ and $\Delta_K = \mathrm{diag}(\sqrt{K_{ii}})$. Then compute $H_s = \Delta_H^{-1} H \Delta_H^{-1}$, $K_s = \Delta_K^{-1} K \Delta_K^{-1}$, $S_1 = \Delta_H^{-1} S \Delta_K^{-1}$.

Step 2 Compute the LU factorization with complete pivoting of S_1,

$$\Pi_1 S_1 \Pi_2 = LU, \quad L_{ii} = 1, \quad 1 \leq i \leq \rho = \mathrm{rank}(L) = \mathrm{rank}(U).$$

Also, compute the Cholesky factorizations (with pivoting) $\Pi_3^\tau H_s \Pi_3 = (R_H)_c^\tau (R_H)_c$,
and
$\Pi_4^\tau K_s \Pi_4 = (R_K)_c^\tau (R_K)_c$.

Step 3 Compute $M^\tau = (R_H)_c^{-\tau} (\Pi_1 \Pi_3)^\tau L$, $N = U(\Pi_2^\tau \Pi_4)(R_K)_c^{-1}$.

Step 4 Apply Algorithm 2.1 to the product $M^\tau N$. (Note that $M^\tau N = R_H^{-\tau} S R_K^{-1}$.)

Output Let $W^\tau (M^\tau N) V = \Sigma$ be the computed SVD of $M^\tau N$. Compute $Y = R_H^\tau W$ and $Z = R_K^{-1} V$. The (H^{-1}, K)–SVD of S is $Y^{-1} S Z = \Sigma$.

Note that Step 1 implicitly replaces the covariance matrix C with the scaled covariance matrix $C_1 = \Delta_C^{-1} C \Delta_C^{-1}$, $\Delta_C = \mathrm{diag}(\sqrt{C_{ii}})$. In Step 4, C_1 is implicitly replaced with $\left[\begin{smallmatrix} I & M^\tau N \\ N^\tau M & I \end{smallmatrix}\right]$ and the problem is reduced to the SVD of $M^\tau N$. If the input data are given as observation data matrices X and Y (instead of the covariance matrix C), then the canonical correlations (the cosines of the principal angles between the spans of X and Y) are best computed using an algorithm of Björck and Golub [3], [19]. In the Björck–Golub algorithm, the problem is reduced to the SVD of $Q_x^\tau Q_y$, where $X = Q_x R_x$ and $Y = Q_y R_y$ are the QR factorizations. This rather elegant reduction to the ordinary SVD is possible because of the special choice of the weighting matrices. On the other hand, in Algorithm 4.1, as well as in the algorithm of Ewerbring and Luk [23], the covariance matrices C_{xx} and C_{yy} may be replaced with general positive definite weighting matrices.

Remark 4.2. If X and Y are normally scaled, that is, if $X^\tau Y = I_p = I_q$, then the canonical correlation problem reduces to the SVD computation of YX^τ. Indeed, if $X = Q_x R_x$ and $Y = Q_y R_y$ are the QR factorizations of X and Y, then R_x and R_y are nonsingular and the covariance matrix

$$\begin{bmatrix} X^\tau X & I_p \\ I_p & Y^\tau Y \end{bmatrix} \text{ can be replaced with } \begin{bmatrix} I & (R_y R_x^\tau)^{-1} \\ (R_x R_y^\tau)^{-1} & I \end{bmatrix}.$$

(This corresponds to the change of bases in span(X) and span(Y).) Hence, the cosines of the canonical angles between span(X) and span(Y) are the inverses of the singular values of $R_y R_x^\tau$. The latter are the nonzero singular values of YX^τ. (Golub and Zha [28] derived the same result using the GSVD of (X, Y).)

It can be easily shown that the accuracy of Algorithm 4.2 is nearly the same as that of Algorithm 4.1. We omit the details for the sake of brevity.

4.5. An algorithm based on the GSVD. In this section we show that computation of the (H, K)-SVD of a full column rank matrix S can be accurately reduced to computation of the GSVD of certain regular matrix pair. The GSVD-based algorithm uses a tangent GSVD-algorithm from [16].

ALGORITHM 4.3. (H, K)-SVD(S).

Input $H \in \mathbf{R}^{p\times p}$, $K \in \mathbf{R}^{q\times q}$, $S \in \mathbf{R}^{p\times q}$. H and K positive definite.

Step 1 Compute $\Delta_H = \mathrm{diag}(\sqrt{H_{ii}})$ and $\Delta_K = \mathrm{diag}(\sqrt{K_{ii}})$. Then compute $H_s = \Delta_H^{-1} H \Delta_H^{-1}$, $K_s = \Delta_K^{-1} K \Delta_K^{-1}$, $S_1 = \Delta_H S \Delta_K^{-1}$.

Step 2 Compute the LU factorization with complete pivoting of S_1,

$$\Pi_1 S_1 \Pi_2 = LU, \;\; L_{ii} = 1, \;\; 1 \le i \le q = \mathrm{rank}(L) = \mathrm{rank}(U).$$

Also, compute the Cholesky factorizations $H_s = (R_H)_c^\tau (R_H)_c$, $K_s = (R_K)_c^\tau (R_K)_c$.

Step 3 Compute $M^\tau = (R_H)_c \Pi_1^\tau L$, $N = (R_K)_c \Pi_2 U^{-1}$.

Step 4 Apply GSVD algorithm to the pair (M^τ, N). (Note that GSVD(M^τ, N) means the SVD of $M^\tau N^{-1} = R_H S R_K^{-1}$.)

Output Let $W^\tau (M^\tau N^{-1}) V = \Sigma$ be the computed SVD of $M^\tau N^{-1}$ (via the GSVD of (M^τ, N)). Compute $Y = R_H^{-1} W$ and $Z = R_K^{-1} V$. The (H, K)-SVD of S is $Y^{-1} S Z = \Sigma$.

Implementation and error analysis of Algorithm 4.3 are similar to those in Algorithm 4.1. Here we compute N using BLAS 3 procedure `STRSM()` instead of `STRMM()`, and the computation completes with GSVD(M^τ, N) instead of PSVD(M, N). For an analysis of GSVD(M^τ, N) see [16]. We omit the details for the sake of brevity.

5. Numerical examples. In this section, we present the software that implements algorithms from this paper. Since all operations in our algorithms are standard matrix operations, most of them can be efficiently implemented using BLAS and LAPACK libraries. The only two exceptions are the LU factorization with complete pivoting and the Jacobi SVD algorithm. Our procedure for the LU factorization with complete pivoting is a modification of the LAPACK procedure `SGETF2()`. The Jacobi SVD algorithm is implemented following the analyses from [13], [18].

5.1. Test matrix generation. The main characteristic of the algorithms presented in this paper is that their accuracy depends on the condition numbers of equilibrated matrices. For instance, the accuracy of the algorithm PSVD(B, C) (Algorithm 2.1) depends on $\kappa_2(B_r)$ and $\kappa_2(C_r)$, where $B = \mathrm{diag}(\|B^\tau e_i\|_2) B_r$, $C =$

$\mathrm{diag}(\|C^\tau e_i\|_2)C_r$. In order to show that our estimates are sharp and that they are almost attainable in the practice, we need to generate test examples with prescribed values of all relevant condition numbers. Thus, to test PSVD(B,C), we need pairs (B,C) with specified values of $\kappa_2(B_r)$ and $\kappa_2(C_r)$.

To generate test matrices, we use an algorithm $X = \mathrm{RANDOM}(p,m,\omega_1,\mu_1,\omega_2,\mu_2)$ which generates a random full row rank matrix $X = \Delta_X X_r$, where $\Delta_X = \mathrm{diag}(\|X^\tau e_i\|_2)$, $\kappa_2(X_r) = \omega_1$, $\kappa_2(\Delta_X) = \omega_2$, and with distributions of the singular values of X_r and Δ_X determined by the parameters μ_1 and μ_2, respectively. The values of μ_1 and μ_2 are from the set of admissible values of the parameter `MODE` in the procedure `DLATM1()` from [12]. The algorithm RANDOM() is based on [27, Problems P.8.5.3 and P.8.5.4], and it has been frequently used in recent years in connection with the design and analysis of accurate Jacobi-type methods [13], [44], [42], [17].

It is difficult to generate an S with prescribed accuracy of the LU factorization with complete pivoting of S. Let $P_1 S P_2 = D_1 Z D_2 \in \mathbf{R}^{p\times q}$, $p \geq q$, and let $Z = L_Z U_Z$, where L_Z is unit lower trapezoidal and U_Z is $q \times q$ upper triangular and nonsingular. Then

$$\|Z^\dagger\|_2 \leq \|L_Z^\dagger\|_2 \|U_Z^\dagger\|_2 \leq \|L_Z\|_2 \|U_Z\|_2 \|Z^\dagger\|_2^2. \tag{55}$$

On the other hand, let $S = \Delta_1 S_0 \Delta_2$, where S_0 has equilibrated rows and columns and Δ_1, Δ_2 are diagonal scalings. Then $P_1 S P_2 = D_1 Z D_2 = (P_1 \Delta_1 P_1^\tau) P_1 S_0 P_2 (P_2^\tau \Delta_2 P_2)$, and we hope that an estimate similar to (55) holds with Z replaced with S_0. Therefore, in our tests we use S of the form $S = \Delta_1 S_0 \Delta_2$. In this way, we nearly control the impact of the matrix S to the accuracy of the SVD of $B^\tau S C$ and $B^\tau S^{-1} C$ (in the case of square and nonsingular S). Since $B^\tau S C = B_r^\tau \Delta_B \Delta_1 S_0 \Delta_2 \Delta_C C_r^\tau$, we simplify our generator of test examples by setting $\Delta_B := \Delta_B \Delta_1$, $\Delta_C := \Delta_2 \Delta_C$, $B^\tau S C = (\Delta_B B_r)^\tau S_0 (\Delta_C C_r)$, and by controlling the values of $\kappa_2(B_r)$, $\kappa_2(\Delta_B)$, $\kappa_2(S_0)$, $\kappa_2(C_r)$, $\kappa_2(\Delta_C)$.

5.2. Error measures. We measure the forward error in the computed singular values. As usual, we test the single precision procedure and we use the double precision procedure as a reference. However, to ensure that the double precision approximations are good enough to be used as reference values for the single precision procedure, we use the following strategy: First, each test matrix triplet (B,S,C) is generated in double precision arithmetic and with the value of $\mathrm{cond}(B,S,C) \equiv \max\{\kappa_2(B_r), \kappa_2(C_r), \kappa_2(S_0)\}$ below 10^7. This ensures that the double precision procedure can compute the singular values with nearly seven or eight correct digits. Then, we test the *consistency* of the double precision procedure. The consistency is in the framework of the theory of section 3 defined as follows: If we use the same procedure to compute the singular values of $B^\tau S C$ and $C^\tau S^\tau B$ in floating point arithmetic with the round-off unit $\boldsymbol{\varepsilon}$, then the computed approximations match in roughly $-\log_{10}(\boldsymbol{\varepsilon} \cdot \mathrm{cond}(B,S,C))$ decimal places. In all tests, the double precision procedure passed the consistency test. Therefore, if $\sigma_1' \geq \cdots \geq \sigma_\rho'$ and $\tilde\sigma_1 \geq \cdots \geq \tilde\sigma_\rho$ are the double and the single precision approximations of the singular values of $B^\tau S C$, we measure the accuracy of the single precision procedure by

$$\epsilon(B,S,C) = \max_{1\leq i\leq \rho} \frac{|\tilde\sigma_i - \sigma_i'|}{\sigma_i'}. \quad (\rho = \min\{p,q\}). \tag{56}$$

(The input to the single precision procedure is (B,S,C) rounded to single precision.) Theoretical prediction is that the value of $\mathbf{e}(B,S,C) = \frac{\epsilon(B,S,C)}{\mathrm{cond}(B,S,C)}$ is, up to a factor

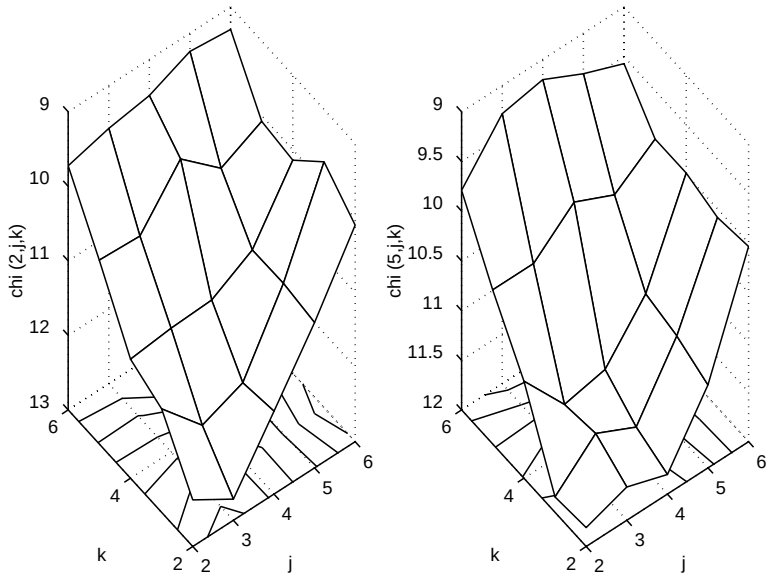


FIG. 3. *The values of $\chi(2,j,k)$ and $\chi(5,j,k)$, $2 \le j,k \le 6$. The double precision procedure has passed the consistency test and it can be used with confidence as reference for the single precision procedure.*

of dimensionality, of the order of the machine precision $\boldsymbol{\varepsilon}$. For the purpose of the test, $\mathrm{cond}(B,S,C)$ is computed in double precision arithmetic, using the LAPACK procedure `DGESVD()`.

5.3. Results. We now present the numerical results. All experiments in Example 5.1 were done in Sun FORTRAN on a Sun 4 workstation. The experiments in Example 5.2 were done on a DEC Alpha workstation. The figures are produced using MATLAB.

Example 5.1. In this example, we set the dimensions to be $m = 80$, $p = 50$, $q = 40$, and $n = 100$. The triplet $(\kappa_2(B_r), \kappa_2(S_0), \kappa_2(C_r))$ of condition numbers takes all values from the set $\{10^2, 10^3, 10^4, 10^5, 10^6\}^3$, while the condition numbers $(\kappa_2(\Delta_B), \kappa_2(\Delta_C))$ of diagonal scalings are chosen from the set $\{10^8, 10^{12}, 10^{16}\} \times \{10^9, 10^{13}, 10^{15}\}$. The modes of distribution of the singular values of B_r, Δ_B, C_r, Δ_C, S_0, respectively, are chosen from the set $\{5,3\}\times\{5,4\}\times\{5,-4\}\times\{5,3\}^2$. Combining all values of the above described parameters we obtain 36,000 test triplets, divided into 125 classes, where each class $\mathcal{C}_{ijk}$ has fixed $(\kappa_2(B_r), \kappa_2(S_0), \kappa_2(C_r)) \approx (10^i, 10^j, 10^k)$, $2 \le i,j,k \le 6$.

We first show the results of the consistency test for the double precision procedure. In Figure 3, we display some of the values of

$$\chi(i,j,k) = \log_{10}\left(\max_{(B,S,C)\in\mathcal{C}_{ijk}} \max_{1\le l\le\rho} \frac{|\sigma_l'(B,S,C) - \sigma_l'(C,S^\tau,B)|}{\sigma_l'(B,S,C)} \right), \tag{57}$$

where $(\sigma_l'(B,S,C))_{l=1}^{\rho}$ and $(\sigma_l'(C,S^\tau,B))_{l=1}^{\rho}$ are the ordered double precision approximations of the singular values of $B^\tau SC$ and $C^\tau S^\tau B$, respectively. In Figure 4, we display the values of $\mathbf{e}(B,S,C)$.

Next, we plot the relative error $\epsilon(B,S,C)$ versus $(\kappa_2(B_r), \kappa_2(S_0), \kappa_2(C_r))$. More precisely, we plot the values of

$$\epsilon(i,j,k) = \max_{(B,S,C)\in\mathcal{C}_{ijk}} \epsilon(B,S,C). \tag{58}$$

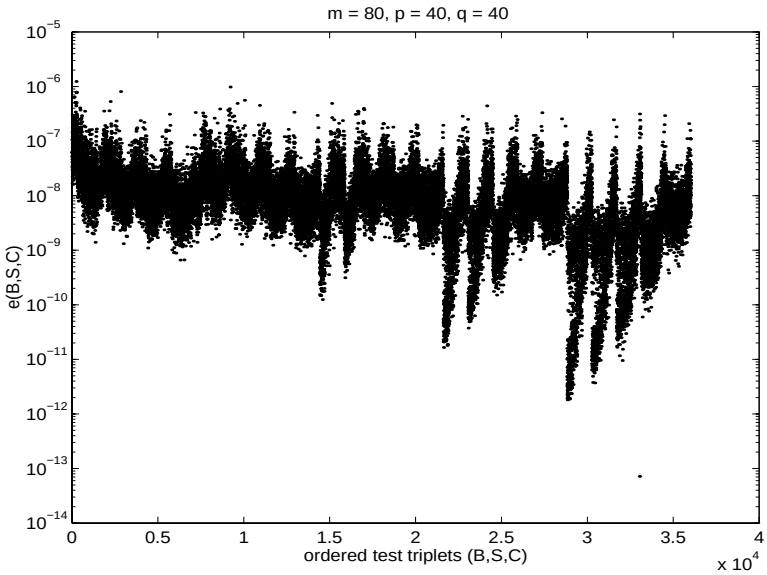


FIG. 4. *The values of* $\mathbf{e}(B,S,C)$ *for all test triplets. (Test triplets are generated in a sequence of nested loops, and the order of generation defines their ordering in the figure.)*

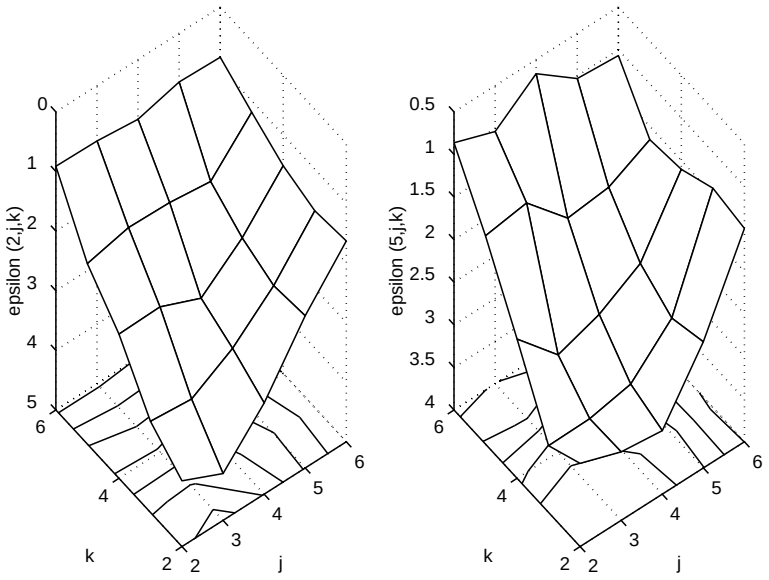


FIG. 5. *The values of* $\epsilon(2,j,k)$ *and* $\epsilon(5,j,k)$, $2 \le j,k \le 6$. *Note that the measured relative errors behave as predicted by the theory.*

Note the similarities between the behaviors of the errors shown in Figure 3 and Figure 5.

Example 5.2. In this example, we test Algorithm 4.1 and Algorithm 4.3. We set the dimensions to $m = p = 200$, $n = q = 100$. We follow the same strategy as in Example 5.1: We generate 9,000 test problems, divided into 125 classes, where each class has fixed $(\kappa_2(H_c), \kappa_2(S_0), \kappa_2(K_c)) \approx (10^i, 10^j, 10^k) \in \{10^2, 10^3, 10^4, 10^5, 10^6\}^3$.

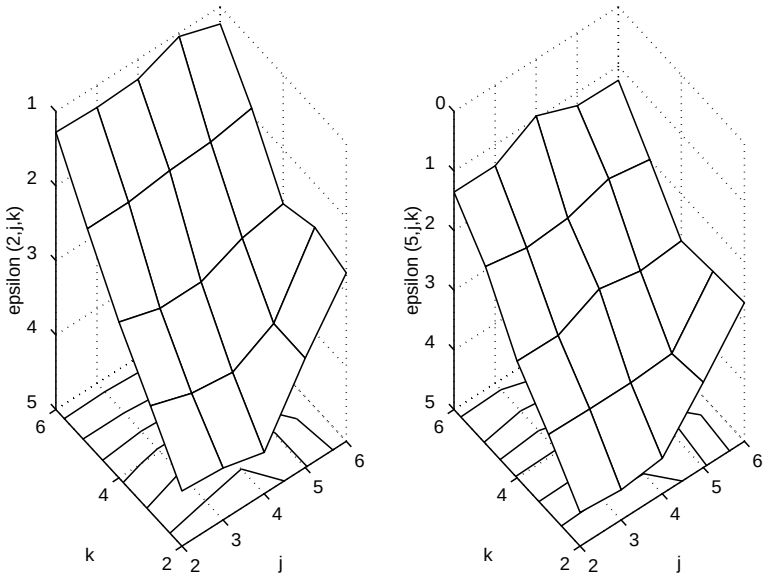


FIG. 6. *The values of* $\epsilon(2,j,k)$ *and* $\epsilon(5,j,k)$, $2 \le j,k \le 6$, *in an* (H,K)*-SVD algorithm based on the PSVD*$(\cdot,\cdot)$ *computation.*

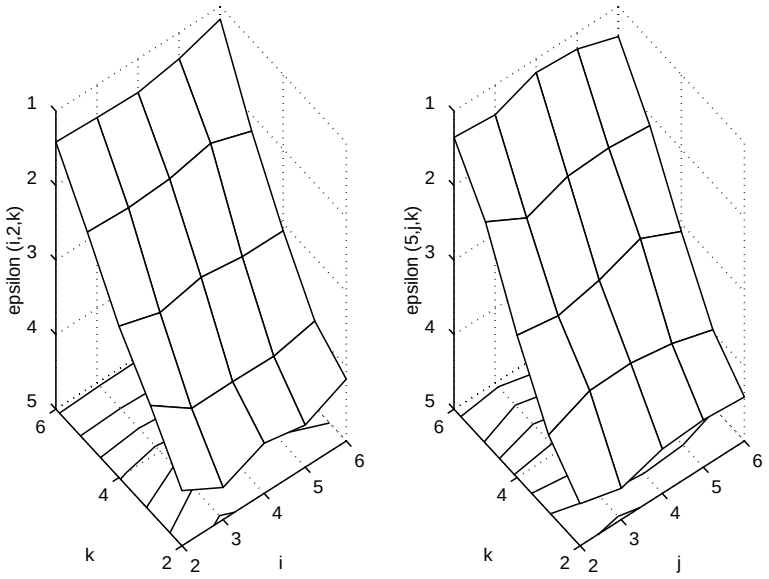


FIG. 7. *The values of* $\epsilon(i,2,k)$ *and* $\epsilon(5,j,k)$, $2 \le j,k \le 6$, *in an* (H,K)*-SVD algorithm based on the GSVD*$(\cdot,\cdot)$ *computation. Note that the measured accuracy is slightly better than predicted by the theory.*

For fixed class, we choose diagonal scalings Δ_H, Δ_K with $(\kappa_2(\Delta_H), \kappa_2(\Delta_K))$ chosen from $\{10^8, 10^{12}, 10^{16}\} \times \{10^9, 10^{13}, 10^{15}\}$. The modes of distributions (cf. Example 5.1) are chosen from $\{5,3\} \times \{5\} \times \{5,-4\} \times \{5\} \times \{5,3\}$. We first compare the results of double precision computations using Algorithm 4.1 and Algorithm 4.3. Since in all test examples the double precision eigenvalues computed by the two algorithms agree

to more that 10 decimal places, we use double precision Algorithm 4.1 as a reference for single precision computation. The measured error bounds, shown in Figures 6 and 7, are slightly better than the theoretical predictions from section 4. This is probably related to the fact that the main part of the error is due to the Cholesky factorization, which is usually more accurate than the theory predicts (cf. [6]).

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen, *LAPACK Users' Guide,* 2nd ed., SIAM, Philadelphia, PA, 1995.

[2] J. Barlow, *Stability analysis of the G-algorithm and a note on its application to sparse least squares problems*, BIT, 25 (1985), pp. 507–520.

[3] A. Björck and G. H. Golub, *Numerical methods for computing angles between linear subspaces*, Math. Comp., 27 (1973), pp. 579–594.

[4] L. S. Blackford, J. Choi, A. Cleary, E. D'Azevedo, J. Demmel, I. Dhillon, J. Dongarra, S. Hammarling, G. Henry, A. Petitet, K. Stanley, D. Walker, and R. C. Whaley, ScaLAPACK Users' Guide, SIAM, Philadelphia, PA, 1997.

[5] A. W. Bojanczyk, L. M. Ewerbring, F. T. Luk, and P. van Dooren, *An accurate product SVD algorithm*, Signal Process., 25 (1991), pp. 189–201.

[6] X.-W. Chang, C. C. Paige, and G. W. Stewart, *New perturbation analyses for the Cholesky factorization*, IMA J. Numer. Anal., 16 (1996), pp. 457–484.

[7] J. Choi, J. Dongarra, S. Ostrouchov, A. Petitet, D. Walker, and R. C. Whaley, *A Proposal for a Set of Parallel Basic Linear Algebra Subprograms*, Technical Report CS-95-292, Department of Computer Science, University of Tennessee, Knoxville.

[8] A. J. Cox and N. J. Higham, *Stability of Householder QR factorization for weighted least squares problems*, in Numerical Analysis 1997, D. F. Griffiths, D. J. Higham, and G. A. Watson, eds., Pitman Res. Notes Math. Ser. 380, Longman Scientific and Technical, Harlow, UK, 1998, pp. 57–73.

[9] J. C. Davis, *Statistics and Data Analysis in Geology*, 2nd ed., John Wiley, New York, 1986.

[10] J. Demmel, *On Floating Point Errors in Cholesky*, LAPACK Working Note 14, Computer Science Department, University of Tennessee, Knoxville, 1989.

[11] J. Demmel, M. Gu, S. Eisenstat, I. Slapničar, K. Veselić, and Z. Drmač, *Computing the singular value decomposition with high relative accuracy*, Linear Algebra Appl., 299 (1999), pp. 21–80.

[12] J. Demmel and A. McKenney, *A Test Matrix Generation Suite*, LAPACK Working Note 9, Courant Institute, New York, 1989.

[13] J. Demmel and K. Veselić, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1204–1245.

[14] J. J. Dongarra, J. J. Du Croz, I. Duff, and S. Hammarling, *A set of level* 3 *basic linear algebra subprograms*, ACM Trans. Math. Software, 16 (1990), pp. 1–17.

[15] Z. Drmač, *Accurate computation of the product-induced singular value decomposition with applications*, SIAM J. Numer. Anal., 35 (1998), pp. 1969–1994.

[16] Z. Drmač, *A tangent algorithm for computing the generalized singular value decomposition*, SIAM J. Numer. Anal., 35 (1998), pp. 1804–1832.

[17] Z. Drmač, *Computing the Singular and the Generalized Singular Values*, Ph.D. thesis, Lehrgebiet Mathematische Physik, Fernuniversität Hagen, Hagen, Germany, 1994.

[18] Z. Drmač, *Implementation of Jacobi rotations for accurate singular value computation in floating point arithmetic*, SIAM J. Sci. Comput., 18 (1997), pp. 1200–1222.

[19] Z. Drmač, *On principal angles between subspaces of Euclidean space*, SIAM J. Matrix Anal. Appl., to appear.

[20] Z. Drmač and E. R. Jessup, *On accurate generalized singular value computation in floating-point arithmetic*, SIAM J. Matrix Anal. Appl., to appear.

[21] Z. Drmač, M. Omladič, and K. Veselić, *On the perturbation of the Cholesky factorization*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1319–1332.

[22] S. C. Eisenstat and I. C. F. Ipsen, *Relative perturbation techniques for singular value problems*, SIAM J. Num. Anal., 32 (1995), pp. 1972–1988.

[23] L. M. Ewerbring and F. T. Luk, *Canonical correlations and generalized SVD: Applications and new algorithms*, J. Comput. Appl. Math., 27 (1989), pp. 37–52.

[24] K. V. Fernando and S. Hammarling, *A product induced singular value decomposition*

(ΠSVD) for two matrices and balanced realization, in Linear Algebra in Signals, Systems, and Control, SIAM, Philadelphia, 1988, pp. 128–140.

[25] S. K. Godunov, A. G. Antonov, O. P. Kirilyuk, and V. I. Kostin, *Garantirovannaya tochnost resheniya sistem lineĭnykh uravneniĭ v evklidovykh prostranstvakh*, Novosibirsk Nauka, Sibirskoe Otdelenie, 1988.

[26] G. H. Golub, *Numerical methods for solving linear least squares problems*, Numer. Math., 7 (1965), pp. 206–216.

[27] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.

[28] G. H. Golub and H. Zha, *Perturbation analysis of the canonical correlations of matrix pairs*, Linear Algebra Appl., 210 (1994), pp. 3–28.

[29] M. Gu and S. C. Eisenstat, *A divide-and-conquer algorithm for the bidiagonal SVD*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 79–92.

[30] M. T. Heath, A. J. Laub, C. C. Paige, and R. C. Ward, *Computing the singular value decomposition of a product of two matrices*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1147–1159.

[31] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.

[32] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[33] E. G. Kogbetliantz, *Solution of linear equations by diagonalization of coefficient matrix*, Quart. Appl. Math., 13 (1955), pp. 123–132.

[34] W. E. Larimore and F. T. Luk, *System identification and control using SVDs on systolic arrays*, in High Speed Computing, Proc. SPIE 880, SPIE, Bellingham, WA, 1988, pp. 37–48.

[35] A. J. Laub, M. T. Heath, C. C. Paige, and R. C. Ward, *Computation of system balancing transformations and other applications of simultaneous diagonalization algorithms*, IEEE Trans. Automat. Control, AC–32 (1987), pp. 115–122.

[36] C. L. Lawson, R. J. Hanson, D. R. Kincaid, and F. T. Krogh, *Algorithm* 539. *Basic linear algebra subprograms for Fortran usage*, ACM Trans. Math. Software, 5 (1979), pp. 324–325.

[37] C. L. Lawson, R. J. Hanson, D. R. Kincaid, and F. T. Krogh, *Basic linear algebra subprograms for Fortran usage*, ACM Trans. Math. Software, 5 (1979), pp. 308–323.

[38] R.-C. Li, *Relative Perturbation Theory:* (I) *Eigenvalue and Singular Value Variations*, Technical Report, Mathematical Science Section, Oak Ridge National Laboratory, Oak Ridge, TN, 1996.

[39] R.-C. Li, *Relative Perturbation Theory:* (II) *Eigenspace and Singular Subspace Variations*, Technical Report, Mathematical Science Section, Oak Ridge National Laboratory, Oak Ridge, TN, 1996.

[40] C. C. Paige, *Computing the generalized singular value decomposition*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1126–1146.

[41] C. C. Paige and M. A. Saunders, *Towards a generalized singular value decomposition*, SIAM J. Numer. Anal., 18 (1981), pp. 398–405.

[42] E. Pietzsch, *Genaue Eigenwertberechnung nichtsingulärer schiefsymmetrischer Matrizen mit einem Jacobi-änlichen Verfahren*, Ph.D. thesis, Lehrgebiet Mathematische Physik, Fernuniversität Hagen, Hagen, Germany, 1993.

[43] M. J. D. Powell and J. K. Reid, *On applying Householder transformations to linear least squares problems*, in Information Processing 68, North–Holland, Amsterdam, 1969, pp. 122–126.

[44] I. Slapničar, *Accurate Symmetric Eigenreduction by a Jacobi Method*, Ph.D. thesis, Lehrgebiet Mathematische Physik, Fernuniversität Hagen, Hagen, Germany, 1992.

[45] J.-G. Sun, *Componentwise perturbation bounds for some matrix decompositions*, BIT, 32 (1992), pp. 702–714.

[46] C. F. Van Loan, *Generalized Singular Values with Algorithms and Applications*, Ph.D. thesis, University of Michigan, 1973.

[47] C. F. Van Loan, *Generalizing the singular value decomposition*, SIAM J. Numer. Anal., 13 (1976), pp. 76–83.

[48] A. M. Wesselman, *The population-sample decomposition method,* in International Studies in Economics and Econometrics, Vol. 19, Kluwer Academic Publishers, Dordrecht, Netherlands, 1987.

[49] H. Y. Zha, *The restricted singular value decomposition of matrix triplets*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 172–194.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 4, pp. 1051–1078
© 2000 Society for Industrial and Applied Mathematics

LOOK-AHEAD PROCEDURES FOR LANCZOS-TYPE PRODUCT METHODS BASED ON THREE-TERM LANCZOS RECURRENCES*

MARTIN H. GUTKNECHT† AND KLAUS J. RESSEL‡

Abstract. Lanczos-type product methods for the solution of large sparse non-Hermitian linear systems either square the Lanczos process or combine it with a local minimization of the residual. They inherit from the underlying Lanczos process the danger of breakdown. For various Lanczos-type product methods that are based on the Lanczos three-term recurrence, look-ahead versions are presented which avoid such breakdowns or near-breakdowns at the cost of a small computational overhead. Different look-ahead strategies are discussed and their efficiency is demonstrated by several numerical examples.

Key words. Lanczos-type product methods, look-ahead, iterative methods, non-Hermitian matrices, sparse linear systems

AMS subject classifications. 65F15, 65F10

PII. S0895479896306744

1. Introduction.

Lanczos-type product methods (LTPMs) such as (Bi)CGS [42], BiCGStab [44], BiCGStab2 [22], and BiCGStab(ℓ) [38], [41] are among the most efficient methods for solving large systems of linear equations $Ax = b$ with nonsymmetric sparse system matrix $A \in \mathbb{C}^{N\times N}$. Compared to the biconjugate gradient (BiCG) method, they have the advantage of converging roughly twice as fast and of not requiring a routine for applying the adjoint system matrix A^H to a vector. Nevertheless, they inherit from BiCG the short recurrence formulas for generating the approximations x_k and the corresponding residuals $r_k := b - Ax_k$.

As for BiCG, where the convergence can be smoothed by applying the quasi-minimal residual (QMR) method [16] or the local minimum residual process (MR smoothing) [37], [47], product methods can be combined with the same techniques [14], [47] to avoid the likely "erratic" convergence behavior [11], [36]; the benefit of these smoothing techniques is disputed, however.

A well-known problem of all methods that make implicit use of the Lanczos polynomials generated by a non-Hermitian matrix is the danger of breakdown. Although exact breakdowns are very rare in practice, it has been observed that near-breakdowns can slow down or even prevent convergence [15]. Look-ahead techniques for the Lanczos process [15], [21], [23], [34], [35], [43] allow us to avoid this problem when we use variants of the BiCG method with or without the mentioned smoothing techniques. However, the general look-ahead procedures that have so far been proposed for (Bi)CGS or other LTPMs are either limited to exact breakdowns [8], [9] or are based on different look-ahead recursions. In fact, the look-ahead steps proposed by Brezinski and Redivo Zaglia [5] to avoid near-breakdowns in the conjugate gradient squared (CGS) method and those in [6], which are applicable to all LTPMs, are based on the so-called BSMRZ algorithm, which is itself based on a generalization of cou-

*Received by the editors July 17, 1996; accepted for publication (in revised form) by Ž. Strakoš June 4, 1999; published electronically March 21, 2000.
http://www.siam.org/journals/simax/21-4/30674.html

†Seminar for Applied Mathematics, ETH Zürich, ETH-Zentrum HG, CH-8092 Zürich, Switzerland (mhg@sam.math.ethz.ch). Formerly at the Swiss Center for Scientific Computing (CSCS/SCSC).
‡German Aerospace Research Establishment (DLR), German Remote Sensing Data Center (DFD), D-82234 Oberpfaffenhofen, Germany (kjr@dfd.dlr.de). Formerly at CSCS/SCSC.

pled recurrences (different from those of the standard BiOMin and BiODir versions of the BiCG method) for implementing the Lanczos method. For a theoretical comparison of the two approaches we refer to [25, section 19]. Recently, a number of further "look-ahead-like" algorithms for Lanczos-type solvers have been proposed by Ayachour [1], Brezinski, Redivo Zaglia, and Sadok [7], Graves-Morris [19], and Ziegler [48]. Their discussion is beyond the scope of this paper, but it seems that [1], [7], [48] are restricted to exact breakdowns.

In this paper, we start from the approach in [15] and [23, section 9] and derive an alternative look-ahead procedure for LTPM algorithms that make use of the Lanczos three-term recurrences. Compared to the standard coupled two-term recurrences they have the advantage of being simpler to handle with regard to look-ahead, since they are only affected by one type of breakdown. In contrast to the first version of this work [26], we also capitalize upon an enhancement for the look-ahead Lanczos algorithm pointed out by Hochbruck recently (see [28], [25]), which is here adapted to LTPMs. Other improvements help to further reduce the overhead and stabilize the process.

Starting with an initial approximation x_0 and a corresponding initial residual $y_0 := b - Ax_0$, LTPMs generate in n steps a basis of the $2n$-dimensional Krylov space $\mathcal{K}_{2n} := \mathcal{K}_{2n}(A, y_0)$ in such a way that the even indexed basis vectors are of the form

$$\tau_n(A)\rho_n(A)y_0, \tag{1.1}$$

where ρ_n is the nth Lanczos polynomial (see below) and τ_n is another suitably chosen polynomial of exact degree n. In the algorithms we discuss, these vectors will be either the residual of the nth approximation x_n or a scalar multiple of it. By allowing them to be multiples of the residuals, that is, by considering so-called unnormalized [20] or inconsistent [25] Krylov space solvers, we avoid the occurrence of pivot (or ghost) breakdowns. (For the various types of breakdowns and their connection to the block structure of the Padé table see [20], [24], [30], [31]. In the setting of [5], [6], they have been addressed particularly in [4].)

More generally, we define a doubly indexed sequence of *product vectors* w_n^l by

$$w_n^l := \tau_l(A)y_n := \tau_l(A)\rho_n(A)y_0. \tag{1.2}$$

The aim is to find in the $(n+1)$th step of an LTPM an improved approximation x_{n+1} by computing a new product vector w_{n+1}^{n+1} from previously determined ones in a stable way. To visualize the progression of an algorithm and the recurrences it uses we arrange the product vectors w_n^l in a w-table.[1] Its n-axis points downward and its l-axis to the right. We describe then how an algorithm moves in this table from the upper left corner downward to the right.

This paper is organized as follows. In section 2 we review the look-ahead Lanczos process. Versions based on the Lanczos three-term recurrences of various LTPMs are introduced in section 3. In section 4 we present look-ahead procedures for these LTPMs and analyze their computational overhead, and in section 5 several look-ahead strategies are discussed. Our preferred way of applying the look-ahead procedures to obtain the solution of a linear system is presented in section 6. In section 7 the efficiency of the proposed algorithms is demonstrated by numerical examples, and in section 8 we draw some conclusions.

[1] The w-table is different from the scheme introduced by Sleijpen and Fokkema [38] for BiCGStab(ℓ).

2. The look-ahead Lanczos process. The primary aim of the Lanczos process [32] is the construction of a pair of biorthogonal bases for two nested sequences of Krylov spaces. Given $A \in \mathbb{C}^{N\times N}$ and a pair of starting vectors, $(\widetilde{y}_0, y_0)$, the Lanczos biorthogonalization (BiO) algorithm generates a pair of finite sequences, $\{\widetilde{y}_n\}_{n=0}^{\nu}$, $\{y_n\}_{n=0}^{\nu}$ of *left* and *right Lanczos vectors* such that

$$\begin{aligned} y_n \in \mathcal{K}_{n+1} &:= \operatorname{span}\{y_0, Ay_0, \ldots, A^n y_0\}, \\ \widetilde{y}_n \in \widetilde{\mathcal{K}}_{n+1} &:= \operatorname{span}\{\widetilde{y}_0, A^H\widetilde{y}_0, \ldots, (A^H)^n\widetilde{y}_0\}, \end{aligned} \tag{2.1}$$

and

$$\langle \widetilde{y}_m, y_n\rangle = \begin{cases} 0 & \text{if} \quad m \neq n \ \text{ or } \ m = n = \nu, \\ \delta_n \neq 0 & \text{if} \quad m = n < \nu, \end{cases} \tag{2.2}$$

or, equivalently,

$$\widetilde{y}_n \perp \mathcal{K}_n, \quad \widetilde{\mathcal{K}}_n \perp y_n. \tag{2.3}$$

Here, $\langle\cdot,\cdot\rangle$ denotes the inner product in $\mathbb{C}^N$, which we choose to be linear in the second argument, and $\perp$ indicates the corresponding orthogonality.

The sequence of pairs of Lanczos vectors can be constructed by the three-term recursions

$$\begin{aligned} y_{n+1} &= (Ay_n - y_n\alpha_n - y_{n-1}\beta_n)/\gamma_n, \\ \widetilde{y}_{n+1} &= (A^H\widetilde{y}_n - \widetilde{y}_n\overline{\alpha}_n - \widetilde{y}_{n-1}\overline{\beta}_n)/\overline{\gamma}_n, \end{aligned} \tag{2.4}$$

with coefficients α_n and β_n that are determined from the orthogonality condition (2.2) and nonvanishing scale factors γ_n that can be chosen arbitrarily. Choosing $\gamma_n = -\alpha_n - \beta_n$ would allow us to consider the right Lanczos vectors y_n as residuals and to update the iterates x_n particularly simply. But this choice also introduces the possibility of a breakdown due to $\alpha_n + \beta_n = 0$. Therefore, we suggest in section 6 a different way of defining iterates.

From (2.4) it follows directly that the Lanczos vectors can be written in the form

$$y_n = \rho_n(A)y_0, \tag{2.5}$$

$$\widetilde{y}_n = \overline{\rho}_n(A^H)\widetilde{y}_0, \tag{2.6}$$

where ρ_n denotes the nth Lanczos polynomial. Since we aim here at LTPMs, we consider for the Krylov spaces $\widetilde{\mathcal{K}}_n$ more general basis vectors of the form

$$\widetilde{z}_n = \overline{\tau}_n(A^H)\widetilde{z}_0 \tag{2.7}$$

with arbitrary but suitably chosen polynomials τ_n of exact degree n and $\widetilde{z}_0 := \widetilde{y}_0$. In general, $\widetilde{z}_{n+1} \perp \mathcal{K}_{n+1}$ will no longer hold, but $\widetilde{\mathcal{K}}_{n+1} \perp y_{n+1}$ can still be attained by enforcing $\widetilde{z}_{n-1} \perp y_{n+1}$ and $\widetilde{z}_n \perp y_{n+1}$, which means choosing the coefficients α_n and β_n in (2.4) in the following way: if $n > 0$ we need

$$0 = \langle \widetilde{z}_{n-1}, y_{n+1}\rangle = \left(\langle \widetilde{z}_{n-1}, Ay_n\rangle - \langle \widetilde{z}_{n-1}, y_n\rangle\alpha_n - \langle \widetilde{z}_{n-1}, y_{n-1}\rangle\beta_n\right)/\gamma_n,$$

but since $\langle \widetilde{z}_{n-1}, y_n\rangle = 0$, we have, with $\delta_n^\tau := \langle \widetilde{z}_n, y_n\rangle$,

$$\beta_n = \frac{\langle \widetilde{z}_{n-1}, Ay_n\rangle}{\langle \widetilde{z}_{n-1}, y_{n-1}\rangle} = \frac{\langle \widetilde{z}_{n-1}, Ay_n\rangle}{\delta_{n-1}^\tau}. \tag{2.8}$$

When $n = 0$, we set $\beta_0 = 0$. Similarly, from the orthogonality condition $\langle \widetilde{z}_n, y_{n+1} \rangle = 0$ we obtain

$$\alpha_n = \frac{\langle \widetilde{z}_n, Ay_n - y_{n-1}\beta_n \rangle}{\langle \widetilde{z}_n, y_n \rangle} = \frac{\langle \widetilde{z}_n, Ay_n - y_{n-1}\beta_n \rangle}{\delta_n^\tau}. \tag{2.9}$$

Equations (2.7)–(2.9) and the first recursion of (2.4) specify what one might call the *one-sided* Lanczos process, formulated in [36] to derive LTPMs.

Clearly, this recursive process terminates with $y_\nu = 0$ or $\widetilde{z}_\nu = 0$, or it breaks down with $\delta_\nu^\tau = 0$ and $y_\nu \neq 0$, $\widetilde{z}_\nu \neq 0$. The *look-ahead* Lanczos process [23, section 9], [15] overcomes such a breakdown if curable, i.e., if for some k, $\langle \widetilde{z}_\nu, A^k y_\nu \rangle \neq 0$. However, its role is not restricted to treating such exact breakdowns with $\delta_n^\tau = 0$: it allows us to continue the BiO process whenever for stability reasons we choose to enforce the orthogonality condition (2.2) only partially for a couple of steps. We will come back later to the conditions that make us start such a look-ahead phase. In the look-ahead Lanczos process the price we have to pay is that the Gramian matrix $D := (\langle \widetilde{z}_m, y_n \rangle)$ becomes block lower triangular instead of triangular (as in the generic one-sided Lanczos algorithm) or diagonal (as in the generic two-sided Lanczos algorithm). In other words, we replace the conditions $\widetilde{\mathcal{K}}_n \perp y_n$ and $\delta_n^\tau \neq 0$ by

$$\widetilde{\mathcal{K}}_{n_j} \perp y_n \quad \text{if } n \geq n_j \ (j = 0, \ldots, J) \tag{2.10}$$

and

$$D_j := (\langle \widetilde{z}_k, y_l \rangle)_{k,l=n_j}^{n_{j+1}-1} \quad \text{nonsingular if} \quad j < J, \tag{2.11}$$

where $0 = n_0 < n_1 < \cdots < n_J = \nu \leq N$ is the subsequence of indices of (well-conditioned) *regular* Lanczos vectors y_{n_j}; for the other indices the vectors are called *inner* vectors. Likewise, we refer to n as a *regular index* if $n = n_j$ for some j, while n is called an *inner index* otherwise. Note that there is considerable freedom in choosing the subsequence $\{n_j\}_{j=0}^J$: for example, we might request that the smallest singular value $\sigma_{\min}(D_j)$ be sufficiently large.

The sequence $\{y_n\}$ is determined by the condition $\widetilde{\mathcal{K}}_{n_j} \perp y_n$ if $n_j \leq n < n_{j+1}$ and hence it does not depend directly on the sequence $\{\widetilde{z}_n\}$ but only on the spaces $\widetilde{\mathcal{K}}_{n_j}$ that are spanned by the first n_j elements of $\{\widetilde{z}_n\}$. On the other hand, the smallest singular value of each D_j depends on the basis chosen, a fact that one should take into account. Forming the blocks

$$Y_j := \begin{bmatrix} y_{n_j} & y_{n_j+1} & \cdots & y_{n_{j+1}-1} \end{bmatrix}, \quad \widetilde{Z}_j := \begin{bmatrix} \widetilde{z}_{n_j} & \widetilde{z}_{n_j+1} & \cdots & \widetilde{z}_{n_{j+1}-1} \end{bmatrix},$$

we can express (2.10) as

$$\widetilde{Z}_l^H Y_j = \begin{cases} D_j & \text{if } l = j, \\ 0 & \text{if } l < j. \end{cases} \tag{2.12}$$

The look-ahead step size is in the following denoted by $h_j := n_{j+1} - n_j$.

In the look-ahead case the Lanczos vectors $\{y_n\}$ can be generated by the recursion [23, section 9], [15]:

$$y_{n+1} = (Ay_n - \widehat{Y}_j \alpha_n - Y_{j-1}\beta_n)/\gamma_n \qquad (n_j < n+1 \leq n_{j+1}), \tag{2.13}$$

where $\widehat{Y}_j := \left[\begin{array}{ccc} y_{n_j} & \ldots & y_n \end{array}\right]$ denotes the not yet fully completed j-block if $n+1$ is an inner index, while $\widehat{Y}_j = Y_j$ if $n+1$ is regular, that is, if $n+1 = n_{j+1}$. The coefficient vector β_n is determined by the condition

$$0 = \widetilde{Z}_{j-1}^H y_{n+1} = \widetilde{Z}_{j-1}^H A y_n - \widetilde{Z}_{j-1}^H Y_{j-1}\beta_n,$$

thus,

$$\beta_n = D_{j-1}^{-1}\widetilde{Z}_{j-1}^H A y_n. \tag{2.14}$$

The coefficient vector α_n can be chosen arbitrarily if $n+1$ is an inner index. Obviously, the choice $\alpha_n = 0$ yields the cheapest recursion, but we may gain numerical stability by choosing $\alpha_n \neq 0$. On the other hand, if $n+1$ is a regular index, α_n results from the condition

$$0 = \widetilde{Z}_j^H y_{n+1} = \widetilde{Z}_j^H A y_n - \widetilde{Z}_j^H Y_j \alpha_n - \widetilde{Z}_j^H Y_{j-1}\beta_n,$$

that is,

$$\alpha_n = D_j^{-1}\widetilde{Z}_j^H \left(A y_n - Y_{j-1}\beta_n\right). \tag{2.15}$$

Recently, it was pointed out by Hochbruck [28] that the recursion (2.13) can be simplified since the contribution of the older block Y_{j-1} can be represented by multiples of a single *auxiliary vector* y'_{j-1}. This is due to the fact (noticed in [23]) that the matrix made up of the coefficient vectors $\beta_{n_j}, \ldots, \beta_{n_{j+1}-1}$ for block j is of rank one; see also [25, section 19]. This simplification has also been capitalized upon in the look-ahead Hankel solver of Freund and Zha [17], which is closely related to the look-ahead Lanczos algorithm.

In particular, due to (2.10) or (2.12) we have

$$\langle \widetilde{z}_{l+1}, y_n\rangle = 0 \quad \text{and} \quad \langle \widetilde{z}_l, A y_n\rangle = 0 \quad \text{if } l < n_j - 1 < n. \tag{2.16}$$

Therefore, (2.14) simplifies to

$$\beta_n = D_{j-1}^{-1}\widetilde{Z}_{j-1}^H A y_n = D_{j-1}^{-1}\,\ell\,\widetilde{z}_{n_j-1}^H A y_n =: D_{j-1}^{-1}\ell\,\beta'_n, \tag{2.17}$$

where $\ell := [0, \ldots, 0, 1]^T$ and

$$\beta'_n := \langle \widetilde{z}_{n_j-1}, A y_n\rangle. \tag{2.18}$$

Introducing the *auxiliary vector*

$$y'_{j-1} := Y_{j-1}\left(D_{j-1}^{-1}\ell\right), \tag{2.19}$$

we have

$$Y_{j-1}\beta_n = \left(Y_{j-1}D_{j-1}^{-1}\ell\right)\left(\widetilde{z}_{n_j-1}^H A y_n\right) = y'_{j-1}\beta'_n, \tag{2.20}$$

so that the recurrence (2.13) becomes

$$y_{n+1} = \left(A y_n - \widehat{Y}_j\alpha_n - y'_{j-1}\beta'_n\right)/\gamma_n \quad (n_j < n+1 \le n_{j+1}) \tag{2.21}$$

and (2.15) changes to

$$\alpha_n = D_j^{-1}\widetilde{Z}_j^H\left(A y_n - y'_{j-1}\beta'_n\right). \tag{2.22}$$

3. LTPMs. LTPMs are based on two ideas. The first is to derive for a certain sequence of product vectors w_{n+1}^{l+1} recursion formulas that involve only previously computed product vectors, so that there is no need to explicitly compute the vectors $\widetilde{z}_l$ and y_n. By multiplying the three-term recursion (2.4) for the right Lanczos vectors y_n by $\tau_l(A)$ we obtain

$$w_{n+1}^l = \left(Aw_n^l - w_n^l\alpha_n - w_{n-1}^l\beta_n\right)/\gamma_n. \tag{3.1}$$

This recursion can be applied to move forward in the vertical direction in the w-table. To obtain a formula for proceeding in the horizontal direction, the recursion for the chosen polynomials $\tau_l(\zeta)$ is capitalized upon in an analogous way.

The second basic idea is to rewrite the inner products that appear in the Lanczos process in terms of product vectors:

$$\begin{aligned}\langle\widetilde{z}_l, y_n\rangle &= \langle\overline{\tau}_l(A^H)\widetilde{z}_0, y_n\rangle &&= \langle\widetilde{z}_0, \tau_l(A)y_n\rangle &&= \langle\widetilde{z}_0, w_n^l\rangle,\\ \langle\widetilde{z}_l, Ay_n\rangle &= \langle\overline{\tau}_l(A^H)\widetilde{z}_0, Ay_n\rangle &&= \langle\widetilde{z}_0, A\tau_l(A)y_n\rangle &&= \langle\widetilde{z}_0, Aw_n^l\rangle.\end{aligned}$$

For the coefficients α_n and β_n of (2.8) and (2.9) this results in

$$\beta_n = \frac{\langle\widetilde{z}_0, Aw_n^{n-1}\rangle}{\langle\widetilde{z}_0, w_{n-1}^{n-1}\rangle}, \quad \alpha_n = \frac{\langle\widetilde{z}_0, Aw_n^n - w_{n-1}^n\beta_n\rangle}{\langle\widetilde{z}_0, w_n^n\rangle}. \tag{3.2}$$

LTPMs still have short recurrence formulas if the polynomials τ_l have a short one. In addition, they have two advantages over BiCG: first, multiplications with the adjoint system matrix A^H are avoided; second, for an appropriate choice of the polynomials $\tau_l(\zeta)$, smaller new residuals $r_l := w_l^l = \tau_l(A)y_l$ can be expected because of a further reduction of y_l by the operator $\tau_l(A)$. Different choices of the polynomials $\tau_l(\zeta)$ lead to different LTPMs. In the following we briefly review some possible choices for these polynomials. We start with the general class where they satisfy a three-term recursion. Since then both the "left" polynomials τ_l and the "right" polynomials ρ_n fulfill a three-term recursion, we say that this is the class of $(3,3)$-*type LTPMs*.

In the following we briefly review some possible choices for these polynomials. For a more detailed discussion and a pseudocode for the resulting algorithms we refer to [36].

3.1. LTPMs based on a three-term recursion for $\{\tau_l\}$: BiOxCheb and BiOxMR2. Assume the polynomials $\tau_l(\zeta)$ are generated by a three-term recursion of the form [22]

$$\tau_{l+1}(\zeta) = (\xi_l + \eta_l\zeta)\,\tau_l(\zeta) + (1-\xi_l)\,\tau_{l-1}(\zeta) \tag{3.3}$$

with $\tau_{-1}(\zeta) \equiv 0$, $\tau_0(\zeta) \equiv 1$, $\xi_0 = 1$, and $\eta_l \neq 0$ for all l. Note that, by induction, τ_l has exact degree l and $\tau_l(0) = 1$ for all l. Hence, the polynomials qualify as residual polynomials of a Krylov space method.

Multiplying y_n by $\tau_{l+1}(A)$ from the left and applying (3.3) yields the horizontal recursion

$$w_n^{l+1} = Aw_n^l\eta_l + w_n^l\xi_l + w_n^{l-1}(1-\xi_l). \tag{3.4}$$

By applying (3.1) and (3.4), Loop 3.1 produces a new product vector w_{n+1}^{n+1} from previously calculated ones, namely w_m^l, $(m,l) := (n,n),(n-1,n),(n,n-1),(n-1,n-1)$, and the product Aw_n^{n-1} also evaluated before. At the end of the loop,

besides w_{n+1}^{n+1}, also w_n^{n+1}, w_{n-1}^{n+1}, and Aw_{n+1}^n, which will be needed in the next run through the loop, are available.

LOOP 3.1 (general $(3,3)$-type LTPM).
1. *Compute* Aw_n^n *and determine* β_n *and* α_n.
2. *Use* (3.1) *to compute* w_{n+1}^{n-1} *(if* $n > 0$*) and* w_{n+1}^n.
3. *Compute* Aw_{n+1}^n *and determine* ξ_n *and* η_n.
4. *Use* (3.4) *to compute* w_n^{n+1} *and* w_{n+1}^{n+1}.

Next we discuss two ways of choosing the polynomials τ_l, that is, of specifying the recurrence coefficients ξ_l and η_l.

One possibility is to combine the Lanczos process with the Chebyshev method [13], [45] by choosing τ_l as a suitably shifted and scaled Chebyshev polynomial. This combination was suggested in [3], [22], [44]. Let us call the resulting $(3,3)$-type LTPM BiOxCheb. It is well known that these polynomials satisfy a recurrence of the form (3.3). After acquiring some information about the spectrum of the matrix A, for example, by performing a few iterations with another Krylov space method, one can determine the recursion for the scaled and shifted Chebyshev polynomials that correspond to an ellipse surrounding the estimated spectrum [33]. However, the necessity to provide spectral information is often seen as a drawback of this method.

Another idea is to use the coefficients ξ_l and η_l of (3.3) for locally minimizing the norm of the new residual w_{l+1}^{l+1}. This idea is borrowed from the BiCGStab and BiCGStab2 methods [44], [22] reviewed below: ξ_l and η_l are determined by solving the two-dimensional minimization problem

$$\min_{\xi,\eta\in\mathbb{C}} \left\| w_{l+1}^{l-1} + (w_{l+1}^l - w_{l+1}^{l-1})\xi + Aw_{l+1}^l\eta \right\|. \tag{3.5}$$

Introducing the $N \times 2$ matrix $B_{l+1} := \left[\ (w_{l+1}^l - w_{l+1}^{l-1}) \ \middle|\ Aw_{l+1}^l\ \right]$ we can write (3.5) as the least-squares problem

$$\min_{\xi,\eta\in\mathbb{C}} \left\| w_{l+1}^{l-1} + B_{l+1} \begin{bmatrix} \xi \\ \eta \end{bmatrix} \right\|.$$

Therefore, ξ_l and η_l can be computed as the solution of the normal equations

$$\left(B_{l+1}^H B_{l+1}\right) \begin{bmatrix} \xi_l \\ \eta_l \end{bmatrix} = -B_{l+1}^H w_{l+1}^{l-1}. \tag{3.6}$$

In view of the two-dimensional local residual norm minimization performed at every step (except the first one) we call this the BiOxMR2 method.[2] A version based on coupled two-term Lanczos recurrences of this method was introduced by the first author in a talk in Oberwolfach (April 1994). Independently it was as well proposed by Cao [10] and by Zhang [46], whose technical report is dated April 1993. Zhang also considered two-term formulas for τ_l and presented very favorable numerical results.

3.2. LTPMs based on a two-term recursion for $\{\tau_l\}$: BiOStab. In Van der Vorst's BiCGStab [44], the polynomials $\tau_l(\zeta)$ are built up successively as products of polynomials of degree 1:

$$\tau_0(\zeta) \equiv 1, \qquad \tau_{l+1}(\zeta) = (1 - \chi_l\zeta)\tau_l(\zeta) \quad \text{for } l \geq 0. \tag{3.7}$$

[2]The letter x in the name BiOxMR2 reflects the fact that the residual polynomials of this method are the products of the Lanczos polynomials generated by the BiO process with polynomials obtained from a successive two-dimensional minimization of the residual (MR2). Similarly, BiOxCheb means a combination of the BiO process with a Chebyshev process.

Inserting here for ζ the system matrix A and multiplying by y_n from the right yields

$$w_n^{l+1} = w_n^l - Aw_n^l\chi_l. \tag{3.8}$$

The coefficient χ_l is determined by minimizing the norm of $w_{l+1}^{l+1} = w_{l+1}^l - Aw_{l+1}^l\chi_l$, that is, by solving the one-dimensional minimization problem

$$\min_{\chi\in\mathbb{C}} \left\|w_{l+1}^l - Aw_{l+1}^l\chi\right\|, \tag{3.9}$$

which leads to

$$\chi_l = \frac{\langle Aw_{l+1}^l, w_{l+1}^l\rangle}{\langle Aw_{l+1}^l, Aw_{l+1}^l\rangle}. \tag{3.10}$$

Recurrences (3.1) and (3.8) can be used to compute a new product vector w_{n+1}^{n+1} from w_n^n and w_{n-1}^n as described in Loop 3.1 with $\xi_n = 1$ and $\eta_n = -\chi_n$, where χ_n is calculated from (3.10). However, since (3.8) is only a two-term recurrence, there is no need now to compute w_{n+1}^{n-1} in substep 2 of the loop.

We call this algorithm BiOStab since it is a version of BiCGStab that is based on the three-term recurrences of the Lanczos BiO process.[3]

3.3. BiOStab2. BiOStab2 is the version of BiCGStab2 [22] based on the three-term Lanczos recurrences. The polynomials $\tau_l(\zeta)$ satisfy the recursions

$$\begin{array}{lcll} \tau_0(\zeta) & \equiv & 1, & \\ \tau_{l+1}(\zeta) & = & (1-\chi_l\zeta)\,\tau_l(\zeta) & \text{if } l \text{ is even,} \\ \tau_{l+1}(\zeta) & = & (\xi_l+\eta_l\zeta)\,\tau_l(\zeta) + (1-\xi_l)\,\tau_{l-1}(\zeta) & \text{if } l \text{ is odd.} \end{array} \tag{3.11}$$

Here χ_l may be obtained by solving the one-dimensional minimization problem (3.9), so χ_l is given by (3.10). However, if $|\chi_l|$ is small, this choice is dangerous since the vector component needed to enlarge the Krylov space becomes negligible [38]. Then some other value of χ_l should be chosen. Except for roundoff, the choice has no effect on later steps, because ξ_l and η_l are determined by solving the two-dimensional minimization problem (3.5).

Multiplying y_n by $\tau_{l+1}(A)$ and applying (3.11) leads to

$$w_n^{l+1} = \begin{cases} w_n^l - Aw_n^l\chi_l & \text{if } l \text{ is even,} \\ Aw_n^l\eta_l + w_n^l\xi_l + w_n^{l-1}(1-\xi_l) & \text{if } l \text{ is odd.} \end{cases} \tag{3.12}$$

Now, Loop 3.1 applies with $\xi_n = 1$, $\eta_n = -\chi_n$ if n is even, while ξ_n and η_n are chosen as indicated above if n is odd. If n is even, there is no need to compute w_{n+1}^{n-1} in substep 2 of the loop.

3.4. BiO-squared. BiO-squared (BiOS) is obtained by "squaring" the three-term Lanczos process: among the basis vectors generated are those Krylov-space vectors that correspond to the squared Lanczos polynomials. By complementing BiOS with a recursion for Galerkin iterates we will obtain BiOResS, a $(3,3)$-type version of Sonneveld's $(2,2)$-type *CGS method* [42]. The method fits into the framework of LTPMs if we identify

$$\tau_l(\zeta) = \rho_l(\zeta). \tag{3.13}$$

[3] Eijkhout [12] also proposed such a variant of BiCGStab; however, his way of computing the Lanczos coefficients is much too complicated.

The vectors $\widetilde{z}_n = \overline{\tau}_n(A^H)\widetilde{z}_0 = \overline{\rho}_n(A^H)\widetilde{y}_0 = \widetilde{y}_n$ are then exactly the left Lanczos vectors so that now $y_n \perp \widetilde{\mathcal{K}}_n$ as well as $\widetilde{z}_n \perp \mathcal{K}_n$ is fulfilled. Thus, the coefficient α_n in (3.2) simplifies to

$$\alpha_n = \frac{\langle \widetilde{z}_0, Aw_n^n \rangle}{\langle \widetilde{z}_0, w_n^n \rangle}, \tag{3.14}$$

and the w-table becomes symmetric since

$$w_n^l = \rho_l(A)\rho_n(A)y_0 = \rho_n(A)\rho_l(A)y_0 = w_l^n. \tag{3.15}$$

Consequently, we have in analogy to (3.1)

$$w_n^{l+1} = \left(Aw_n^l - w_n^l\alpha_l - w_n^{l-1}\beta_l\right)/\gamma_l. \tag{3.16}$$

These two recursions lead to the following loop.

LOOP 3.2 (BiOS).
1. *Compute* Aw_n^n *and determine* β_n *and* α_n.
2. *Use* (3.1) *to compute* w_{n+1}^{n-1} *(if* $n > 0$*) and* w_{n+1}^n.
3. *Compute* Aw_{n+1}^n.
4. *Use* (3.16) *to compute* w_{n+1}^{n+1}.

Note that we exploit in substep 2 the symmetry of the w-table: the product vector w_{n-1}^n, which is needed for the calculation of w_{n+1}^n by (3.1), is equal to w_n^{n-1} and need not be stored. We also point out that (3.16) is not of the form (3.4); in particular, the coefficients of w_n^{l-1} and w_n^l need not sum up to 1.

4. Look-ahead procedures for LTPMs. Look-ahead steps in an LTPM serve to stabilize the Lanczos process, the vertical movement in the w-table. Except for BiOS, the recursion formulas for the horizontal movement remain the same. The vertical movement is now in general based on the recurrence formula (2.21) for the Lanczos vectors, but we need to replace these vectors by product vectors w_n^l; see (1.2).

We introduce the blocks of product vectors

$$\widehat{W}_j^l := \left[\begin{array}{ccc} w_{n_j}^l & \dots & w_n^l \end{array}\right], \quad W_{j-1}^l := \left[\begin{array}{ccc} w_{n_{j-1}}^l & \dots & w_{n_j-1}^l \end{array}\right]$$

and the *auxiliary product vector* $w_{j-1}'^l$ defined by

$$w_{j-1}'^l := \tau_l(A)y'_{j-1} = \tau_l(A)Y_{j-1}\left(D_{j-1}^{-1}\ell\right) = W_{j-1}^l\left(D_{j-1}^{-1}\ell\right). \tag{4.1}$$

Again $\widehat{W}_j^l = W_j^l$ if $n+1 = n_{j+1}$ is a regular index, while $\widehat{W}_j^l$ denotes the not yet completed jth block if $n+1$ is an inner index. Now, multiplying (2.21) by $\tau_l(A)$ from the left, we obtain

$$w_{n+1}^l = \left(Aw_n^l - \widehat{W}_j^l\alpha_n - w_{j-1}'^l\beta'_n\right)/\gamma_n \qquad (n_j < n+1 \le n_{j+1}). \tag{4.2}$$

As in section 3, the coefficient vectors β'_n and, in the regular case, α_n can be expressed in terms of the product vectors by rewriting all inner products in such a way that the part $\overline{\tau}_l(A^H)$ of $\widetilde{z}_l$ on the left side is transferred to the right side of the inner product. For the diagonal blocks D_j of the Gramian this yields

$$D_j = \left[\langle \widetilde{z}_0, w_k^i \rangle\right]_{i,k=n_j}^{n_{j+1}-1}, \tag{4.3}$$

inner case

	n_{j-1}			n_j ... l
n_{j-1}	d	d	d	
	d	d	d	
	d	d	d	v′
n_j				V
⋮				V
n				V / A
$n+1$				∗

regular case

	n_{j-1}			n_j ... l n				n_{j+1}
n_{j-1}	d	d	d					
	d	d	d					
	d	d	d	α′	α′	v′	α′	
n_j				D	D	V	D	
⋮				D	D	V	D	
⋮				D	D	V	D	
n				D / A	D / A	V / A	D / A	
n_{j+1}						∗		

FIG. 4.1. *Entries in the w-table needed for the construction of a new inner or regular vector marked by ∗ by recurrence formula* (4.2). *Here, v′ indicates entries needed for* w'^{l}_{j-1}, *V those for* $\widehat{W}^{l}_{j}$, *d those for* D_{j-1}, *D those additionally required for* D_j. *Moreover,* α' *marks auxiliary vectors* w'^{k}_{j-1} $(n_j \le k < n_{j+1})$ *in the formula* (4.5) *for* α_n, *and A stands for a matrix-vector product (MV) with A in this formula. One such product also appears explicitly in* (4.2).

and β'_n of (2.18) becomes

$$\beta'_n = \widetilde{z}_0^H A w_n^{n_j-1} \quad (n_j < n+1 \le n_{j+1}). \tag{4.4}$$

Likewise, α_n of (2.22) turns into

$$\alpha_n = D_j^{-1} \begin{bmatrix} \langle \widetilde{z}_0, A w_n^{n_j} - w'^{\,n_j}_{j-1} \beta'_n \rangle \\ \vdots \\ \langle \widetilde{z}_0, A w_n^{n_{j+1}-1} - w'^{\,n_{j+1}-1}_{j-1} \beta'_n \rangle \end{bmatrix} \quad (n+1 = n_{j+1}). \tag{4.5}$$

For advancing in the horizontal direction of the w-table we will still use (3.4), (3.8), the combination (3.12), or (3.16). Substituting $A w_n^{n_j-1}$ in (4.4) according to these formulas and letting $\delta_n^{n_j} = \langle \widetilde{z}_0, w_n^{n_j} \rangle$ be the (n_j, n)-element of D, (4.4) simplifies to

$$\beta'_n = \begin{cases} \delta_n^{n_j}/\eta_{n_j-1} & \text{if (3.4) holds for } l = n_j - 1, \\ -\delta_n^{n_j}/\chi_{n_j-1} & \text{if (3.8) holds for } l = n_j - 1, \\ \delta_n^{n_j}\gamma_{n_j-1} & \text{if (3.16) holds for } l = n_j - 1. \end{cases} \tag{4.6}$$

We remark that by using (4.6) instead of $\langle \widetilde{z}_0, A w_n^{n_j-1} \rangle$ we avoid computing and storing the complete W_j^{j-1} block. Now only a few elements of this block are needed; see Figure 4.1. In this figure we display those entries w_n^l and products $A w_n^l$ in the w-table that are needed to compute by (4.2) a new vector w_{n+1}^l marked by $*$.

Note that the horizontal recursions are valid for the auxiliary product vectors $w'^{\,l}_{j-1}$ too: for example, (3.4) and (4.1) imply

$$w'^{\,l+1}_{j-1} = A w'^{\,l}_{j-1} \eta_l + w'^{\,l}_{j-1} \xi_l + w'^{\,l-1}_{j-1} (1 - \xi_l), \tag{4.7}$$

while (3.8) and (4.1) provide

$$w'^{\,l+1}_{j-1} = w'^{\,l}_{j-1} - A w'^{\,l}_{j-1} \chi_l \,. \tag{4.8}$$

Thus, the auxiliary vector can be updated in the horizontal direction at the cost of one matrix-vector product (MV) per step.

Combining the recursions for the vertical movement with those for the horizontal movement leads to the look-ahead version of an LTPM. (Actually, we will also need to compute approximations to the solution $A^{-1}b$ of the linear system; but we defer this until section 6.) Because of the additional vectors involved in the above recurrence formulas, it seems that such a look-ahead LTPM requires much more computational work and storage than its unstable, standard version. However, the number of look-ahead steps as well as the size of their blocks is small in practice, so that the overhead is moderate. Moreover, we describe in the following for various choices of polynomials $\tau_n(\zeta)$ how MVs that are needed can be computed indirectly by applying the recurrence formulas. In the same way also the values of inner products can be obtained indirectly at nearly no cost.

4.1. Look-ahead LTPMs based on a three-term recursion for $\{\tau_l\}$: LABiOxCheb and LABiOxMR2. For methods incorporating a horizontal three-term recurrence, such as BiOxCheb and BiOxMR2, applying the above principles for look-ahead LTPMs leads to the general Loop 4.1. Note that the first four substeps are identical in both cases. We will see in section 5 that the decision between a regular and an inner loop is made during substep 5.

In Figure 4.2 we display the action of this loop in the w-table, but now we use a different format than before, which specifies what has been known before the current sweep through the loop and what is being computed in this sweep. In particular, the following symbols and indicators are used:

- V indicates that the corresponding product vector is already known;
- A as denominator indicates that the product of the vector represented by V with the matrix A was needed;
- a solid box around a fraction means that this product by A required (or requires) an MV;
- no box around a fraction means that this product can be obtained by applying a recurrence formula;
- a number as entry specifies in which substep of the current sweep this entry is calculated;
- a prime indicates that the vector is an auxiliary one, as defined in (4.1) (these vectors are displayed in the last row of the corresponding block of the w-table);
- A$'$ as denominator indicates that the product of the vector represented by V$'$ with the matrix A was needed;
- double primes will indicate that the vector is an auxiliary one of the type defined in (4.11) used in LABiCGS (these vectors are displayed in the lower-right corner of the corresponding block of the w-table);
- S will indicate that the vector or the MV is obtained for free due to the symmetry of the w-table of LABiCGS.

In the first two w-tables of Figure 4.2 we display what happens in a first inner step (at left) and in a regular step (at right) that follows directly a regular step (so that $h_j = 1$). In the second pair of w-tables of Figure 4.2 we consider an inner and a regular step in the case of a large block size. Note that substeps 2–4 of the loop do not appear in the first pair, and that substep 4 would still not be active in the regular loop at the end of a look-ahead step of length $h_j = 3$, while, when $h_j \geq 4$, it becomes active.

For such a look-ahead step of length h_j, *the cost in terms of MVs is* $4h_j - 3$ MVs if $h_j > 1$. In fact, in the first substep of the $h_j - 1$ inner and the one regular

LOOP 4.1 (look-ahead for $(3,3)$-type LTPM). *Let* $\mu := \min\{n_j, n-1\}$.

Inner Loop: $(n_j < n+1 < n_{j+1})$

1. *Compute* Aw_n^n.
2. *If* $n > n_j$, *use* (4.2) *to compute indirectly* $Aw_{n_j}^n, \ldots, Aw_{n-1}^n$.
3. *If* $n \geq n_j + 2$, *compute* Aw_n^μ.
4. *If* $n \geq n_j + 3$, *use* (3.4) *to compute indirectly* $Aw_n^{\mu+1}, \ldots, Aw_n^{n-2}$.
5. *Use* (4.2) *to compute* $w_{n+1}^n, w_{n+1}^{n-1}, \ldots, w_{n+1}^\mu$.
6. *Compute* Aw_{n+1}^n *and* ξ_n, η_n.
7. *Use* (3.4) *to compute* $w_{n_j}^{n+1}, \ldots, w_{n+1}^{n+1}$.
8. *Compute* $Aw_{j-1}'^n$ *and use* (4.7) *to compute* $w_{j-1}'^{n+1}$.

Regular Loop: $(n+1 = n_{j+1})$

1. *Compute* Aw_n^n.
2. *If* $n > n_j$, *use* (4.2) *to compute indirectly* $Aw_{n_j}^n, \ldots, Aw_{n-1}^n$.
3. *If* $n \geq n_j + 2$, *compute* Aw_n^μ.
4. *If* $n \geq n_j + 3$, *use* (3.4) *to compute indirectly* $Aw_n^{\mu+1}, \ldots, Aw_n^{n-2}$.
5. *Use* (4.2) *to compute* w_{n+1}^n *and* w_{n+1}^{n-1}.
6. *Compute* Aw_{n+1}^n *and* ξ_n, η_n.
7. *Use* (3.4) *to compute* $w_{n_j}^{n+1}, \ldots, w_{n+1}^{n+1}$.
8. *Compute* $w_j'^n$ *and* $w_j'^{n+1}$ *according to definition* (4.1).

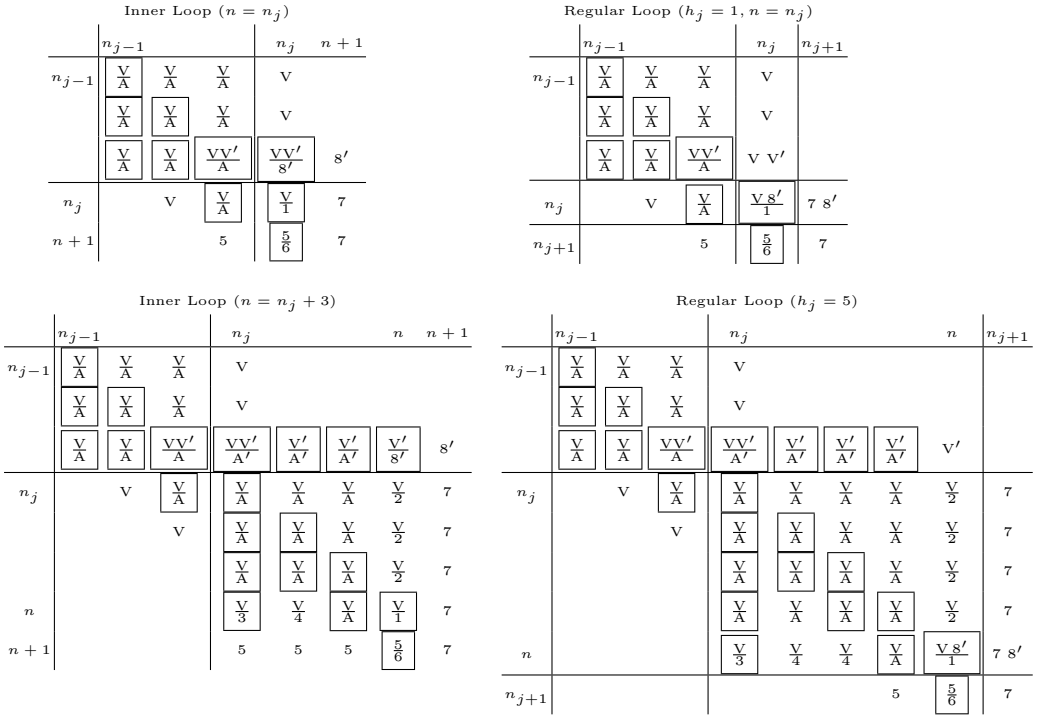


FIG. 4.2. *Action of Loop* 4.1 *(look-ahead for* $(3,3)$*-type LTPM) in the w-table. For blocks of different sizes an inner step (at left) and the following regular step (at right) are shown.*

loops h_j MVs are consumed; another $h_j - 2$ MVs are needed in substep 3, a further h_j MVs in substep 6, and the remaining $h_j - 1$ MVs are used in substep 6 of the inner loops. If no look-ahead is needed, that is, if $h_j = 1$, then 2 MVs are required, both in our procedure and in the standard one that does not allow for look-ahead. Hence, in a step of length 2 we have 25% overhead, in a step of length 3 there is 50% overhead, and for even longer steps, which are very rare in practice, the overhead grows gradually toward 100%.

4.2. Look-ahead for BiOStab and BiOStab2. Since the horizontal recurrence (3.8) for BiOStab is only a two-term one, there is no need to compute elements of the second subdiagonal of the w-table as long as we are not in a look-ahead step. In the case of a look-ahead step, this remains true for those elements that lie in a subdiagonal block, but, of course, not for those in a diagonal block. For Loop 4.1 this means that μ simplifies to $\mu := n_j$ and that in substep 5 of a regular loop there is no need to compute w_{n+1}^{n-1}. All the other changes refer to equation numbers or the coefficients ξ_n and η_n. In summary, we obtain Loop 4.2. Again the first four substeps are the same in both cases, and the choice between them will be made in substep 5. In Figure 4.3 we display for this loop the two sections of w-tables that correspond to the second pair in Figure 4.2.

Since those product vectors from Loop 4.1 that are no longer needed in Loop 4.2 were found without an extra MV before, the overhead in terms of MVs remains the same here.

Look-ahead for BiOStab2 could be defined along the same lines by alternating between steps of Loop 4.1 and Loop 4.2. At this point it also becomes clear how to obtain a look-ahead version of BiOStab(ℓ), an algorithm analogous to BiCGStab(ℓ) of Sleijpen and Fokkema [38], but based on the three-term Lanczos process instead of coupled two-term BiCG formulas.

4.3. Look-ahead (bi)conjugate gradient squared: LABiOS. For BiOS, which will be the underlying process for our BiOResS version of BiCGS, the horizontal recurrence (3.4) has to be substituted by the Lanczos recurrence given in (3.16), which may need to be replaced by the look-ahead formula that is analogous to (4.2) with l and n exchanged and with suitably defined auxiliary vectors and blocks of vectors. But since the w-table is symmetric, we can build it up by vertical recursions only and reflections at the diagonal.

We only formulate the recursion for $w'^{\,l}_{j-1}$ as a horizontal one that replaces (4.7): if $n_k < l + 1 \le n_{k+1}$,

$$w'^{\,l+1}_{j-1} = \left(A w'^{\,l}_{j-1} - \widehat{W}'^{\,k}_{j-1} \alpha_l - W'^{\,k-1}_{j-1} \beta_l \right) / \gamma_l \,, \tag{4.9}$$

where now

$$\widehat{W}'^{\,k}_{j-1} := \left[\begin{array}{ccc} w'^{\,n_k}_{j-1} & \dots & w'^{\,l}_{j-1} \end{array} \right], \quad W'^{\,k-1}_{j-1} := \left[\begin{array}{ccc} w'^{\,n_{k-1}}_{j-1} & \dots & w'^{\,n_k-1}_{j-1} \end{array} \right]. \tag{4.10}$$

Recurrence (4.9) follows from the horizontal Lanczos look-ahead recurrences for computing $w^{l+1}_{n_{j-1}}, \dots, w^{l+1}_{n_j-1}$, derived from (2.13) instead of (2.21), which can be gathered into one recurrence for these h_{j-1} columns of W^{l+1}_{j-1} and then postmultiplied by $D^{-1}_{j-1}\ell$, as in (4.1).

Again, (4.9) can be simplified: if we define for block $k-1$ and the needed values of j the auxiliary vector

$$w''^{\,k-1}_{j-1} := W'^{\,k-1}_{j-1} D^{-1}_{k-1} \ell, \tag{4.11}$$

LOOP 4.2 (look-ahead for BiOStab).

Inner Loop: $(n_j < n+1 < n_{j+1})$

1. *Compute* Aw_n^n.
2. *If* $n > n_j$, *use* (4.2) *to compute indirectly* $Aw_{n_j}^n, \ldots, Aw_{n-1}^n$.
3. *If* $n \geq n_j + 2$, *compute* $Aw_n^{n_j}$.
4. *If* $n \geq n_j + 3$, *use* (3.8) *to compute indirectly* $Aw_n^{n_j+1}, \ldots, Aw_n^{n-2}$.
5. *Use* (4.2) *to compute* $w_{n+1}^n, w_{n+1}^{n-1}, \ldots, w_{n+1}^{n_j}$.
6. *Compute* Aw_{n+1}^n *and* χ_n.
7. *Use* (3.8) *to compute* $w_{n_j}^{n+1}, \ldots, w_{n+1}^{n+1}$.
8. *Compute* $Aw_{j-1}'^n$ *and use* (4.8) *to compute* $w_{j-1}'^{n+1}$.

Regular Loop: $(n+1 = n_{j+1})$

1. *Compute* Aw_n^n.
2. *If* $n > n_j$, *use* (4.2) *to compute indirectly* $Aw_{n_j}^n, \ldots, Aw_{n-1}^n$.
3. *If* $n \geq n_j + 2$, *compute* $Aw_n^{n_j}$.
4. *If* $n \geq n_j + 3$, *use* (3.8) *to compute indirectly* $Aw_n^{n_j+1}, \ldots, Aw_n^{n-2}$.
5. *Use* (4.2) *to compute* w_{n+1}^n.
6. *Compute* Aw_{n+1}^n *and* χ_n.
7. *Use* (3.8) *to compute* $w_{n_j}^{n+1}, \ldots, w_{n+1}^{n+1}$.
8. *Compute* $w_j'^n$ *and* $w_j'^{n+1}$ *according to definition* (4.1).

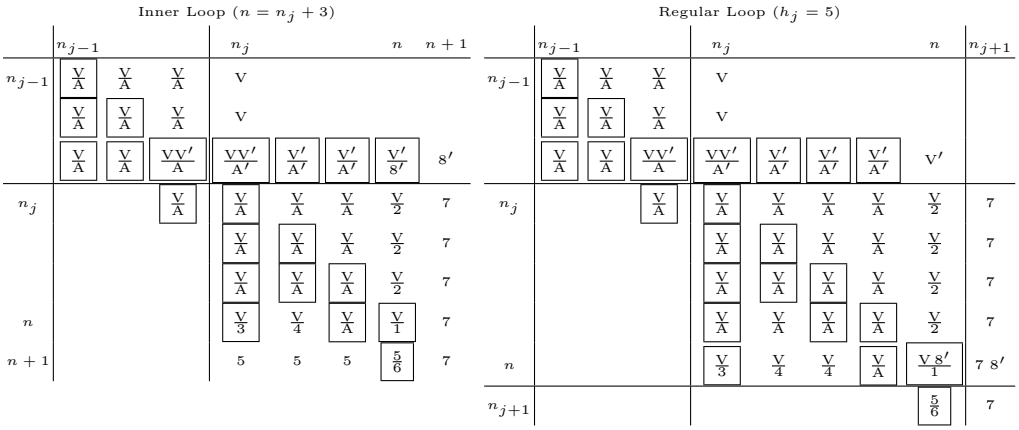


FIG. 4.3. *Action of Loop* 4.2 *(look-ahead for BiOStab) in the w-table. An inner step (at left) and the following regular step (at right) are shown.*

and for each l with $n_k \leq l < n_{k+1}$ the same coefficient $\beta_l' := \delta_l^{n_k}\gamma_{n_k-1}$ as in (4.6), then in view of (2.17) and (4.6),

$$W_{j-1}'^{k-1}\beta_l = W_{j-1}'^{k-1}D_{k-1}^{-1}\ell\beta_l' = w_{j-1}''^{k-1}\beta_l'\,. \tag{4.12}$$

Consequently, (4.9) simplifies to

$$w_{j-1}'^{l+1} = \left(Aw_{j-1}'^{l} - \widehat{W}_{j-1}'^{k}\alpha_l - w_{j-1}''^{k-1}\beta_l'\right)/\gamma_l\,. \tag{4.13}$$

A step of the resulting LABiOS algorithm is summarized as Loop 4.3.

Here, the decision between a regular and an inner loop is made in substep 3 (which corresponds to substep 5 in Loops 4.1 and 4.2). Now even the first five substeps of the two versions of the loop are identical. In Figure 4.4 we display for LABiOS the same pair of loops as we did in Figure 4.3 for LABiOStab.

LOOP 4.3 (look-ahead for BiOS).

Inner Loop: ($n_j < n+1 < n_{j+1}$)

1. *Compute* Aw_n^n.
2. *If* $n \geq n_j+2$, *use* (4.2) *to compute indirectly* $Aw_{n_j}^n, \ldots, Aw_{n-2}^n$.
3. *Use* (4.2) *to compute* $w_{n+1}^n, w_{n+1}^{n-1}, \ldots, w_{n+1}^{n_j}$.
4. *Compute* Aw_{n+1}^n *and* $Aw_{j-1}'^{n}$.
5. *Use* (4.13) *to compute* $w_{j-1}'^{n+1}$.
6. *Use* (4.2) *to compute* w_{n+1}^{n+1}.

Regular Loop: ($n+1 = n_{j+1}$)

1. *Compute* Aw_n^n.
2. *If* $n \geq n_j+2$, *use* (4.2) *to compute indirectly* $Aw_{n_j}^n, \ldots, Aw_{n-2}^n$.
3. *Use* (4.2) *to compute* $w_{n+1}^n, w_{n+1}^{n-1}, \ldots, w_{n+1}^{n_j}$.
4. *Compute* Aw_{n+1}^n *and* $Aw_{j-1}'^{n}$.
5. *Use* (4.13) *to compute* $w_{j-1}'^{n+1}$.
6. *Use definition* (4.1) *to compute* $w_j'^{n_j}, \ldots, w_j'^{n_{j+1}-1}$ *and use definition* (4.11) *to compute* $w_j''^{j}$.
7. *Use* (4.2) *to compute* w_{n+1}^{n+1}.
8. *Use definition* (4.1) *to compute* $w_j'^{n+1}$.

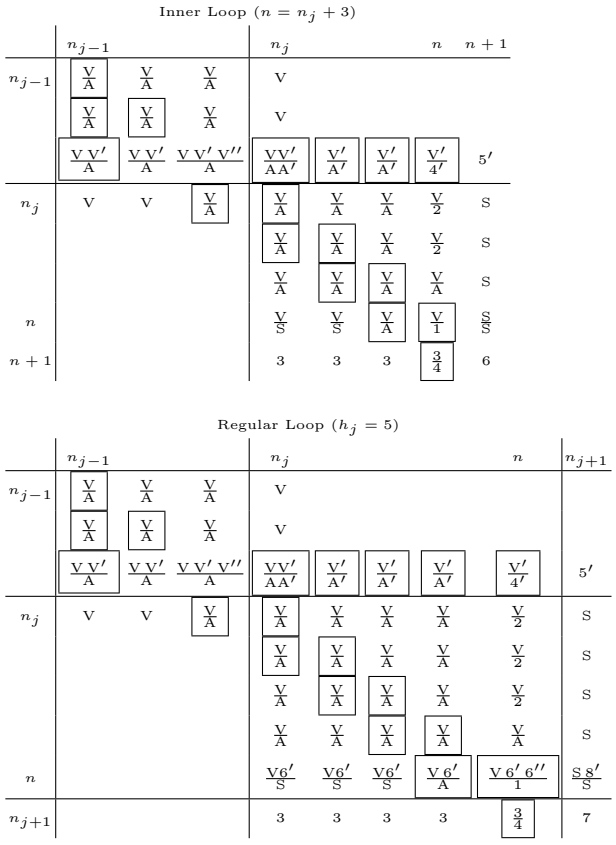


FIG. 4.4. *Action of Loop* 4.3 *(look-ahead for BiOS) in the w-table. An inner loop (at the top) and the following regular loop (at the bottom) are shown.*

TABLE 4.1

Cost and overhead of look-ahead LTPMs when constructing a look-ahead step of length $h_j > 1$. The construction of iterates is not yet included. By further capitalizing upon storage locations that become available during a look-ahead step, the storage overhead could be reduced by roughly 50%.

Method	Total cost	Cost overhead		Storage overhead
	(MVs, IPs)	(MVs, IPs)	relative	(N-vectors)
$(3,3)$-type LA-LTPM	$4h_j - 3$	$2h_j - 3$	25%–100%	$2h_j^2 + 3h_j + 5$
LABiOStab	$4h_j - 3$	$2h_j - 3$	25%–100%	$2h_j^2 + 3h_j + 3$
LABiOS (if $h_{j-1} > 1$)	$3h_j$	h_j	50%	$h_j^2 + 4h_j + 4$
LABiOS (if $h_{j-1} = 1$)	$3h_j - 1$	$h_j - 1$	25%–50%	$h_j^2 + 4h_j + 4$

For a look-ahead step of length h_j, *the cost in terms of MVs is now* $3h_j$ MVs if $h_j > 1$ and $h_{j-1} > 1$, while it is only $3h_j - 1$ if $h_{j-1} = 1$. As the standard, non-look-ahead algorithm requires 2 MVs per step, this means that the overhead is at most 50%. Here, in the first substep of the $h_j - 1$ inner and the one regular loops h_j MVs are consumed, and another $2h_j$ MVs are needed in substep 4.

4.4. Overhead for look-ahead. In this subsection we further discuss the overhead of the look-ahead process in terms of MVs, inner products, and the necessary storage of N-vectors.

First we note that all inner products required in the look-ahead algorithms have the fixed initial vector $\widetilde{z}_0$ as their first argument. Therefore, if the second argument of an inner product was computed by a recurrence formula, then also the inner product of this vector with $\widetilde{z}_0$ can be computed indirectly by applying the same recurrence formula. Thus, only inner products of the form $\langle \widetilde{z}_0, Aw_n^l \rangle$ for which Aw_n^l is computed directly need to be computed explicitly. This means that in all algorithms the number of required inner products is equal to the number of required MVs. We must admit, however, that such recursively computed inner products, as well as the recursively computed MVs, may be the source of additional roundoff, which may cause instability.

Actually, to understand how to compute all the inner products needed, the reader may want to introduce a δ-table and a σ-table with entries $\delta_n^l := \langle \widetilde{z}_0, w_n^l \rangle$ and $\sigma_n^l := \langle \widetilde{z}_0, Aw_n^l \rangle$, respectively. Our loops and figures for generating the w-table then hold as well for these two tables. Note that the δ-table just contains the transpose of the matrix D.

In terms of MVs, the cost of a look-ahead step of length $h_j > 1$ has been specified in the previous subsection. By subtracting the cost for h_j non-look-ahead steps, that is, $2h_j$ MVs, we obtain the overhead summarized in Table 4.1, where IP stands for inner product. We stress that when $h_j = 1$ our algorithm has no overhead except for the necessary test of regularity, which, if it fails, would reveal an upcoming instability and initiate a look-ahead step. The table lists additionally the overhead in storage of N-vectors in a straightforward implementation that is not optimized with respect to memory usage.

For comparison, we cite from page 60 of [5] or page 180 of [6] that the CGS look-ahead procedure of Brezinski and Redivo Zaglia requires $6h_j - 3$ MVs if $h_j \leq n_j + 1$ (the typical case) and $5h_j + n_j - 2$ MVs if $h_j > n_j + 1$ (which means that a relatively large look-ahead step is needed in one of the first few iterations). Therefore, compared to our numbers, the overhead in terms of MVs is about four times (if h_j is large, but $h_j \leq n_j + 1$) to five times (if $h_j = 2$) larger than in our LABiOS (assuming as the basis a standard CGS implementation requiring 2 MVs per ordinary iteration). However,

we also note that, according to the above numbers, for a step without look-ahead ($h_j = 1$), the methods of [5] and [6] need 3 MVs instead of 2 MVs.

5. Look-ahead strategies. In this section we address the delicate issue of when to perform a look-ahead step, which means in an LTPM to decide whether the new vertical index $n+1$ is a regular index or an inner index, i.e., whether the required product vectors w_{n+1}^l in the $(n+1)$th row of the w-table should be computed as regular or as inner vectors. Therefore, the look-ahead procedure in LTPMs serves to stabilize the underlying Lanczos method, the vertical movement of the product method in the w-table. Consequently, the criterion of when to carry out a look-ahead step in an LTPM can be based on the criterion given in [15] for the Lanczos algorithm. However, since the Lanczos vectors $\widetilde{y}_n$ and y_n are not computed explicitly in a product method, we need to rewrite the conditions of this criterion in terms of the product vectors w_n^l. Let us first motivate these conditions.

In the case of an *exact breakdown*, where $0 = \delta_n^\tau = \langle \widetilde{z}_0, w_n^n \rangle = \langle \widetilde{z}_n, y_n \rangle$ and $\widetilde{z}_n \neq 0, y_n \neq 0$, a division by zero would occur in the next Lanczos step. The first task of the look-ahead process is to circumvent these exact breakdowns without the necessity of restarting the Lanczos process and losing its superlinear convergence.

In finite-precision arithmetic, exact breakdowns are very unlikely. However, *near-breakdowns*, where $|\delta_n^\tau|$ is very small, may occur and cause large relative roundoff errors in the Lanczos coefficients α_n and β_n given by (3.2). To be more precise, we recall that the relative roundoff error in the computation of the inner product $\delta_n^\tau = \langle \widetilde{z}_0, w_n^n \rangle$ is bounded by [18, p. 64]

$$\frac{|\mathrm{fl}(\delta_n^\tau) - \delta_n^\tau|}{|\delta_n^\tau|} \leq 1.01N\varepsilon \frac{\langle |\widetilde{z}_0|, |w_n^n| \rangle}{|\delta_n^\tau|} \leq 1.01N\varepsilon \frac{\|\widetilde{z}_0\| \, \|w_n^n\|}{|\delta_n^\tau|},$$

where ε denotes the roundoff unit. Thus, a small value of $|\delta_n^\tau|$ leads in finite-precision arithmetic to a big relative roundoff error in the computation of the inner product δ_n^τ, which also causes a perturbation of the Lanczos coefficients α_n and β_n, since they depend on δ_n^τ and δ_{n-1}^τ, respectively. The second task of the look-ahead process is therefore to avoid a convergence deterioration due to perturbed Lanczos coefficients. Of course, similar roundoff effects may come up in the numerators of the formulas (3.2) for α_n and β_n, but large relative errors in those will only be harmful if the denominators are small too.

We would like to point out that in an LTPM the inner products δ_n^τ can be enlarged to a certain extent by an appropriate adaptive choice of the polynomials τ_n [39], [40] as long as $\langle \left(A^H\right)^n \widetilde{z}_0, y_n \rangle \neq 0$. For example, considering the BiOStab case, where $w_n^n = w_n^{n-1} - Aw_n^{n-1}\chi_{n-1}$, we obtain, since $\langle \widetilde{z}_0, w_n^{n-1} \rangle = 0$,

$$\frac{\|\widetilde{z}_0\| \, \|w_n^n\|}{|\delta_n^\tau|} = \frac{\left\|w_n^{n-1} - Aw_n^{n-1}\chi_{n-1}\right\|}{|\chi_{n-1}|} \frac{\|z_0\|}{\langle \widetilde{z}_0, Aw_n^{n-1} \rangle}.$$

Thus, minimizing the relative roundoff error in the calculation of δ_n^τ is equivalent to choosing χ_{n-1} such that it minimizes $\left\|w_n^{n-1} - Aw_n^{n-1}\chi_{n-1}\right\| / |\chi_{n-1}|$. This leads to $\chi_{n-1}^{OR} := \left\|w_n^{n-1}\right\|^2 / \langle w_n^{n-1}, Aw_n^{n-1} \rangle$, which corresponds to the use of orthogonal residual polynomials of degree 1 instead of minimal residual polynomials of degree 1 in the recursive definition of $\tau_l(\zeta)$. Therefore, minimizing the relative roundoff error in the inner product δ_n^τ conflicts often with the objective of avoiding large intermediate residuals in order to prevent the recursive residual from drifting apart from the true residual [40]. Performing a look-ahead step is then the only possible remedy.

We need now to find a criterion for deciding when a look-ahead step should be performed so that both the above objectives can be attained. In view of the recursion (4.2) for the product vectors in the case of look-ahead, it follows that a block can be closed and a new product vector can be computed as a regular vector only if the diagonal blocks D_{j-1} and D_j of the Gramian are numerically nonsingular. Thus, the first condition that needs to be fulfilled in order to compute a new product vector w_{n+1}^l as a regular vector ($n+1 = n_{j+1}$) is given by

$$\sigma_{\min}(D_j) \geq \varepsilon, \tag{5.1}$$

where $\sigma_{\min}(D_j)$ denotes the minimal singular value of the block D_j. Note that this does not mean that D_j is well conditioned; it just guarantees numerical nonsingularity. In practice, (5.1) can be replaced by some other condition that implies this nonsingularity, for example by one implemented in a linear solver used for computing $D_{j-1}^{-1}\ell$ in (2.19) and α_n in (2.22).

Freund et al. [15] use a second condition to guarantee that the Krylov space is stably extended in the next Lanczos step in the sense that the basis of Lanczos vectors is sufficiently well conditioned. In terms of product vectors, this second condition amounts to computing, for any l, the new product vector w_{n+1}^l as a regular vector if in addition to (5.1) the following conditions for the coefficients α_n and β_n are fulfilled:

$$\|\alpha_n\|_1 \leq n(A), \qquad \|\beta_n\|_1 \leq n(A). \tag{5.2}$$

Here $\|\cdot\|_1$ denotes the ℓ_1-norm and $n(A)$ is an estimate for $\|A\|$ which is updated dynamically to ensure that the blocks W_j^n do not become larger than a user-specified maximal size [15]. The motivation for (5.2) was to ensure that in the new regular vector

$$w_{n+1}^n = \left(Aw_n^n - W_j^n\alpha_n - W_{j-1}^n\beta_n\right)/\gamma_n = \left(Aw_n^n - W_j^n\alpha_n - w_{j-1}'^n\beta_n'\right)/\gamma_n$$

(obtained from (4.2) with $\widehat{W}_j^n = W_j^n$ since $n+1 = n_{j+1}$ is regular) the component in the new direction Aw_n^n is sufficiently large, which will be the case if

$$\|Aw_n^n\| \geq \mathrm{tol}_2\ \|w_t\|, \tag{5.3}$$

where $w_t := W_j^n\alpha_n + W_{j-1}^n\beta_n = W_j^n\alpha_n - w_{j-1}'^n\beta_n'$ and tol_2 is a chosen tolerance.

Compared to (5.2), condition (5.3) costs two additional inner products and the calculation of w_t, which can only be reused in the regular case for the computation of the new product vector w_{n+1}^n. Since we have

$$\|w_t\| \leq C\left(\|\alpha_n\|_1 + |\beta_n'|\right) \quad \text{with} \quad C := \max\left\{\left\|w_{j-1}'^n\right\|;\ \|w_k^n\| : n_j \leq k < n_{j+1}\right\}, \tag{5.4}$$

we could replace (5.3) by the less expensive condition

$$\|Aw_n^n\| \geq \mathrm{tol}_2\, C\left(\|\alpha_n\|_1 + |\beta_n'|\right). \tag{5.5}$$

However, in an LTPM it is not possible to normalize all product vectors w_n^l (see section 6), so C is not equal to 1. Moreover, (5.5) is less strict than (5.3) and (5.2).

Since a look-ahead step is more expensive than regular steps providing the same increase of the Krylov space dimension, a tight look-ahead criterion can save overall computational cost. Therefore, it is reasonable to spend extra effort for it. For this reason we favor criterion (5.3).

A drawback of (5.3) is that it does not take the angle between Aw_n^n and w_t into account. If $C_c := 1 - |\langle Aw_n^n, w_t\rangle| / (\|Aw_n^n\| \, \|w_t\|)$ is small, tol_2 in (5.3) should be chosen larger than in the case where it is nearly 1. This motivates the choice

$$\mathrm{tol}_2 = \frac{C_1}{1 - (1 - C_2)\left|\frac{\langle Aw_n^n, w_t\rangle}{\|Aw_n^n\|\|w_t\|}\right|} \tag{5.6}$$

with suitably chosen constants $C_1 = C_1(\varepsilon)$ and $C_2 = C_2(\varepsilon)$ depending on the roundoff unit ε. This criterion requires an extra inner product and an appropriate choice for C_1 and C_2. For many small problems $C_1 = 10^{-3}$ and $C_2 = 10^{-2}$ worked well, but for larger problems we observed that C_c decays dramatically with the block size. Therefore, the probability that (5.3) with tol_2 as defined in (5.6) will be fulfilled decreases with the block length and leads very often (especially in BiOS) to situations where the maximal user-specified block size was reached. Further investigations are needed to see if this problem can be solved by a better choice of C_1 and C_2 or by a more appropriate selection for α_n in the inner case, instead of using, as in [15], $\alpha_n^n = 1, \alpha_n^{n-1} = 1$, $\alpha_n^k = 0$ for $k = 1, \ldots, n-2$.

6. Obtaining the solution of $Ax = b$. So far we have only introduced various algorithms for constructing a sequence of product vectors w_n^l that provide a basis for $\mathcal{K}_m$. However, our goal is to solve the linear system $Ax = b$. We now describe how to accomplish this with these algorithms. There are several basic approaches to constructing approximate solutions of linear systems from a Krylov-space basis. Ours is related to the Galerkin method, but avoids the difficulty that arises when the Galerkin solution does not exist. (This difficulty causes, for example, the so-called *pivot breakdown* of the BiCG method.) Our approach is a natural generalization of the one that led to "unnormalized BiORes" introduced in [20] and renamed "inconsistent BiORes" in [25]. In contrast to the BiOMin, BiODir, and BiORes versions of the BiCG method, the inconsistent BiORes variant is not endangered by pivot breakdowns. An alternative would be to construct approximate solutions based on the QMR approach [16]. For a combination of this approach with LTPMs we refer the reader to [36].

Let the doubly indexed sequence of scalars $\dot\rho_n^l$ be given by

$$\dot\rho_n^l := \tau_l(0)\rho_n(0). \tag{6.1}$$

We define a doubly indexed sequence of *product iterates* x_n^l as follows. Starting with an arbitrary $x_0^0 \in \mathbb{C}^N$, we choose the initial product vector w_0^0 so that

$$w_0^0 = b\dot\rho_0^0 - Ax_0^0. \tag{6.2}$$

Here, $\dot\rho_0^0 = 1$ since $\tau_0(\zeta) = \rho_0(\zeta) \equiv 1$. For $l, n > 0$ the product iterates are now implicitly defined by

$$b\dot\rho_n^l - Ax_n^l := w_n^l \qquad \text{so that} \quad b - A\frac{x_n^l}{\dot\rho_n^l} = \frac{w_n^l}{\dot\rho_n^l} \ \text{ if } \ \dot\rho_n^l \neq 0. \tag{6.3}$$

Of course, x_n^l will be constructed only when w_n^l is. If $\dot\rho_n^l \neq 0$, it follows from (6.3) that $x_n^l/\dot\rho_n^l$ can be considered as an approximate solution of $Ax = b$, whose corresponding residual is $w_n^l/\dot\rho_n^l$. In order to derive recursions for the scalars $\dot\rho_n^l$ and the product iterates x_n^l, we introduce the blocks

$$\begin{aligned} X_{j-1}^l &:= \left[\, x_{n_{j-1}}^l \ \cdots \ x_{n_j-1}^l \,\right], & \widehat{X}_j^l &:= \left[\, x_{n_j}^l \ \cdots \ x_n^l \,\right], \\ P_{j-1}^l &:= \left[\, \dot\rho_{n_{j-1}}^l \ \cdots \ \dot\rho_{n_j-1}^l \,\right], & \widehat{P}_j^l &:= \left[\, \dot\rho_{n_j}^l \ \cdots \ \dot\rho_n^l \,\right], \end{aligned}$$

as well as the *auxiliary product iterates* x'^l_{j-1} and *auxiliary scalars* $\dot\rho'^l_{j-1}$ defined by

$$x'^l_{j-1} := X^l_{j-1}D^{-1}_{j-1}\ell, \qquad \dot\rho'^l_{j-1} := P^l_{j-1}D^{-1}_{j-1}\ell. \tag{6.4}$$

Again, $\widehat X^l_j = X^l_j$ and $\widehat P^l_j = P^l_j$ if $n+1$ is a regular index.

Then, by (6.3), (4.1), and (6.4),

$$\widehat W^l_j = b\widehat P^l_j - A\widehat X^l_j, \quad w'^l_{j-1} = b\dot\rho'^l_{j-1} - Ax'^l_{j-1}. \tag{6.5}$$

Next, using (4.2) we conclude that

$$\begin{aligned} b\dot\rho^l_{n+1} - Ax^l_{n+1} &= w^l_{n+1} = \left(Aw^l_n - \widehat W^l_j\alpha_n - w'^l_{j-1}\beta'_n\right)/\gamma_n \\ &= \left(Aw^l_n - \left[b\widehat P^l_j - A\widehat X^l_j\right]\alpha_n - \left[b\dot\rho'^l_{j-1} - Ax'^l_{j-1}\right]\beta'_n\right)/\gamma_n \\ &= b\left[-\widehat P^l_j\alpha_n - \dot\rho'^l_{j-1}\beta'_n\right]/\gamma_n + A\left[w^l_n + \widehat X^l_j\alpha_n + x'^l_{j-1}\beta'_n\right]/\gamma_n. \end{aligned}$$

This shows that

$$x^l_{n+1} = -\left[w^l_n + \widehat X^l_j\alpha_n + x'^l_{j-1}\beta'_n\right]/\gamma_n, \quad \dot\rho^l_{n+1} = -\left[\widehat P^l_j + \dot\rho'^l_{j-1}\beta'_n\right]/\gamma_n. \tag{6.6}$$

If we arrange the product iterates x^l_n and the scalars $\dot\rho^l_n$ in two tables analogous to the w-table (with the n-axis pointing downward and the l-axis to the right), these two recursions can be used to proceed in the vertical direction.

To obtain recursions for a horizontal movement, we assume first that the polynomials $\tau_n(\zeta)$ are given by the normalized three-term recurrence (3.3). This covers all algorithms described in this paper except LABiOS. Using (3.4) and (6.3) we see that the product iterates satisfy

$$x^{l+1}_n = -w^l_n\eta_l + x^l_n\xi_l + x^{l-1}_n(1-\xi_l). \tag{6.7}$$

For the scalars $\dot\rho^l_n$ the recursion $\dot\rho^{l+1}_n = \dot\rho^l_n\xi_l + \dot\rho^{l-1}_n(1-\xi_l)$ is valid, but since the polynomials τ_l are normalized (that is, $\tau_l(0) = 1$ for all l), the scalars $\dot\rho^l_n$ do not change with the index l, and we have simply $\dot\rho^l_n = \dot\rho^0_n = \rho_n(0)$.

In LABiOS only one horizontal movement is explicitly computed per step, namely in substep 5 of Loop 4.3 based on the recurrence (4.13). If we define in analogy to (4.10) and (4.11)

$$\begin{aligned} \widehat X'^k_{j-1} &:= \left[\begin{array}{ccc} x'^{n_k}_{j-1} & \dots & x'^l_{j-1}\end{array}\right], & X'^{k-1}_{j-1} &:= \left[\begin{array}{ccc} x'^{n_{k-1}j-1} & \dots & x'^{n_k-1}_{j-1}\end{array}\right], \\ \widehat P'^k_{j-1} &:= \left[\begin{array}{ccc} \dot\rho'^{n_k}_{j-1} & \dots & \dot\rho'^l_{j-1}\end{array}\right], & P'^{k-1}_{j-1} &:= \left[\begin{array}{ccc} \dot\rho'^{n_{k-1}j-1} & \dots & \dot\rho'^{n_k-1}_{j-1}\end{array}\right], \end{aligned}$$

and

$$x''^{k-1}_{j-1} := X'^{k-1}_{j-1}D^{-1}_{k-1}\ell, \quad \dot\rho''^{k-1}_{j-1} := P'^{k-1}_{j-1}D^{-1}_{k-1}\ell, \tag{6.8}$$

recurrences for the auxiliary iterates and the corresponding scalars are given by

$$\begin{aligned} x'^{l+1}_{j-1} &= -\left[w'^l_{j-1} + \widehat X'^k_{j-1}\alpha_l + x''^{k-1}_{j-1}\beta'_l\right]/\gamma_l, \\ \dot\rho'^{l+1}_{j-1} &= -\left[\widehat P'^k_{j-1}\alpha_l + \dot\rho''^{k-1}_{j-1}\beta'_l\right]/\gamma_l. \end{aligned} \tag{6.9}$$

Using for the scaling parameter γ_n in (4.2) the special choice

$$\gamma_n = \begin{cases} -\mathbf{1}_{h_j}^T \alpha_n - \mathbf{1}_{h_{j-1}}^T \beta_n & \text{if } n+1 = n_{j+1}, \\ -\mathbf{1}_{n-n_j+1}^T \alpha_n - \mathbf{1}_{h_{j-1}}^T \beta_n & \text{if } n_j < n+1 = n_{j+1}, \end{cases} \tag{6.10}$$

with $\mathbf{1}_m := [1 \cdots 1]_{m\times 1}^T$, the Lanczos polynomials ρ_n also could be normalized ($\rho_n(0) = 1$) so that $\dot{\rho}_n^l = 1$ for all $n, l \geq 0$. However, as we mentioned before, some γ_n might turn out to be zero, which would lead to a so-called pivot breakdown. Moreover, to avoid overflow or underflow, in the Lanczos process the scaling parameter γ_n is often used to normalize the Lanczos vectors y_n. But since the Lanczos vectors y_n are not explicitly computed in an LTPM, we cannot base the choice of γ_n here on their norm. However, independent of the size of the blocks generated by the look-ahead process, it is always necessary to compute the product vectors w_{n+1}^n in an LTPM. Therefore, we choose here γ_n to normalize w_{n+1}^n, that is,

$$\gamma_n := \left\| A w_n^n - \widehat{W}_j^n \alpha_n - W_{j-1}^n \beta_n \right\| \quad \text{if } n_j < n+1 \leq n_{j+1}. \tag{6.11}$$

7. Numerical examples. In this section we demonstrate the practical performance of our look-ahead versions of LTPMs in numerical examples. The tests are restricted to BiOStab, BiOxMR2, and BiOS. The look-ahead versions of these LTPMs are denoted by LABiOStab, LABiOxMR2, and LABiOS, respectively. For all tests the initial iterate $x_0 = 0$ is used, and the iteration is terminated when the norm of the recursive residual is less than $\sqrt{\varepsilon}$, the square root of the roundoff unit. The test programs were written in FORTRAN90/95 and run on workstations with 64-bit IEEE arithmetic. We start with small, artificially constructed model problems and move gradually to large real-world problems.

Example 1. The following small test example was proposed by Joubert [29] and also used by Brezinski and Redivo Zaglia [6]:

$$A := \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 3 & -1 \\ 0 & 0 & 1 & 3 \end{bmatrix}, \quad b := \begin{bmatrix} 0 \\ 2 \\ 2 \\ 4 \end{bmatrix}, \tag{7.1}$$

$x_0 = 0$, and $\widetilde{z}_0 = [1, 1, 1, 1]^T$. The Lanczos process and hence BiOStab, BiOxMR2, and BiOS without look-ahead break down at step 2. On the contrary, all look-ahead versions avoid this breakdown and converge after four iterations as shown in our plots of the true residual norms $\|b - Ax_n^l \frac{1}{\dot{\rho}_n^l}\|$ in Figure 7.1.

Example 2. Our second example is taken from [6, Example 5.2], the matrix

$$A := \begin{bmatrix} 2 & 1 & & & \\ 0 & 2 & 1 & & \\ 1 & 0 & 2 & 1 & \\ & 1 & 0 & 2 & \ddots \\ & & \ddots & \ddots & \ddots \end{bmatrix} \tag{7.2}$$

of order 400 and the right-hand side $b = (3, 3, 4, \ldots, 4, 3, 3)^T$ imply that the solution is $x = (1, 1, \ldots, 1)^T$. With $x_0 = 0$, $\widetilde{z}_0 = (0, 0, 0, -1, 1, \ldots, 0)^T$ the Lanczos process, and thus the LTPMs, break down in the first iteration. Our look-ahead versions perform

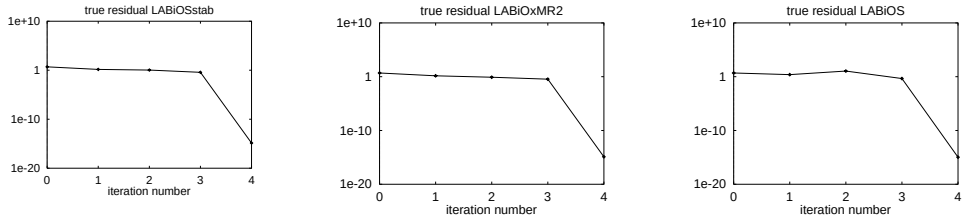


FIG. 7.1. *The true residual norm history (i.e.,* $\log(\|b - Ax_n^l \frac{1}{\rho_n^l}\|)$ *vs. n) for the linear system defined in* (7.1) *solved by different LTPMs with look-ahead.*

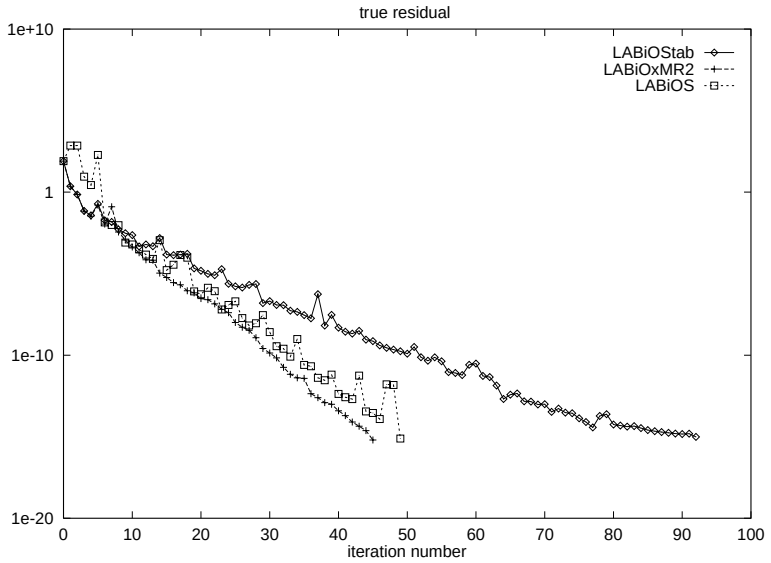


FIG. 7.2. *The true residual norm history (i.e.,* $\log(\|b - Ax_n^l \frac{1}{\rho_n^l}\|)$ *vs. n) for the linear system defined in* (7.2) *solved by different LTPMs with look-ahead.*

two inner steps at the first and second iterations. All following iterations are regular steps. This is a typical behavior: there are only few and short look-ahead steps, and therefore the resulting mean overhead per iteration is nearly negligible. The residual norm history is shown in Figure 7.2.

In the next set of examples we consider p-cyclic matrices of the form

$$A := \begin{bmatrix} I_1 & & & B_1 \\ B_2 & I_2 & & \\ & \ddots & \ddots & \\ & & B_p & I_p \end{bmatrix}. \tag{7.3}$$

Hochbruck [27] showed that the computational work for solving a linear system with a p-cyclic system matrix by QMR with look-ahead can be reduced by approximately a factor $1/p$ (compared to a straightforward implementation using sparse matrix-vector multiplications with A) if the initial Lanczos vectors have only one nonzero block conforming to the block structure of A, if the inner vectors are chosen so that the nonzero structure of $(A - I)y_n$ is not destroyed, and if the blocks B_k are used for generating only possibly nonzero components of the Krylov space basis. Then it can

TABLE 7.1
Indices of regular steps in LTPMs for three problems with a p-cyclic system matrix.

Example	LABiOStab	LABiOxMR2	LABiOS
3	1,5,6,10,11,15	1,5,6,10,11,15	1,5,6,10,11,15
4	1,4,5,8,9,12,13,16	1,4,5,8,9,12,13	1,4,5,8,9,12,13,16
5	1,8,9,16,17	1,8,9,16,17	1,8,9,16,17

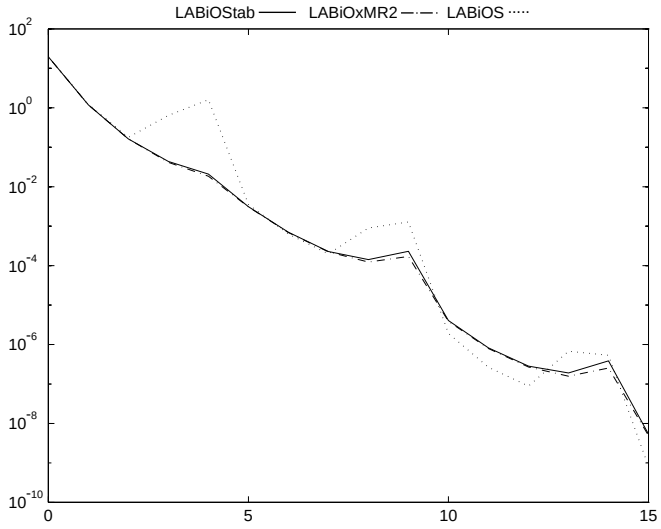


FIG. 7.3. *The true residual norm history (i.e.,* $\log(\|b - Ax_n^l \frac{1}{\rho_n^l}\|)$ *vs. n) for the linear system with the* 5*-cyclic system matrix defined in Example* 3 *solved by different LTPMs with look-ahead.*

be proved that in each cycle of p steps there are at least $p-2$ consecutive exact breakdowns for $p > 2$. But when using directly the system matrix A to generate the Krylov subspace, we have only in the first cycle of p steps $p-2$ consecutive exact breakdowns, while in the following cycles these will, in general, no longer persist, but must be expected to become near-breakdowns. Therefore, such problems provide good test examples for look-ahead algorithms.

Example 3. In this example we consider a 5-cyclic matrix with $B_k = B$ for $k = 1, \ldots, 5$, where B is a randomly generated 10×10 matrix. As right-hand side and as initial left Lanczos vector we choose

$$b = \begin{bmatrix} f_1 \\ 0 \end{bmatrix}, \qquad \widetilde{z}_0 = \begin{bmatrix} g_1 \\ 0 \end{bmatrix}, \tag{7.4}$$

where f_1, g_1 have random entries. The convergence history for the different LTPMs applied to this problem are shown in Figure 7.3, and the indices of the regular steps are listed in Table 7.1. Those are found exactly where predicted.

Example 4. We move now to a bigger 4-cyclic system matrix with $B_k = B$ for $k = 1, \ldots, 4$, where B is a 100×100 matrix with random entries. The results for this problem are shown in Figure 7.4, and the indices of the regular steps are depicted

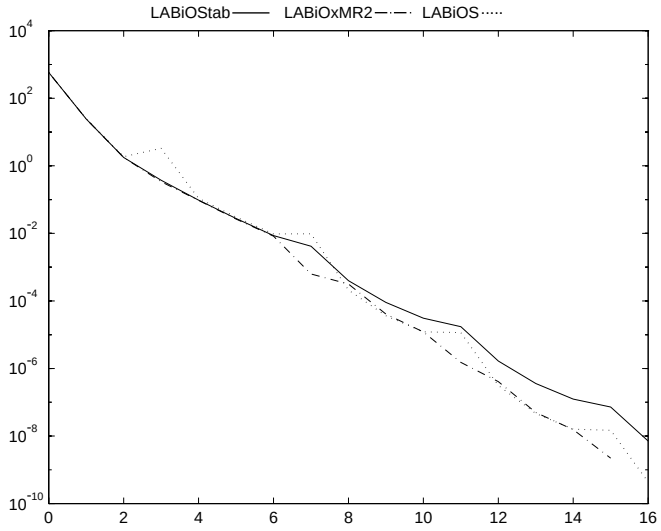


FIG. 7.4. *The true residual norm history (i.e.,* $\log(\|b - Ax_n^l \frac{1}{\rho_n^l}\|)$ *vs. n) for the linear system with the* 4*-cyclic system matrix defined in Example* 4 *solved by different LTPMs with look-ahead.*

also in Table 7.1. Again, they occur where predicted.

Example 5. Finally, we consider an 8-cyclic system matrix with B defined as in Example 4. The convergence history plotted in Figure 7.5 shows oscillations in the residual norm history of LABiOS, but overall LABiOS needs one iteration step fewer than LABiOStab and LABiOxMR2 to fulfill the convergence condition. For all methods the same look-ahead criterion ((5.3) with tol_2 defined as in (5.6) and $C_1 = 10^{-3}$, $C_2 = 10^{-2}$) is used. Especially for LABiOS the correct choice of the look-ahead criterion seems to be crucial. While with the above values of C_1 and C_2 the breakdowns occurred only where expected, we discovered for this larger problem that in BiOS the maximal user-specified block length of 10 was reached very often for $C_1 \gg 10^{-3}$ and $C_2 \ll 10^{-2}$, which indicates that the constructed inner vectors become more and more linearly dependent. Therefore, further investigations are needed to figure out a better choice for the coefficient vector α_n in the inner case. For example, following the proposal in [21], Hochbruck used in [27] a Chebyshev iteration for the generation of the inner vectors instead of $\alpha_n^n = 1, \alpha_n^{n-1} = 1$, and $\alpha_n^k = 0$ for $k = 1, \ldots, n-1$, which we adapted from [15].

Example 6. In our last example we take a real-world problem from the Harwell-Boeing Sparse Matrix Collection, namely SHERMAN1, a matrix of order 1000 with 3750 nonzero entries. The right-hand side b and the initial left Lanczos vector $\widetilde{z}_0$ were generated as different unit vectors with random entries. Without look-ahead, all LTPMs introduced here break down (BiOStab at step 186, BiOStab2 at step 176, BiOxMR2 at step 145, and BiOS at step 352). On the contrary, the look-ahead versions in combination with the look-ahead criterion (5.3) with tol_2 defined as in (5.6) and $C_1 = 10^{-3}$ and $C_2 = 10^{-2}$ converge as shown in Figure 7.6. Due to the real spectrum of the system matrix A, there is only a slight difference in the convergence of LABiOStab and LABiOxMR2. It was reported to us that BiCGStab(ℓ) with $\ell = 8$ and no look-ahead can handle this problem in about the same number of MVs as our BiCGStab with look-ahead.

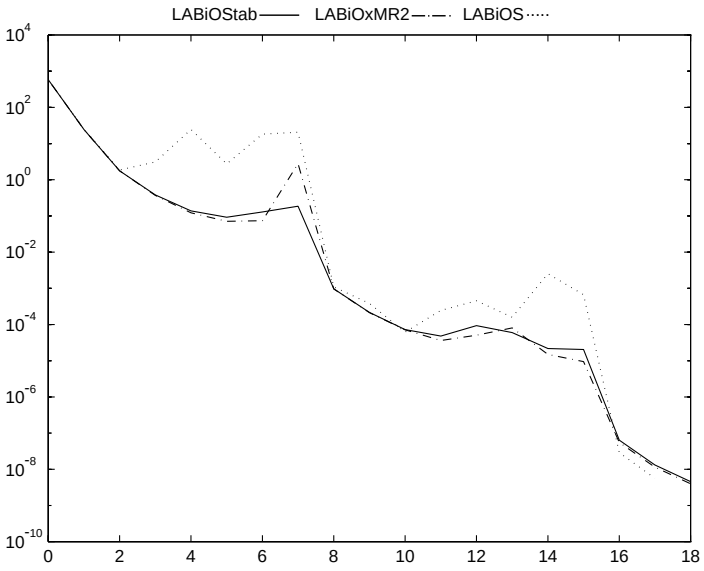


FIG. 7.5. *The true residual norm history (i.e.,* $\log(\|b - Ax_n^l \frac{1}{\hat{\rho}_n^l}\|)$ *vs. n) for the linear system with the* 8*-cyclic system matrix defined in Example* 5 *solved by different LTPMs with look-ahead.*

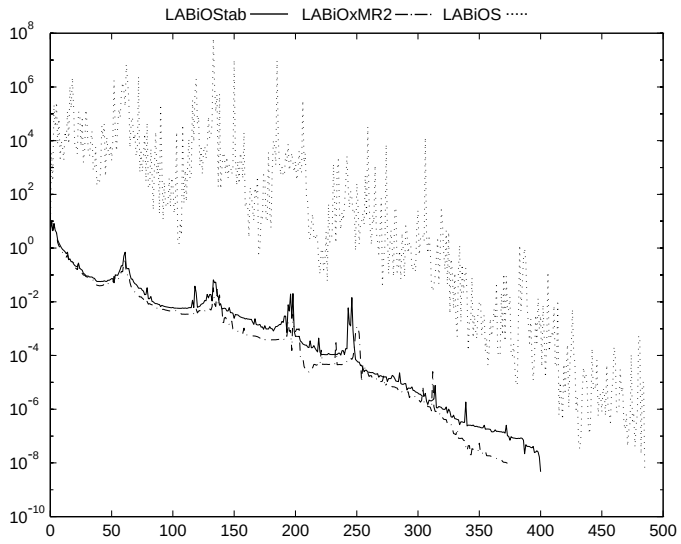


FIG. 7.6. *The true residual norm history (i.e.,* $\log(\|b - Ax_n^l \frac{1}{\hat{\rho}_n^l}\|)$ *vs. n) for a real-world problem with the SHERMAN*1 *matrix from the Harwell-Boeing collection solved by different LTPMs with look-ahead.*

8. Conclusions. We have proposed look-ahead versions for various Lanczos-type product methods that make use of the Lanczos three-term recurrences. Since they are based on the Lanczos look-ahead version of Gutknecht [23] and Freund et al. [15], they can handle look-ahead steps of any length and avoid steps that are longer than needed. The algorithms proposed in this work should be easy to understand due

to the introduction of an array of product vectors, symbolically displayed in the w-table, and the visualization of the progress in this w-table. Furthermore, the w-table proved to be a useful tool to derive optimal variants for which the computational work in terms of MVs is minimized.

A variety of numerical examples demonstrate the practical performance of the proposed algorithms. However, larger problems indicate that further work should be directed to finding an improved look-ahead criterion that more reliably avoids critical perturbations of the Lanczos coefficients by roundoff errors. Moreover, one should investigate if there is a better way of constructing inner vectors than the choice adapted from [15]. Alternatives would be to orthogonalize them within each block [2] or to construct them by Chebyshev iteration [21], but it is not clear if the additional cost involved pays off.

The look-ahead process in an LTPM stabilizes primarily the vertical movement in the w-table, except in the BiOS algorithm where the w-table is symmetric. For the horizontal movement it is also important to generate the Krylov space stably, and both BiOStab2 and BiOxMR2 (in particular when suitably modified) do that more reliably than BiCGStab, since the two-dimensional steps offer more flexibility. This has also a lasting positive effect on the roundoff in the vertical movement. To stabilize the horizontal movement further, a local minimal residual polynomial of degree $\ell \geq 1$ with an adaptive choice of ℓ as in BiCGStab(ℓ) could be used. A further possibility is to adapt ℓ to the size h_j of the current Lanczos block, which would mean to perform in each regular step an h_j-dimensional local minimization of the residual. An alternative is to trade in the local residual minimization for a more stable Krylov-space generation whenever the former causes a problem. Yet another possibility, indicated in section 4.2, is to combine the Lanczos process with a hybrid Chebyshev iteration.

It is known that in finite-precision arithmetic BiORes is usually more affected by roundoff than the standard BiOMin version of BiCG, at least with regard to the gap between recursively and explicitly computed residuals. Therefore, we are in the process of extending this work to look-ahead procedures for LTPMs that are based on coupled two-term recurrences.

REFERENCES

[1] E. Ayachour, *Avoiding the look-ahead in the Lanczos method*, Appl. Math. (Warsaw), 26 (1999), pp. 33–62.

[2] D. L. Boley, S. Elhay, G. H. Golub, and M. H. Gutknecht, *Nonsymmetric Lanczos and finding orthogonal polynomials associated with indefinite weights*, Numer. Algorithms, 1 (1991), pp. 21–43.

[3] C. Brezinski, *CGM: A Whole Class of Lanczos-Type Solvers for Linear Systems*, Tech. Report ANO-253, Université Lille Flandres Artois, France, 1991.

[4] C. Brezinski and M. Redivo Zaglia, *Breakdowns in the computation of orthogonal polynomials*, in Nonlinear Numerical Methods and Rational Approximation II, A. Cuyt, ed., Kluwer, Dordrecht, The Netherlands, 1994, pp. 49–59.

[5] C. Brezinski and M. Redivo Zaglia, *Treatment of near-breakdown in the CGS algorithms*, Numer. Algorithms, 7 (1994), pp. 33–73.

[6] C. Brezinski and M. Redivo Zaglia, *Look-ahead in Bi-CGSTAB and other product methods for linear systems*, BIT, 35 (1995), pp. 169–201.

[7] C. Brezinski, M. Redivo Zaglia, and H. Sadok, *New look-ahead Lanczos-type algorithms for linear systems*, Numer. Math., 83 (1999), pp. 53–85.

[8] C. Brezinski and H. Sadok, *Avoiding breakdown in the CGS algorithm*, Numer. Algorithms, 1 (1991), pp. 199–206.

[9] Z.-H. Cao, *Avoiding breakdown in variants of the BI-CGSTAB algorithm*, Linear Algebra

Appl., 263 (1997), pp. 113–132.
[10] Z.-H. Cao, *On the QMR approach for iterative methods including coupled three-term recurrences for solving nonsymmetric linear systems*, Appl. Numer. Math., 27 (1998), pp. 123–140.
[11] T. F. Chan, E. Gallopoulos, V. Simoncini, T. Szeto, and C. H. Tong, *A quasi-minimal residual variant of the Bi-CGSTAB algorithm for nonsymmetric systems*, SIAM J. Sci. Comput., 15 (1994), pp. 338–347.
[12] V. Eijkhout, *LAPACK Working Note* 78: *Computational Variants of the CGS and BiCGstab Methods*, Tech. Report UT-CS-94-241, Computer Science Department, University of Tennessee, Knoxville, TN, 1994.
[13] D. A. Flanders and G. Shortly, *Numerical determination of fundamental modes*, J. Appl. Phys., 21 (1950), pp. 1326–1332.
[14] R. W. Freund, *A transpose-free quasi-minimal residual algorithm for non-Hermitian linear systems*, SIAM J. Sci. Comput., 14 (1993), pp. 470–482.
[15] R. W. Freund, M. H. Gutknecht, and N. M. Nachtigal, *An implementation of the look-ahead Lanczos algorithm for non-Hermitian matrices*, SIAM J. Sci. Comput., 14 (1993), pp. 137–158.
[16] R. W. Freund and N. M. Nachtigal, *QMR: A quasi-minimal residual method for non-Hermitian linear systems*, Numer. Math., 60 (1991), pp. 315–339.
[17] R. W. Freund and H. Zha, *A look-ahead algorithm for the solution of general Hankel systems*, Numer. Math., 64 (1993), pp. 295–321.
[18] G. H. Golub and C. F. van Loan, *Matrix Computations*, 2nd ed., Johns Hopkins University Press, Baltimore, MD, 1989.
[19] P. R. Graves-Morris, *A "look-around Lanczos" algorithm for solving a system of linear equations*, Numer. Algorithms, 15 (1997), pp. 247–274.
[20] M. H. Gutknecht, *The unsymmetric Lanczos algorithms and their relations to Padé approximation, continued fractions, and the qd algorithm*, in Preliminary Proceedings of the Copper Mountain Conference on Iterative Methods, Copper Mountain, CO, 1990.
[21] M. H. Gutknecht, *A completed theory of the unsymmetric Lanczos process and related algorithms*, I, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 594–639.
[22] M. H. Gutknecht, *Variants of BiCGStab for matrices with complex spectrum*, SIAM J. Sci. Comput., 14 (1993), pp. 1020–1033.
[23] M. H. Gutknecht, *A completed theory of the unsymmetric Lanczos process and related algorithms*, II, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 15–58.
[24] M. H. Gutknecht, *The Lanczos process and Padé approximation*, in Proceedings of the Cornelius Lanczos International Centenary Conference, J. D. Brown, M. T. Chu, D. C. Ellison, and R. J. Plemmons, eds., SIAM, Philadelphia, PA, 1994, pp. 61–75.
[25] M. H. Gutknecht, *Lanczos-type solvers for nonsymmetric linear systems of equations*, Acta Numer., 6 (1997), pp. 271–397.
[26] M. H. Gutknecht and K. J. Ressel, *Look-Ahead Procedures for Lanczos-Type Product Methods Based on Three-Term Recurrences*, Tech. Report TR-96-19, Swiss Center for Scientific Computing, Manno, Switzerland, 1996.
[27] M. Hochbruck, *Lanczos- und Krylov-Verfahren für nicht-Hermitesche lineare Systeme*, Ph.D. thesis, Fakultät für Mathematik, Universität Karlsruhe, Karlsruhe, Germany, 1992.
[28] M. Hochbruck, *The Padé table and its relation to certain numerical algorithms*, Habilitationsschrift, Universität Tübingen, Germany, 1996.
[29] W. D. Joubert, *Generalized Conjugate Gradient and Lanczos Methods for the Solution of Nonsymmetric Systems of Linear Equations*, Ph.D. thesis, Center for Numerical Analysis, University of Texas at Austin, 1990.
[30] W. D. Joubert, *Iterative Methods for the Solution of Nonsymmetric Systems of Linear Equations*, Tech. Report CNA-239, Center for Numerical Analysis, University of Texas at Austin, 1990.
[31] W. Joubert, *Lanczos methods for the solution of nonsymmetric systems of linear equations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 926–943.
[32] C. Lanczos, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Research Nat. Bur. Standards, 45 (1950), pp. 255–281.
[33] T. A. Manteuffel, *The Tchebyshev iteration for nonsymmetric linear systems*, Numer. Math., 28 (1977), pp. 307–327.
[34] B. N. Parlett, *Reduction to tridiagonal form and minimal realizations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 567–593.
[35] B. N. Parlett, D. R. Taylor, and Z. A. Liu, *A look-ahead Lanczos algorithm for unsymmetric matrices*, Math. Comp., 44 (1985), pp. 105–124.

[36] K. J. Ressel and M. H. Gutknecht, *QMR smoothing for Lanczos-type product methods based on three-term recurrences*, SIAM J. Sci. Comput., 19 (1998), pp. 55–73.
[37] W. Schönauer, *Scientific Computing on Vector Computers*, Elsevier, Amsterdam, 1987.
[38] G. L. G. Sleijpen and D. R. Fokkema, *BiCGstab(l) for linear equations involving unsymmetric matrices with complex spectrum*, Electron. Trans. Numer. Anal., 1 (1993), pp. 11–32.
[39] G. L. G. Sleijpen and H. A. van der Vorst, *Maintaining convergence properties of BiCGstab methods in finite precision arithmetic*, Numer. Algorithms, 10 (1995), pp. 203–223.
[40] G. L. G. Sleijpen and H. A. van der Vorst, *An overview of approaches for the stable computation of hybrid BiCG methods*, Appl. Numer. Math., 19 (1995), pp. 235–254.
[41] G. L. G. Sleijpen, H. A. van der Vorst, and D. R. Fokkema, *BiCGstab(l) and other hybrid Bi-CG methods*, Numer. Algorithms, 7 (1994), pp. 75–109.
[42] P. Sonneveld, *CGS, a fast Lanczos-type solver for nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 36–52.
[43] D. R. Taylor, *Analysis of the Look Ahead Lanczos Algorithm*, Ph.D. thesis, Department of Mathematics, University of California, Berkeley, 1982.
[44] H. A. van der Vorst, *Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 631–644.
[45] H. E. Wrigley, *Accelerating the Jacobi method for solving simultaneous equations by Chebyshev extrapolation when the eigenvalues of the iteration matrix are complex*, Comput. J., 6 (1963), pp. 169–176.
[46] S.-L. Zhang, *GPBI-CG: Generalized product-type methods based on Bi-CG for solving nonsymmetric linear systems*, SIAM J. Sci. Comput., 18 (1997), pp. 537–551.
[47] L. Zhou and H. F. Walker, *Residual smoothing techniques for iterative methods*, SIAM J. Sci. Comput., 15 (1994), pp. 297–312.
[48] M. Ziegler, *Generalized biorthogonal bases and tridiagonalisation of matrices*, Numer. Math., 77 (1977), pp. 407–421.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 4, pp. 1079–1101
© 2000 Society for Industrial and Applied Mathematics

RESIDUAL-MINIMIZING KRYLOV SUBSPACE METHODS FOR STABILIZED DISCRETIZATIONS OF CONVECTION-DIFFUSION EQUATIONS*

OLIVER G. ERNST†

Abstract. We discuss the behavior of the minimal residual method applied to stabilized discretizations of one- and two-dimensional model problems for the stationary convection-diffusion equation. In the one-dimensional case, it is shown that eigenvalue information for estimating the convergence rate of the minimal residual method is highly misleading due to the strong nonnormality of these operators for large grid Péclet numbers. It is also shown that the field of values is a more reliable tool for assessing the convergence rate. In the two-dimensional model problems considered, we observe two distinct phases in the convergence of the iterative method: the first determined by the field of values and the second by the spectrum. We conjecture that the first phase lasts as long as the longest streamline takes to traverse the grid with the flow.

Key words. convection-diffusion equation, Krylov subspace methods, streamline diffusion, field of values

AMS subject classifications. 65F10, 65N30, 65N22

PII. S0895479897325761

1. Introduction.

It is well known [12] that when convection-diffusion problems are discretized using centered schemes such as central differences or the Galerkin finite element method, nonphysical oscillations can occur in the discrete solution whenever convection is the dominating term. In the context of finite element discretizations, an approach to remedy this convective instability has been developed by Brooks and Hughes [2] and Johnson, Nävert, and Pitkäranta [16]. These techniques are called the *streamline upwind Petrov Galerkin* (SUPG) and the *Galerkin least-squares* (GLS) methods and are based on adding a term to the variational formulation of the problem which is proportional to the residual of the discrete solution on each element. Extensive discussions of these stabilization techniques can be found in the recent monographs of Morton [18] and Roos, Stynes, and Tobiska [22].

Little attention, however, has been devoted to the effect of this stabilization on the resulting discrete linear system of equations and its solution by iterative methods. In this paper, we consider one- and two-dimensional model problems, compare the properties of the resulting discretization matrices, and analyze the convergence of residual minimizing Krylov subspace methods applied to these linear systems. We discover two things: First, for the one-dimensional Dirichlet problem, the stabilized discrete operators are highly nonnormal. As a result, using spectral information to estimate the convergence rate of Krylov subspace methods is potentially misleading (cf. [19, 24]) and it is shown that this is indeed so. While most of the examples of highly nonnormal matrices in, e.g., [24] are contrived mathematical pathologies, the one-dimensional stabilized convection-diffusion discretization arises very naturally in applications. We also show that, for this class of problems, the field of values (cf. [7]) is a more reliable tool for assessing the convergence rate. Our second observation con-

*Received by the editors August 8, 1997; accepted for publication (in revised form) by M. Eiermann December 18, 1998; published electronically March 21, 2000.
http://www.siam.org/journals/simax/21-4/32576.html
†Institut für Angewandte Mathematik II, TU Bergakademie Freiberg, Germany (ernst@math.tu-freiberg.de).

cerns the two-dimensional model problem. It is observed that the convergence of the minimal residual method consists of two distinct phases of linear convergence whose rates seem to be determined by the field of values and the spectrum, respectively. Thus, in this case the nonnormality only affects convergence in the first phase, whose duration we conjecture to be the number of iteration steps it takes for information to traverse the underlying grid along the longest streamline.

The paper is organized as follows. Section 2 introduces the continuous problem, its standard Galerkin discretization, and two variants of finite element stabilization techniques known as the SUPG and GLS methods. Section 3 introduces residual minimizing Krylov subspace techniques and reviews classical and more recent convergence results. In section 4 we consider a one-dimensional model problem and show that the field of values gives a much better estimate of the rate of convergence than the spectrum. Section 5 performs a computational study of two model problems in two dimensions.

2. The convection-diffusion equation and its stabilized discretization. This section briefly reviews the boundary value problem under consideration and its discretization by stabilized finite element methods.

2.1. The continuous problem. We consider a bounded domain $\Omega \subset \mathbb{R}^d$ on which a given solenoidal velocity field $\boldsymbol{a} : \Omega \to \mathbb{R}^d$ and a diffusivity tensor $\boldsymbol{\kappa} : \Omega \to \mathbb{R}^{d\times d}$ are defined. We seek a function $u : \Omega \to \mathbb{R}$ which satisfies the differential equation

$$-\nabla\cdot(\boldsymbol{\kappa}\nabla u - \boldsymbol{a}u) = f \tag{2.1}$$

in Ω with a given source term $f : \Omega \to \mathbb{R}$. Appropriate boundary conditions for this problem are the specification of the total flux $-\boldsymbol{n}\cdot\nabla u + \boldsymbol{a}u$ along the inflow boundary (where $\boldsymbol{a}\cdot\boldsymbol{n} < 0$, $\boldsymbol{n}$ being the exterior unit normal), the convective flux $\boldsymbol{a}u$ along the outflow boundary, or Dirichlet conditions on either part of the boundary. We restrict ourselves to Dirichlet conditions

$$u = g \quad \text{ on } \Gamma = \partial\Omega. \tag{2.2}$$

2.2. Stabilized finite element methods. The variational formulation of problem (2.1), (2.2) is to find a function $u \in V$ which satisfies

$$a(u, v) = \ell(v) \qquad \forall v \in V$$

with a bilinear form $a : V \times V \to \mathbb{R}$ and a linear functional $\ell : V \to \mathbb{R}$ given by

$$a(u, v) = \int_\Omega \nabla v \cdot (\boldsymbol{\kappa}\nabla u - \boldsymbol{a}u)\, dx \quad \text{ and } \quad \ell(v) = \int_\Omega vf\, dx,$$

along with a suitable trial/test space V depending on where essential boundary conditions are imposed. By choosing trial and weighting space to coincide, we make the assumption that all essential boundary conditions have been made homogeneous. For a pure Dirichlet problem we may choose the usual Sobolev space $V = H_0^1(\Omega)$.

Given a finite-dimensional subspace V_h of V, the Galerkin finite element method computes an approximate solution $u_h \in V_h$ determined by

$$a(u_h, v) = \ell(v) \qquad \forall v \in V_h. \tag{2.3}$$

The SUPG and GLS stabilizations of the Galerkin discretization lead to modified bilinear forms and right-hand side functionals, which we shall denote by $a_j^h : V_h \times V_h \to \mathbb{R}, j = 1, 2$, and $\ell_j^h : V^h \to \mathbb{R}, j = 1, 2$, respectively. When referring to both methods we shall omit the index j. To define these quantities we introduce the diffusive part $L^D u = -\nabla\cdot(\boldsymbol{\kappa}\nabla u)$ and the advective part $L^A u = \boldsymbol{a} \cdot \nabla u$ of the advection-diffusion operator $L = L^D + L^A$. The SUPG method is defined by

$$a_1^h(u, v) = a(u, v) + \sum_{K\in\mathcal{T}_h} (Lu, \tau L^A v)_K, \quad u, v \in V_h, \tag{2.4}$$

$$\ell_1^h(v) = \ell(v) + \sum_{K\in\mathcal{T}_h} (f, \tau L^A v)_K, \quad v \in V_h, \tag{2.5}$$

where K denotes an arbitrary element in the finite element mesh $\mathcal{T}_h$, $(\cdot, \cdot)_K$ denotes the L^2 inner product on K, and τ denotes an appropriately chosen stability parameter. The corresponding terms for GLS are given by

$$a_2^h(u, v) = a(u, v) + \sum_{K\in\mathcal{T}_h} (Lu, \tau Lv)_K, \quad u, v \in V_h, \tag{2.6}$$

$$\ell_2^h(v) = \ell(v) + \sum_{K\in\mathcal{T}_h} (f, \tau Lv)_K, \quad v \in V_h, \tag{2.7}$$

i.e., the stabilization term weights with the full operator instead of just the advective part. This makes it applicable to more general problems.

Following [3], we choose the stabilization parameter τ as

$$\tau = \frac{h}{2|\boldsymbol{a}|}\xi(\alpha) \qquad \text{with} \qquad \xi(\alpha) = \coth(\alpha) - \frac{1}{\alpha}, \tag{2.8}$$

where $|\cdot|$ denotes the Euclidean length of a vector and the parameter $\alpha = ah/2\kappa$ is the *grid Péclet number*, which measures the strength of convection versus diffusion relative to the mesh size. This choice leads to nodally exact solutions for the one-dimensional constant-coefficient problem and has been shown to converge with order $\mathcal{O}(h^{p+1/2})$ in the L^2-norm for higher dimensions, where p denotes the maximal degree of complete polynomials used in the finite element approximation (cf. [26]).

For the following, we make the assumption that the diffusivity tensor $\boldsymbol{\kappa}$ is isotropic, i.e., diagonal and elementwise constant. Moreover, we assume that the finite element space consists of either piecewise linear or piecewise bilinear functions. In this case we have

$$(Lu, \tau Lv)_K = (Lu, \tau L^A v)_K = (L^A u, \tau L^A v)_K = (\tau \boldsymbol{a}\boldsymbol{a}^T \nabla u, \nabla v)_K,$$

i.e., the stabilization term has the form of an additional diffusivity tensor given by $\tau\boldsymbol{a}\boldsymbol{a}^T$, which acts only in the direction of the flow. For this reason this type of stabilization scheme is also known as the *streamline diffusion* method.

3. Iterative solution of the discrete system. In this section we briefly review some well-known facts about residual-minimizing Krylov subspace methods—as implemented, e.g., by the popular GMRES algorithm of Saad and Schultz [23]—for solving the discrete linear system. The setting is the complex vector space $\mathbb{C}^n$ endowed with an arbitrary (not necessarily the Euclidean) inner product $(\cdot, \cdot)$. We denote the associated vector norm and induced matrix norm by $\|\cdot\|$.

3.1. Minimal residual methods. Minimal residual methods belong to the family of Krylov subspace algorithms, in which approximations to the solution $\boldsymbol{u}$ of the matrix equation $A\boldsymbol{u} = \boldsymbol{f}$ are sought in the sequence of *shifted Krylov spaces* $V_m := \boldsymbol{u}_0 + K_m(A, \boldsymbol{r}_0)$, where

$$K_m(A, \boldsymbol{r}_0) := \operatorname{span}\{\boldsymbol{r}_0, A\boldsymbol{r}_0, \ldots, A^{m-1}\boldsymbol{r}_0\}, \quad m = 1, 2, \ldots,$$

is the mth Krylov space with respect to A and $\boldsymbol{r}_0 = \boldsymbol{f} - A\boldsymbol{u}_0$, the residual vector of the initial guess $\boldsymbol{u}_0$. The vectors $\boldsymbol{v}_m \in V_m$ are of the form $\boldsymbol{v}_m = \boldsymbol{u}_0 + q_{m-1}(A)\boldsymbol{r}_0$ with a polynomial $q_{m-1} \in \Pi_{m-1}$, the space of polynomials of degree not exceeding $m-1$. The residual of a vector in this space can then be written in the form $\boldsymbol{r}_m = \boldsymbol{f} - A\boldsymbol{v}_m = p_m(A)\boldsymbol{r}_0$ with a polynomial p_m in the set $\Pi_m^* := \{p \in \Pi_m : p(0) = 1\}$. The difference $\boldsymbol{e}_m = \boldsymbol{u} - \boldsymbol{v}_m$ between $\boldsymbol{v}_m \in V_m$ and the solution $\boldsymbol{u}$ is then given by $\boldsymbol{e}_m = p_m(A)\boldsymbol{e}_0$, where $\boldsymbol{e}_0 = \boldsymbol{u} - \boldsymbol{u}_0$ is the error in the initial guess. In particular, for the given vector norm and its induced operator norm,

$$\|\boldsymbol{e}_m\| = \|p_m(A)\boldsymbol{e}_0\|, \quad \text{implying} \quad \frac{\|\boldsymbol{e}_m\|}{\|\boldsymbol{e}_0\|} \le \|p_m(A)\|$$

with the analogous inequalities holding for $\boldsymbol{r}_m$. Thus, for a given matrix A, a Krylov subspace method will succeed in reducing error and residual for all right-hand sides and initial vectors to the extent it is able to find residual polynomials $p_m \in \Pi_m^*$ which make $\|p_m(A)\|$ small.

Different Krylov subspace algorithms result from different strategies for selecting a particular approximation $\boldsymbol{u}_m \in V_m$ or, equivalently, selecting a polynomial $p_m \in \Pi_m^*$. One possible and well-defined approach is to choose $\boldsymbol{u}_m$ to minimize the norm of the residual, i.e., that

$$\|\boldsymbol{r}_m\| = \min_{\boldsymbol{v} \in V_m} \|\boldsymbol{f} - A\boldsymbol{v}\|.$$

We will refer to algorithms which implement this strategy as *minimal residual methods*. The GMRES algorithm of Saad and Schultz [23] is an implementation of this method for general nonsingular linear systems of equations which employs an orthonormal basis of $K_m(A, \boldsymbol{r}_0)$ which is augmented by one basis vector in each iteration step using the Arnoldi process. Two other well-known algorithms implementing the minimal residual approach are MINRES by Paige and Saunders [21] for Hermitian indefinite systems and the conjugate residual (CR) method for Hermitian positive definite systems, which was first described in [14] by Hestenes and Stiefel on the conjugate gradient algorithm.

3.2. Error bounds. The most well-known convergence result, which follows immediately from the relations discussed above, is based on spectral properties of the matrix A.

THEOREM 3.1. *Let $A = V\Lambda V^{-1}$ be diagonalizable and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ contain the eigenvalues on its diagonal. Then the residual $\boldsymbol{r}_m$ after m steps of the GMRES algorithm satisfies*

$$\frac{\|\boldsymbol{r}_m\|}{\|\boldsymbol{r}_0\|} \le \|Vp_m(\Lambda)V^{-1}\| \le \operatorname{cond}(V) \min_{p_m \in \Pi_m^*} \|p_m\|_{\Lambda(A)}, \tag{3.1}$$

where $\|\phi\|_D$ denotes the maximum value of the function ϕ on the compact set D and $\operatorname{cond}(V) = \|V\|\,\|V^{-1}\|$ is the condition number of the eigenvector matrix with respect to the norm $\|\cdot\|$.

If A is a normal matrix, then $\mathrm{cond}(V) = 1$, and hence in this case the bound (3.1) for the convergence rate of the minimal residual method depends entirely on how rapidly the quantity $\max_{p\in\Pi_m^*} |p|_{\Lambda(A)}$ decreases with m, i.e., the issue is reduced to a polynomial approximation problem on the discrete set $\Lambda(A)$ of eigenvalues of the matrix A. The class of normal matrices includes the important cases of Hermitian, skew-Hermitian, and circulant matrices. In the general nonnormal case, however, $\mathrm{cond}(V)$ may become sufficiently large to have a profound influence on the convergence rate. This is particularly important when addressing the question of the convergence rate for a whole sequence of parameter-dependent matrices such as those belonging to the discretization of a convection-diffusion problem on successively refined meshes or of problems with increasing Péclet numbers on one given mesh. Although often ignored in applications, this effect may be strong enough to make any available eigenvalue information completely useless (cf. [24]).

One approach for addressing this difficulty is due to Trefethen [25], who has derived residual bounds based on *pseudospectra* of the matrix A. For a positive number ϵ, the associated ϵ-pseudospectrum of A is the set in the complex plane defined by $\Lambda_\epsilon(A) = \{z : \|(zI - A)^{-1}\| \geq 1/\epsilon\}$. This set contains the spectrum of A, hence its boundary Γ_ϵ may be used in the resolvent integral representation of any analytic function of A, in particular of any polynomial of A. This results in the bound

$$\|p_m(A)\| \leq \frac{\ell(\Gamma_\epsilon)}{2\pi\epsilon}\|p_m\|_{\Gamma_\epsilon}$$

for the norm of a residual polynomial p_m applied to A, where $\ell(\Gamma_\epsilon)$ denotes the length of the curve Γ_ϵ, which in turn implies the bound

$$\frac{\|\boldsymbol{r}_m\|}{\|\boldsymbol{r}_0\|} \leq \frac{\ell(\Gamma_\epsilon)}{2\pi\epsilon} \min_{p_m\in\Pi_m^*} \|p_m\|_{\Gamma_\epsilon}$$

for the residual reduction.

Pseudospectra can sometimes result in much more realistic bounds than (3.1) but are expensive to compute. Moreover, it is not always clear which value of ϵ leads to the most useful information. In this paper we turn to another set associated with a matrix A for predicting the convergence rate of minimum residual methods, namely, its *field of values*

$$W(A) = \left\{ \frac{(A\boldsymbol{z}, \boldsymbol{z})}{(\boldsymbol{z}, \boldsymbol{z})} : \quad 0 \neq \boldsymbol{z} \in \mathbb{C}^n \right\} = \{(A\boldsymbol{z}, \boldsymbol{z}) : \|\boldsymbol{z}\| = 1\},$$

sometimes also called its *numerical range*. The field of values of a matrix is a convex and compact set in the complex plane which contains the eigenvalues. For normal matrices $W(A)$ coincides with the convex hull of $\Lambda(A)$, whereas for nonnormal matrices it may be considerably larger. A measure for the size of $W(A)$ is the *numerical radius* $\mu(A) := \max\{|\zeta| : \zeta \in W(A)\}$. The numerical radius is related to the norm associated with the underlying inner product by $\frac{1}{2}\|A\| \leq \mu(A) \leq \|A\|$. Further properties of $W(A)$ can be found in [15, Chapter 1]. The bound (3.1) relies on the fact that the eigenvalues of $p(A)$ are related to those of A via the spectral mapping theorem. An analogous mapping theorem for the field of values does not hold in general. However, a useful result recently obtained by Eiermann [8] does hold for convex sets and a special sequence of polynomials, the Faber polynomials associated with this set.

THEOREM 3.2. *If $\{F_m\}_{m=0}^\infty$ denotes the sequence of Faber polynomials of the field of values $W(A)$ of a matrix $A \in \mathbb{C}^{n\times n}$, then the numerical radius $\mu(F_m(A))$ satisfies*

$$\mu(F_m(A)) \leq \|F_m\|_{W(A)}. \tag{3.2}$$

Proof. See [8] for the proof. ∎

The crucial point of this result is that it relates the field of values of $F_m(A)$ to the size of the polynomial F_m on the set $W(A)$. To use Faber polynomials for estimating the convergence rate of Krylov subspace methods, we turn to the normalized Faber polynomials $\hat{F}_m(z) := F_m(z)/F_m(0)$, which are admissible as residual polynomials. Hence we must require $F_m(0) \neq 0$, which is assured if we assume $0 \notin W(A)$. The asymptotic behavior of these polynomials as $m \to \infty$ is well understood: The normalized Faber polynomials $\{\hat{F}_m\}$ of a convex bounded set $0 \notin D \subset \mathbb{C}$ satisfy $\|\hat{F}_m\|_D \leq c_m \gamma^m$ with $0 < \gamma < 1$ and $c_m < 2/(1-\gamma^m)$. The number γ is known as the asymptotic convergence factor of the set D. In fact, if $p_m^* \in \Pi_m^*$ minimizes $\|p\|_D$ over all $p \in \Pi_m^*$, then

$$\gamma = \lim_{m\to\infty} \|p_m^*\|_D^{1/m}.$$

If p_m denotes the mth residual polynomial selected in the mth step of the minimal residual method, then the minimization property implies

$$\|\boldsymbol{r}_m\| = \|p_m(A)\boldsymbol{r}_0\| \leq \|\hat{F}_m(A)\boldsymbol{r}_0\| \leq \|\hat{F}_m(A)\|\,\|\boldsymbol{r}_0\|$$

for the residual after m steps of the algorithm. Theorem 3.2, together with the asymptotic properties of normalized Faber polynomials, thus yields the bound

$$\frac{\|\boldsymbol{r}_m\|_2}{\|\boldsymbol{r}_0\|_2} \leq \|\hat{F}_m(A)\|_2 \leq 2\mu(\hat{F}_m(A)) \leq 2\|\hat{F}_m\|_{W(A)} \leq 2c_m\gamma^m.$$

3.3. A special case. The one-dimensional model problem with constant coefficients and Dirichlet boundary conditions introduced in the next section is particularly easy to analyze since in this case the field of values is an ellipse, for which both the asymptotic convergence factor γ as well as the Faber polynomials are known explicitly. In particular, the Faber polynomials are just suitably scaled and shifted Chebyshev polynomials of the first kind.

The asymptotic convergence factor γ for domains D not containing the origin whose complement with respect to the extended complex plane $\hat{\mathbb{C}}$ is simply connected may be determined using conformal mapping: In this case there exists a conformal map Φ of the complement $\hat{\mathbb{C}}\backslash D$ of D to the exterior of the unit disk. The asymptotic convergence factor $\gamma = \gamma(D)$ is then given by (cf. [9])

$$\gamma = \frac{1}{|\Phi(0)|}. \tag{3.3}$$

For an ellipse $\mathcal{E}_\rho(\sigma,\tau)$ with foci at $\sigma \pm \tau$ defined by

$$\mathcal{E}_\rho(\sigma,\tau) = \left\{ z \in \mathbb{C} : |z-\sigma+\tau| + |z-\sigma-\tau| \leq |\tau|\left(\rho+\rho^{-1}\right)\right\}, \tag{3.4}$$

we obtain

$$\gamma(\mathcal{E}_\rho(\sigma,\tau)) = \rho \cdot \left| -\frac{\sigma}{\tau} + \sqrt{\left(\frac{\sigma}{\tau}\right)^2 - 1} \right|^{-1},$$

in which that branch of the square root is selected which results in $\gamma < 1$. Provided $0 \notin \mathcal{E}_\rho(\sigma,\tau)$, the Faber polynomials are given in terms of the first-kind Chebyshev polynomials T_m by

$$F_m(z) = \frac{T_m(\zeta(z))}{T_m(\zeta(0))}, \quad \zeta(z) = \frac{z-\sigma}{\tau}, \qquad m = 0, 1, \ldots.$$

4. The one-dimensional case. It is instructive to look at the one-dimensional case, since, at least for constant coefficients, all the important quantities associated with the discrete problem can be computed analytically. If we focus on the Dirichlet problem,

$$-\kappa u'' + au' = f \text{ on } (0, L), \qquad u(0) = u(L) = 0,$$

then a Galerkin discretization using piecewise linear elements on a grid with uniform spacing h and N interior mesh points leads to the discrete linear system of equations $A\boldsymbol{u} = \boldsymbol{f}$, in which, after scaling by h/κ, the coefficient matrix A is given by

$$A = \text{ tridiag}(-1-\alpha, 2, -1+\alpha) \in \mathbb{R}^{N\times N}.$$

The stabilized scheme leads to a linear system $\tilde{A}\boldsymbol{u} = \tilde{\boldsymbol{f}}$ with modified right-hand side $\tilde{\boldsymbol{f}}$ and modified coefficient matrix $\tilde{A}$, which, after scaling by $h/(\kappa+\tilde{\kappa})$, is given by

$$\tilde{A} = \text{ tridiag}(-1-\tilde{\alpha}, 2, 1+\tilde{\alpha}) \in \mathbb{R}^{N\times N}.$$

The parameter $\tilde{\alpha} = ah/(2(\kappa+\tilde{\kappa})), \kappa+\tilde{\kappa} = \kappa(1+\alpha\xi)$ may be interpreted as the effective Péclet number of the stabilized scheme.

4.1. Eigensystems. Both discretization matrices A and $\tilde{A}$ are of the form $T = \text{tridiag}(-1-t, 2, -1+t)$ and, using the results on the eigenvalues of tridiagonal Toeplitz matrices collected in the appendix, the eigenvalues of these matrices are given by

$$\tau_j = 2\left[1 + \sqrt{|(1+t)(1-t)|}e^{\frac{i}{2}[\pi+\arg(t-1)]}\cos\left(\frac{j\pi}{N+1}\right)\right], \qquad j = 1, \ldots, N.$$

For the Galerkin discretization matrix A this results in eigenvalues $\{\lambda_j\}_{j=1}^N$ given by

$$\lambda_j = 2\begin{cases} 1 - \sqrt{1-\alpha^2}\cos(j\pi h), & 0 \le \alpha \le 1, \\ 1 + i\sqrt{\alpha^2-1}\cos(j\pi h), & \alpha \ge 1, \end{cases} \qquad j = 1 \ldots, N.$$

The eigenvalues for the stabilized discretization are obtained by replacing the Péclet number α with $\tilde{\alpha} = \tanh(\alpha)$. The effective Péclet number $\tilde{\alpha}$ is a strictly increasing function of α: It equals α in the diffusion-dominated limit $\alpha = 0$ and approaches unity in the convection-dominated limit $\alpha \to \infty$. Hence $\tilde{\alpha} \in [0, 1)$ so that the eigenvalues $\{\tilde{\lambda}_j\}_{j=1}^N$ of $\tilde{A}$ are always real and given by

$$\tilde{\lambda}_j = 2\left(1 - \sqrt{1-\tilde{\alpha}^2}\cos(j\pi h)\right), \qquad j = 1, \ldots, N.$$

This results in the following qualitative behavior of the eigenvalues for varying Péclet number: In the diffusion-dominated limit both spectra lie in the real interval $[0, 4]$ with a slight clustering at the endpoints. As α approaches the critical value of unity, the eigenvalues of A coalesce at the value 2 and, as α grows beyond 1, the eigenvalues of A lie on the complex interval $[2 - i\alpha, 2 + i\alpha]$ parallel to the imaginary axis. In contrast, the eigenvalues of $\tilde{A}$ always lie on a real interval symmetric with respect to two which shrinks monotonically as α increases. The diameter of this interval is

$$\tilde{\lambda}_{\max} - \tilde{\lambda}_{\min} = \frac{4\cos\pi h}{\cosh\alpha},$$

which decreases at an exponential rate as the Péclet number α increases.

The GMRES convergence bound (3.1) is the product of the condition number of the eigenvector matrix and the maximum norm of the GMRES polynomial on the spectrum of the matrix. By the minimization property of GMRES, we obtain an upper bound by replacing the mth GMRES polynomial with the shifted and scaled Chebyshev polynomial

$$p_m(z) = \frac{T_m(\zeta(z))}{T_m(\zeta(0))}, \qquad \zeta(z) = \frac{z-\sigma}{\tau},$$

where T_m is the first-kind Chebyshev polynomial of degree m and $\sigma = (\lambda_1 + \lambda_N)/2$ is the center of the spectrum, which in our case is a line segment centered at $z = 2$ parallel to either the real or imaginary axis. By employing the bound (A.8) on the polynomials p_m from the appendix, we obtain the bounds

$$\mathrm{cond}_2(V)\,\|p_m\|_{\Lambda(A)} \le \left|\frac{\alpha+1}{\alpha-1}\right|^{(N-1)/2} \frac{2\gamma^m}{1-\gamma^{2m}}$$

for the matrix A and

$$\mathrm{cond}_2(\tilde{V})\,\|p_m\|_{\Lambda(\tilde{A})} \le e^{(N-1)\alpha}\frac{2\tilde{\gamma}^m}{1-\tilde{\gamma}^{2m}}$$

for the matrix $\tilde{A}$, respectively, where

$$\gamma = |\Phi(\zeta(0))|^{-1}, \qquad \Phi(\zeta) = \zeta + \sqrt{\zeta^2-1}, \;\; \zeta \notin [-1,1],$$

is the asymptotic convergence factor of $\Lambda(A)$, and $\tilde{\gamma}$ is the corresponding quantity for $\Lambda(\tilde{A})$.

If only the spectral distributions of A and $\tilde{A}$ are considered, the residual bound (3.1) would indicate that GMRES should converge much more rapidly for $\tilde{A}$ than for A in the convection-dominated case, since then the spectrum of the former rapidly shrinks to a point, hence low order polynomials are sufficient to yield small values on $\Lambda(\tilde{A})$. Numerical experience, however, results in almost identical behavior of GMRES for both systems.

An indication that something is going wrong is obtained from looking at the second term in (3.1), the condition number of the eigenvector matrix. Drawing again from the results in the appendix, the eigenvector matrix V of the tridiagonal Toeplitz matrix T has the form $V = DU$ with an orthogonal matrix U and a diagonal matrix

$$D = \mathrm{diag}(\delta, \ldots, \delta^N), \quad \text{where} \quad \delta = \delta(t) = \begin{cases} \sqrt{\frac{1+t}{1-t}}, & 0 < t < 1, \\ i\sqrt{\frac{t+1}{t-1}}, & t > 1. \end{cases}$$

Hence, the spectral condition number of the eigenvector matrix V is given by

$$\mathrm{cond}_2(V) = \|V\|_2\|V^{-1}\|_2 = \|D\|_2\|D^{-1}\|_2 = |\delta|^{N-1},$$

which, written in terms of the grid Péclet number α, yields

$$\mathrm{cond}_2(V) = \left|\frac{\alpha+1}{\alpha-1}\right|^{(N-1)/2} \qquad \text{and} \qquad \mathrm{cond}_2(\tilde{V}) = e^{(N-1)\alpha}$$

for the matrices A ($t = \alpha$) and $\tilde{A}$ ($t = \tilde{\alpha} = \tanh\alpha$), respectively. We see that $\mathrm{cond}_2(V)$ is bounded in the limit $\alpha \to \infty$ while $\mathrm{cond}(\tilde{V})$ grows exponentially.

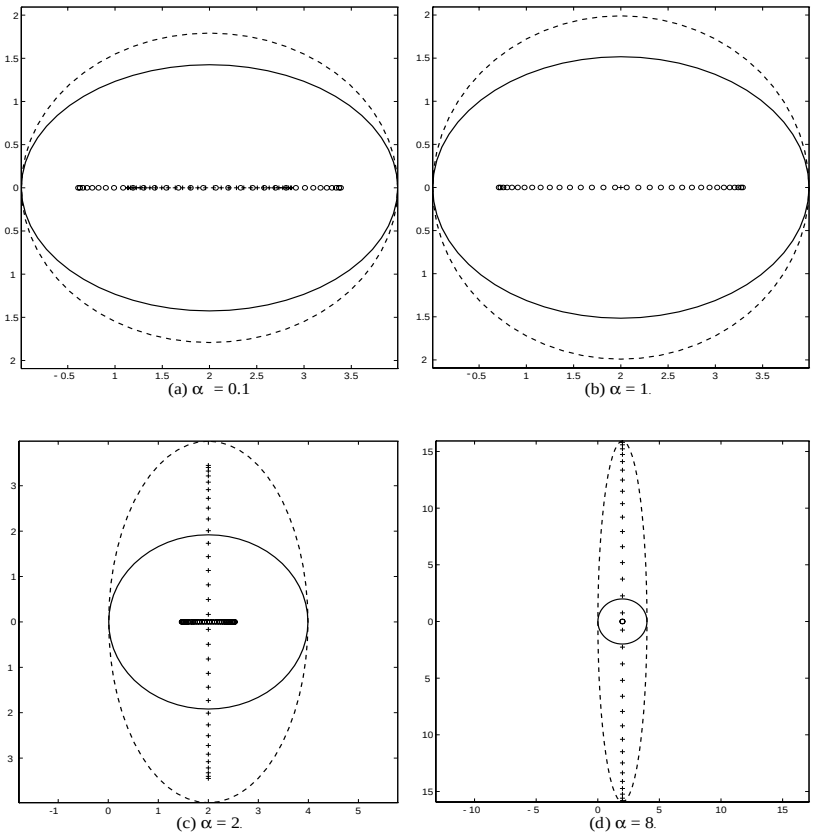


FIG. 4.1. $\Lambda(A)$ *(crosses),* $\Lambda(\tilde{A})$ *(circles),* $W(A)$ *(dashed line), and* $W(\tilde{A})$ *(solid line) for* $N = 32$ *and* $\alpha = 0.1$*,* 1*,* 2*, and* 8*.*

4.2. Field of values. From the results collected in the appendix, the field of values of the tridiagonal Toeplitz matrix $T = \text{tridiag}(-1-t, 2, -1+t)$ is the ellipse $\mathcal{E}_\rho(\sigma, \tau)$ with

$$\sigma = 2, \qquad \tau = 2\cos(\pi h)\sqrt{1-t^2}, \qquad \rho = \left|\frac{1+t}{1-t}\right|^{1/2}.$$

In particular, for $t < 1$ the focal line lies on the real axis, while for $t > 1$ it lies parallel to the imaginary axis. Thus, since $\tilde{\alpha} \in [0, 1)$ for all values of the grid Péclet number, the field of values of the stabilized matrix $\tilde{A}$ is always an ellipse with a real focal line. As derived in the appendix, the field of values of the one-dimensional convection-diffusion operator is a parabola, and for decreasing values of h the elliptical fields of values of A and $\tilde{A}$, suitably rescaled, approximate this parabola near its intersection with the real axis. In the convection-dominated limit $\alpha \to \infty$, the field of values degenerates to a circle centered at $z = 2$ with radius $r = 2\cos(\pi h)$. Figure 4.1 shows the spectra and fields of values for the discretization of the one-dimensional model problem on a mesh with 32 interior points and $\alpha = 0.1, 1, 2$, and 8.

4.3. Numerical experiments. The asymptotic convergence factors of the field of values and the convex hull of the spectrum of A and $\tilde{A}$ of dimension $N = 255$ for

TABLE 4.1
Asymptotic convergence factors, $\alpha = 0.1, 1, 2,$ *and* 8.

	$\alpha = 0.1$	$\alpha = 1$	$\alpha = 2$	$\alpha = 8$
$W(A)$	0.9992	0.9999	1	1
$CH(\Lambda(A))$	0.9038	0	0.5773	0.8819
$W(\tilde{A})$	0.9992	0.9999	0.9999	0.9999
$CH(\Lambda(\tilde{A}))$	0.9041	0.3678	0.1353	3.354e-04

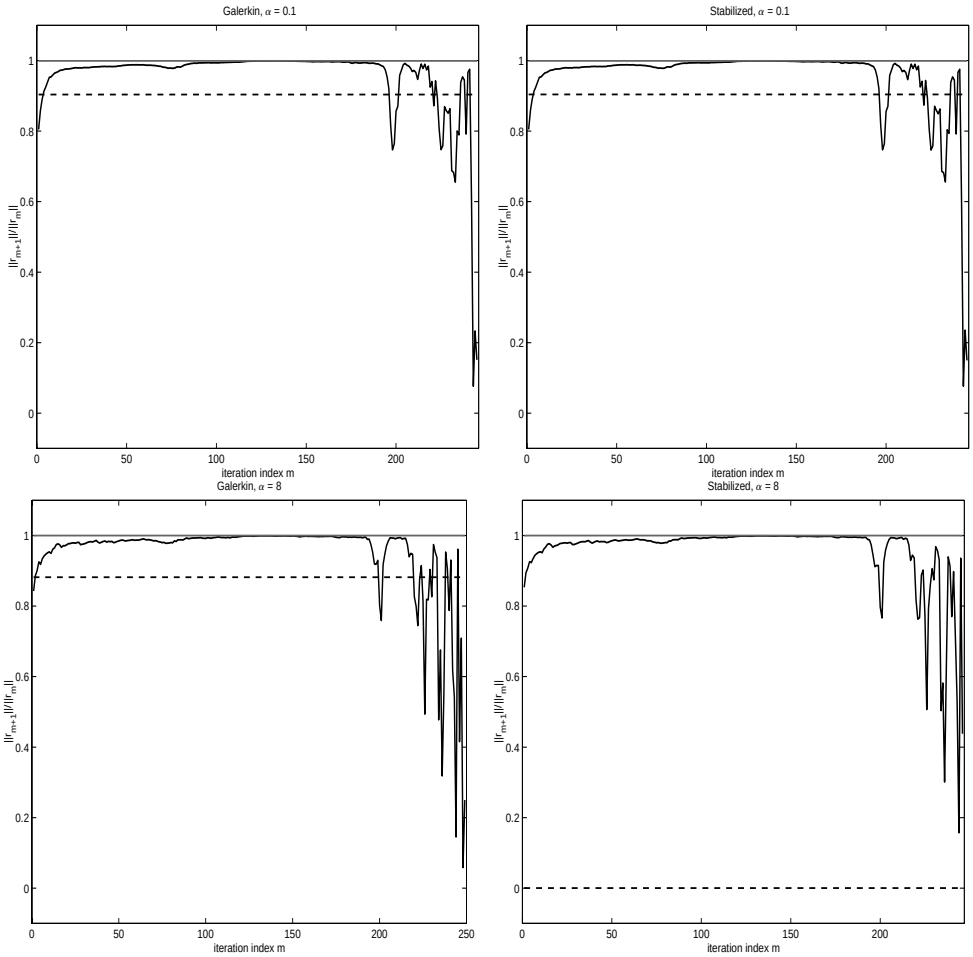


FIG. 4.2. *GMRES "lifespan" curves for* A *(left) and* $\tilde{A}$ *(right) for the cases* $\alpha = 0.1$ *and* $\alpha = 8$. *The solid vertical line represents the asymptotic convergence factor predicted by the field of values and the dashed vertical line represents that predicted by the spectrum.*

$\alpha = 0.1, 1, 2,$ and 8 are given in Table 4.1 to four digits. This table shows that, while the spectral information alone would predict considerably faster convergence for the stabilized discretization as soon as $\alpha > 1$, the fields of values of both discretization matrices indicate the same slow convergence rate of almost unity.

The behavior of GMRES for the Galerkin and the stabilized discretizations for the cases $\alpha = 0.1$ and $\alpha = 8$ can be seen in Figure 4.2. Here GMRES was applied to both systems, again of size $N = 255$, with zero initial guess and a random right-hand side.

While right-hand sides arising from the discretization of the boundary value problem resulted in similar behavior, we chose nonetheless to use a random right-hand side to make sure we were observing generic behavior of GMRES for these matrices. Rather than the usual plot of normalized residual norms $\|\boldsymbol{r}_m\|/\|\boldsymbol{r}_0\|$ in a logarithmic scale, we instead plot quotients $\|\boldsymbol{r}_{m+1}\|/\|\boldsymbol{r}_m\|$ of successive residual norms throughout the iteration history. Such *lifespan curves* were introduced by Nevanlinna [20] and give a more accurate view of the rate of convergence in different phases of the iteration. The solid vertical line represents the asymptotic rate of convergence predicted by the field of values and the dashed line represents that predicted by the spectrum. We observe that, after an initial phase of sublinear convergence (cf. [20]), the rate of convergence of GMRES during the linear phase is predicted remarkably well by the field of values, whereas the spectrum would have indicated a much too optimistic rate. For the case $\alpha = 8$, which is a very low Péclet number with regard to applications, the matrix $\tilde{A}$ is basically a Jordan block and the spectrum gives no information whatsoever regarding the convergence rate. Moreover, in all cases GMRES behaves almost identically for the Galerkin and the stabilized discretizations.

5. The two-dimensional case. In this section we consider two model problems on a rectangular domain. The first is a constant-coefficient problem with velocity at an angle to the coordinate axes; the second involves a semicircular velocity field. The numerical experiments will focus only on the stabilized problem since, for interesting grid Péclet numbers, it yields the only physically meaningful discretization.

5.1. First model problem. We consider the Dirichlet problem (2.1), (2.2) on the unit square $\Omega = (0,1) \times (0,1)$ with the constant coefficients $\boldsymbol{\kappa} = \kappa I_2$ and $\boldsymbol{a} = a(\cos\theta, \sin\theta)^T$ with $f = 0$. We discretize the problem using bilinear elements on a uniform rectangular mesh. The resulting stiffness matrices A and $\tilde{A}$ can then be written as the sum of Kronecker products

$$A = M \otimes \left(\kappa_{11} K + \frac{a_1 h}{2} C\right) + \left(\kappa_{22} K + \frac{a_2 h}{2} C\right) \otimes M - \frac{\kappa_{12} + \kappa_{21}}{4} C \otimes C$$

in terms of the matrices

$$M = \frac{1}{6}\ \text{tridiag}(1, 4, 1), \quad K = \ \text{tridiag}(-1, 2, -1), \quad C = \ \text{tridiag}(-1, 0, 1),$$

which are recognized as the mass, stiffness, and gradient matrices of the discretization of the one-dimensional constant-coefficient/uniform mesh model problem using linear elements. The diffusivity tensor $\boldsymbol{\kappa}$ is given by $\boldsymbol{\kappa} = \kappa I$ in the Galerkin case and by $\boldsymbol{\kappa} = \kappa I + \tau \boldsymbol{a}\boldsymbol{a}^T$ in the stabilized case. Using the definition (2.8) again for the stabilization parameter τ, we obtain, after scaling the system by $(\kappa(1 + \alpha\xi))^{-1}$, the coefficient matrix

$$\begin{aligned}\tilde{A} = M \otimes &\left(\frac{1 + \alpha\xi c^2}{1 + \alpha\xi} K + c \tanh(\alpha) C\right) \\ &+ \left(\frac{1 + \alpha\xi s^2}{1 + \alpha\xi} K + s \tanh(\alpha) C\right) \otimes M + \frac{sc}{2} \xi \tanh(\alpha) C \otimes C\end{aligned}$$

for the stabilized discretization, where $c = \cos\theta$, $s = \sin\theta$. The corresponding matrix for the Galerkin case is obtained by setting $\xi = 0$.

Although the Kronecker product representation of $\tilde{A}$ seems simple enough, we were unable to find a closed form representation for the eigenvalues, eigenvectors,

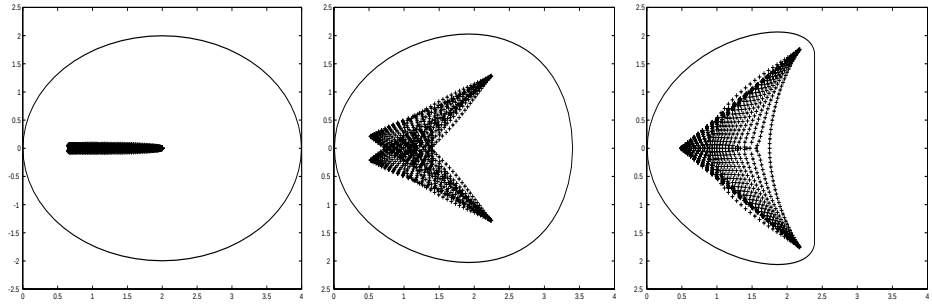


FIG. 5.1. $W(\tilde{A})$ *and* $\Lambda(\tilde{A})$ *for* $N = 31^2$, $\alpha = 1000$, *and* $\theta = 0$, $\pi/8$, *and* $\pi/4$.

and field of values of the discretization matrices of even this simple model problem without making some simplifications. One such simplification results from setting the flow angle θ to zero, which corresponds to flow in the direction of the x-axis (cf. also [11]). In this case $s = 0$, $c = 1$, and the resulting matrix $\tilde{A}_0$ is the sum of only two Kronecker products

$$\tilde{A}_0 = M \otimes (K + \tanh(\alpha)C) + \frac{\tanh(\alpha)}{\alpha} K \otimes M.$$

The first factors of the two terms are both symmetric tridiagonal Toeplitz matrices, and hence they share a common system of orthogonal eigenvectors $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$ such that

$$M\boldsymbol{u}_j = \mu_j \boldsymbol{u}_j \quad \text{and} \quad K\boldsymbol{u}_j = \nu_j \boldsymbol{u}_j, \qquad j = 1, \ldots, n,$$

where the eigenvalues $\{\mu_j\}_{j=1}^n$ and $\{\nu_j\}_{j=1}^n$ and eigenvectors $\{\boldsymbol{u}_j\}_{j=1}^n$ are given in terms of the well-known trigonometric formulas

$$\begin{aligned} \mu_j &= \frac{1}{3}(2 + \cos(j\pi h), \\ \nu_j &= 2(1 - \cos(j\pi h)), \\ \text{and } (\boldsymbol{u}_j)_k &= \sin(jk\pi h), \qquad k = 1, \ldots, n. \end{aligned}$$

By using the properties of Kronecker products, we obtain that a vector $\boldsymbol{u}_j \otimes \boldsymbol{v}_j \in \mathbb{R}^{n^2}$ is an eigenvector of $\tilde{A}_0$ if the vector $\boldsymbol{v}_j \in \mathbb{R}^n$ is an eigenvector of the matrix

$$\mu_j \, (K + \tanh(\alpha)C) + \nu_j \frac{\tanh(\alpha)}{\alpha} M, \quad j = 1, \ldots n.$$

Each value of j yields a nonsymmetric tridiagonal Toeplitz matrix whose eigenvalues and eigenvectors can be determined using the results cited in the appendix.

For other flow angles, we resort to numerical computation. Figure 5.1 shows the computed eigenvalues and the boundary of $W(\tilde{A})$ for $N = 31^2$, $\alpha = 1000$, $\kappa = 1$ and flow angles $\theta = 0$, $\pi/8$, and $\pi/4$. In the second and third cases the eigenvalues were computed with MATLAB's `eig` function and the boundary of the field of values was computed by finding the extremal eigenvalues of the symmetric part of rotated matrices using the Lanczos algorithm combined with Chebyshev acceleration (cf. [1]). We note that, in terms of distance of the boundary $\partial W(\tilde{A})$ to the convex hull of the

TABLE 5.1
Asymptotic convergence factors of $W(\tilde{A})$ and $CH(\Lambda(\tilde{A}))$.

	$\theta = 0$	$\theta = \pi/8$	$\theta = \pi/4$
$W(\tilde{A})$	0.999	0.995	0.997
$CH(\Lambda(\tilde{A}))$	0.308	0.631	0.658

spectrum, the matrix seems farthest from normal in the case of flow along one of the coordinate axes.

Next, we compute the asymptotic convergence factors γ of both the convex hull $CH(\Lambda(\tilde{A}))$ of the spectrum and the field of values for the three cases depicted in Figure 5.1. Since these sets are both simply connected and do not contain the origin, γ may be calculated via the conformal mapping connection (3.3). To this end, we have used the Schwarz–Christoffel Toolbox (version 2.0) of Driscoll [5] to evaluate the exterior mapping Φ of $CH(\Lambda(\tilde{A}))$ and a polygonal approximation of $W(\tilde{A})$, respectively, at the origin. The results are shown in Table 5.1. As expected from the eigenvalue plot, the convergence factors of $CH(\Lambda(\tilde{A}))$ are much smaller than those of $W(\tilde{A})$, the latter being very close to unity due to the proximity of the field of values to the origin.

The convergence behavior of GMRES applied to the linear systems belonging to the three flow angles, $\alpha = 1000$ and $N = 31^2$ is shown in Figure 5.2. In all cases, a zero initial vector and a random right-hand side were used. The figures on the left show the normalized residual norms $\|\boldsymbol{r}_k\|/\|\boldsymbol{r}_0\|$ of GMRES applied to each of the three systems (solid line); the dashed lines indicate the linear convergence rates predicted by the convex hull of the spectrum and by the field of values, respectively. To make sure that the convex hull of the spectrum isn't overestimating the convergence rate, we have also included the residual curve of GMRES applied to the same system with $\tilde{A}$ replaced by a diagonal matrix with the same eigenvalues as $\tilde{A}$ (dotted line). The three plots on the right show the corresponding lifespan curves $\|\boldsymbol{r}_{k+1}\|/\|\boldsymbol{r}_k\|$. We observe two distinct phases of linear convergence in the residual curves of all three cases. The convergence rates of the first phase are slightly below the rate predicted by the field of values. Those of the second phase seem to approach the rate the spectrum would predict, in the absence of nonnormality, although the lifespan curves lie slightly below and above this rate in the first and third examples. The transition between the two phases occurs at iteration steps 35 and 44 for flow angles $\theta = \pi/8$ and $\theta = \pi/4$, respectively, which are upper bounds for the number of steps information would take to traverse the underlying finite element grid. For the case $\theta = 0$, however, this transition does not occur until step 52. Moreover, the transition is much more gradual than in the other two cases.

5.2. Second model problem. Our second model problem is a slight modification of a widely used test problem for discretizations of convection-diffusion equations (cf. [18, p. 10]). The domain is the rectangle $\Omega = (-1, 1) \times (0, 1)$, the diffusion tensor $\boldsymbol{\kappa}$ is the identity, and the incompressible velocity field is given by

$$\boldsymbol{a}(x, y) = 2a \begin{bmatrix} y(1 - x^2) \\ -x(1 - y^2) \end{bmatrix},$$

the semicircular flow pattern of which is shown in Figure 5.3. The parameter a can be used to vary the Péclet number. Along the resulting inflow boundary we impose

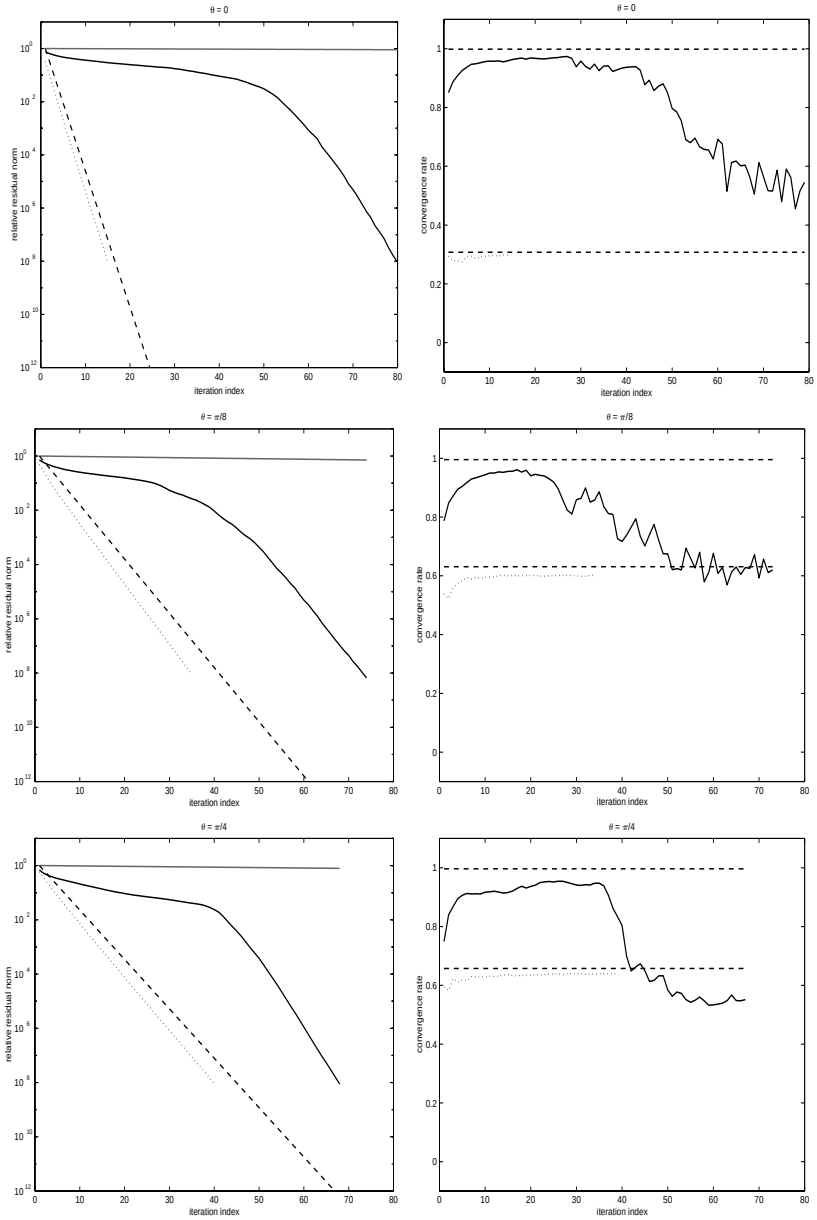


FIG. 5.2. *GMRES residual curves for* $\theta = 0, \pi/8,$ *and* $\pi/4$, $\alpha = 1000$ *and* $N = 31^2$.

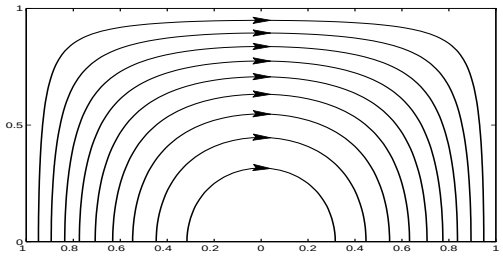


FIG. 5.3. *Velocity field of second model problem.*

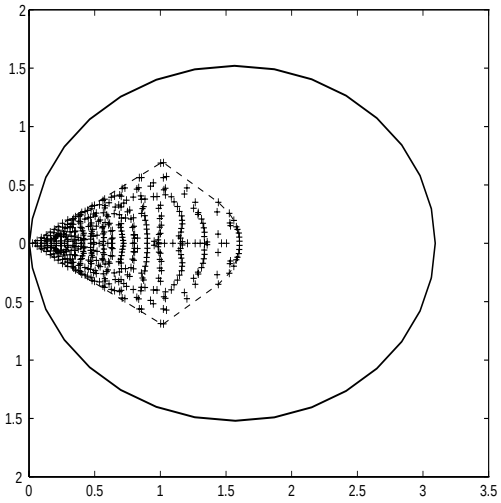


FIG. 5.4. *Field of values, spectrum, and the convex hull of the spectrum for the stabilized discretization of the second model problem on a* 32×16 *grid.*

the Dirichlet condition

$$u(x,0) = 1 + \tanh[\nu(2x+1)], \qquad -1 \le x \le 0,$$

in which another parameter ν determines the sharpness of the inflow profile. The remaining Dirichlet conditions are given by

$$\begin{aligned} &u(x,0) = 0, \qquad 0 \le x \le 1, \\ &u(x,y) = 1 + \tanh(\nu) \qquad \text{on the remaining portion of } \partial\Omega. \end{aligned}$$

Figure 5.4 shows the field of values, the spectrum and the convex hull of the spectrum for this problem with $a = 10^5$ discretized on a uniform mesh of 32×16 bilinear rectangular elements, which corresponds to a grid Péclet number of $\alpha = 6250$. The field of values is more than twice the size in diameter than the spectrum, so some nonnormality effects can be expected. Again, we have scaled the problem by $(\kappa(1 + \alpha\xi))^{-1}$.

In Figure 5.5, the solid line represents the GMRES residual norm curve for this problem. As before, the upper and lower dashed lines show the linear rates of convergence predicted by the asymptotic convergence factors of $W(\tilde{A})$ and $CH(\Lambda(\tilde{A})$, respectively. The dotted line is the residual curve of GMRES applied to a diagonal matrix $\tilde{D}$ with the same eigenvalues as $\tilde{A}$ using a zero initial guess and a random right-hand side. Since $\tilde{D}$ is a normal matrix, the convergence of GMRES is completely determined by $\Lambda(\tilde{D}) = \Lambda(\tilde{A})$. Again we observe two distinct phases of linear convergence. The rate in the first phase is somewhat overestimated by the field of values but noticeably smaller than the rate predicted by the spectrum. The convex hull of the spectrum also overestimates the rate in the second phase, but the residual curve of the diagonal matrix is seen to have the same rate as that observed in the second phase. The transition between these two phases takes place at iteration step 40, which is the roughly the number of iteration steps required for the profile prescribed at the inlet boundary to propagate across the mesh to the outflow boundary.

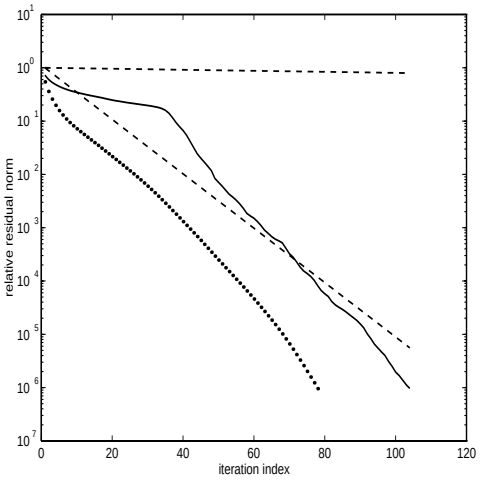


FIG. 5.5. *GMRES residual norms for the solution of the second model problem (solid), linear convergence rate predicted by $W(\tilde{A}$ (upped dashed) and $CH(\Lambda(\tilde{A}))$ (lower dashed), and residual norms for GMRES applied to diagonal matrix D with $\Lambda(D) = \Lambda(\tilde{A})$.*

6. Conclusions. We have tried to gain insight into the convergence behavior of residual minimizing Krylov subspace methods for stabilized discretizations of convection-diffusion problems by studying various model problems. The one-dimensional results show that, while the stabilization results in a better discretization of the boundary value problem, it also results in a highly nonnormal discrete operator. The nonnormality can be characterized by the condition number of the eigenvector matrix, which grows exponentially with both the grid Péclet number and the grid size. A consequence of the high degree of nonnormality is that spectral information is virtually useless for assessing the convergence rate of Krylov subspace methods. We have also seen that the field of values is a viable alternative in this case. In the two-dimensional examples, we observed that, while less pronounced, nonnormality is still an issue. Its effect, as observed in computational experiments, is an initial rate of convergence governed essentially by the field of values, after which convergence governed by the spectrum takes over. It is conjectured that the duration of the initial phase is governed by the time it takes for boundary information to pass from the inflow boundary across the domain following the streamlines of the velocity field.

Appendix A. Ellipses and Chebyshev polynomials. In this section we collect some results pertaining to Chebyshev polynomials on ellipses in the complex plane. Chebyshev polynomials are often used to bound the convergence rate of Krylov subspace methods. For real intervals not containing the origin (and for ellipses with real foci "far enough away" from the origin [10]), scaled Chebyshev polynomials *are* the polynomials of least maximum modulus normalized at the origin. For general ellipses, the Chebyshev polynomials are still asymptotically optimal, i.e., they satisfy $\lim_{m\to\infty} \|p_m\|_\Omega^{1/m} = \gamma(\Omega)$. These results are well known [4, 17, 6] and we include them only for convenient reference.

We parameterize an ellipse in the complex plane by two complex numbers σ and τ such that the former represents the midpoint of the focal line and the two foci lie at $\sigma \pm \tau$. A third parameter $\rho > 1$ is used to parameterize the family of ellipses with these two foci ($0 < \rho < 1$ yields the same family again). The closed interior of an

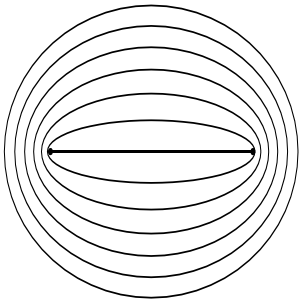


FIG. A.1. *Ellipses in the confocal family* $\mathcal{E}_\rho = \mathcal{E}_\rho(0,1)$, $\rho = 1, 1.25, \ldots, 2.5$.

ellipse can then be characterized in terms of the sum of the distances to the foci as the set

$$\mathcal{E}_\rho(\sigma,\tau) = \left\{ z \in \mathbb{C} : |z-\sigma+\tau| + |z-\sigma-\tau| \le |\tau| \left(\rho + \rho^{-1}\right) \right\}.$$

The boundary of the ellipse may be parameterized as

$$\partial\mathcal{E}_\rho(\sigma,\tau) = \left\{ z = \sigma + \frac{\tau}{2}\left(\rho e^{i\theta} + \rho^{-1} e^{-i\theta}\right) : \theta \in [0, 2\pi) \right\},$$

from which we immediately see that the two semiaxes, i.e., the largest and smallest distances from σ to the boundary of the ellipse, are $|\tau|(\rho \pm \rho^{-1})/2$. It is convenient to consider Chebyshev polynomials on the family of confocal ellipses $\mathcal{E}_\rho = \mathcal{E}_\rho(0,1)$ with foci located at ± 1. A given ellipse $\mathcal{E}_\rho(\sigma,\tau)$ is mapped onto the associated ellipse from this family having the same eccentricity (which is $|\tau|$) by the linear transformation $\zeta = (z-\sigma)/\tau$. The ellipses $\mathcal{E}_\rho$ for $\rho = 1, 1.25, \ldots, 2.5$ are shown in Figure A.1. For $\zeta \in [-1,1]$, the first-kind Chebyshev polynomials can be defined by

$$T_m(\zeta) = \cos(m \arccos\zeta), \quad m = 0, 1, \ldots.$$

These polynomials satisfy the recurrence relation

$$T_0 \equiv 1, \quad T_1(\zeta) = \zeta, \quad T_{m+1}(\zeta) = 2\zeta T_m(\zeta) - T_{m-1}(\zeta) \quad (m > 1) \tag{A.1}$$

and their zeros are all contained in the interval $(-1,1)$. The analysis of the Chebyshev polynomials is facilitated by introducing the Joukowski map

$$\zeta = \Psi(w) := \frac{1}{2}\left(w + \frac{1}{w}\right), \quad w \neq 0,$$

which maps the exterior $|w| > 1$ of the unit circle one-to-one onto $\mathbb{C}\backslash[-1,1]$. Each circle $|w| = \rho > 1$ is mapped to the ellipse $\mathcal{E}_\rho$. The inverse map Φ is given by

$$w = \Phi(\zeta) = \zeta + \sqrt{\zeta^2 - 1}, \quad \zeta \notin [-1,1],$$

in which that branch of the square root is chosen which results in $|\Phi(\zeta)| > 1$. (A method for selecting the proper branch is described in [13, p. 296].) Using the recurrence relation (A.1), one easily verifies that the Chebyshev polynomials satisfy

$$T_m(\zeta) = \frac{1}{2}(w^m + w^{-m}) = \frac{1}{2}[\Phi(\zeta)^m + \Phi(\zeta)^{-m}], \quad \zeta \notin [-1,1].$$

This also holds for $z \in [-1,1]$, but then the w associated with ζ is no longer unique. Setting $w = \rho e^{i\theta}$, this yields

$$T_m(\zeta) = \frac{1}{2}(\rho^m e^{im\theta} + \rho^{-m} e^{-im\theta}), \tag{A.2}$$

from which we conclude that the mth Chebyshev polynomial maps the ellipse $\mathcal{E}_\rho$ to the ellipse $\mathcal{E}_{\rho^m}$, which is covered m times. This implies that

$$\frac{1}{2}(\rho^m - \rho^{-m}) \le |T_m(\zeta)| \le \frac{1}{2}(\rho^m + \rho^{-m}), \quad \zeta \in \mathcal{E}_\rho,$$

or, equivalently, $|T_m(\zeta)| = (\rho^m + t_m \rho^{-m})/2$ with $|t_m| < 1$.

We obtain normalized Chebyshev polynomials p_m on arbitrary ellipses $\mathcal{E}_\rho(\sigma,\tau)$ not containing the origin by shifting and scaling the Chebyshev polynomials on $\mathcal{E}_\rho$:

$$p_m(z) := \frac{T_m(\zeta(z))}{T_m(\zeta(0))} = \frac{T_m((z-\sigma)/\tau)}{T_m(-\sigma/\tau)}, \quad m = 0, 1, \ldots. \tag{A.3}$$

Since $0 \notin \mathcal{E}_\rho(\sigma,\tau)$ implies $\zeta(0) \notin \mathcal{E}_\rho$ and since the zeros of T_m lie in $(-1,1)$, we have $T_m(\zeta(0)) \neq 0$.

We also define

$$\gamma = \gamma(\sigma,\tau) := \frac{1}{|\Phi(\zeta(0))|} = \frac{1}{|\Phi(\frac{-\sigma}{\tau})|}. \tag{A.4}$$

We can now describe the asymptotic properties of the residual polynomials defined by (A.3).

THEOREM A.1. *If σ and τ are arbitrary complex numbers such that $0 \notin \mathcal{E}_\rho(\sigma,\tau)$, then the residual polynomials defined in* (A.3) *satisfy*

$$\|p_m\|_{\mathcal{E}_\rho(\sigma,\tau)} \le (\rho^m + \rho^{-m}) \frac{\gamma^m}{1-\gamma^{2m}}. \tag{A.5}$$

The corresponding bound for $\|p_m\|_{[\sigma-\tau,\sigma+\tau]}$ is obtained by setting $\rho = 1$.

Proof. By their definition, the polynomials p_m satisfy

$$|p_m(z)| \le \frac{1}{2}(\rho^m + \rho^{-m}) \frac{1}{|T_m(\zeta(0))|}. \tag{A.6}$$

It is thus only necessary to examine $|T_m(\zeta(0))| = |T_m(-\sigma/\tau)|$. Since $\gamma^{-1} = |\Phi(\zeta(0))|$ lies outside the unit circle, its image under Ψ—i.e., $\zeta(0)$—lies on the boundary of the ellipse $\mathcal{E}_{\gamma^{-1}}$ and hence, by the mapping properties of the Chebyshev polynomials, $T_m(\zeta(0))$ lies in the ellipse $\mathcal{E}_{\gamma^{-m}}$. This implies

$$\frac{1}{2}(\gamma^{-m} - \gamma^m) \le |T_m(\zeta(0))| \le \frac{1}{2}(\gamma^{-m} + \gamma^m), \tag{A.7}$$

and, together with (A.6), yields (A.5). The assertion for the complex line segment is obtained in the same way by using the fact that $|T_m(\zeta)| \le 1|$ for $\zeta \in [-1,1]$. ∎

Two special cases allow the bounds in Theorem A.1 to be improved, as follows.

THEOREM A.2. *If, in addition to the assumptions of Theorem* A.1*, the foci of $\mathcal{E}_\rho(\sigma,\tau)$ and the origin are colinear, there holds*

$$\|p_m\|_{\mathcal{E}_\rho(\sigma,\tau)} \le (\rho^m + \rho^{-m}) \frac{\gamma^m}{1+\gamma^{2m}}. \tag{A.8}$$

If, in addition to the assumptions of Theorem A.1*, the line connecting* σ *and the origin is perpendicular to the focal line of* $\mathcal{E}_\rho(\sigma,\tau)$*, then*

$$\|p_m\|_{\mathcal{E}_\rho(\sigma,\tau)} \le (\rho^m + \rho^{-m}) \begin{cases} \frac{\gamma^m}{1+\gamma^{2m}} & \text{for } m \text{ even,} \\ \frac{\gamma^m}{1-\gamma^{2m}} & \text{for } m \text{ odd.} \end{cases} \tag{A.9}$$

The corresponding bounds for $\|p_m\|_{[\sigma-\tau,\sigma+\tau]}$ *are obtained by setting* $\rho = 1$*.*

Proof. Since $\gamma^{-1} = |\Phi(\zeta(0))|$, we can write $\Phi(\zeta(0)) = \gamma^{-1}e^{i\alpha}$, with α being the argument of $\Phi(\zeta(0))$. Using the relation (A.2), we obtain

$$|T_m(\zeta(0))| = \left|\frac{1}{2}(\gamma^{-m}e^{im\alpha} + \gamma^m e^{-im\alpha})\right|.$$

Thus,

$$|T_m(\zeta(0))| \le \frac{1}{2}\begin{cases} \gamma^{-m} + \gamma^m & \text{for } \arg(\Phi(\zeta(0))) = \frac{k}{m}\pi, \\ \gamma^{-m} - \gamma^m & \text{for } \arg(\Phi(\zeta(0))) = \frac{2k+1}{2m}\pi, \end{cases} \qquad k \in \mathbb{Z}.$$

If the foci are colinear or, equivalently, $-\sigma/\tau$ is real, this implies that also $\Phi(-\sigma/\tau)$ is real and hence its argument can be written in the form $k\pi/m$ for some suitable $k \in \{0,\dots,2m-1\}$. In the second special case $-\sigma/\tau$, and hence also $\Phi(\sigma/\tau)$, is pure imaginary. For even m, $\arg(\Phi(-\sigma/\tau))$ is equal to $\pi/2$ or $3\pi/2$ and may be written in the form $k\pi/m$. For odd m it can be written in the form $(2k+1)\pi/(2m)$. ∎

As a consequence of Theorem A.1, the asymptotic convergence factor of the residual polynomials p_m on the ellipse $\mathcal{E}_\rho(\sigma,\tau)$ is given by

$$\lim_{m\to\infty} \|p_m\|^{1/m}_{\mathcal{E}_\rho(\sigma,\tau)} = \rho\gamma.$$

Appendix B. Spectral properties and the field of values of tridiagonal Toeplitz matrices. For convenient reference, we include a derivation of the eigenvalues, eigenvectors and field of values of tridiagonal Toeplitz matrices. The presentation follows [6].

B.1. Eigensystem. It is sufficient to consider tridiagonal Toeplitz matrices with zero diagonal

$$T = \text{ tridiag}(\alpha, 0, \beta) \in \mathbb{C}^{n\times n},$$

since otherwise the spectrum is simply shifted by the diagonal term. For $\alpha\beta = 0$ the spectrum consists of one zero eigenvalue of algebraic multiplicity n and, unless $\alpha = \beta = 0$, geometric multiplicity one. Assuming $\alpha\beta \neq 0$, the matrix T may be symmetrized by the diagonal similarity transformation $D^{-1}TD = \eta T_1$, where

$$\eta = \eta(\alpha,\beta) = \sqrt{|\alpha\beta|}e^{\frac{i}{2}[\arg(\alpha)+\arg(\beta)]}$$

and $T_1 = \text{ tridiag}(1,0,1)$ and $D = \text{diag}(\delta, \delta^2, \dots, \delta^n)$ with

$$\delta = \delta(\alpha,\beta) = \sqrt{|\alpha/\beta|}e^{\frac{i}{2}[\arg(\alpha)-\arg(\beta)]}.$$

As is easily verified using basic trigonometric identities, the eigenvalues of T_1 are

$$\mu_j = 2\cos\left(\frac{j\pi}{n+1}\right) \qquad (j = 1,\dots,n) \tag{B.1}$$

with corresponding normalized orthogonal eigenvectors $\boldsymbol{u}_j = [u_{1j}, \ldots, u_{nj}]^T$ given by

$$u_{kj} = \sqrt{\frac{2}{n+1}} \sin\left(\frac{kj\pi}{n+1}\right) \qquad (k, j = 1, \ldots, n).$$

For T this results in the eigenvectors

$$\boldsymbol{t}_j = D\boldsymbol{u}_j = [\delta u_{1j}, \ldots, \delta^n u_{nj}]^T \qquad (j = 1, \ldots, n)$$

and corresponding eigenvalues $\tau_j = \eta\mu_j$ $(j = 1, \ldots, n)$. The eigenvector matrix is $V = [\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n] = DU$.

B.2. Field of values.

LEMMA B.1. *The field of values of the tridiagonal Toeplitz matrix* $T = \mathrm{tridiag}(\alpha, 0, \beta) \in \mathbb{C}^{n\times n}$ *consists of the closed interior of the ellipse*

$$z = c_n(\alpha e^{i\theta} + \beta e^{-i\theta}), \quad c_n = \cos(\pi/(n+1)), \quad 0 \le \theta < 2\pi. \tag{B.2}$$

Proof. We begin by showing that the field of values of the $n \times n$ Jordan block $J = \mathrm{tridiag}(0, 0, 1)$, which is connected with T by $T = \alpha J^T + \beta J$, is a closed disk centered at the origin. If $\zeta \in W(J)$, then there exists a vector $\boldsymbol{z} \in \mathbb{C}^n$ of unit norm such that $\zeta = \boldsymbol{z}^H J\boldsymbol{z}$. If a new vector $\boldsymbol{w}$ is defined by $w_j = e^{ij\phi}z_j$ $(j = 1, \ldots, n)$ with an arbitrary angle ϕ, then we have $\|\boldsymbol{w}\|_2 = 1$ and

$$\boldsymbol{w}^H J\boldsymbol{w} = \sum_{j=1}^{n-1} \overline{w_j} w_{j+1} = e^{i\phi} \sum_{j=1}^{n-1} \overline{z_j} z_{j+1} = e^{i\phi}\zeta.$$

This shows that, for any $\zeta \in W(J)$, the entire circle $|z| = |\zeta|$ must also belong to $W(J)$. Since the field of values is convex, this means that $W(J)$ must be a disk centered at the origin. The radius of this disk is obtained as the largest eigenvalue of $(J + J^H)/2$. This follows from the characterization of the field of values of a matrix A as the intersection of strips S_θ in the complex plane:

$$W(A) = \bigcap_{\theta\in[0,\pi)} S_\theta, \qquad S_\theta = \{z \in \mathbb{C} : \lambda_{\min}(H_\theta) \le \mathrm{Re}\ (e^{i\theta}z) \le \lambda_{\max}(H_\theta)\},$$

where H_θ is the Hermitian part $H_\theta := (A_\theta + A_\theta^H)/2$ of the rotated matrix $A_\theta = e^{i\theta}A$. By (B.1), the largest eigenvalue of $J + J^T$ is $c_n = \cos(\pi/(n+1))$, hence $W(J) = S(0, c_n)$.

Now assume $\omega \in W(T)$. This means there exists $\boldsymbol{w} \in \mathbb{C}^n$, $\|\boldsymbol{w}\|_2 = 1$, such that

$$\omega = \boldsymbol{w}^H T\boldsymbol{w} = \alpha\boldsymbol{w}^H J^H\boldsymbol{w} + \beta\boldsymbol{w}^H J\boldsymbol{w} = \alpha\overline{\zeta} + \beta\zeta$$

for some $\zeta \in W(J_n)$. Thus, each point in $W(T)$ has the form (B.2) and, conversely, any complex number of the form (B.2) lies in $W(T)$. ∎

We note that, in the notation defined in (3.4), Lemma B.1 states that $W(T) = \mathcal{E}_\rho(\sigma, \tau)$ with

$$\sigma = 0, \quad \tau = 2\cos(\pi/(n+1))\sqrt{|\alpha\beta|}e^{\frac{i}{2}[\arg(\alpha)-\arg(\beta)]}, \quad \text{and} \quad \rho = \sqrt{|\alpha/\beta|}.$$

Appendix C. Field of values and spectrum of the one-dimensional differential operator. For the one-dimensional case, we determine the field of values

$$W(L) = \left\{ \frac{a(u,u)}{(u,u)} : 0 \neq u \in V \right\}$$

of the differential operator with constant coefficients

$$L : V \to V', \quad Lu = -(\kappa u')' + au', \quad V = H_0^1(0,1)$$

associated with the Dirichlet problem considered as an unbounded linear operator on $L^2(0,1)$ with associated bilinear form $a(u,v) = (Lu,v)$. Of course, for an unbounded operator, the field of values will be an unbounded set in the complex plane. Since L^{-1} is a compact operator, its spectrum consists only of eigenvalues and these accumulate at zero. To determine $W(L)$, we thus proceed similarly as in the finite-dimensional case. We consider the rotated operators $L_\theta = e^{i\theta}L, \theta \in (-\pi/2, \pi/2)$ and determine the minimal eigenvalues of their Hermitian parts

$$H_\theta = \frac{1}{2}(L_\theta + L_\theta^*),$$

where the L^2-adjoint operator is characterized by

$$(L_\theta u, v) = (u, L_\theta^* v) \quad \forall u, v \in V.$$

Integration by parts yields $L_\theta^* v = e^{-i\theta}[-(\kappa v')' - av']$, which results in

$$H_\theta u = -\cos(\theta)\kappa u'' + i\sin(\theta)au'$$

for the Hermitian part. The eigenvalues of H_θ are determined by those values of $\lambda = \lambda(\theta)$ for which the boundary value problem

$$\begin{gathered} H_\theta u = -\cos(\theta)\kappa u'' + i\sin(\theta)au' = \lambda u, \quad x \in (0,1), \\ u(0) = u(1) = 0 \end{gathered}$$

possesses nontrivial solutions. Solutions of the form $u(x) = e^{\mu x}$ are determined by the solutions $\mu_\pm$ of the characteristic equation

$$-\cos(\theta)\kappa\mu^2 + ia\sin(\theta)\mu - \lambda = 0.$$

After imposing the homogeneous boundary condition, this results in

$$\lambda_k(\theta) = k^2\pi^2\kappa\cos(\theta) - \frac{a^2\sin^2(\theta)}{4\kappa\cos(\theta)}, \quad k \in \mathbb{Z}.$$

Since $k = 0$ results in a constant (zero) eigenfunction, and $\cos(\theta) > 0$ for $\theta \in (-\pi/2, \pi/2)$, the smallest eigenvalue of H_θ is thus obtained for $k = 1$ as

$$\lambda_{\min}(\theta) = \pi^2\kappa\cos(\theta) - \frac{a^2\sin^2(\theta)}{4\kappa\cos(\theta)}.$$

We obtain the field of values $W(L)$ as the intersection of all half-planes

$$S_\theta = \{z \in \mathbb{C} : \lambda_{\min}(\theta) \leq \mathrm{Re}(e^{i\theta}z)\}, \qquad \theta \in (-\pi/2, \pi/2).$$

By intersecting the boundaries of two half-planes H_{θ_1} and H_{θ_2} and taking the limit $\theta_2 \to \theta_1$, we obtain a parametrization of the boundary of $W(L)$, which turns out to be the parabola

$$x = \kappa \left(\pi^2 + \frac{y^2}{a^2} \right).$$

The eigenvalues of the operator L itself may be obtained analogously and are given by

$$\lambda_k = \kappa \left[\left(\frac{a}{2\kappa} \right)^2 + k^2 \pi^2 \right], \quad k = 1, 2, \ldots.$$

Acknowledgment. The author would like to thank Michael Eiermann for many helpful discussions.

REFERENCES

[1] T. Braconnier and N. J. Higham, *Computing the field of values and pseudospectra using the Lanczos method with continuation*, BIT, 36 (1996), pp. 422–440.
[2] A. N. Brooks and T. J. R. Hughes, *A multidimensional upwind scheme with no crosswind diffusion*, in Finite Element Methods for Convection Dominated Flows, T. J. R. Hughes, ed., ASME, New York, 1979, pp. 199–259.
[3] A. N. Brooks and T. J. R. Hughes, *Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 32 (1982), pp. 199–259.
[4] P. J. Davis, *Interpolation and Approximation*, 2nd ed., Dover Books on Advanced Mathematics 14, Dover, New York, 1975.
[5] T. A. Driscoll, *A MATLAB toolbox for Schwarz-Christoffel mapping*, ACM Trans. Math. Software, 22 (1996), pp. 168–186.
[6] M. Eiermann, *Semiiterative Verfahren für nichtsymmetrische lineare Gleichungssysteme*, Habilitationsschrift, Universität Karlsruhe, Karlsruhe, Germany, 1989.
[7] M. Eiermann, *Fields of values and iterative methods*, Linear Algebra Appl., 180 (1993), pp. 167–197.
[8] M. Eiermann, *Field of Values and Iterative Methods*, 1996, manuscript.
[9] M. Eiermann, W. Niethammer, and R. S. Varga, *A study of semiiterative methods for nonsymmetric systems of linear equations*, Numer. Math., 47 (1985), pp. 505–533.
[10] B. Fischer and R. Freund, *Chebyshev polynomials are not always optimal*, J. Approx. Theory, 65 (1991), pp. 261–272.
[11] B. Fischer, A. Ramage, D. J. Silvester, and A. J. Wathen, *On parameter choice and iterative convergence for stabilised discretisations of advection-diffusion problems*, Comput. Methods Appl. Mech. Engrg., 179 (1999). pp. 185–202.
[12] P. M. Gresho and R. L. Lee, *Don't suppress the wiggles—they're telling you something*, in Finite Element Methods for Convection Dominated Flows, Vol. 34, AMD/ASME, New York, 1979, pp. 37–61.
[13] P. Henrici, *Applied and Computational Complex Analysis* I, Wiley Classics Library, Wiley-Interscience, New York, 1988.
[14] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–436.
[15] R. Horn and C. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.
[16] C. Johnson, U. Nävert, and J. Pitkäranta, *Finite element methods for linear hyperbolic problems*, Comput. Methods Appl. Mech. Engrg., 45 (1984), pp. 285–312.
[17] T. A. Manteuffel, *Adaptive procedure for estimating parameters for the nonsymmetric Tchebychev iteration*, Numer. Math., 31 (1978), pp. 183–208.
[18] K. W. Morton, *Numerical Solution of Convection-Diffusion Problems*, Appl. Math. Math. Comput., Chapman and Hall, London, 1996.

[19] N. M. Nachtigal, S. C. Reddy, and L. N. Trefethen, *How fast are nonsymmetric matrix iterations?*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 778–795.
[20] O. Nevanlinna, *Convergence of Iterations for Linear Equations*, Birkhäuser, Basel, 1993.
[21] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.
[22] H.-G. Roos, M. Stynes, and L. Tobiska, *Numerical Methods for Singularly Perturbed Differential Equations*, Springer Ser. Comput. Math. 24, Springer-Verlag, Berlin, 1996.
[23] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
[24] Z. Strakoš, V. Pták, and A. Greenbaum, *Any nonincreasing convergence curve is possible for GMRES*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 465–469.
[25] L. N. Trefethen, *Pseudospectra of matrices*, in Numerical Analysis 1991, D. F. Griffiths and G. A. Watson, eds., Longman Press, London, 1992.
[26] G. Zhou, *How accurate is the streamline diffusion finite element method?*, Math. Comp., 66 (1997), pp. 31–44.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 4, pp. 1102–1111
© 2000 Society for Industrial and Applied Mathematics

CONVERGENCE OF EIGENVALUES IN STATE-DISCRETIZATION OF LINEAR STOCHASTIC SYSTEMS*

JOSÉ A. DE DONÁ†, GRAHAM C. GOODWIN†, RICHARD H. MIDDLETON†, AND IAIN RAEBURN‡

Abstract. The transition operator that describes the time evolution of the state probability distribution for continuous-state linear systems is given by an integral operator. A state-discretization approach is proposed, which consists of a finite rank approximation of this integral operator. As a result of the state-discretization procedure, a Markov chain is obtained, in which case the transition operator is represented by a transition matrix. Spectral properties of the integral operator for the continuous-state case are presented. The relationships between the integral operator and the finite rank approximation are explored. In particular, the limiting properties of the eigenvalues of the transition matrices of the resulting Markov chains are studied in connection to the eigenvalues of the original continuous-state integral operator.

Key words. state-discretization, Markov chains, integral operators, finite rank approximation

AMS subject classifications. 60J10, 47N30

PII. S0895479897329874

1. Introduction.

The understanding and use of links between continuous and discrete features of systems has proven to be a fruitful area. In some cases, knowledge of properties from discrete processes gives insight into the continuous counterpart. On the other hand, the mathematical methods associated with the continuum (e.g., differential equations, integrals) very often turn out to be easier for analytical treatment than those associated with discrete coordinate axes (e.g., difference equations, summations) [2].

The algorithms available for solving discrete-state problems, for example, parameter estimation in Markov chains, have a very high computational complexity when the number of states is large. Typically the number of computations required is proportional to the square of the number of states, for example, the Viterbi algorithm or the probability grid filter (see, e.g., [13]). On the other hand, the tools available for solving linear continuous-state problems (e.g., the celebrated Kalman filter) are well known for their simplicity.

With these advantages of continuous-state algorithms over discrete-state ones in mind, we are interested in looking at continuous-state approximations to discrete-state problems. This line of research is aimed at identifying the appropriate geometric structures in discrete-state models which are amenable to linear continuous-state approximations. As part of this program of characterization of the class of discrete-state systems suitable for continuous-state approximation, we are studying the inverse process,

*Received by the editors November 10, 1997; accepted for publication (in revised form) by A. C. M. Ran February 17, 1999; published electronically March 21, 2000. This work was supported by the Centre for Integrated Dynamics and Control, The University of Newcastle, Australia. Part of this work has been published, in a preliminary form, by the same authors in *Consistent state-discretization of linear systems*, in Proceedings of the 36th IEEE Conference on Decision and Control, San Diego, CA, 1997, pp. 878–883.

http://www.siam.org/journals/simax/21-4/32987.html

†Department of Electrical and Computer Engineering, The University of Newcastle, Newcastle, NSW 2308, Australia (eejose@ee.newcastle.edu.au, eegcg@cc.newcastle.edu.au, rick@ee.newcastle.edu.au).

‡Department of Mathematics, The University of Newcastle, Newcastle, NSW 2308, Australia (iain@maths.newcastle.edu.au).

namely, the properties of discrete-state systems that arise in the state-discretization of continuous-state linear systems. In particular, this paper examines the asymptotic eigenstructure of the discrete-state systems (Markov chains) generated by the state-discretization of linear continuous-state systems. Related work can be found in [4, 6, 10] in a general setting. However, since we are restricting our study to linear continuous-state systems, we are able to provide more specific information about the structure of the discrete-state approximation.

This paper is organized as follows. In section 2 we present some properties of state space formulations of linear systems that are relevant for their state-discretization. An operator which provides the time evolution of continuous-state probability density functions, together with its eigenvalues and eigenfunctions, is discussed in that section. In section 3 a state-discretization approach for linear systems is presented, and properties of the resulting discrete-state Markov chains are discussed. In particular, it is proved that the eigenvalues of the transition matrices obtained converge to a subset of the eigenvalues of the original continuous-state operator. Section 4 is dedicated to an example in order to illustrate the material presented.

2. Discrete-time linear state space models. Let a linear discrete-time system be described by the state space equation

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{v}_k, \tag{1}$$

where $\mathbf{x}_k, \mathbf{x}_{k+1} \in \mathbb{R}^n$ are the states at discrete-times k and $k+1$, respectively, $\mathbf{v}_k \in \mathbb{R}^m$ is an independent and identically distributed (iid) sequence of random vectors with finite variance, and the initial state $\mathbf{x}_0$ is assumed to have an arbitrary distribution with finite variance and independent of $\mathbf{v}_k$ for all k. $\mathbf{A}$ and $\mathbf{B}$ are real matrices of appropriate dimension, and we assume $\mathbf{A}$ such that all its eigenvalues belong to the interior of the unit disk. For simplicity of presentation we will further assume that the matrix $\mathbf{A}$ is invertible (see, e.g., (2)). This condition is naturally satisfied in the context of time-discretization of a continuous time system (see, e.g., [1]).

We will denote the joint probability density function (PDF) of a vector of n random variables $\mathbf{y}$, evaluated at the n-tuple $(x_1, x_2, \dots, x_n)$, by $f_{\mathbf{y}}(\mathbf{x}) \triangleq f_{\mathbf{y}}(x_1, x_2, \dots, x_n)$, where $\mathbf{x}$ denotes the vector $\mathbf{x} = (x_1, \dots, x_n)^T \in \mathbb{R}^n$.

We assume here that all the PDFs of interest have compact support and belong to the Hilbert space $\mathcal{L}^2([a_1, b_1] \times \cdots \times [a_n, b_n])$ (e.g., the uniform distribution). This assumption is necessary in the proofs below, and from a computational point of view, in most interesting applications densities do not significantly differ from zero outside a suitable compact set.

The PDF of $\mathbf{x}_{k+1}$ can then be expressed in terms of those of $\mathbf{x}_k$ and $\mathbf{v}_k$ (for example, by using the auxiliary variables method; see, e.g., [8]) as follows:

$$\begin{aligned} f_{\mathbf{x}_{k+1}}(\mathbf{x}) &= \frac{1}{|\det(\mathbf{A})|} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f_{\mathbf{x}_k}\left(\mathbf{A}^{-1}(\mathbf{x} - \mathbf{B}\boldsymbol{\xi})\right) f_{\mathbf{v}}(\boldsymbol{\xi}) d\xi_1 \dots d\xi_m \\ &\triangleq (T f_{\mathbf{x}_k})(\mathbf{x}), \end{aligned} \tag{2}$$

where $\mathbf{x} \in \mathbb{R}^n, \boldsymbol{\xi} \in \mathbb{R}^m$ and (2) defines a convex operator on the convex space of PDFs. Since we are interested in spectral properties, we work with the extension of this operator to the linear space $\mathcal{L}^2([a_1, b_1] \times \cdots \times [a_n, b_n])$. Note that this extension is the linear operator T on $\mathcal{L}^2([a_1, b_1] \times \cdots \times [a_n, b_n])$ defined by (2). In (2) we omitted k in $f_{\mathbf{v}_k}$ since $\mathbf{v}_k$ is iid.

In what follows we use the *n-dimensional characteristic function* (which exists for all distributions; see, e.g., [9]) and denote

$$\Phi_{\mathbf{x}_k}(i\mathbf{w}) = \mathcal{F}\{f_{\mathbf{x}_k}(\mathbf{x})\} = E\exp\{-i\mathbf{w}^T\mathbf{x}_k\}, \tag{3}$$

$$\Phi_{\mathbf{v}}(i\mathbf{w}) = \mathcal{F}\{f_{\mathbf{v}}(\mathbf{x})\} = E\exp\{-i\mathbf{w}^T\mathbf{v}\}, \tag{4}$$

where $\mathcal{F}\{\cdot\}$ stands for the Fourier transform and $i \triangleq \sqrt{-1}$. In expression (3) $\mathbf{w} \in \mathbb{R}^n$, and in (4) $\mathbf{w} \in \mathbb{R}^m$.

In the Fourier transform domain (2) becomes

$$\Phi_{\mathbf{x}_{k+1}}(i\mathbf{w}) = \Phi_{\mathbf{x}_k}(\mathbf{A}^T i\mathbf{w})\Phi_{\mathbf{v}}(\mathbf{B}^T i\mathbf{w}) \triangleq (\tilde{T}\Phi_{\mathbf{x}_k})(i\mathbf{w}), \tag{5}$$

where $\mathbf{w} = (w_1, \ldots, w_n)^T \in \mathbb{R}^n$ and $\tilde{T}$ is the linear operator on Fourier transforms (which corresponds to the linear operator T).

Since $\mathbf{v}_k \in \mathbb{R}^m$ is iid and the eigenvalues of matrix $\mathbf{A}$ are less than unity in modulus, there exists a *stationary PDF*, denoted $f_{\mathbf{x}_\infty}$, such that satisfies $f_{\mathbf{x}_\infty} = Tf_{\mathbf{x}_\infty}$ in (2) (see, for example, Theorems 2.3 and 2.7 in [7] for a proof of the more general case where the coefficients in the autoregressive model are random).

We let $\Phi_{\mathbf{x}_\infty}$ denote the Fourier transform of the stationary distribution, satisfying

$$\Phi_{\mathbf{x}_\infty}(i\mathbf{w}) = (\tilde{T}\Phi_{\mathbf{x}_\infty})(i\mathbf{w}) = \Phi_{\mathbf{x}_\infty}(\mathbf{A}^T i\mathbf{w})\Phi_{\mathbf{v}}(\mathbf{B}^T i\mathbf{w}). \tag{6}$$

For simplicity, the following theorem is formulated in the Fourier domain, although an equivalent result can be obtained in the state domain (see Remark 2.1).

THEOREM 2.1. *The complex numbers $\lambda_{\tilde{T}}$ and the functions $V_{\tilde{T}}$ given by*

$$\lambda_{\tilde{T}} = (\lambda_1)^{l_1}(\lambda_2)^{l_2}\ldots(\lambda_r)^{l_r}, \tag{7}$$

$$V_{\tilde{T}}(i\mathbf{w}) = (\mathbf{v}_1^T i\mathbf{w})^{l_1}(\mathbf{v}_2^T i\mathbf{w})^{l_2}\ldots(\mathbf{v}_r^T i\mathbf{w})^{l_r}\Phi_{\mathbf{x}_\infty}(i\mathbf{w}) \tag{8}$$

are eigenvalues and eigenfunctions of the operator $\tilde{T}$, where $\lambda_1, \ldots, \lambda_r$ are the eigenvalues corresponding to each block of the Jordan canonical form of matrix $\mathbf{A}$, $r \leq n$ is the number of blocks, $\mathbf{v}_1, \ldots, \mathbf{v}_r$ are the corresponding eigenvectors, and $l_1, \ldots, lr$ are arbitrary nonnegative integers.

Proof. Substituting expression (8) into (5), we get

$$\begin{aligned}(\tilde{T}V_{\tilde{T}})(i\mathbf{w}) &\triangleq V_{\tilde{T}}(\mathbf{A}^T i\mathbf{w})\Phi_{\mathbf{v}}(\mathbf{B}^T i\mathbf{w})\\ &= (\mathbf{v}_1^T\mathbf{A}^T i\mathbf{w})^{l_1}(\mathbf{v}_2^T\mathbf{A}^T i\mathbf{w})^{l_2}\ldots(\mathbf{v}_r^T\mathbf{A}^T i\mathbf{w})^{l_r}\Phi_{\mathbf{x}_\infty}(\mathbf{A}^T i\mathbf{w})\Phi_{\mathbf{v}}(\mathbf{B}^T i\mathbf{w}).\end{aligned}$$

Now, by using (6) and the fact that λ_i and $\mathbf{v}_i$ are eigenvalues and eigenvectors of $\mathbf{A}$, we have

$$(\tilde{T}V_{\tilde{T}})(i\mathbf{w}) = (\lambda_1)^{l_1}(\lambda_2)^{l_2}\ldots(\lambda_r)^{l_r}(\mathbf{v}_1^T i\mathbf{w})^{l_1}(\mathbf{v}_2^T i\mathbf{w})^{l_2}\ldots(\mathbf{v}_r^T i\mathbf{w})^{l_r}\Phi_{\mathbf{x}_\infty}(i\mathbf{w}),$$

which, from (7) and (8), can be written as

$$(\tilde{T}V_{\tilde{T}})(i\mathbf{w}) = \lambda_{\tilde{T}}V_{\tilde{T}}(i\mathbf{w}). \qquad \square$$

Remark 2.1. The eigenvalues of $\tilde{T}$ given by (7) are also eigenvalues of T (denoted λ_T). The eigenfunctions of T can be evaluated by noting that $(iw_j)^l\Phi_{\mathbf{x}_\infty}(i\mathbf{w})$ is the Fourier transform of $\partial^l f_{\mathbf{x}_\infty}(\mathbf{x})/\partial x_j^l$ (see, e.g., [14]). Hence, from (8) it can be seen that the eigenfunctions of T corresponding to the eigenvalues λ_T are linear combinations of high order partial derivatives of $f_{\mathbf{x}_\infty}(\mathbf{x})$ with respect to the components of $\mathbf{x} = (x_1, \ldots, x_n)^T$.

3. State-discretization of linear state space models. The linear operator T defined by (2) provides the time evolution of the continuous-state probability density functions. The state-discretization of the system (1) is performed by approximating this operator and the PDFs by discrete-state counterparts. We will restrict our study to a bounded region of the state space: an n-dimensional hypercube $\mathcal{X} = [a_1, b_1] \times \cdots \times [a_n, b_n]$ of volume $V_{\mathcal{X}}$. As explained before, we assume that all the PDFs of interest have compact support inside this region and belong to $\mathcal{L}^2([a_1, b_1] \times \cdots \times [a_n, b_n])$.

To discretize the region $\mathcal{X}$ of the state space we use the idea of *refinements* [6, 10], i.e., we divide $\mathcal{X}$ into N^n subsets of size $V_{\mathcal{X}}/N^n$ denoted $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{N^n}$, and define a Markov chain on the resulting discrete state space.

The probability vectors of the Markov chain are given by

$$\mathbf{p}_k \triangleq (P\{\mathbf{x}_k \in \mathbf{e}_1\}, \ldots, P\{\mathbf{x}_k \in \mathbf{e}_{N^n}\})^T, \tag{9a}$$

where

$$P\{\mathbf{x}_k \in \mathbf{e}_j\} = \int_{\mathbf{e}_j} \cdots \int f_{\mathbf{x}_k}(\mathbf{x}) dx_1 \ldots dx_n, \tag{9b}$$

and the conditional PDF of $\mathbf{x}_{k+1}$ is computed according to (2) as

$$\begin{aligned} f_{\mathbf{x}_{k+1}}(\mathbf{x}|\mathbf{x}_k \in \mathbf{e}_j) = \frac{1}{|\det(\mathbf{A})|} \int_{-\infty}^{+\infty} \cdots \\ \cdots \int_{-\infty}^{+\infty} f_{\mathbf{x}_k}\left(\mathbf{A}^{-1}(\mathbf{x} - \mathbf{B}\boldsymbol{\xi})|\mathbf{x}_k \in \mathbf{e}_j\right) f_{\mathbf{v}}(\boldsymbol{\xi}) d\xi_1 \ldots d\xi_m. \end{aligned} \tag{10}$$

In (10), $f_{\mathbf{x}_k}(\,\cdot\,|\mathbf{x}_k \in \mathbf{e}_j)$ is the conditional PDF of the random variable $\mathbf{x}_k$ evaluated at $(\cdot)$, given the condition that $\mathbf{x}_k$ belongs to the set $\mathbf{e}_j$ and is defined as (see, e.g., [8])

$$f_{\mathbf{x}_k}(\mathbf{y}|\mathbf{x}_k \in \mathbf{e}_j) = \begin{cases} \dfrac{f_{\mathbf{x}_k}(\mathbf{y})}{\int_{\mathbf{e}_j} \cdots \int f_{\mathbf{x}_k}(\mathbf{x}) dx_1 \ldots dx_n} & \text{if } \mathbf{y} \in \mathbf{e}_j, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

The transition matrix between time k and $k+1$, $\mathbf{Q}_k = \{q_{ij}^{(k)}\}$, $1 \le i, j \le N^n$, is then defined as

$$q_{ij}^{(k)} \triangleq P\{\mathbf{x}_{k+1} \in \mathbf{e}_i | \mathbf{x}_k \in \mathbf{e}_j\} = \int_{\mathbf{e}_i} \cdots \int f_{\mathbf{x}_{k+1}}(\mathbf{x}|\mathbf{x}_k \in \mathbf{e}_j) dx_1 \ldots dx_n. \tag{12}$$

Notice that we have made explicit the dependence of $\mathbf{Q}_k$ on k. If this is not the case, we say that the Markov chain has stationary or homogeneous transition probabilities, and we denote by $\mathbf{Q}$ the transition matrix. Notice also that the matrix $\mathbf{Q}_k$ is defined by (10) and (12), and these depend on the PDF $f_{\mathbf{x}_k}$ (see (11)). In order to simplify the computation of the matrix $\mathbf{Q}_k$, we approximate the PDFs $f_{\mathbf{x}_k}$ used in (11) by functions that assume piecewise-constant values in each subset $\mathbf{e}_i$ (see the discussion preceding (24)). This approach has the advantage that the transition matrices are stationary (see the independence from $f_{\mathbf{x}_k}$ in the elements q_{ij} of $\mathbf{Q}$ in (24)).

We denote by λ_T the eigenvalues of the operator T and by $\lambda_{\mathbf{Q}}(N)$ the eigenvalues of the transition matrix $\mathbf{Q}$ obtained with a partition of N^n cells. Notice that the spectrum of $\mathbf{Q}$, $\sigma(\mathbf{Q})$, has N^n points. We are interested in the accumulation (or limit) points of these sets as $N \to \infty$. The following theorem provides a connection between these accumulation points and the eigenvalues of the operator T.

THEOREM 3.1. *The set of nonzero limit points, as $N \to \infty$, of eigenvalues $\lambda_{\mathbf{Q}}(N)$ is a subset of the set of eigenvalues λ_T of the operator T.*

Proof. To maintain the notation as simple as possible, we will prove the theorem for the case of scalar systems ($n = 1$). The proof for the multidimensional case is similar, but with far more complicated notation.

We consider a scalar linear system described by

$$x_{k+1} = ax_k + bv_k, \tag{13}$$

where $0 \neq |a| < 1$ and $b \neq 0$. As explained before, it is assumed that all distributions have compact support. In particular we suppose that $\operatorname{supp} f_v \subseteq [-B, B]$ is fixed.

In the scalar case, the operator T defined by (2) can be written as

$$(Tf_{x_k})(x) = \int_{-\infty}^{+\infty} k(x, y) f_{x_k}(y) dy,$$

where the kernel is given by $k(x, y) = f_v(\frac{x-ay}{b})/|b|$ and satisfies $\int_{-\infty}^{+\infty} k(x, y)dx = 1$. Here we choose $A \geq |b|B/(1 - |a|)$ to ensure that the support of Tf_{x_k} lies in $[-A, A]$ whenever $\operatorname{supp} f_{x_k} \subseteq [-A, A]$ (and, in a practical situation, we also choose A large enough to ensure that the support of the initial state distribution satisfies $\operatorname{supp} f_{x_0} \subseteq [-A, A]$). Then T becomes the integral operator on $\mathcal{L}^2([-A, A])$:

$$(Tf)(x) = \int_{-A}^{A} k(x, y) f(y) dy \tag{14}$$

with $k \in \mathcal{L}^2([-A, A] \times [-A, A])$.

To perform the state-discretization of this system we divide the interval $\mathcal{X} = [-A, A]$ into $N^1 = N$ subsets of length $\frac{2A}{N}$, denoted $e_1, \ldots, e_N$, and we define the operator P_N on $\mathcal{L}^2([-A, A])$:

$$P_N \triangleq \sum_{j=1}^{N} I_{N,j}, \tag{15}$$

where the operators $I_{N,j}$ are defined as follows:

$$(I_{N,j} f)(x) = \begin{cases} \frac{N}{2A} \int_{e_j} f(z)dz & \text{if } x \in e_j, \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

for $f \in \mathcal{L}^2([-A, A])$.

Notice that the operator P_N is a projection onto the N-dimensional subspace of $\mathcal{L}^2([-A, A])$ consisting of functions with N piecewise constant values, and that $P_N f = E[f(U)|\mathcal{G}_N]$, where $\mathcal{G}_N$ is the sigma-algebra generated by the e_i, U is the identity mapping, and the probability space $(\Omega, \mathcal{F}, P)$ is such that $\Omega = \mathbb{R}$ and P is the uniform probability measure on $[-A, A]$.

Direct computation shows that the image of the operator $P_N T P_N$ is constant over each cell e_i, taking the value

$$(P_N T P_N f)(x) = \frac{N}{2A} \sum_{j=1}^{N} \left[\left(\int_{e_i} \int_{e_j} k(x,y) dy dx \right) \left(\frac{N}{2A} \int_{e_j} f(z) dz \right) \right] \tag{17}$$

for $x \in e_i$.

We need to prove now that the operator $P_N T P_N$ converges, as $N \to \infty$, to T in the natural norm of operators on $\mathcal{L}^2([-A, A])$. For this purpose we use $k_{i,j}$ to denote the scaled integral of $k(x,y)$ over the cell $e_i \times e_j$:

$$k_{i,j} = \left(\frac{N}{2A} \right)^2 \int_{e_i} \int_{e_j} k(x,y) dy dx, \tag{18}$$

and we define a *piecewise-constant* function on $[-A, A] \times [-A, A]$ by

$$n_N^k(x,y) = \sum_{i=1}^{N} \sum_{j=1}^{N} k_{i,j} \mathcal{X}_{e_i}(x) \mathcal{X}_{e_j}(y), \tag{19}$$

where $\mathcal{X}_{e_i}(\cdot)$ is the indicator function $\mathcal{X}_{e_i}(x) = 1$ if $x \in e_i$ and 0 otherwise.

From (17), (18), and (19) it can be seen that the operator $P_N T P_N$ is given by an integral operator with kernel $n_N^k(\cdot, \cdot)$, i.e.,

$$(P_N T P_N f)(x) = \int_{-A}^{A} n_N^k(x,y) f(y) dy. \tag{20}$$

Then, by using Hölder's inequality, it can be shown from (14) and (20) that

$$\| T - P_N T P_N \| \leq \left(\int_{-A}^{A} \int_{-A}^{A} |k(x,y) - n_N^k(x,y)|^2 dy dx \right)^{\frac{1}{2}}. \tag{21}$$

If the function k is continuous, it is uniformly continuous in the compact interval $[-A, A] \times [-A, A]$. It follows that the functions n_N^k converge uniformly to k. Thus, in the case of continuous k, it follows from the inequality $\| \cdot \|_2 \leq 2A \| \cdot \|_\infty$ that the right-hand side of (21) converges to zero as the number of divisions N tends to infinity.

It is a standard result from measure theory that continuous functions on $[-A, A] \times [-A, A]$ constitute a dense subspace of $\mathcal{L}^2([-A, A] \times [-A, A])$ (see, e.g., [11]). Hence, for an arbitrary $k \in \mathcal{L}^2([-A, A] \times [-A, A])$, we can approximate k in the 2-norm by a continuous function g (i.e., $\|k - g\|_2 \to 0$). For any N, the corresponding discrete approximations n_N^k and n_N^g, given by (18) and (19), satisfy

$$\begin{aligned} \| n_N^k - n_N^g \|_2^2 &= \int_{-A}^{A} \int_{-A}^{A} \sum_{i,j} \mathcal{X}_{e_i}(x) \mathcal{X}_{e_j}(y) |k_{i,j} - g_{i,j}|^2 dy dx \\ &= \sum_{i,j} \left(\frac{2A}{N} \right)^2 \left| \left(\frac{N}{2A} \right)^2 \int_{-A}^{A} \int_{-A}^{A} (k(x,y) - g(x,y)) \mathcal{X}_{e_i}(x) \mathcal{X}_{e_j}(y) dy dx \right|^2 \\ &\leq \sum_{i,j} \left(\frac{N}{2A} \right)^2 \| (k - g) \mathcal{X}_{e_i} \mathcal{X}_{e_j} \|_1^2. \end{aligned}$$

Now, writing the identity $(k-g)\mathcal{X}_{e_i}\mathcal{X}_{e_j} \equiv [(k-g)\mathcal{X}_{e_i}\mathcal{X}_{e_j}][\mathcal{X}_{e_i}\mathcal{X}_{e_j}]$ and applying Hölder's inequality gives

$$\begin{aligned}\|n_N^k - n_N^g\|_2^2 &\le \sum_{i,j}\left(\frac{N}{2A}\right)^2 \|(k-g)\mathcal{X}_{e_i}\mathcal{X}_{e_j}\|_2^2 \|\mathcal{X}_{e_i}\mathcal{X}_{e_j}\|_2^2 \\ &= \sum_{i,j}\|(k-g)\mathcal{X}_{e_i}\mathcal{X}_{e_j}\|_2^2 = \|k-g\|_2^2 \end{aligned} \tag{22}$$

(or we could observe that $k \mapsto n_N^k$ is the orthogonal projection on the span of $\{\mathcal{X}_{e_i}(x)\mathcal{X}_{e_j}(y)\}$, and hence is norm-decreasing).

Since we have already seen that $\|n_N^g - g\|_2 \to 0$ for the continuous function g, and since from (22) we have

$$\begin{aligned}\|n_N^k - k\|_2 &\le \|n_N^k - n_N^g\|_2 + \|n_N^g - g\|_2 + \|g-k\|_2 \\ &\le 2\|k-g\|_2 + \|n_N^g - g\|_2, \end{aligned} \tag{23}$$

we deduce that $\|n_N^k - k\|_2 \to 0$. Therefore, we conclude from (21) that the operator $P_N T P_N$ converges to T in norm.

The transition matrix $\mathbf{Q}$ of the resulting Markov chain is defined by (10) and (12). As explained before, we simplify the computation by taking a piecewise-constant approximation of f_{x_k} (i.e., of the form $\tilde{f}_{x_k} = P_N f_{x_k}$). Then, the elements of the matrix $\mathbf{Q}$ are given by $q_{ij} = \frac{N}{2A}\int_{e_i}\int_{e_j} k(x,y)dydx$. Therefore, the ith element of the product of the matrix $\mathbf{Q}$ and a vector $\mathbf{p} = \{p_j\}$; $1 \le j \le N$, is given by

$$(\mathbf{Qp})_i = \frac{N}{2A}\sum_{j=1}^{N}\left[\left(\int_{e_i}\int_{e_j} k(x,y)dydx\right)p_j\right]. \tag{24}$$

From (17) and (24) it can be seen that the matrix $\mathbf{Q}$ (acting on vectors) and the operator $P_N T P_N$ (acting on piecewise-constant functions) are conjugate. To see this, consider a function $f = \sum_i v_i \mathcal{X}_{e_i}$ and a vector $\mathbf{v} = \{v_i\}$; $1 \le i \le N$. Then, by noticing that $\frac{N}{2A}\int_{e_j} f(z)dz = v_j$, it can be seen from (17) and (24) that $P_N T P_N f = \sum_i (\mathbf{Qv})_i \mathcal{X}_{e_i}$. It follows immediately that the eigenvalues of the operator $P_N T P_N$ and those of the matrix $\mathbf{Q}$ are the same, and that the eigenfunctions of $P_N T P_N$ are given by the *piecewise-constant held versions* $f_{\mathbf{v_Q}} = \sum_i v_{\mathbf{Q}i}\mathcal{X}_{e_i}$ of the eigenvectors $\mathbf{v_Q} = \{v_{\mathbf{Q}i}\}_{i=1}^N$ of $\mathbf{Q}$.

It remains to prove that the set of nonzero limit points of the eigenvalues of the operator $P_N T P_N$ is a subset of the set of eigenvalues of the operator T. For this purpose, let's suppose that $\lambda \neq 0$ is a limit point of a sequence of eigenvalues λ_N of $P_N T P_N$; then we have to show that λ is an eigenvalue of T. As we proved before, the operator $P_N T P_N$ converges to T in the natural norm of operators on $\mathcal{L}^2([-A,A])$. Hence, $(\lambda_N I - P_N T P_N)$ converges to $(\lambda I - T)$ in norm. Since the set of invertible operators is open [12, Theorem 10.12], if $(\lambda_N I - P_N T P_N)$ is not invertible (i.e., λ_N belongs to the spectrum of $P_N T P_N$) for all N then $(\lambda I - T)$ is not invertible (i.e., λ belongs to the spectrum of T). Since the operator T is compact [3], and for a compact linear operator every spectral value, with the possible exclusion of zero, is an eigenvalue [5, Theorem 8.4–4], the theorem is proved. □

4. Example. In order to illustrate the convergence of eigenvalues and eigenvectors we consider the following scalar state space model:

$$x_{k+1} = 0.5x_k + v_k, \tag{25}$$

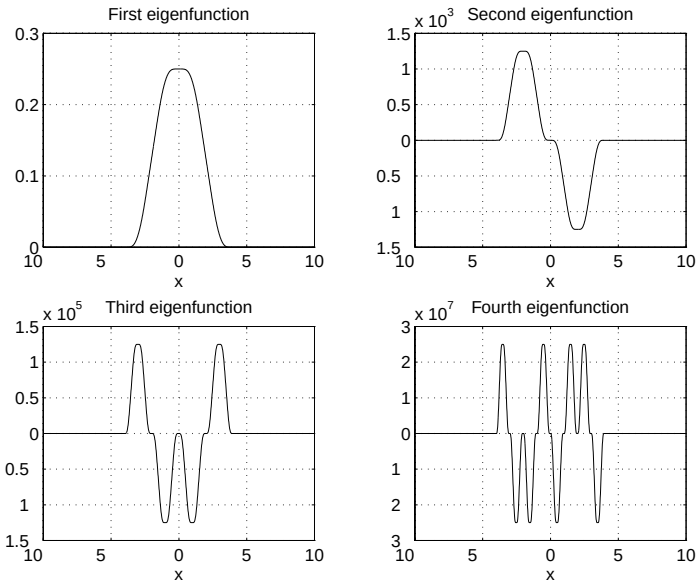


FIG. 1. *Eigenfunctions of the operator T corresponding to the four eigenvalues of larger magnitude* $(1, 0.5, 0.25,$ *and* $0.125,$ *respectively).*

where $v_k \sim \text{Uniform}[-2, 2]$ is an iid sequence, and $x_0 \sim \text{Uniform}[-10, 10]$.

Notice that, by (7), (8), and Remark 2.1, the eigenvalues of the operator T defined by (2) for the scalar case are the nonnegative integer powers of $\lambda = 0.5$, and the eigenfunctions of T are the stationary PDF f_{x_∞} and its derivatives. In Figure 1 the stationary PDF f_{x_∞} and its three first derivatives are shown. The eigenfunctions are ordered according to the magnitude of the corresponding eigenvalue (i.e., the first eigenfunction corresponds to the larger eigenvalue, etc.). It is easy to see, for the scalar system (13) and $\text{supp}\, f_v \equiv [-B, B]$, that f_{x_∞} has support equal to $[\ -|b|B/(1-|a|)\ ,\ |b|B/(1-|a|)\ ]$. For this example, $\text{supp}\, f_{x_\infty} = [-4, 4]$.

Since all distributions in this example have compact support equal or contained in the interval $[-10, 10]$ of the real line, the region in the state space to be discretized is chosen as the interval $\mathcal{X} = [-10, 10]$. This interval is divided into N subsets of size $20/N$ and a Markov chain is defined on the discrete state space $\{e_1, e_2, \ldots, e_N\}$. The transition probability matrix $\mathbf{Q}$ is computed using (10) and (12).

In Figure 2 the eight larger eigenvalues of the matrices $\mathbf{Q}$ corresponding to different values of N (denoted $\lambda_{\mathbf{Q}}(N)$) are shown on a logarithmic scale. Notice that they converge to the eigenvalues corresponding to the continuous state operator T:

$$\frac{\log(\lambda_{\mathbf{Q}}(N))}{\log(0.5)} \to \frac{\log(0.5^l)}{\log(0.5)} \ \equiv\ l\ ;\quad \text{for } l = 0, 1, \ldots, 7 \tag{26}$$

as $N \to \infty$. Notice also that the number of points N needed for convergence of the successive eigenvalues is exponentially increasing. The reason for this can be found in Figure 1. The successive eigenfunctions have peaks, the number of which increases proportionally to the powers of two. Therefore, in order to have a faithful representation of them with the eigenvectors of the $\mathbf{Q}$ matrices, the number of turning points of the eigenvectors increases by powers of two. In general, for a particular N,

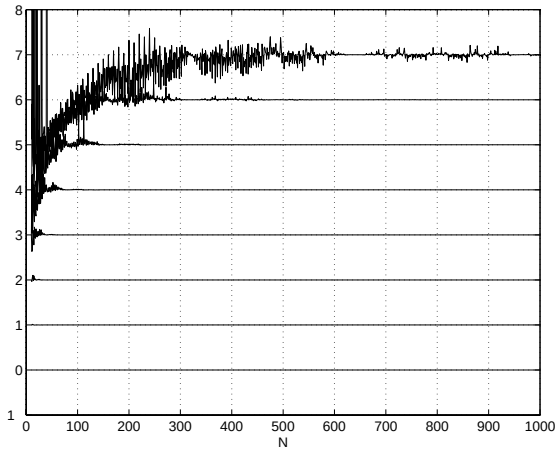


FIG. 2. *Eight larger eigenvalues of matrices* $\mathbf{Q}$ *on logarithmic scale (*$\log(\lambda_{\mathbf{Q}}(N))/\log(0.5)$*) as a function of the number of subintervals* N.

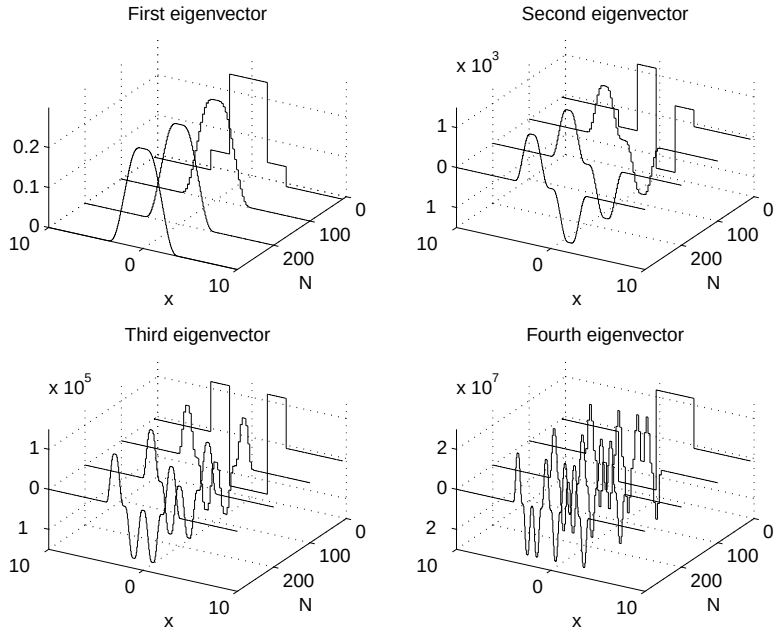


FIG. 3. *Eigenfunctions of* $P_N T P_N$ *corresponding to the four eigenvalues of larger magnitude for* $N = 10$, 100, 200, *and* 300 *subintervals.*

unless an eigenfunction of $P_N T P_N$ approximates one of T, we cannot expect the corresponding eigenvalues to be close approximations.

In Figure 3 the eigenfunctions corresponding to the four eigenvalues of larger modulus of the operators $P_N T P_N$ for $N = 10$, 100, 200, and 300 are shown to illustrate their convergence to the eigenfunctions of T depicted in Figure 1. The eigenfunctions are ordered, as in Figure 1, according to the magnitude of the corresponding eigenvalue.

5. Conclusions. Initial results in the characterization of the class of discrete-state systems amenable to a linear continuous-state approximation have been presented. In particular, we have explored in this paper the properties of discrete-state systems that arise in the inverse process, namely, in the state-discretization of linear continuous-state systems. The method to perform the state-discretization presented consists of a finite rank discrete approximation of the transition integral operator corresponding to the linear continuous-state system. As a result of this state-discretization procedure a Markov chain is obtained. In this work we have shown that the eigenvalues of the resulting Markov chain transition matrices converge to those of the transition operator of the underlying continuous-state linear model. This convergence has been illustrated by a scalar example.

Acknowledgments. The authors wish to thank Professor Alex Novikov and Dr. Nino Kordzakhia from the Department of Statistics at the University of Newcastle for valuable discussions.

REFERENCES

[1] K. Åström, *Introduction to Stochastic Control Theory*, Academic Press, New York, 1970.
[2] D. R. Cox and H. D. Miller, *The Theory of Stochastic Processes*, Chapman and Hall, London, 1965.
[3] I. Gohberg and S. Goldberg, *Basic Operator Theory*, Birkhäuser, Boston, 1980.
[4] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer-Verlag, Berlin, New York, Heidelberg, 1992.
[5] E. Kreyszig, *Introductory Functional Analysis with Applications*, John Wiley & Sons, New York, London, Sydney, 1978.
[6] H. J. Kushner and P. G. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, Springer-Verlag, Berlin, New York, Heidelberg, 1992.
[7] D. F. Nicholls and B. G. Quinn, *Random Coefficient Autoregressive Models: An Introduction*, Springer-Verlag, Berlin, New York, Heidelberg, 1982.
[8] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, McGraw–Hill, New York, Toronto, London, 1965.
[9] M. B. Priestley, *Spectral Analysis and Time Series*, Academic Press, New York, 1981.
[10] J. S. Rosenthal, *Markov chain convergence: From finite to infinite*, Stochastic Process. Appl., 62 (1996), pp. 55–72.
[11] W. Rudin, *Real and Complex Analysis*, McGraw–Hill, London, 1970.
[12] W. Rudin, *Functional Analysis*, McGraw–Hill, New York, Toronto, London, 1973.
[13] P. Terwiesch and M. Agarwal, *A discretized nonlinear state estimator for batch processes*, Comput. Chem. Engrg., 19 (1995), pp. 155–169.
[14] K. B. Wolf, *Integral Transforms in Science and Engineering*, Plenum Press, New York, 1979.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 4, pp. 1112–1135
© 2000 Society for Industrial and Applied Mathematics

IMPLICITLY RESTARTED GMRES AND ARNOLDI METHODS FOR NONSYMMETRIC SYSTEMS OF EQUATIONS*

RONALD B. MORGAN†

Abstract. The generalized minimum residual method (GMRES) is well known for solving large nonsymmetric systems of linear equations. It generally uses restarting, which slows the convergence. However, some information can be retained at the time of the restart and used in the next cycle. We present algorithms that use implicit restarting in order to retain this information. Approximate eigenvectors determined from the previous subspace are included in the new subspace. This deflates the smallest eigenvalues and thus improves the convergence. The subspace that contains the approximate eigenvectors is itself a Krylov subspace, but not with the usual starting vector. The implicitly restarted FOM algorithm includes standard Ritz vectors in the subspace. The eigenvalue portion of its calculations is equivalent to Sorensen's IRA algorithm. The implicitly restarted GMRES algorithm uses harmonic Ritz vectors. This algorithm also gives a new approach to computing interior eigenvalues.

Key words. GMRES, implicit restarting, iterative methods, nonsymmetric systems, harmonic Ritz values

AMS subject classifications. 65F10, 15A06

PII. S0895479897321362

1. Introduction.

Large systems of linear equations arise in many different scientific applications. Notably, partial differential equations discretized with finite difference or finite element methods yield systems of equations. Large systems can be solved with either sparse factorization techniques or with iterative methods. These two approaches can be combined into a method that uses approximate factorization preconditioning for an iterative method. Here we consider only iterative methods.

For symmetric systems of linear equations, the most popular iterative method is the conjugate gradient method [18, 13, 14, 17]. The conjugate gradient method has several nice properties: it has a Krylov subspace for good convergence, it can be preconditioned, it has a short recurrence for efficiency, and it extracts an optimal solution from the subspace when the matrix is positive definite. Even for indefinite problems, it can be implemented so that the minimum residual solution is found [34].

In the nonsymmetric case, the choice for the iterative method is not so clear. Many methods have been proposed [2, 41, 51], but two of the best known are the generalized minimum residual method (GMRES) [43] and the quasi-minimal residual method (QMR) [9, 12]. These methods are generalizations of the conjugate gradient method. However, neither method retains all of the good properties of the conjugate gradient method. QMR has the Lanczos short recurrence [21, 22, 46, 54] but it has stability problems, is not guaranteed to extract the best solution from the subspace, and uses two matrix-vector products per iteration. Look-ahead [11, 12, 36] can be used to improve stability, the quasi-minimum property is designed to give smoother convergence, and transpose-free versions attempt to take advantage of the extra matrix-vector product. GMRES is stable and has the reassuring minimum

*Received by the editors May 12, 1997; accepted for publication by L. Reichel (in revised form) September 11, 1998; published electronically March 21, 2000. This work was supported by the National Science Foundation under grant NSF-CCR-9522612.

http://www.siam.org/journals/simax/21-4/32136.html

†Department of Mathematics, Baylor University, P.O. Box 97328, Waco, TX 76798-7398 (morganr @baylor.edu).

residual property. However, because the basis vectors for the Krylov subspace are explicitly orthogonalized, GMRES becomes increasingly expensive and requires more storage as the iteration proceeds. It can be restarted, but with the dimension of the subspace limited, the convergence slows down. GMRES often exhibits steady convergence, while QMR convergence curves are characterized by plateaus and sudden drops. In spite of somewhat mysterious behavior and significant complications for stability control, QMR is appealing because its use of the Lanczos recurrence allows for larger subspaces. Restarting is not needed. Large subspaces can be much better for difficult problems.

We propose a method that maintains the good qualities of GMRES but is more competitive with QMR for difficult problems. The new method is mathematically equivalent to a method in [28], but it has a more efficient implementation. The new implementation combines ideas from Sorensen's implicitly restarted Arnoldi algorithm [47] with harmonic Ritz values [27, 33, 45, 30].

We next review GMRES and several related methods that are needed in developing the new version of GMRES. Section 3 discusses how approximate eigenvectors can improve convergence. Section 4 gives a method that uses approximate eigenvectors and implicit restarting with the Arnoldi method for linear equations or full orthogonalization method (FOM) [38]. In section 5, approximate eigenvectors are added to GMRES using implicit restarting. Section 6 has numerical examples, and the related eigenvalue problem is mentioned in section 7.

2. Background on related methods. Here we discuss two methods for solving linear equations and also two eigenvalue methods. The methods are related in that they all use Krylov subspaces and the Arnoldi iteration [1, 37, 39]. With starting vector v_1, the Krylov subspace of dimension $m+1$ is $Span\{v_1, Av_1, A^2v_1, \ldots, A^mv_1\}$. The Arnoldi iteration generates an orthonormal basis for this subspace.

The Arnoldi iteration.

(1) *Start:* Choose a starting vector and normalize for v_1.

(2) *Iterate:* For $j = 1, 2, \ldots, m$ compute:

$h_{ij} = (v_i, Av_j), i = 1, 2, \ldots, j,$

$w_j = Av_j - \sum_{i=1}^{j} h_{ij}v_i,$

$h_{j+1,j} = ||w_j||$, and

$v_{j+1} = w_j/h_{j+1,j}$.

Let H_m be the $m \times m$ upper-Hessenberg matrix whose nonzero entries are defined in the Arnoldi iteration. Let $\bar{H}_m$ be the corresponding $m+1$ by m matrix with last row having only the one nonzero element $h_{m+1,m}$. Let V_m be the $n \times m$ matrix whose columns are v_1 through v_m. These relations hold [41, p. 148]:

$$AV_m = V_m H_m + h_{m+1,m} v_{m+1} e_m^T \tag{2.1}$$

$$= V_{m+1}\bar{H}_m, \tag{2.2}$$

$$V_m^T A V_m = H_m. \tag{2.3}$$

All four of the methods that follow have a residual vector usually denoted by r. We add some different accents to distinguish them, except for the GMRES residual which is denoted by plain r.

2.1. Arnoldi for linear equations (FOM). The Arnoldi method for linear equations or FOM uses the Arnoldi iteration to solve nonsymmetric systems of linear equations [16, 38, 43, 41]. We call the Arnoldi residual vector $\acute{r}_0$. Suppose the problem

$Ax = b$ has been recast into form $A(x - x_0) = \acute{r}_0$. The starting vector for the Arnoldi iteration is $\acute{r}_0$. The small problem

$$H_m d = \beta e_1 \tag{2.4}$$

is solved, where $\beta = ||\acute{r}_0||$. Then the approximate solution of the original problem is $x_m = x_0 + V_m d$. It can be shown that the new residual vector satisfies

$$\acute{r} \equiv b - Ax_m \tag{2.5}$$

$$= -d_m h_{m+1,m} v_{m+1}, \tag{2.6}$$

where d_m is the last entry of d. So $\acute{r}$ is a multiple of v_{m+1}, and it becomes the new $\acute{r}_0$ at the restart. Therefore v_1 for the new Krylov subspace is the old v_{m+1}.

Restarted FOM.

(1) *Start:* Choose x_0 and compute $\acute{r}_0 = b - Ax_0$ and $v_1 = \acute{r}_0/||\acute{r}_0||$. Let $\beta = ||\acute{r}_0||$.
(2) *Iterate:* Apply m steps of the Arnoldi iteration.
(3) *Form the approximate solution:* $x_m = x_0 + V_m d$, where d is the solution of $H_m d = \beta e_1$.
(4) *Restart:* Compute the residual norm $||\acute{r}||$; if satisfied then stop, else let $x_0 = x_m$, $\beta = ||\acute{r}||$, $v_1 = \acute{r}/\beta$, $\acute{r}_0 = \acute{r}$, and go to 2.

2.2. GMRES. Most steps of the GMRES method [43] are the same as FOM. Let the recast problem be $A(x - x_0) = r_0$ and the starting vector for the Krylov subspace be r_0. Instead of (2.4), the small least squares problem

$$min||\beta e_1 - \bar{H}_m d||$$

is solved. GMRES finds the approximate solution that minimizes the residual norm.

2.3. Arnoldi for eigenvalue problems. To find eigenvalues of A using the Arnoldi method [1, 37, 39], we solve the small eigenvalue problem

$$H_m g_i = \theta_i g_i.$$

The θ_i's are approximate eigenvalues of A and are called Ritz values. The approximate eigenvectors or Ritz vectors are $y_i = V_m g_i$. The residual vector for each Ritz pair (θ_i, y_i) is defined to be $\dot{r}_i \equiv Ay_i - \theta_i y_i$. It is easy to derive [39, p. 175] that

$$\dot{r}_i = h_{m+1,m} e_m^T g_i v_{m+1}.$$

Defining

$$\beta_{mi} = h_{m+1,m} e_m^T g_i$$

gives

$$\dot{r}_i = Ay_i - \theta_i y_i = \beta_{mi} v_{m+1} \tag{2.7}$$

and

$$||\dot{r}_i|| = |\beta_{mi}|.$$

Note that every residual vector is a scalar multiple of the same vector, namely v_{m+1}. Furthermore, they are all multiples of the residual vector developed by FOM while solving linear equations, assuming the same starting vectors are used. They are not multiples of the GMRES residual vector.

Arnoldi for eigenvalues.

(1) *Start:* Choose an initial vector for the Krylov subspace and normalize for v_1.
(2) *Iterate:* Apply m steps of the Arnoldi iteration.
(3) *Find approximate eigenpairs:* Find eigenpairs (θ_i, g_i) of H_m as desired. The θ_i's are Ritz values. Compute Ritz vectors, $y_i = V_m g_i$, as desired.
(4) *Restart:* Residual norms can be checked for convergence. If needed, choose a new starting vector for v_1 and go to 2.

The restarting is difficult because one new starting vector must be chosen, while several Ritz vectors may need to be passed on to the new subspace. If not carefully chosen, a combination of Ritz vectors can be a very poor choice for the new starting vector [29]. Sorensen's implicitly restarted Arnoldi (IRA) algorithm [47, 23, 24] elegantly solves this problem. A QR iteration with H_m is used. The right combination of Ritz vectors is formed for the new starting vector [29].

We give a few of the details of the IRA method which will be needed later. Suppose k is the number of Ritz vectors to be saved for the next subspace. Let $p = m - k$. Then p shifts $\tau_1, \ldots, \tau_p$ are selected. Let $H^{(1)} \equiv H_m$. The main steps of the QR method at iteration i are first an orthogonal or unitary factorization of a shifted $H^{(i)}$,

$$H^{(i)} - \tau_i I = Q^{(i)} R^{(i)},$$

and then formation of $H^{(i+1)}$ using

$$R^{(i)} Q^{(i)} = H^{(i+1)} - \tau_i I.$$

In practice, a double shift is used to avoid unitary factorization in the case of a complex τ_i. Define $Q \equiv Q^{(1)} Q^{(2)} \cdots Q^{(p)}$ and $R \equiv R^{(p)} R^{(p-1)} \cdots R^{(1)}$. It is a standard result for the QR iteration [23, p. 24] that

$$QR = \prod_{l=1}^{p} (H_m - \tau_l I). \tag{2.8}$$

At the end of the QR phase, the new Krylov basis V^+ is

$$V^+ = V_m Q, \tag{2.9}$$

and the first k columns are used to start the next Arnoldi iteration. The $(k+1)$st column of the new V can also be calculated. For the case of the shifts being Ritz values, it is the same as the old v_{m+1}.

2.4. Interior or harmonic Arnoldi. The Arnoldi algorithm applies the Rayleigh–Ritz procedure [35, 39] to a Krylov subspace. It has trouble finding interior eigenvalues mainly because the Krylov subspace usually does not contain good approximations to the corresponding eigenvectors, but also due to properties of the Rayleigh–Ritz procedure. For symmetric matrices, Rayleigh–Ritz is optimal at extracting exterior eigenvalues [35, p. 215], but not for interior eigenvalues [27]. With nonsymmetric problems, there is still a tendency for Rayleigh–Ritz to be more reliable at extracting exterior eigenpairs ([30] has some discussion of this). An interior Arnoldi algorithm using a modified Rayleigh–Ritz procedure is given in [30] as a generalization of methods in [27]; see also [33, 45, 55]. Here we look only at the case where eigenvalues near zero are desired. For this case, this method has been mentioned in other recent papers in connection with GMRES [8, 10, 25, 26, 31, 44].

The small eigenvalue problem that needs to be solved is the generalized eigenvalue problem

$$V_m^T A^T V_m \tilde{g}_i = \frac{1}{\tilde{\theta}_i} V_m^T A^T A V_m \tilde{g}_i.$$

Using (2.2) and (2.3), this becomes

$$H_m^T \tilde{g}_i = \frac{1}{\tilde{\theta}_i} \bar{H}_m^T \bar{H}_m \tilde{g}_i. \tag{2.10}$$

This can be simplified with an orthogonal factorization $\bar{H}_m = QR$ to

$$H_m^T \tilde{g}_i = \frac{1}{\tilde{\theta}_i} R^T R \tilde{g}_i.$$

This formula has been stable in experiments, but a possibly better form from [33] is

$$(H_m + h_{m+1,m}^2 f e_m^*)\tilde{g}_i = \tilde{\theta}_i \tilde{g}_i,$$

where $f = H^{-T} e_m$. The $\tilde{\theta}_i$'s are called harmonic Ritz values in [33]. The harmonic Ritz vectors are

$$\tilde{y}_i = V_m \tilde{g}_i.$$

There are eigenvalue approximations available that are more accurate than the harmonic Ritz values. They are the Rayleigh quotients of the $\tilde{y}_i$'s, called ρ values in [27] and [30]. However, in this paper, the harmonic Ritz values are more important. Define the harmonic residual vectors to be

$$\tilde{r}_i \equiv A\tilde{y}_i - \tilde{\theta}_i \tilde{y}_i. \tag{2.11}$$

See [30] for formulas for finding the residual norms in the interior Arnoldi method, but note the residual using the Rayleigh quotient ρ_i is different from the harmonic residual defined in (2.11).

The eigenvalue methods described in this and the previous subsection will be used to improve the convergence of the linear equations methods in the first two subsections. The basic idea is that we want to add approximate eigenvectors to the subspaces for the linear equations methods. This idea is discussed in the next section.

3. Using eigenvectors to improve convergence of GMRES. Restarting generally slows the convergence of GMRES. Useful information is discarded at the restart. In this section we consider saving some vectors from the previous subspace and adding them to the new subspace to deflate eigenvalues. This discussion is a quick recap of [28]; see also [4, 42]. For other approaches to deflating eigenvalues for GMRES, see [3, 7, 20]. Kharchenko and Yeremin [20] modify the matrix A with approximate eigenvectors and thus deflate some eigenvalues. Erhel, Burrage, and Pohl [7] use approximate eigenvectors to build a preconditioner that shifts eigenvalues. Baglama, Calvetti, Golub, and Reichel [3] improve upon this last approach by using implicit restarting. The eigenvectors can be kept in a subspace until they are accurate enough.

We mention some other related work. The natural deflation of unrestarted conjugate gradient type methods has been studied; see, for example, [32, 52]. The GMRESR [53] and FGMRES [40] methods use a double loop to keep a part of the Krylov subspace. Adding an outside loop could perhaps benefit the methods in this paper, but we do not investigate this here. An improvement of GMRESR is given in [50], and another approach to what should be kept after a restart is in [49].

3.1. An augmented subspace. At the time of a restart, let $u_1, u_2, \ldots, u_k$ be vectors chosen from the subspace before it is discarded. They will be used in the next subspace that is developed. Specifically, the augmented subspace

$$Span\{r_0, Ar_0, A^2r_0, A^3r_0, \ldots, A^{m-k-1}r_0, u_1, u_2, \ldots, u_k\} \tag{3.1}$$

is generated in the next cycle of GMRES. The approximate solution for the linear equations problem is extracted from this subspace. The subspace has a Krylov portion and an added portion. Many choices are possible for the added vectors. We chose these vectors to be approximate eigenvectors corresponding to the approximate eigenvalues nearest zero. However for some problems, a few large eigenvalues should also be targeted.

3.2. Reasons for using eigenvectors. For symmetric positive definite matrices, slow convergence of the conjugate gradient method is caused by the presence of some relatively small eigenvalues. For nonsymmetric matrices, there can be other factors such as negative and complex eigenvalues. And even with a good eigenvalue distribution, convergence can be poor [15]. Nevertheless, in many difficult linear equations problems, there are small eigenvalues that detract from the convergence. If eigenvectors are added to the subspace, then the corresponding eigenvalues are essentially deflated from the spectrum of the matrix. And if approximate eigenvectors are added, they do not need to be extremely accurate before they are helpful. This is shown in [28, Theorem 2] and has also been observed experimentally.

As an example of how beneficial this deflation of eigenvalues can be, consider a symmetric matrix with eigenvalue distribution $1, 2, 3, \ldots, n-1, n$. If the three smallest eigenvalues are removed from the spectrum, then $\sqrt{\kappa}$, the square root of the condition number, changes from $\sqrt{n}$ to $\frac{\sqrt{n}}{2}$. Using this to estimate convergence gives that convergence should be twice as fast with the eigenvalues removed. If GMRES is unrestarted, removing some eigenvalues may not affect convergence this much, since eigenvalues naturally deflate anyway. But when restarting is used, the convergence will improve about as predicted.

3.3. A few details. The approximate eigenvectors can be chosen to be Ritz vectors from the Arnoldi method or harmonic Ritz vectors from the interior Arnoldi method. In [28], harmonic Ritz vectors are used. For indefinite matrices, they give better approximations for the desired eigenvectors that correspond to eigenvalues near zero. So the subspace is

$$Span\{r_0, Ar_0, A^2r_0, A^3r_0, \ldots, A^{m-k-1}r_0, \tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_k\}, \tag{3.2}$$

where $\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_k$ are the harmonic Ritz vectors corresponding to the harmonic Ritz values nearest to the origin.

The implementation for this augmented subspace is more complicated than for standard GMRES [28]. There is added storage, but often a reduction of expense per iteration, because a matrix-vector product can be avoided when an approximate eigenvector is used instead of another Arnoldi vector; see Table 1 in section 5. For the case of inexpensive matrix-vector product, DAXPYs and DDOTs can be reduced to $m^2 + 2km$, instead of saving on the matrix-vector products. This augmented GMRES method from [28] will subsequently be referred to as GMRES-E (GMRES with eigenvectors).

4. Implicitly restarted FOM. We now reconsider the subspace (3.1) discussed in the previous section. With the GMRES-E method, the approximate eigenvectors do not naturally fit into the subspace, but are forced on at the end. It is not even necessary that the extra vectors be approximate eigenvectors. In this and the next section, we look at how the approximate eigenvectors can be selected to correspond to the linear equations method. Then they fit more naturally into the subspace. In fact, the whole subspace is a Krylov subspace, though not with the same starting vector $\acute{r}_0$ or r_0. With FOM, we need the standard Arnoldi Ritz vectors.

4.1. Adding approximate eigenvectors to the subspace. The desired subspace is

$$Span\{\acute{r}_0, A\acute{r}_0, A^2\acute{r}_0, A^3\acute{r}_0, \ldots, A^{m-k-1}\acute{r}_0, y_1, y_2, \ldots, y_k\}. \tag{4.1}$$

Before considering the general case, we look at how just one Ritz vector can be added to the subspace. This turns out to be fairly simple. We want a subspace consisting of the usual Krylov subspace for FOM, but of one less degree and with one Ritz vector as an additional basis vector. So the desired subspace is

$$Span\{\acute{r}_0, A\acute{r}_0, A^2\acute{r}_0, A^3\acute{r}_0, \ldots, A^{m-2}\acute{r}_0, y_1\}. \tag{4.2}$$

We will show that in spite of appearing otherwise, this subspace is itself a Krylov subspace. But it is a Krylov subspace generated with starting vector y_1 instead of $\acute{r}_0$. The key is that the residual vectors for the eigenvalue problem and the linear equations problem are both multiples of the same vector, v_{m+1}.

PROPOSITION 4.1. *Subspace* (4.2) *is the same subspace as*

$$Span\{y_1, Ay_1, A^2y_1, A^3y_1, \ldots, A^{m-1}y_1\}. \tag{4.3}$$

Proof. From (2.7),

$$Ay_1 = \theta_1 y_1 + \beta_{m1} v_{m+1}.$$

Thus Ay_1 is a combination of y_1 and v_{m+1} and

$$Span\{y_1, Ay_1\} = Span\{y_1, v_{m+1}\}.$$

Next, using (2.7) twice,

$$\begin{aligned} A^2y_1 &= \theta_1 Ay_1 + \beta_{m1} Av_{m+1} \\ &= \theta_1^2 y_1 + \theta_1\beta_{m1} v_{m+1} + \beta_{m1} Av_{m+1} \end{aligned}$$

and

$$Span\{y_1, Ay_1, A^2y_1\} = Span\{y_1, v_{m+1}, Av_{m+1}\}.$$

Continuing this, subspace (4.3) is the same as

$$Span\{y_1, v_{m+1}, Av_{m+1}, A^2v_{m+1}, A^3v_{m+1}, \ldots, A^{m-2}v_{m+1}\}. \tag{4.4}$$

Subspace (4.4) is the same as (4.2) because the residual vector $\acute{r}_0$ is a multiple of v_{m+1}. This establishes the claim. ∎

For adding two Ritz vectors to the subspace, the right combination of the two vectors is needed for the new starting vector. Let $s = \beta_{m2}y_1 - \beta_{m1}y_2$. Then it can be shown that

$$Span\{s, As, A^2s, \ldots, A^{m-1}s\} = Span\{\acute{r}_0, A\acute{r}_0, A^2\acute{r}_0, \ldots, A^{m-3}\acute{r}_0, y_1, y_2\}.$$

See Proposition 5.7 for a similar proof.

For adding k Ritz vectors, there is again a way to combine them into one vector s, so that the Krylov subspace generated with s is the desired subspace (4.1) (see [29, section 3] and also section 7.1). However, Sorensen's implicit restarting [47] in the IRA method provides a better way to implement the formation of subspace (4.1).

4.2. Generating the subspace with implicit restarting. The following theorem establishes the equivalence between augmenting the FOM subspace with Ritz vectors and using implicit restarting. "Exact shifts" [47] should be used for the implicit restarting (the shifts are the unwanted Ritz values).

THEOREM 4.2. *Suppose the IRA method is implemented with the unwanted Ritz values as shifts, and with the initial starting vector the same as the initial $\acute{r}_0$ for FOM. Then IRA generates the subspaces*

$$Span\{\acute{r}_0, A\acute{r}_0, A^2\acute{r}_0, A^3\acute{r}_0, \ldots, A^{m-k-1}\acute{r}_0, y_1, y_2, \ldots, y_k\}, \tag{4.5}$$

where $y_1, \ldots, y_k$ are Ritz vectors and $\acute{r}_0$ is the residual vector for an FOM method using these augmented subspaces.

Proof. The proof has been mostly given in other papers; we will give a sketch. Suppose we have an Arnoldi recurrence $AV_m = V_{m+1}\bar{H}_m$ and apply implicit restarting with the p unwanted Ritz values $\theta_{k+1}, \ldots, \theta_m$ as shifts. Then the first k Arnoldi basis vectors in the next subspace span the subspace $Span\{y_1, \ldots, y_k\}$. This is [29, Theorem 2]; the main step in proving this is that for $j \le k$, $Qe_j = \psi(H_m)t$, where ψ is a polynomial with roots at the undesired Ritz values, and t is some vector (this follows from (2.8), if we let t be the solution of $Rt = e_j$). These unwanted Ritz values are eigenvalues of H_m, so $\psi(H_m)$ purges the unwanted eigencomponents from t and we have that $\psi(H_m)t = Qe_j$ is a combination of only the desired eigenvectors of H_m. Then $V^+e_j = V_mQe_j$ is a combination of the desired Ritz vectors; see [48] for a complete proof.

As observed in the proof of [29, Theorem 3], the v_{k+1} Arnoldi basis vector in the new subspace is the same as the v_{m+1} vector for the previous subspace, so IRA has subspace

$$Span\{y_1, y_2, \ldots y_k, v_{m+1}, Av_{m+1}, A^2v_{m+1}, A^3v_{m+1}, \ldots, A^{m-k-1}v_{m+1}\}.$$

Since the FOM residual vector $\acute{r}_0$ is a multiple of v_{m+1}, this is equivalent to subspace (4.5). ∎

Next, the algorithm is listed for implicitly restarted FOM (FOM-IR). Added to the standard FOM algorithm is the solution of a small eigenvalue problem and the implicit restarting. Also, the reduced problem in step 3 is slightly different. We derive the formula for step 3 using (2.6). Note the superscript *old* denotes a quantity from the previous cycle of FOM. The reduced problem is

$$\begin{aligned} V_m^TAV_md &= V_m^T\acute{r}_0, \\ H_md &= V_m^T(-d_m^{old}h_{m+1,m}^{old}v_{m+1}^{old}), \\ H_md &= -d_m^{old}h_{m+1,m}^{old}e_{k+1}. \end{aligned}$$

The first k entries in the right-hand side are zero because $\acute{r}_0$, being a multiple of v_{m+1}^{old}, is orthogonal to all of the Ritz vectors (which are also from the previous cycle), and as just mentioned in the proof, $v_1, \ldots, v_k$ are combinations of the Ritz vectors. The last $m-k-1$ entries in the right-hand side are zero because $v_{k+1} = v_{m+1}^{old}$ and v_j is orthogonal to v_{k+1} if $j > k+1$.

FOM-IR.

(1) *Start:* Choose m, the maximum size of the subspace, and k, the desired number of approximate eigenvectors. Choose x_0 and compute $\acute{r}_0 = b - Ax_0$. The recast problem is $A(x - x_0) = \acute{r}_0$. Let $v_1 = \acute{r}_0/||\acute{r}_0||$ and $\beta = ||\acute{r}_0||$.
(2) *Iterate:* Apply the Arnoldi iteration out to step m.
(3) *Form the approximate solution:* $x_m = x_0 + V_m d$, where d comes from a reduced problem. In the first cycle, before any restarts, d is the solution of $H_m d = \beta e_1$. In later cycles, d is the solution of $H_m d = -d_m^{old} h_{m+1,m}^{old} e_{k+1}$. Check $||\acute{r}||$ for convergence, and proceed if not satisfied.
(4) *Eigenvalue problem:* Compute the eigenvalues of H_m. These are the Arnoldi–Ritz values, $\theta_1, \ldots, \theta_m$.
(5) *Restart:* Let $x_0 = x_m$. Apply implicit restarting with the unwanted Ritz values as shifts. Both parts of a complex conjugate pair should be included. Go to 2, and resume the Arnoldi iteration from step $k+1$.

FOM-IR solves linear equations and at the same time implements Sorensen's implicitly restarted Arnoldi method for eigenvalues. In spite of the similarity of subspace (4.1) to (3.2), the FOM-IR method is not equivalent to the GMRES-E method discussed in section 3. The residual vectors and approximate eigenvectors are both different.

It is interesting that even though FOM does not use a minimum residual solution, the minimum residual norm can be monitored. It is this minimum residual norm that is given in the examples in section 6. However, when restarting, the approximate solution and residual vector must be found with the FOM approach instead of minimum residual. This is the reason for the jumps in the residual norm at the time of restarts in Figure 1.

5. Implicitly restarted GMRES. Now we include approximate eigenvectors in the subspace for GMRES. For the approximate eigenvectors to fit into a Krylov subspace, we must use the harmonic Ritz vectors mentioned in section 2.4. The subspace is

$$Span\{r_0, Ar_0, A^2r_0, A^3r_0, \ldots, A^{m-k-1}r_0, \tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_k\}, \tag{5.1}$$

which is the same subspace used in the GMRES-E algorithm. The first subsection looks at why this subspace is a Krylov subspace, and the second proves that implicit restarting can be used.

5.1. The whole subspace is a Krylov subspace. We assume that A is a nonsingular matrix. We will deal mainly with the case of distinct harmonic Ritz values, but we first discuss briefly the special situation of multiple values.

An upper-Hessenberg matrix is called unreduced if there are no zero elements on the subdiagonal. The matrix H_m produced by the Arnoldi iteration is generally unreduced. Otherwise, the Krylov subspace at the time the zero occurs is an invariant subspace of A. We first give a standard result for Arnoldi–Ritz pairs. For a proof, see, for instance, [23].

LEMMA 5.1. *If the Arnoldi iteration produces an unreduced Hessenberg matrix H_m, then multiple eigenvalues of H_m have only one eigenvector associated with them. So multiple Arnoldi–Ritz values have only one corresponding Ritz vector.*

We can give a similar result for harmonic Ritz values.

THEOREM 5.2. *Assume that the Arnoldi–Hessenberg matrix is unreduced. Then multiple harmonic Ritz values have only one harmonic Ritz vector associated with them.*

Proof. We consider a problem equivalent to the harmonic Ritz value problem. The equivalent problem is applying the Rayleigh–Ritz procedure with matrix A^{-1} to the subspace spanned by the columns of the matrix AV_m [27, 30, 33]. The reciprocals of the Ritz values from this problem are the harmonic Ritz values. The Ritz vectors of this equivalent problem are $AV_m\tilde{g}_i$, where the $\tilde{g}_i$ are the same as defined earlier in (2.10) for the harmonic Ritz vectors $V_m\tilde{g}_i$.

This equivalent Rayleigh–Ritz problem can be viewed as being based on an Arnoldi iteration, although different from the one we have been considering. To see this, rewrite the subspace for this equivalent problem, $Span\{Av_1, A^2v_1, \ldots, A^mv_1\}$, as $Span\{A^mv_1, A^{-1}A^mv_1, A^{-2}A^mv_1, \ldots A^{-m+1}A^mv_1\}$. This is a Krylov subspace with starting vector A^mv_1 and with matrix A^{-1}. Thus, there is some Arnoldi iteration that would generate this subspace. The corresponding Hessenberg matrix would also be unreduced (if the first j columns of AV_m spanned an invariant subspace, then the first j columns of V_m would also, and this violates an assumption). Thus, from Lemma 5.1, this different Hessenberg matrix has only one eigenvector associated with each eigenvalue. Going back to the original Arnoldi iteration, there is only one harmonic Ritz vector with each harmonic Ritz value. ∎

In the special case of multiple harmonic Ritz values, there is not a full set of harmonic Ritz vectors. For the rest of the paper, we assume the harmonic Ritz values are distinct and that H_m is unreduced. After two lemmas, we establish a connection between the harmonic residual vectors and the residual vector for GMRES.

LEMMA 5.3. *The GMRES residual vector r is orthogonal to the subspace spanned by the columns of AV_m.*

See [41, chapter 5] for a proof. An orthogonality condition can also be given for the harmonic residual vectors [45, 30].

LEMMA 5.4. *Let $\tilde{y}_j$ be a harmonic Ritz vector with associated harmonic residual vector $\tilde{r}_j = A\tilde{y}_j - \tilde{\theta}_j\tilde{y}_j$. Then $\tilde{r}_j$ is orthogonal to the subspace spanned by the columns of AV_m.*

THEOREM 5.5. *Assuming the same underlying Arnoldi iteration is used, the harmonic residual vectors are multiples of the GMRES residual vector.*

Proof. This follows quickly from the two lemmas. The harmonic residual vectors and the GMRES residual vector all reside in the Krylov subspace of dimension $m+1$ and are orthogonal to the same subspace of dimension m, so they must be multiples of each other. ∎

This relationship between the harmonic and GMRES residual vectors can be used to include harmonic Ritz vectors in a Krylov subspace for GMRES. This is similar to how standard Ritz vectors were included in FOM in section 4.1. At the time of a restart for GMRES, the new r_0 is set equal to the r for the current cycle. Let

$$u = \frac{r_0}{||r_0||}.$$

Then for each i, $\tilde{r}_i = \gamma_i u$, for some scalar γ_i. So

$$A\tilde{y}_i - \tilde{\theta}_i \tilde{y}_i = \gamma_i u. \tag{5.2}$$

Using (5.2), we can see that generating a Krylov subspace with starting vector $\tilde{y}_1$ builds a subspace that contains a Krylov subspace with r_0 as starting vector. The proof of the following proposition is like that for Proposition 4.1, except that it uses (5.2) instead of (2.7).

PROPOSITION 5.6. *For $\tilde{y}_1$ a harmonic Ritz vector and r_0 the GMRES residual vector,*

$$Span\{\tilde{y}_1, A\tilde{y}_1, A^2\tilde{y}_1, \ldots, A^{m-1}\tilde{y}_1\} = Span\{\tilde{y}_1, r_0, Ar_0, A^2r_0, \ldots, A^{m-2}r_0\}.$$

Next, for including two harmonic Ritz vectors in the subspace, we use the starting vector

$$s = \gamma_2 \tilde{y}_1 - \gamma_1 \tilde{y}_2. \tag{5.3}$$

The subspace that is generated contains a Krylov subspace of dimension $m-2$ with r_0 as starting vector, along with the two harmonic Ritz vectors.

PROPOSITION 5.7. *With starting vector s as in* (5.3)*,*

$$Span\{\tilde{s}, A\tilde{s}, A^2\tilde{s}, \ldots, A^{m-1}\tilde{s}\} = Span\{\tilde{y}_1, \tilde{y}_2, r_0, Ar_0, A^2r_0, \ldots, A^{m-3}r_0\}.$$

Proof. After multiplying (5.3) through by A and using (5.2),

$$\begin{aligned} As &= \gamma_2(\tilde{\theta}_1\tilde{y}_1 + \gamma_1 u) - \gamma_1(\tilde{\theta}_2\tilde{y}_2 + \gamma_2 u) \\ &= \gamma_2\tilde{\theta}_1\tilde{y}_1 - \gamma_1\tilde{\theta}_2\tilde{y}_2. \end{aligned}$$

The special choice of s caused the u term to drop out. We have that $Span\{s, As\} = Span\{\tilde{y}_1, \tilde{y}_2\}$.

Multiplying by A and again using (5.2) yields $A^2 s$ as a combination of $\tilde{y}_1$, $\tilde{y}_2$, and u. Continuing this, we get that A^{m-1} is a combination of $\tilde{y}_1$, $\tilde{y}_2$, u, Au, $\ldots$, $A^{m-3}u$, and

$$Span\{s, As, A^2s, \ldots, A^{m-1}s\} = Span\{\tilde{y}_1, \tilde{y}_2, u, Au, A^2u, \ldots, A^{m-3}u\}.$$

We are done, since u is a multiple of r_0. ∎

For adding more harmonic Ritz vectors to the subspace, a careful combination of the harmonic Ritz vectors can be used as starting vector (see section 7.2). However, we will concentrate on using implicit restarting.

5.2. Generating the subspace with implicit restarting. We will develop an implicitly restarted GMRES method, called GMRES-IR, that includes harmonic Ritz vectors in the subspace. The approach is like that for FOM-IR, except the unwanted *harmonic* Ritz values are used as shifts during the QR phase of the algorithm. We need to prove that this approach generates the desired subspace (5.1). This is done in Theorem 5.14, but the proof for GMRES-IR is more difficult than was the proof of Theorem 4.2, the corresponding theorem for FOM. As in Theorem 4.2, we use that $Qe_j = \psi(H_m)t$, but here ψ has roots at harmonic Ritz values instead of standard Ritz values, and we need t to have nonzeros entries only in its first k positions. Then it

can be shown that $\psi(H_m)t$ has components only in the directions of the desired $\tilde{g}_i$'s, but it takes several lemmas. We now start on these.

The GMRES residual vector can be written in terms of a polynomial with roots at the harmonic Ritz values: $r_m = \pi(A)v_1$, where π is a polynomial such that $\pi(\lambda) = \alpha \prod_{l=1}^{m}(\lambda - \tilde{\theta}_l)$, and α is a scalar; see [26] and [8]. We need a similar expression for harmonic Ritz vectors.

LEMMA 5.8. *The harmonic Arnoldi–Ritz vectors can be written as* $\tilde{y}_i = \phi_i(A)v_1$, *where* ϕ_i *is the polynomial*

$$\phi_i(\lambda) = \alpha_i \prod_{\substack{l=1 \\ l \neq i}}^{m} (\lambda - \tilde{\theta}_l),$$

and α_i *is a scalar.*

For a proof of this lemma; see [30]. The next lemma relates polynomials of A with polynomials of H_m; see [5, 33, 23] for similar results and proofs.

LEMMA 5.9. *For* p *any polynomial of degree less than* m,

$$p(A)v_1 = V_m p(H_m)e_1.$$

For $\tilde{g}_i$ a solution of (2.10), there is an expression in terms of the same polynomials ϕ_i used for $\tilde{y}_i$ in Lemma 5.8.

LEMMA 5.10. *For* $\tilde{g}_i$ *from* (2.10) *and* ϕ_i *as defined in Lemma* 5.8,

$$\tilde{g}_i = \phi_i(H_m)e_1.$$

Proof. We will write $\tilde{g}_i$ in terms of $\tilde{y}_i$, then use Lemmas 5.8 and 5.9.

$$\begin{aligned} \tilde{g}_i &= V_m^T V_m \tilde{g}_i \\ &= V_m^T \tilde{y}_i \\ &= V_m^T \phi_i(A)v_1 \\ &= V_m^T V_m \phi_i(H_m)e_1 \\ &= \phi_i(H_m)e_1. \qquad \square \end{aligned}$$

Lemma 5.9 is also used in proving the next lemma.

LEMMA 5.11. *For* $j \leq m$, *there are some scalars* γ *and* $\beta_1, \beta_2, \ldots, \beta_{j-1}$ *such that*

$$e_j = \gamma \prod_{l=1}^{j-1} (H_m - \beta_l I)e_1.$$

Proof. Because v_j is a member of the Krylov subspace of degree j with v_1 as a starting vector, there are scalars γ and $\beta_1, \beta_2, \ldots, \beta_{j-1}$ such that

$$v_j = \gamma \prod_{l=1}^{j-1} (A - \beta_l I)v_1.$$

Thus,

$$\begin{aligned} e_j &= V_m^T v_j \\ &= \gamma V_m^T \prod_{l=1}^{j-1}(A-\beta_l I)v_1 \\ &= \gamma V_m^T V_m \prod_{l=1}^{j-1}(H_m-\beta_l I)e_1 \\ &= \gamma \prod_{l=1}^{j-1}(H_m-\beta_l I)e_1, \end{aligned}$$

where we used Lemma 5.9 to convert from a polynomial of A to a polynomial of H_m. ∎

In what follows, we define

$$\psi(\lambda) \equiv \prod_{l=k+1}^{m} (\lambda - \tilde{\theta}_l). \tag{5.4}$$

LEMMA 5.12. *For $t \in Span\{e_1, e_2, \ldots, e_k\}$,*

$$\psi(H_m)t \in Span\{\tilde{g}_1, \tilde{g}_2, \ldots \tilde{g}_k\}.$$

Proof. We will show that for $j \leq k$, $\psi(H_m)e_j \in Span\{\tilde{g}_1, \tilde{g}_2, \ldots \tilde{g}_k\}$, from which the desired result quickly follows.

Using Lemma 5.11,

$$\begin{aligned} \psi(H_m)e_j &= \gamma\psi(H_m)\prod_{l=1}^{j-1}(H_m-\beta_l I)e_1 \\ &= \gamma\Big(\prod_{l=1}^{j-1}(H_m-\beta_l I)\Big)\psi(H_m)e_1. \end{aligned}$$

We need to show that this vector is a combination of the vectors $\tilde{g}_1, \ldots, \tilde{g}_k$, which from Lemma 5.10 can be written as

$$\begin{aligned} \tilde{g}_i &= \phi_i(H_m)e_1 \\ &= \alpha_i \prod_{\substack{l=1 \\ l\neq i}}^{m}(H_m-\tilde{\theta}_l)e_1 \\ &= \alpha_i\Big(\prod_{\substack{l=1 \\ l\neq i}}^{k}(H_m-\tilde{\theta}_l)\Big)\psi(H_m)e_1. \end{aligned}$$

So expressing $\psi(H_m)e_j$ as a combination of the vectors $\tilde{g}_1, \ldots, \tilde{g}_k$ is equivalent to expressing the polynomial

$$\prod_{l=1}^{j-1}(\lambda-\beta_l),$$

with $j \leq k$, as a combination of the polynomials

$$\alpha_i \prod_{\substack{l=1 \\ l \neq i}}^{k} (\lambda - \tilde{\theta}_l),$$

for i from 1 to k. These last polynomials are Lagrange interpolating polynomials except for possibly different normalization, and they are linearly independent since the $\tilde{\theta}_l$'s are distinct. They can be combined to form any polynomial of degree $k-1$ or less, so we are finished. ∎

We now give the last lemma which deals with the QR iteration [14, 23].

LEMMA 5.13. *Let the matrix Q and shifts τ_i be from the QR iteration. Then for $j \leq k$,*

$$Qe_j = \prod_{l=1}^{p} (H_m - \tau_l I)t,$$

where $t \in Span\{e_1, \ldots, e_k\}$.

Proof. Recall from section 2.3 that $H^{(1)} = H_m$ and $H^{(1)} - \tau_i I = Q^{(1)} R^{(1)}$. We assumed earlier that H_m is unreduced, so $H^{(1)} - \tau_1 I$ is an unreduced Hessenberg matrix. Therefore the first $m-1$ columns of $R^{(1)}$ are linearly independent. And $H^{(i)} - \tau_i I$ has its $m-i+1$ by $m-i+1$ leading principle submatrix unreduced (a zero in the subdiagonal can only move one position during each QR iteration). Therefore $R^{(i)}$ has its first $m-i$ columns linearly independent. Since each $R^{(i)}$ is also upper triangular and $R = R^{(p)} R^{(p-1)} \cdots R^{(1)}$, R has its first $m-p = k$ columns linear independent. With this and the fact that R is upper triangular, we have that for $j \leq k$, there is a solution to

$$Rt = e_j, \tag{5.5}$$

such that

$$t \in Span\{e_1, \ldots, e_k\}. \tag{5.6}$$

Let t be the solution of (5.5) such that (5.6) holds. Then $Qe_j = QRt = \prod_{l=1}^{p} (H_m - \tau_l I)t$, from (2.8). ∎

The GMRES-IR method uses implicit restarting with unwanted harmonic Ritz values as shifts. The theorem establishes that in every cycle, between restarts, it generates subspace (5.1).

THEOREM 5.14. *The GMRES-IR method generates the subspace*

$$Span\{r_0, Ar_0, A^2 r_0, A^3 r_0, \ldots, A^{m-k-1} r_0, \tilde{y}_1, \tilde{y}_2, \ldots \tilde{y}_k\}.$$

Proof. The shifts for the QR iteration are the unwanted harmonic Ritz values, so the $\tau_1, \ldots, \tau_p$ in Lemma 5.13 are $\tilde{\theta}_{k+1}, \tilde{\theta}_{k+2}, \ldots, \tilde{\theta}_m$. Therefore $\prod_{l=1}^{p} (H_m - \tau_l I) = \psi(H_m)$, for ψ as defined in (5.4). Lemma 5.13 becomes

$$Qe_j = \psi(H_m)t,$$

where $t \in Span\{e_1, \ldots, e_k\}$. Applying Lemma 5.12 gives

$$Qe_j \in Span\{\tilde{g}_1, \tilde{g}_2, \ldots \tilde{g}_k\}. \tag{5.7}$$

The new Arnoldi basis matrix is V^+. Using (2.9),

$$V^+ e_j = V_m Q e_j.$$

With (5.7), if $j \leq k$, then

$$V^+ e_j \in Span\{V_m \tilde{g}_1, V_m \tilde{g}_2, \ldots, V_m \tilde{g}_k\}.$$

And since the harmonic Ritz vectors are defined as $\tilde{y}_i = V_m \tilde{g}_i$,

$$V^+ e_j \in Span\{\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_k\}.$$

We have that after the implicit restarting, the first k columns of V^+ span the same subspace as $Span\{\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_k\}$.

So $\tilde{y}_1, \ldots, \tilde{y}_k$ are in the new subspace. We need to show that this subspace also contains the vectors $r_0, Ar_0, \ldots, A^{m-k-1} r_0$. Using (5.2), $A\tilde{y}_i$ is a combination of $\tilde{y}_i$ and r_0. Since v_k is a combination of $\tilde{y}_1, \ldots, \tilde{y}_k$, Av_k is a combination of r_0 and $\tilde{y}_1, \ldots, \tilde{y}_k$. Thus v_{k+1} is also a combination of r_0 and $\tilde{y}_1, \ldots, \tilde{y}_k$. Next, using a similar argument, v_{k+2} is a combination of r_0, Ar_0 and $\tilde{y}_1, \ldots, \tilde{y}_k$. This can be continued until we arrive at the desired result. □

So we have established that implicit restarting can be used with GMRES. The changes in the algorithm from FOM-IR are listed.

GMRES-IR. GMRES-IR is the same as FOM-IR, except replace $\acute{r}$ with r and substitute the following:

(3) d is the solution of $\min ||V_{m+1}^T r_0 - \bar{H}_m d||$.

(4) and (5) Compute and use harmonic Ritz values.

In step 3 of the algorithm, the vector $V_{m+1}^T r_0$ has zeros except in its first $k+1$ entries. It can also be shown that after Givens rotations are applied during solution of the least squares, the first k entries of this vector become zero.

Having harmonic Ritz vectors in the subspace for GMRES-IR causes the portion of the subspace that is Krylov with starting vector r_0 to have dimension $m-k$ instead of m as for standard GMRES. This slows convergence, but the k approximate eigenvectors generally improve convergence. So there is a tradeoff in choosing k.

We next compare GMRES-IR with some other methods. While GMRES-IR is mathematically equivalent to GMRES-E for a completed cycle, it can give better results at the start and middle of a cycle. This is because eigenvector information is included in the subspace from the beginning instead of being appended at the end of the cycle. For problems using only a moderate number of iterations, this may be of some significance. GMRES-E can be effective for nonlinear problems, while GMRES-IR is not designed for situations with changing equations. The main advantage of GMRES-IR compared to GMRES-E is in expense and storage. We compare these next. Table 1 has a rough count for one cycle of standard GMRES, GMRES-IR, and GMRES-E. Only the major expenses and storage vectors of length n are considered. The solution of the small eigenvalue problem in GMRES-IR and GMRES-E is not included because it is not of order n, however there are cases where it is a significant expense. The cost and storage depends on the implementation. For GMRES-IR, the formation of the new Krylov basis in (2.9) is assumed to be done in place using a

TABLE 1
Expenses and storage for GMRES methods.

	GMRES(m)	GMRES-IR(m,k)	GMRES-E(m,k)
DAXPYs, DDOTs, etc.	m^2	$m^2 + km - \frac{k^2}{2}$	$m^2 + 3km$
matrix-vector products	m	$m - k$	$m - k$
length n vectors	$m + 2$	$m + 2$	$m + k + 2$

Householder factorization of a portion of Q. We see that GMRES-IR can save matrix-vector products without additional storage. Standard GMRES uses fewer length n vector operations (DAXPYs, DDOTs, etc.) per cycle than GMRES-IR, but uses more matrix-vector products. The main advantage of GMRES-IR over GMRES is in reducing the number of iterations.

The next section continues the comparison of methods with some computational experiments.

6. Examples. We compare the new implicitly restarted GMRES and FOM methods with standard GMRES and also with transpose-free QMR (TFQMR) [9]. Some examples also include full, unrestarted GMRES. Full GMRES is not a practical method, but it is interesting to see how closely the other methods can match its performance. FOM-IR(25,6) has $m = 25$ and $k = 6$ (subspaces are dimension 25, including six approximate eigenvectors). Also used are GMRES-IR(25,6) and GMRES-IR(25,10). GMRES(25) refers to standard GMRES with subspaces of size 25. Residual norms are plotted against the number of matrix-vector products.

Example 6.1. The matrix is bidiagonal with entries 0.01, 0.1, 1, 2, 3, 4, ..., 997, 998 on the main diagonal and 1's on the super diagonal. The right-hand side has all 1's. The matrix has small eigenvalues that slow the convergence. No method can give an accurate solution until it develops approximations to the smallest few eigenvalues. GMRES(25) can never develop these approximations because of the restarts. The convergence stagnates; see Figure 1. TFQMR is much more successful because it generates a large subspace. However, the implictly restarted methods are even better. Comparing the number of matrix-vector products required to reduce the residual norm by a factor of 10^{-6} (from 31.6 to below $3.16 * 10^{-5}$), GMRES(25,10) uses 231 and TFQMR needs 378. This shows potential for GMRES-IR, particularly for problems with expensive matrix-vector products.

GMRES-IR(25,10) competes well with full GMRES, even though it uses subspaces of dimension 25. And the portion of these subspaces that is Krylov with starting vector r_0 is of dimension only 15. Meanwhile, full GMRES generates a subspace of size over 200. Yet the two methods have similar convergence. This is surprising, because this problem is fairly difficult. For easy problems, small subspaces can be almost as effective as large subspaces; for tougher problems, the dimension of the Krylov subspace is very important. However, the deflation of eigenvalues in GMRES-IR turns the difficult problem into an easy one. Once approximate eigenvectors have developed for the 10 smallest eigenvalues, the deflated spectrum is $9, 10, 11, \ldots, 998$. We can compare the square roots of the ratios of largest to smallest eigenvalues (similar to comparing the square roots of condition numbers in section 3.2, although the importance is less obvious). The square root of the ratio of largest to smallest

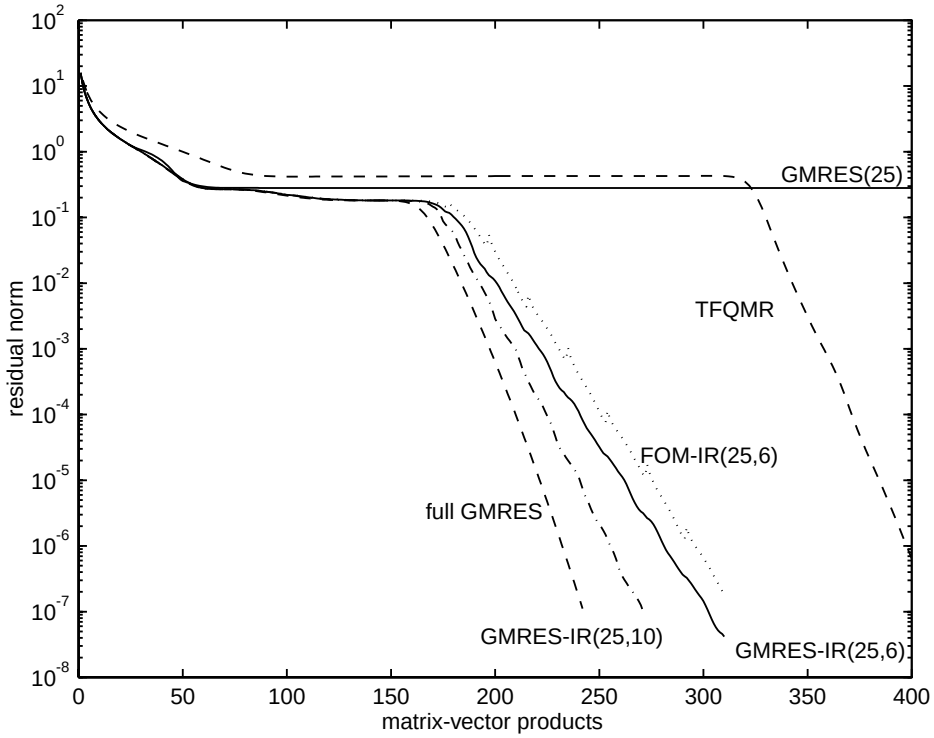


FIG. 1. *The comparison for Example* 6.1.

eigenvalue is 10.5 for the deflated spectrum, compared to the original ratio of 316. This makes the problem considerably easier. Krylov subspaces of dimension 15 are almost as effective for this easier problem as is a single Krylov subspace of dimension 200.

Example 6.2. The matrix is SHERMAN4 from the Harwell–Boeing sparse matrix collection [6]. The right-hand side is the one provided in the collection. This matrix has dimension 1104. All of the eigenvalues are positive. The smallest 11 eigenvalues are 0.031, 0.085, 0.28, 0.40, 0.43, 0.59, 0.63, 0.89, 0.92, 1.00, and 1.13, and the largest is 66.5. See Figure 2 for convergence plots. This problem is not as difficult for standard GMRES as Example 6.1. Convergence of GMRES(25) is slow but steady. TFQMR is erratic, but considerably better than GMRES(25). However, the implicitly restarted GMRES and FOM methods converge both faster and more consistently than TFQMR. Results are very different when comparing length n vector operations. Because the matrix is sparse (3.43 nonzeros per row), GMRES-IR uses about four times as many vector operations per matrix-vector product than does TFQMR, approximately 43 versus 10.4 (7 for DAXPYs and DDOTs, though this depends on the implementation, and 3.4 for the matrix). To improve the residual norm by 10^{-6}, GMRES-IR(25,6) requires about 2.6 times as many vector operations as TFQMR (5.5 versus 2.1 thousands).

Table 2 gives results for various choices of m and k. First, m is fixed at 25 and k varies from $k = 0$ (standard GMRES) up to $k = 10$. The number of matrix-vector products does not change much for $k \geq 4$. Next, m varies with $k = 4$. With $m = 15$,

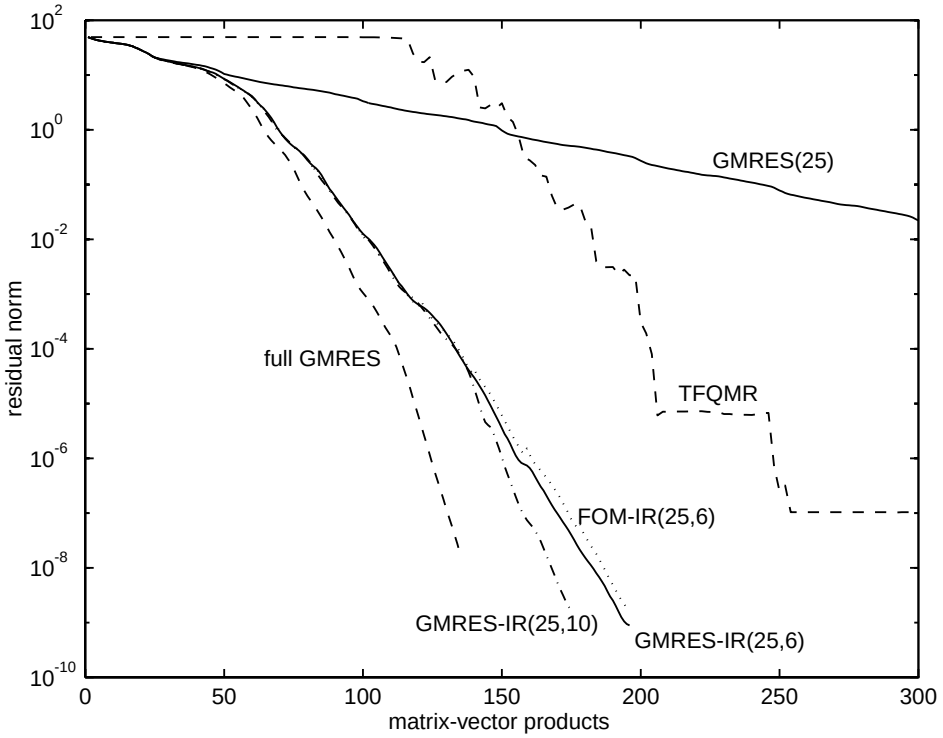


FIG. 2. *SHERMAN4 matrix.*

the number matrix-vector products is 157, only 15 more than with $m = 25$.

The next matrix is difficult for any restarted method.

Example 6.3. The Harwell-Boeing matrix SHERMAN5 is used. The dimension is 3312, and there are eigenvalues with positive and negative real parts. The eight eigenvalues nearest the origin are 0.047, 0.13, 0.40, 0.58, 0.62, 0.85, 0.91, and 1.0. The largest positive and negative eigenvalues are 595 and −189. Without preconditioning this is a difficult problem. GMRES-IR(25,6) stalls out and makes little progress in 350 cycles (over 6000 matrix-vector products). Good approximations to the smallest eigenvalues do not develop. GMRES-IR(30,8) works better; the slightly larger Krylov subspaces make a difference and good eigenvalue approximations are eventually found. Improving the residual norm by a factor of 10^{-6} takes 146 cycles or about 3221 matrix-vector products. Meanwhile, from of [9, Figure 6.2], TFQMR requires about 2700 matrix-vector products. The extremely large subspace generated by TFQMR is needed for this difficult indefinite problem. Finally we note that FOM-IR takes the same number of cycles as GMRES-IR for this indefinite problem. Further comparisons between the two seem warranted.

For more experiments, using the mathematically equivalent GMRES-E approach, see [28, 4, 42].

7. Eigenvalue problems. FOM-IR and GMRES-IR can also be thought of as eigenvalue methods. We briefly look at a few aspects of these methods.

TABLE 2
Comparison with different k and m values.

m	k	mat.-vec. products	thousands of vect.ops
25	0	526	14.9
25	1	298	8.9
25	2	166	5.3
25	3	147	4.9
25	4	142	5.0
25	5	140	5.3
25	6	137	5.5
25	7	138	6.0
25	8	137	6.3
25	9	136	6.7
25	10	137	7.4
7	4	323	8.9
10	4	198	4.9
15	4	157	4.3
20	4	146	4.6
30	4	134	5.4
35	4	135	6.0
40	4	134	6.4

7.1. With FOM-IR. As mentioned in section 5, FOM-IR implements the IRA eigenvalue method. It is interesting how simple it is to add the solution of linear equations on top of IRA. The subspace (4.1) for FOM-IR is not only a Krylov subspace itself, but contains several other Krylov subspaces that are interesting from an eigenvalue viewpoint. The subspaces are $\{y_i, Ay_i, A^2y_i, A^3y_i, \ldots, A^py_i\}$, for each i from 1 to k [29, Theorem 3]. Since these are Krylov subspaces with each Ritz vector as a starting vector, this shows why good eigenvector approximations can develop.

In this paragraph, we give another way of looking at the IRA method, and thus at FOM-IR. It is mentioned in [29] that the starting vector s at the beginning of a new cycle of IRA can be found without the implicit restarting. The vector s is a combination of the k Ritz vectors. The coefficients for this combination are from the solution of the $k-1$ by k homogeneous system of equations with jth row $[\theta_1^{j-1}\beta_{m1}, \theta_2^{j-1}\beta_{m2}, \ldots, \theta_k^{j-1}\beta_{mk}]$. Here we give this solution. Let ω_i be a polynomial with roots at all of the Ritz values except θ_i. Specifically,

$$\omega_i(\lambda) = \prod_{\substack{l=1 \\ l \neq i}}^{k} (\theta_l - \lambda).$$

Then

$$s = \sum_{i=1}^{k} \frac{1}{\beta_{mi}\omega_i(\theta_i)} y_i = V_m \sum_{i=1}^{k} \frac{1}{\beta_{mi}\omega_i(\theta_i)} g_i. \tag{7.1}$$

Formulas can also be given for the vectors $As, A^2s, \ldots, A^{k-1}s$. For $j < k$,

$$A^j s = \sum_{i=1}^{k} \frac{\theta_i^j}{\beta_{mi}\omega_i(\theta_i)} y_i = V_m \sum_{i=1}^{k} \frac{\theta_i^j}{\beta_{mi}\omega_i(\theta_i)} g_i. \tag{7.2}$$

This actually gives a way of implementing an algorithm mathematically equivalent to IRA that does not use the QR iteration (see [29] for yet another mathematically equivalent method). The vectors $s, As, A^2s, \ldots, A^{k-1}s$ need to be orthogonalized in order to form the first k columns of the new V matrix. Just as in IRA, k matrix-vector products are saved. This algorithm is not recommended, because of questionable numerical properties.

7.2. With GMRES-IR. GMRES-IR also solves both linear equations and the associated eigenvalue problem. It finds eigenvalues using the harmonic approach. This is equivalent to the interior Arnoldi with eigenvectors method that is given in [30] (with $\sigma = 0$, but other shifts could be incorporated). The implementation is more efficient. It can also be viewed as an interior eigenvalue version of the IRA method.

One application for the eigenvalue portion of GMRES-IR is with iterative methods for linear equations that require eigenvalue estimates [19, 26, 31]. GMRES-IR can find the eigenvalue estimates and at the same time begin solving the linear equations.

GMRES-IR also develops an interesting polynomial. Since subspace (5.1) is Krylov, there is a polynomial associated with it. And this polynomial has some zeros approximating eigenvalues near the origin. It may be of use for polynomial preconditioning [19].

The subspace for GMRES-IR contains Krylov subspaces with the harmonic Ritz vectors as starting vectors.

THEOREM 7.1. *The GMRES-IR subspace* (5.1) *contains the subspaces* $\{\tilde{y}_i, A\tilde{y}_i, A^2\tilde{y}_i, \ldots, A^p\tilde{y}_i\}$, *for each* i *from* 1 *to* k.

Proof. From (5.2),

$$A\tilde{y}_i = \tilde{\theta}_i\tilde{y}_i + \gamma_i u,$$

where u is a multiple of r_0. So $Span\{\tilde{y}_i, A\tilde{y}_i\} = Span\{\tilde{y}_i, r_0\}$. Next, again using (5.2),

$$A^2\tilde{y}_i = \tilde{\theta}_i^2\tilde{y}_i + \gamma_i\tilde{\theta}_i u + \gamma_i Au.$$

So $Span\{\tilde{y}_i, A\tilde{y}_i, A^2\tilde{y}_i\} = Span\{\tilde{y}_i, r_0, Ar_0\}$. This process can be continued until we reach

$$Span\{\tilde{y}_i, A\tilde{y}_i, \ldots, A^p\tilde{y}_i\} = Span\{\tilde{y}_i, r_0, Ar_0, \ldots, A^{p-1}r_0\}.$$

The right-hand side of this last equation is a subspace of (5.1), so we have the desired result. ∎

Theorem 7.1 shows why good approximations to eigenvectors can be expected in GMRES-IR. We next look at the eigenvalues that are found in Example 6.1 by GMRES-IR and note a correspondence between when the eigenvalues and linear equations converge.

Example 7.1. We continue Example 6.1 and look at the accuracy of the harmonic Ritz vectors that are developed by GMRES-IR(25,6). Figure 3 plots the norms of the residual vectors $\tilde{r}_i = A\tilde{y}_i - \rho_i\tilde{y}_i$, where ρ_i is the Rayleigh quotient of y_i. Figure 3 also has the residual norm for the linear equations. These linear equations residual norms are calculated at every matrix-vector product, while the eigenvalue residuals are computed only every 19 iterations. The linear equations residual does not begin to improve until after the approximations to the two smallest eigenvalues have made significant progress.

A version of GMRES-IR (equivalently GMRES-E) can be given that uses combinations of harmonic Ritz vectors instead of the QR-iteration, similar to (7.1) and (7.2)

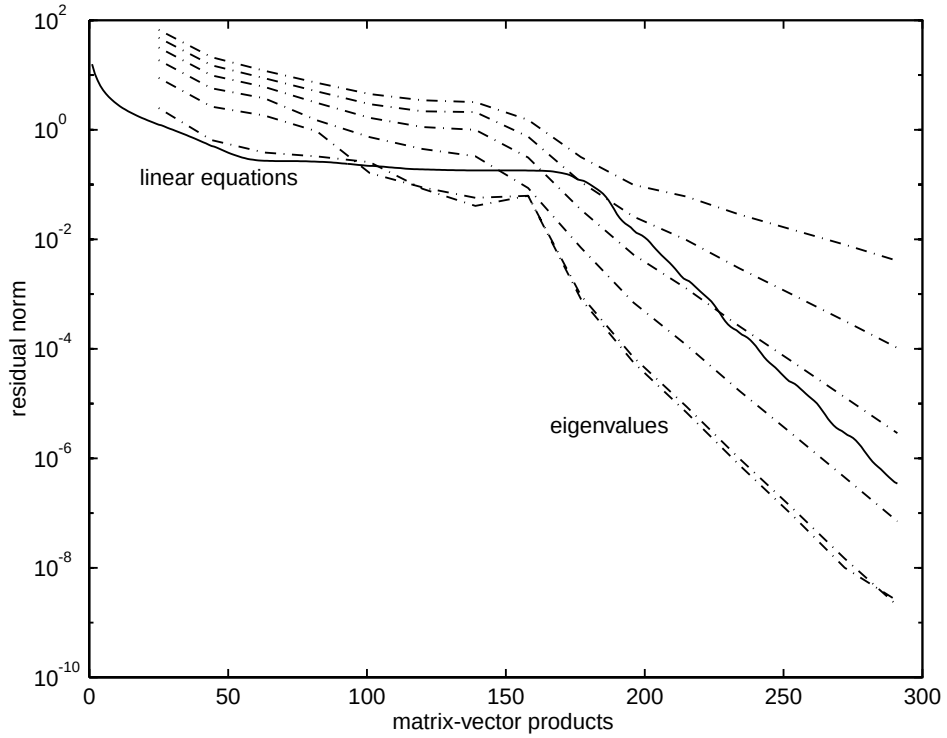


FIG. 3. *Convergence for the linear equations and six eigenvalues.*

for FOM-IR. For s the starting vector of the new GMRES cycle, the γ_i's from (5.2), and

$$\omega_i(\lambda) = \prod_{\substack{l=1 \\ l \neq i}}^{k} (\tilde{\theta}_l - \lambda),$$

if $0 \leq j < k$, then

$$A^j s = \sum_{i=1}^{k} \frac{\tilde{\theta}_i^j}{\gamma_i \omega_i(\tilde{\theta}_i)} \tilde{y}_i = V_m \sum_{i=1}^{k} \frac{\tilde{\theta}_i^j}{\gamma_i \omega_i(\tilde{\theta}_i)} \tilde{h}_i.$$

8. Conclusion. Implicit restarting can be used with the FOM and GMRES methods. A subspace is generated that contains both approximate eigenvectors and the necessary Krylov subspace with the linear equations residual vector as starting vector. Remarkably, the whole subspace is itself a Krylov subspace, though with a different starting vector. With the FOM-IR method, standard Ritz vectors are included in the subspace. With GMRES-IR, harmonic Ritz vectors are used.

Examples show that the implicitly restarted methods can be much better than standard GMRES, and also can be competitive with TFQMR, especially when the matrix-vector product is expensive. These new methods compute eigenvalues at the same time that they solve the linear equations.

We list some topics for future work. Most important is a study of the stability of the implicit restarting. We have seen stability problems when there are eigenvalues that stand out in the spectrum and thus converge rapidly. Such eigenvalues are not uncommon when preconditioning is used. Similar problems have been dealt with for the IRA method [23, 24].

Other possible topics include the possibility that eigenvectors could be released once they have converged to a certain point. An automatic procedure would be desirable for choosing the number of approximate eigenvectors and for selecting exterior Ritz values if they are outstanding; see [48] for related work.

Also, the polynomial preconditioning mentioned in the previous section is currently being investigated and there are possible applications to systems with multiple right-hand sides. The relation

$$AV_k = V_{k+1}\bar{H}_k,$$

that is formed after the implicit restarting, gives both approximate eigenvectors and their matrix-vector products with A in a compact form. With this relation, eigenvalues can be deflated in the solution of subsequent right-hand sides.

Acknowledgments. The author wishes to thank Rich Lehoucq for helpful discussions and the referees for their helpful suggestions.

REFERENCES

[1] W. E. Arnoldi, *The principle of minimized iterations in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.
[2] S. F. Ashby, T. A. Manteuffel, and P. E. Saylor, *A taxonomy for conjugate gradient methods*, SIAM J. Numer. Anal., 27 (1990), pp. 1542–1568.
[3] J. Baglama, D. Calvetti, G. H. Golub, and L. Reichel, *Adaptively preconditioned GMRES algorithms*, SIAM J. Sci. Comput., 20 (1999), pp. 243–269.
[4] A. Chapman and Y. Saad, *Deflated and augmented Krylov subspace techniques*, Numer. Linear Algebra Appl., 4 (1997), pp. 43–66.
[5] V. L. Druskin and L. A. Knizhnerman, *Two polynomial methods of calculating functions of a symmetric matrices*, Comput. Math. Math. Phys., 29 (1989), pp. 112–121.
[6] I. S. Duff, R. G. Grimes, and J. G. Lewis, *Sparse matrix test problems*, ACM Trans. Math. Software, 15 (1989), pp. 1–14.
[7] J. Erhel, K. Burrage, and B. Pohl, *Restarted GMRES preconditioned by deflation*, J. Comput. Appl. Math., 69 (1996), pp. 303–318.
[8] R. W. Freund, *Quasi-kernel polynomials and their use in non-Hermitian matrix iterations*, J. Comput. Appl. Math., 43 (1992), pp. 135–158.
[9] R. W. Freund, *A transpose-free quasi-minimal residual algorithm for non-Hermitian linear systems*, SIAM J. Sci. Comput., 14 (1993), pp. 470–482.
[10] R. W. Freund, G. H. Golub, and N. M. Nachtigal, *Iterative solution of linear systems*, Acta Numerica, 1 (1992), pp. 57–100.
[11] R. W. Freund, M. H. Gutknecht, and N. M. Nachtigal, *An implementation of the look-ahead Lanczos algorithm for non-Hermitian matrices*, SIAM J. Sci. Comput., 14 (1993), pp. 137–158.
[12] R. W. Freund and N. M. Nachtigal, *QMR: a quasi-minimal residual method for non-Hermitian linear systems*, Numer. Math., 60 (1991), pp. 315–339.
[13] G. H. Golub and D. P. O'Leary, *Some history of the conjugate gradient and Lanczos algorithms:* 1948–1976, SIAM Rev., 31 (1989), pp. 50–102.
[14] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., Johns Hopkins University Press, Baltimore, MD, 1989.
[15] A. Greenbaum, V. Ptak, and Z. Strakos, *Any nonincreasing convergence curve is possible for GMRES*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 465–469.
[16] A. Greenbaum and L. N. Trefethen, *GMRESICR and Arnoldi Lanczos as matrix approximation problems*, SIAM J. Sci. Comput., 15 (1994), pp. 359–368.

[17] L. A. Hageman and D. M. Young, *Applied Iterative Methods*, Academic Press, New York, 1981.
[18] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–436.
[19] W. Joubert, *A robust GMRES-based adaptive polynomial preconditioning algorithm for nonsymmetric linear systems*, SIAM J. Sci. Comput., 15 (1994), pp. 427–439.
[20] S. A. Kharchenko and A. Y. Yeremin, *Eigenvalue translation based preconditioners for the GMRES(k) method*, Numer. Linear Algebra Appl., 2 (1995), pp. 51–77.
[21] C. Lanczos, *An iterative method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Research Nat. Bur. Standards, 45 (1950), pp. 255–282.
[22] C. Lanczos, *Solution of systems of linear equations by minimized iterations*, J. Research Nat. Bur. Standards, 49 (1952), pp. 33–53.
[23] R. B. Lehoucq, *Analysis and Implementation of an Implicitly Restarted Arnoldi Iteration*, Ph.D. thesis, Rice University, Houston, TX, 1995.
[24] R. B. Lehoucq and D. C. Sorensen, *Deflation techniques for an implicitly restarted Arnoldi iteration*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 789–821.
[25] T. A. Manteuffel and J. Otto, *On the roots of orthogonal polynomials and residual polynomials associated with a conjugate gradient method*, Numer. Linear Algebra Appl., 1 (1994), pp. 449–475.
[26] T. A. Manteuffel and G. Starke, *On hybrid iterative methods for nonsymmetric systems of linear equations*, Numer. Math., 73 (1996), pp. 489–506.
[27] R. B. Morgan, *Computing interior eigenvalues of large matrices*, Linear Algebra Appl., 154-156 (1991), pp. 289–309.
[28] R. B. Morgan, *A restarted GMRES method augmented with eigenvectors*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1154–1171.
[29] R. B. Morgan, *On restarting the Arnoldi method for large nonsymmetric eigenvalue problems*, Math. Comp., 65 (1996), pp. 1213–1230.
[30] R. B. Morgan and M. Zeng, *Harmonic projection methods for large non-symmetric eigenvalue problems*, Numer. Linear Algebra Appl., 5 (1998), pp. 33–55.
[31] N. M. Nachtigal, L. Reichel, and L. N. Trefethen, *A hybrid GMRES algorithm for nonsymmetric linear systems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 796–825.
[32] R. A. Nicolaides, *Deflation of conjugate gradients with applications to boundary value problems*, SIAM J. Numer. Anal., 24 (1987), pp. 355–365.
[33] C. C. Paige, B. N. Parlett, and H. A. van der Vorst, *Approximate solutions and eigenvalue bounds from Krylov subspaces*, Numer. Linear Algebra Appl., 2 (1995), pp. 115–133.
[34] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.
[35] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.
[36] B. N. Parlett, D. R. Taylor, and Z. A. Liu, *A look-ahead Lanczos algorithm for unsymmetric matrices*, Math. Comp., 44 (1985), pp. 105–124.
[37] Y. Saad, *Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices*, Linear Algebra Appl., 34 (1980), pp. 269–295.
[38] Y. Saad, *Krylov subspace methods for solving large unsymmetric linear systems*, Math. Comp., 37 (1981), pp. 105–126.
[39] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Halsted Press, New York, NY, 1992.
[40] Y. Saad, *A flexible inner-outer preconditioned GMRES algorithm*, SIAM J. Sci. Comput., 14 (1993), pp. 461–469.
[41] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing, Boston, MA, 1995.
[42] Y. Saad, *Analysis of augmented Krylov subspace techniques*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 435–449.
[43] Y. Saad and M. H. Schultz, *GMRES: a generalized minimum residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
[44] P. E. Saylor and D. C. Smolarski, *Implementation of an adaptive algorithm for Richardson's method*, Linear Algebra Appl., 154/156 (1991), pp. 615–646.
[45] G. L. G. Sleijpen and H. A. van der Vorst, *A Jacobi-Davidson iteration method for linear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 401–425.
[46] P. Sonneveld, *CGS, a fast Lanczos-type solver for nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 36–52.
[47] D. C. Sorensen, *Implicit application of polynomial filters in a k-step Arnoldi method*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 357–385.

[48] A. Stathopoulos, Y. Saad, and K. Wu, *Dynamic thick restarting of the Davidson, and the implicitly restarted Arnoldi methods*, SIAM J. Sci. Comput., 19 (1998), pp. 227–245.

[49] E. De Sturler, *Truncation strategies for optimal Krylov subspace methods*, Technical report TR-96-38, Swiss Center for Scientific Computing (ETH Zurich), Zurich, Switzerland, 1996.

[50] E. De Sturler and D. R. Fokkema, *Nested Krylov methods and preserving the orthogonality*, in Proceedings Sixth Copper Mountain Conference on Multigrid Methods, NASA Con. Publ. 3224, Part 1, S. F. McCormick, N. D. Melson, and T. A. Manteuffel, eds., NASA Langley Research Center, Hampton, VA, 1993.

[51] C. H. Tong, *A comparative study of preconditioned lanczos methods for nonsymetric linear systems*, Technical report SAND91-8240, Sandia National Laboratories, Albuquerque, NM, 1992.

[52] H. A. van der Vorst and C. Vuik, *The superlinear convergence behaviour of GMRES*, J. Comput. Appl. Math., 48 (1993), pp. 327–341.

[53] H. A. van der Vorst and C. Vuik, *GMRESR: A family of nested GMRES methods*, Numer. Linear Algebra Appl., 1 (1994), pp. 369–386.

[54] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, New York, 1965.

[55] M. Zeng, *Finding Interior Eigenvalues of Large Nonsymmetric Matrices*, Ph.D. thesis, University of Missouri, Columbia, MO, 1996.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 4, pp. 1136–1155
© 2000 Society for Industrial and Applied Mathematics

STRUCTURE OF EXPANSION-CONTRACTION MATRICES IN THE INCLUSION PRINCIPLE FOR DYNAMIC SYSTEMS*

LUBOMÍR BAKULE†, JOSÉ RODELLAR‡, AND JOSEP M. ROSSELL§

Abstract. This paper characterizes a general structure of complementary matrices involved in the input-state-output inclusion principle for linear time invariant dynamic systems. Aggregations and restrictions are adopted as two practically important classes within this scheme. Further, contractibility conditions for feedback controllers are presented, including the decentralized control design. The identified structure enables the formulation of conditions on the complementary matrices which can be useful for analysis and synthesis of control problems.

Key words. inclusion principle, large scale systems, decentralized control, decomposition, overlapping

AMS subject classifications. 15A04, 93A15, 93A14, 93B17, 93C35, 34H05, 34A30

PII. S0895479898340548

1. Introduction.

The *inclusion principle* was proposed in the early eighties in the context of analysis and control of complex and large-scale systems. Essentially, the inclusion principle establishes a mathematical framework in which two dynamic systems with different dimensions may have an equivalent behavior. This means that (1) a "big" system can be built from a "small" one, through an *expansion* process, in such a way that the "big" system contains the essential information about the behavior of the "small" system; and (2) this information can be extracted from the "big" system by a *contraction* process. The inclusion principle has been introduced by Šiljak [10], Ikeda and Šiljak [4], [5], [6], and Ikeda, Šiljak, and White [7]. Particular attention has been given to the design of overlapping decentralized controllers within the above-mentioned framework. Basically, the idea is to expand the control system with overlapped components into a larger dimensional system that appears decomposed into a number of disjoint subsystems. Then, decentralized controllers are designed for the expanded system and contracted to be implemented in the original system. This control design approach has been further elaborated in, for instance, [1], [2], [3], [8]. A general systematic characterization of the expansion-contraction process has been recently presented in [9]. In the following we summarize the background results of the inclusion principle.

Consider a pair of linear systems,

$$\mathbf{S}: \quad \begin{aligned} \dot{x} &= Ax + Bu, \\ y &= Cx \end{aligned} \tag{1.1}$$

*Received by the editors June 8, 1998; accepted for publication (in revised form) by L. Reichel August 12, 1999; published electronically March 21, 2000. The research of the second and third authors was supported by National Research and Development Plan of Spain through project HID96-1417.

http://www.siam.org/journals/simax/21-4/34054.html

†Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, 182 08 Prague 8, Czech Republic (bakule@utia.cas.cz). The research of this author was supported by the Academy of Sciences of the Czech Republic through grant A2075802.

‡Department of Applied Mathematics III, Technical University of Catalunya (UPC), Campus Nord, C-2, 08034 Barcelona, Spain (jose.rodellar@upc.es).

§Department of Applied Mathematics III, Technical University of Catalunya (UPC), 08240 Manresa, Spain (rossell@bages.eupm.upc.es).

and

$$\tilde{\mathbf{S}}: \quad \begin{aligned} \dot{\tilde{x}} &= \tilde{A}\tilde{x} + \tilde{B}\tilde{u}, \\ \tilde{y} &= \tilde{C}\tilde{x}, \end{aligned} \tag{1.2}$$

where $x(t) \in \mathbb{R}^{n}$, $u(t) \in \mathbb{R}^{m}$, $y(t) \in \mathbb{R}^{l}$ are the state, input, and output of $\mathbf{S}$ at time $t \in \mathbb{R}^{+}$, and $\tilde{x}(t) \in \mathbb{R}^{\tilde{n}}$, $\tilde{u}(t) \in \mathbb{R}^{\tilde{m}}$, $\tilde{y}(t) \in \mathbb{R}^{\tilde{l}}$ are those of $\tilde{\mathbf{S}}$. A, B, C and $\tilde{A}$, $\tilde{B}$, $\tilde{C}$ are constant matrices of dimensions $n\times n$, $n\times m$, $l\times n$ and $\tilde{n}\times\tilde{n}$, $\tilde{n}\times\tilde{m}$, $\tilde{l}\times\tilde{n}$, respectively. Suppose that the dimensions of the state, input, and output vectors x, u, y of $\mathbf{S}$ are smaller than (or at most equal to) those of $\tilde{x}$, $\tilde{u}$, $\tilde{y}$ of $\tilde{\mathbf{S}}$. Denote $x(t; x_0, u)$ and $y[x(t)]$ the state behavior and the corresponding output of $\mathbf{S}$ for a fixed input $u(t)$ and for an initial state $x(0) = x_0$, respectively. Similar notations $\tilde{x}(t; \tilde{x}_0, \tilde{u})$ and $\tilde{y}[\tilde{x}(t)]$ are used for the state behavior and output of system $\tilde{\mathbf{S}}$.

Let us consider the following transformations:

$$\begin{aligned} V &: \mathbb{R}^{n} \longrightarrow \mathbb{R}^{\tilde{n}}, & U &: \mathbb{R}^{\tilde{n}} \longrightarrow \mathbb{R}^{n}, \\ R &: \mathbb{R}^{m} \longrightarrow \mathbb{R}^{\tilde{m}}, & Q &: \mathbb{R}^{\tilde{m}} \longrightarrow \mathbb{R}^{m}, \\ T &: \mathbb{R}^{l} \longrightarrow \mathbb{R}^{\tilde{l}}, & S &: \mathbb{R}^{\tilde{l}} \longrightarrow \mathbb{R}^{l}, \end{aligned} \tag{1.3}$$

where $\operatorname{rank}(V) = n$, $\operatorname{rank}(R) = m$, $\operatorname{rank}(T) = l$ and

$$UV = I_n, \qquad QR = I_m, \qquad ST = I_l, \tag{1.4}$$

where I_n, I_m, I_l are identity matrices of indicated dimensions.

DEFINITION 1.1 (the inclusion principle). *We say that the system $\tilde{\mathbf{S}}$ includes the system $\mathbf{S}$, that is, $\mathbf{S}$ is included by $\tilde{\mathbf{S}}$, if there exists a quadruplet (U, V, R, S) such that, for any initial state x_0 and any fixed input $u(t)$ of $\mathbf{S}$, the choice*

$$\begin{aligned} \tilde{x}_0 &= Vx_0, \\ \tilde{u}(t) &= Ru(t) \ \text{ for all } t \geq 0 \end{aligned} \tag{1.5}$$

of the initial state $\tilde{x}_0$ and input $\tilde{u}(t)$ of the system $\tilde{\mathbf{S}}$ implies

$$\begin{aligned} x(t; x_0, u) &= U\tilde{x}(t; \tilde{x}_0, \tilde{u}), \\ y[x(t)] &= S\tilde{y}[\tilde{x}(t)] \ \text{ for all } t \geq 0. \end{aligned} \tag{1.6}$$

DEFINITION 1.2. *If the system $\tilde{\mathbf{S}}$ includes the system $\mathbf{S}$, then $\tilde{\mathbf{S}}$ is said to be an expansion of $\mathbf{S}$ and $\mathbf{S}$ is a contraction of $\tilde{\mathbf{S}}$.*

DEFINITION 1.3. *The control law $\tilde{u} = -\tilde{K}\,\tilde{x}$ for $\tilde{\mathbf{S}}$ is contractible to the control law $u = -K\,x$ for $\mathbf{S}$ if the choice $\tilde{x}_0 = Vx_0$ and $\tilde{u} = Ru$ implies*

$$Kx(t; x_0, u) = Q\tilde{K}\tilde{x}(t; \tilde{x}_0, \tilde{u}) \tag{1.7}$$

for all $t \geq 0$ for any initial state x_0 and any fixed input $u(t)$ of $\mathbf{S}$.

Definition 1.2 implies that the system $\tilde{\mathbf{S}}$ contains all the necessary information about the behavior of the system $\mathbf{S}$. We can extract properties of $\mathbf{S}$ from $\tilde{\mathbf{S}}$. Suppose that the pairs of matrices (U, V), (Q, R), and (S, T) are given; then the matrices $\tilde{A}$, $\tilde{B}$, $\tilde{C}$, and $\tilde{K}$ can be expressed as

$$\tilde{A} = VAU + M, \quad \tilde{B} = VBQ + N, \quad \tilde{C} = TCU + L, \quad \tilde{K} = RKU + F, \tag{1.8}$$

where M, N, L, and F are complementary matrices of appropriate dimensions. For $\tilde{\mathbf{S}}$ to be an expansion of $\mathbf{S}$, a proper choice of M, N, and L is required, which is provided by the following theorem [6].

THEOREM 1.4. *The system* $\tilde{\mathbf{S}}$ *is an expansion of the system* $\mathbf{S}$ *if and only if*

$$\begin{array}{ll} \text{(a)} & UM^iV = 0, \\ \text{(b)} & UM^{i-1}NR = 0, \\ \text{(c)} & SLM^{i-1}V = 0, \\ \text{(d)} & SLM^{i-1}NR = 0 \end{array} \tag{1.9}$$

for all $i = 1, 2, \ldots, \tilde{n}$.

DEFINITION 1.5. *We say that the system* $\mathbf{S}$ *is a restriction of* $\tilde{\mathbf{S}}$ *if there exists a triplet* (V, R, T) *such that, for any initial state* x_0 *and any fixed input* $u(t)$ *of* $\mathbf{S}$*, the choice*

$$\begin{aligned} \tilde{x}_0 &= Vx_0, \\ \tilde{u}(t) &= Ru(t) \ \textit{for all } t \geq 0 \end{aligned} \tag{1.10}$$

of the initial state $\tilde{x}_0$ *and input* $\tilde{u}(t)$ *of* $\tilde{\mathbf{S}}$ *implies*

$$\begin{aligned} \tilde{x}(t; \tilde{x}_0, \tilde{u}) &= Vx(t; x_0, u), \\ \tilde{y}[\tilde{x}(t)] &= Ty[x(t)] \ \textit{for all } t \geq 0. \end{aligned} \tag{1.11}$$

DEFINITION 1.6. *We say that the system* $\mathbf{S}$ *is an aggregation of* $\tilde{\mathbf{S}}$ *if there exists a triplet* (U, Q, S) *such that, for any initial state* $\tilde{x}_0$ *and any fixed input* $\tilde{u}(t)$ *of* $\tilde{\mathbf{S}}$*, the choice*

$$\begin{aligned} x_0 &= U\tilde{x}_0, \\ u(t) &= Q\tilde{u}(t) \ \textit{for all } t \geq 0 \end{aligned} \tag{1.12}$$

of the initial state x_0 *and input* $u(t)$ *of* $\mathbf{S}$ *implies*

$$\begin{aligned} x(t; x_0, u) &= U\tilde{x}(t; \tilde{x}_0, \tilde{u}), \\ y[x(t)] &= S\tilde{y}[\tilde{x}(t)] \ \textit{for all } t \geq 0. \end{aligned} \tag{1.13}$$

THEOREM 1.7. *The system* $\mathbf{S}$ *is a restriction of the system* $\tilde{\mathbf{S}}$ *if and only if*

$$\begin{array}{ll} \text{(a)} & MV = 0, \\ \text{(b)} & NR = 0, \\ \text{(c)} & LV = 0. \end{array} \tag{1.14}$$

THEOREM 1.8. *The system* $\mathbf{S}$ *is an aggregation of the system* $\tilde{\mathbf{S}}$ *if and only if*

$$\begin{array}{ll} \text{(a)} & UM = 0, \\ \text{(b)} & UN = 0, \\ \text{(c)} & SL = 0. \end{array} \tag{1.15}$$

THEOREM 1.9. *A control law* $\tilde{u} = -\tilde{K}\,\tilde{x}$ *for* $\tilde{\mathbf{S}}$ *is contractible to the control law* $u = -K\,x$ *for* $\mathbf{S}$ *if and only if*

$$\begin{array}{ll} \text{(a)} & QFM^{i-1}V = 0, \\ \text{(b)} & QFM^{i-1}NR = 0 \end{array} \tag{1.16}$$

for all $i = 1, 2, \ldots, \tilde{n}$.

2. Problem statement. Let us quote from [5]: "Our ability to use generalized decompositions depends crucially on the choice of the transformation matrix V and, subsequently, the generalized inverse U and the complementary matrix M. How to use this fact as a recipe in more general cases, is an objective of the future research."

This quotation is in fact extended to the transformations of the input and output spaces and the corresponding complementary matrices N and L in (1.8). The problem is that the necessary and sufficient conditions given by Theorem 1.4 are difficult to verify in practice, which is why, to the authors' knowledge, there is not a systematic procedure for choosing those matrices.

Therefore, the motivation of this work is to extend systematically the class of complementary matrices to obtain a more flexible tool primarily for the control design. The goal is formulated as follows:

(1) to characterize the structure of the matrices that appear in any expansion-contraction process;

(2) to determine the structure of matrices in restrictions and aggregations as two particular important cases of the expansion-contraction process;

(3) to derive the contractibility conditions, including their application on the decentralized control design.

3. Main results. The results included in this section cover mainly the expansion-contraction process at a general level. Subsection 3.1 includes the central result of this paper presented by Theorem 3.9. This theorem gives us a block structure of the complementary matrices as well as the conditions to be satisfied by blocks to guarantee the inclusion principle. The expansion forms in subsection 3.2 explore such conditions to identify possible choices of complementary matrices. Subsection 3.3 characterizes restrictions and aggregations as important particular cases of the expansion forms. Subsection 3.4 includes the contractibility conditions relating the control laws in the original and expanded spaces, while subsection 3.5 includes a direct application of Theorem 3.9 when considering decentralized control design.

3.1. Expansions and contractions. The expansion and contraction between systems $\mathbf{S}$ and $\tilde{\mathbf{S}}$ as considered in [3], [4], [5], [7], [6], [10] can be schematically illustrated in the form

$$\begin{array}{ccccc} \mathbf{S} & \longrightarrow & \tilde{\mathbf{S}} & \longrightarrow & \mathbf{S}, \\ \mathbb{R}^{\mathbf{n}} & \xrightarrow{V} & \mathbb{R}^{\tilde{\mathbf{n}}} & \xrightarrow{U} & \mathbb{R}^{\mathbf{n}}, \\ \mathbb{R}^{\mathbf{m}} & \xrightarrow{R} & \mathbb{R}^{\tilde{\mathbf{m}}} & \xrightarrow{Q} & \mathbb{R}^{\mathbf{m}}, \\ \mathbb{R}^{\mathbf{l}} & \xrightarrow{T} & \mathbb{R}^{\tilde{\mathbf{l}}} & \xrightarrow{S} & \mathbb{R}^{\mathbf{l}}. \end{array} \tag{3.1}$$

Since the inclusion principle summarized in the introduction does not depend on the specific basis used in the state, input, and output spaces for both systems $\mathbf{S}$ and $\tilde{\mathbf{S}}$, we may introduce convenient changes of basis in the above scheme [7]. Having this idea in mind, we consider changes of basis in the spaces that correspond with the expanded system $\tilde{\mathbf{S}}$.

These changes allow us to obtain conditions for the complementary matrices more flexible than those stated in Theorem 1.4 to satisfy the inclusion principle.

In order to simplify the notation, consider the system $\mathbf{S}$ having the following

structure:

$$
\begin{aligned}
\begin{pmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{pmatrix} &= \begin{matrix}n_1\\ n_2\\ n_3\end{matrix}\overset{\begin{matrix}n_1 & n_2 & n_3\end{matrix}}{\begin{pmatrix}A_{11} & A_{12} & A_{13}\\ A_{21} & A_{22} & A_{23}\\ A_{31} & A_{32} & A_{33}\end{pmatrix}}\begin{pmatrix}x_1\\ x_2\\ x_3\end{pmatrix} + \begin{matrix}n_1\\ n_2\\ n_3\end{matrix}\overset{\begin{matrix}m_1 & m_2 & m_3\end{matrix}}{\begin{pmatrix}B_{11} & B_{12} & B_{13}\\ B_{21} & B_{22} & B_{23}\\ B_{31} & B_{32} & B_{33}\end{pmatrix}}\begin{pmatrix}u_1\\ u_2\\ u_3\end{pmatrix},\\
\begin{pmatrix}y_1\\ y_2\\ y_3\end{pmatrix} &= \begin{matrix}l_1\\ l_2\\ l_3\end{matrix}\overset{\begin{matrix}n_1 & n_2 & n_3\end{matrix}}{\begin{pmatrix}C_{11} & C_{12} & C_{13}\\ C_{21} & C_{22} & C_{23}\\ C_{31} & C_{32} & C_{33}\end{pmatrix}}\begin{pmatrix}x_1\\ x_2\\ x_3\end{pmatrix},
\end{aligned}
\tag{3.2}
$$

where n_i, m_i, and l_i indicate the dimensions of the corresponding matrices with $n_1+n_2+n_3=n$, $m_1+m_2+m_3=m$, $l_1+l_2+l_3=l$, $n_1+2n_2+n_3=\tilde{n}$, $m_1+2m_2+m_3=\tilde{m}$, $l_1+2l_2+l_3=\tilde{l}$. This system is composed of two subsystems with an overlapped part, but it can be generalized for any number of interconnected overlapping subsystems. This structure has been extensively adopted as a prototype in the literature.

Let us consider the change of basis for the expanded state space in $\mathbb{R}^{\tilde{\mathbf{n}}}$ defined by the following matrices:

$$
T_A = \begin{pmatrix} I_{n_1} & 0 & 0 & 0\\ 0 & I_{n_2} & 0 & I_{n_2}\\ 0 & I_{n_2} & 0 & -I_{n_2}\\ 0 & 0 & I_{n_3} & 0\end{pmatrix}, \qquad T_A^{-1} = \begin{pmatrix} I_{n_1} & 0 & 0 & 0\\ 0 & \frac{1}{2}I_{n_2} & \frac{1}{2}I_{n_2} & 0\\ 0 & 0 & 0 & I_{n_3}\\ 0 & \frac{1}{2}I_{n_2} & -\frac{1}{2}I_{n_2} & 0\end{pmatrix}.
\tag{3.3}
$$

I_{n_i} denote the identity matrices of orders n_i, $i=1,2,3$. Analogously, the matrices T_B, T_B^{-1}, and T_C, T_C^{-1} correspond with the input in $\mathbb{R}^{\tilde{\mathbf{m}}}$ and the output in $\mathbb{R}^{\tilde{\mathbf{l}}}$, respectively.

The expansion-contraction scheme can now be modified in the form

$$
\begin{array}{ccccccccc}
\mathbf{S} & \longrightarrow & \tilde{\mathbf{S}} & \longrightarrow & \bar{\tilde{\mathbf{S}}} & \longrightarrow & \tilde{\mathbf{S}} & \longrightarrow & \mathbf{S},\\
\mathbb{R}^{\mathbf{n}} & \xrightarrow{V} & \mathbb{R}^{\tilde{\mathbf{n}}} & \xrightarrow{T_A^{-1}} & \bar{\mathbb{R}}^{\tilde{\mathbf{n}}} & \xrightarrow{T_A} & \mathbb{R}^{\tilde{\mathbf{n}}} & \xrightarrow{U} & \mathbb{R}^{\mathbf{n}},\\
\mathbb{R}^{\mathbf{m}} & \xrightarrow{R} & \mathbb{R}^{\tilde{\mathbf{m}}} & \xrightarrow{T_B^{-1}} & \bar{\mathbb{R}}^{\tilde{\mathbf{m}}} & \xrightarrow{T_B} & \mathbb{R}^{\tilde{\mathbf{m}}} & \xrightarrow{Q} & \mathbb{R}^{\mathbf{m}},\\
\mathbb{R}^{\mathbf{l}} & \xrightarrow{T} & \mathbb{R}^{\tilde{\mathbf{l}}} & \xrightarrow{T_C^{-1}} & \bar{\mathbb{R}}^{\tilde{\mathbf{l}}} & \xrightarrow{T_C} & \mathbb{R}^{\tilde{\mathbf{l}}} & \xrightarrow{S} & \mathbb{R}^{\mathbf{l}}.
\end{array}
\tag{3.4}
$$

$\bar{\tilde{\mathbf{S}}}$ denotes the expanded system within the new basis. We assume that the matrices V, R, and T are

$$
V = \begin{pmatrix} I_{n_1} & 0 & 0\\ 0 & I_{n_2} & 0\\ 0 & I_{n_2} & 0\\ 0 & 0 & I_{n_3}\end{pmatrix}, \quad R = \begin{pmatrix} I_{m_1} & 0 & 0\\ 0 & I_{m_2} & 0\\ 0 & I_{m_2} & 0\\ 0 & 0 & I_{m_3}\end{pmatrix}, \quad T = \begin{pmatrix} I_{l_1} & 0 & 0\\ 0 & I_{l_2} & 0\\ 0 & I_{l_2} & 0\\ 0 & 0 & I_{l_3}\end{pmatrix}.
\tag{3.5}
$$

From these matrices, we define

$$
U = (V^tV)^{-1}V^t, \qquad Q = (R^tR)^{-1}R^t, \qquad S = (T^tT)^{-1}T^t
\tag{3.6}
$$

as the pseudoinverses of V, R, and T, respectively.

Note. Considering the relations

$$
\begin{array}{lll}
\bar{V} = T_A^{-1}V, & \bar{R} = T_B^{-1}R, & \bar{T} = T_C^{-1}T,\\
\bar{U} = UT_A, & \bar{Q} = QT_B, & \bar{S} = ST_C,
\end{array}
\tag{3.7}
$$

it is easy to verify that $\bar{U}\bar{V} = I_n$ and $\bar{V}\bar{U} = \left(\begin{smallmatrix} I_n & 0 \\ 0 & 0 \end{smallmatrix}\right)$, where each submatrix (0) has appropriate dimensions so that the resulting matrix has an order $\tilde{n}$. Similarly, $\bar{Q}\bar{R} = I_m$, $\bar{R}\bar{Q} = \left(\begin{smallmatrix} I_m & 0 \\ 0 & 0 \end{smallmatrix}\right)$, and $\bar{S}\bar{T} = I_l$, $\bar{T}\bar{S} = \left(\begin{smallmatrix} I_l & 0 \\ 0 & 0 \end{smallmatrix}\right)$.

Notation. Denote the matrices $\tilde{A}$, $\tilde{B}$, $\tilde{C}$, $\tilde{K}$ in the expanded system $\tilde{\mathbf{S}}$. Analogously, $\tilde{\tilde{A}}$, $\tilde{\tilde{B}}$, $\tilde{\tilde{C}}$, $\tilde{\tilde{K}}$ are matrices in $\tilde{\tilde{\mathbf{S}}}$.

Let us pay attention to the pair of linear time invariant systems

$$\begin{aligned} \mathbf{S}: \ \dot{x} &= Ax + Bu, \\ y &= Cx, \end{aligned} \tag{3.8}$$

$$\begin{aligned} \tilde{\tilde{\mathbf{S}}}: \ \dot{\tilde{\tilde{x}}} &= \tilde{\tilde{A}}\tilde{\tilde{x}} + \tilde{\tilde{B}}\tilde{\tilde{u}}, \\ \tilde{\tilde{y}} &= \tilde{\tilde{C}}\tilde{\tilde{x}}, \end{aligned} \tag{3.9}$$

where $\tilde{\tilde{A}}$, $\tilde{\tilde{B}}$, $\tilde{\tilde{C}}$ are constant matrices of appropriate dimensions. The vectors $\tilde{\tilde{x}}$, $\tilde{\tilde{u}}$, $\tilde{\tilde{y}}$ are defined as

$$\tilde{\tilde{x}} = T_{A}^{-1}Vx = \bar{V}x, \qquad \tilde{\tilde{u}} = T_{B}^{-1}Ru = \bar{R}u, \qquad \tilde{\tilde{y}} = T_{C}^{-1}Ty = \bar{T}y. \tag{3.10}$$

Now, denote the relations (1.8) for the open-loop system as

$$\tilde{\tilde{A}} = \bar{V}A\bar{U} + \bar{M}, \qquad \tilde{\tilde{B}} = \bar{V}B\bar{Q} + \bar{N}, \qquad \tilde{\tilde{C}} = \bar{T}C\bar{U} + \bar{L}, \tag{3.11}$$

where the new complementary matrices are

$$\bar{M} = T_{A}^{-1}MT_{A}, \qquad \bar{N} = T_{A}^{-1}NT_{B}, \qquad \bar{L} = T_{C}^{-1}LT_{A}. \tag{3.12}$$

Using these matrices, the conditions given by (1.9) for the inclusion principle are as follows:

$$\begin{array}{ll} \text{(a)} & \bar{U}\bar{M}^{i}\bar{V} = 0, \\ \text{(b)} & \bar{U}\bar{M}^{i-1}\bar{N}\bar{R} = 0, \\ \text{(c)} & \bar{S}\bar{L}\bar{M}^{i-1}\bar{V} = 0, \\ \text{(d)} & \bar{S}\bar{L}\bar{M}^{i-1}\bar{N}\bar{R} = 0 \end{array} \tag{3.13}$$

for all $i = 1, 2, \ldots, \tilde{n}$.

Conditions imposed by the new basis when considering the relations (3.13) are analyzed later. First, we analyze the form of matrices $\bar{M}$, $\bar{N}$, and $\bar{L}$ in the expanded spaces.

Notation. Consider that the complementary matrices have the form

$$M = \left(M_{ij}\right), \quad N = \left(N_{ij}\right), \quad L = \left(L_{ij}\right), \quad i, j = 1, 2, 3, 4, \tag{3.14}$$

where each submatrix has appropriate dimensions. It is convenient to deal with matrix blocks when using the matrices T_{A}, T_{B}, T_{C}, T_{A}^{-1}, T_{B}^{-1}, and T_{C}^{-1}. Therefore, it is useful to consider the matrices

$$\bar{M} = \begin{pmatrix} \bar{M}_{11} & \bar{M}_{12} \\ \bar{M}_{21} & \bar{M}_{22} \end{pmatrix}, \quad \bar{N} = \begin{pmatrix} \bar{N}_{11} & \bar{N}_{12} \\ \bar{N}_{21} & \bar{N}_{22} \end{pmatrix}, \quad \bar{L} = \begin{pmatrix} \bar{L}_{11} & \bar{L}_{12} \\ \bar{L}_{21} & \bar{L}_{22} \end{pmatrix}, \tag{3.15}$$

where $\bar{M}_{11}$, $\bar{M}_{12}$, $\bar{M}_{21}$, $\bar{M}_{22}$ are $n \times n$, $n \times (\tilde{n} - n)$, $(\tilde{n} - n) \times n$, $(\tilde{n} - n) \times (\tilde{n} - n)$ dimensional matrices, respectively. $\bar{N}_{11}$, $\bar{N}_{12}$, $\bar{N}_{21}$, $\bar{N}_{22}$ are $n \times m$, $n \times (\tilde{m} - m)$,

$(\tilde{n}-n)\times m$, $(\tilde{n}-n)\times(\tilde{m}-m)$ dimensional matrices, respectively. $\bar{L}_{11}$, $\bar{L}_{12}$, $\bar{L}_{21}$, $\bar{L}_{22}$ are $l\times n$, $l\times(\tilde{n}-n)$, $(\tilde{l}-l)\times n$, $(\tilde{l}-l)\times(\tilde{n}-n)$ dimensional matrices, respectively. We need to know the form of the submatrices $\bar{M}_{ij}$, $\bar{N}_{ij}$, and $\bar{L}_{ij}$ for all $i,j=1,2$. This is given by the following proposition.

PROPOSITION 3.1. *Consider the matrices* $\bar{M}=\left(\begin{smallmatrix}\bar{M}_{11} & \bar{M}_{12}\\ \bar{M}_{21} & \bar{M}_{22}\end{smallmatrix}\right)$, $\bar{N}=\left(\begin{smallmatrix}\bar{N}_{11} & \bar{N}_{12}\\ \bar{N}_{21} & \bar{N}_{22}\end{smallmatrix}\right)$, *and* $\bar{L}=\left(\begin{smallmatrix}\bar{L}_{11} & \bar{L}_{12}\\ \bar{L}_{21} & \bar{L}_{22}\end{smallmatrix}\right)$ *given in* (3.15)*; then the submatrices* $\bar{M}_{ij}$, $\bar{N}_{ij}$, *and* $\bar{L}_{ij}$ *for* $i,j=1,2$ *have the following form:*

$$
\begin{aligned}
\bar{M}_{11} &= \begin{pmatrix} M_{11} & M_{12}+M_{13} & M_{14}\\ \frac{1}{2}(M_{21}+M_{31}) & \frac{1}{2}(M_{22}+M_{23}+M_{32}+M_{33}) & \frac{1}{2}(M_{24}+M_{34})\\ M_{41} & M_{42}+M_{43} & M_{44}\end{pmatrix},\\
\bar{M}_{12} &= \begin{pmatrix} M_{12}-M_{13}\\ \frac{1}{2}(M_{22}-M_{23}+M_{32}-M_{33})\\ M_{42}-M_{43}\end{pmatrix},\\
\bar{M}_{21} &= \tfrac{1}{2}\begin{pmatrix} M_{21}-M_{31} & M_{22}+M_{23}-M_{32}-M_{33} & M_{24}-M_{34}\end{pmatrix},\\
\bar{M}_{22} &= \tfrac{1}{2}\begin{pmatrix} M_{22}-M_{23}-M_{32}+M_{33}\end{pmatrix}.
\end{aligned}
\tag{3.16}
$$

The submatrices in $\bar{N}$ *and* $\bar{L}$ *have similar structure to those in* $\bar{M}$ *when substituting the submatrices* M_{ij} *by* N_{ij} *and* L_{ij}*, respectively, for all* $i,j=1,2,3,4$.

Proof. To prove (3.16) we use the relation $\bar{M}=T_{A}^{-1}MT_{A}$ given in (3.12), with T_{A} and T_{A}^{-1} given in (3.3) and M in (3.14). Then, identifying the blocks of the matrix resulting from the above product with the matrix $\bar{M}$ defined in (3.15) as $\bar{M}=\left(\begin{smallmatrix}\bar{M}_{11} & \bar{M}_{12}\\ \bar{M}_{21} & \bar{M}_{22}\end{smallmatrix}\right)$, (3.16) is obtained. A similar process follows for $\bar{N}$ and $\bar{L}$. ∎

We are now ready to establish conditions which must be satisfied by the blocks $\bar{M}_{ij}$, $\bar{N}_{ij}$, and $\bar{L}_{ij}$ to verify (3.13).

PROPOSITION 3.2. *Consider the system given by* (3.8) *and the corresponding expanded system* (3.9)*, verifying* (3.10)–(3.13)*; then* $\bar{M}=\left(\begin{smallmatrix}0 & \bar{M}_{12}\\ \bar{M}_{21} & \bar{M}_{22}\end{smallmatrix}\right)$. (0) *denotes a matrix of order* n *in* $\bar{M}$ *and the other blocks satisfy* $\bar{M}_{12}\bar{M}_{22}^{i-2}\bar{M}_{21}=0$ *for all* $i=2,\dots,\tilde{n}$.

Proof. Let $\bar{M}$, $\bar{U}$, $\bar{V}$ be the matrices $\bar{M}=\left(\begin{smallmatrix}\bar{M}_{11} & \bar{M}_{12}\\ \bar{M}_{21} & \bar{M}_{22}\end{smallmatrix}\right)$, $\bar{U}=\begin{pmatrix} I_n & 0\end{pmatrix}$, $\bar{V}=\begin{pmatrix} I_n\\ 0\end{pmatrix}$, where (0) denotes matrices of appropriate dimensions. Imposing the first condition of (3.13) for $i=1$, we get $\bar{U}\bar{M}\bar{V}=0$ so that $\bar{M}_{11}=0$. The power of the matrix $\bar{M}$ for $\mu=2,\dots,\tilde{n}$ results in a general expression

$$
\bar{M}^{\mu}=\begin{pmatrix} \bar{M}_{12}\bar{M}_{22}^{\mu-2}\bar{M}_{21} & \bar{M}_{12}\bar{M}_{22}^{\mu-1}\\ \bar{M}_{22}^{\mu-1}\bar{M}_{21} & \bar{M}_{22}^{\mu}+\sum_{j=0}^{j=\mu-2}\bar{M}_{22}^{j}\bar{M}_{21}\bar{M}_{12}\bar{M}_{22}^{\mu-2-j}\end{pmatrix}.
\tag{3.17}
$$

Then, we get $\bar{U}\bar{M}^{k}\bar{V}=0$ for $i=k\geq 2$. It implies that $\bar{M}_{12}\bar{M}_{22}^{k-2}\bar{M}_{21}=0$. Repeating this process, we get $\bar{U}\bar{M}^{\tilde{n}}\bar{V}=0$ for $i=\tilde{n}$. It leads to $\bar{M}_{12}\bar{M}_{22}^{\tilde{n}-2}\bar{M}_{21}=0$. ∎

COROLLARY 3.3. *Suppose that Proposition* 3.2 *holds; then* $\tilde{\bar{A}}=\left(\begin{smallmatrix}A & \bar{M}_{12}\\ \bar{M}_{21} & \bar{M}_{22}\end{smallmatrix}\right)$.

Proof. From (3.11), $\tilde{\bar{A}}=\bar{V}A\bar{U}+\bar{M}$ and, carrying out these operations, we obtain $\tilde{\bar{A}}=\left(\begin{smallmatrix}A & \bar{M}_{12}\\ \bar{M}_{21} & \bar{M}_{22}\end{smallmatrix}\right)$. ∎

The following propositions have similar proofs.

PROPOSITION 3.4. *Consider the system* (3.8) *and the corresponding expanded*

system (3.9) *verifying* (3.10)–(3.13)*; then* $\bar{N} = \begin{pmatrix} 0 & \bar{N}_{12} \\ \bar{N}_{21} & \bar{N}_{22} \end{pmatrix}$. (0) *denotes an* $n \times m$ *dimensional matrix in* $\bar{N}$ *and the other blocks satisfy* $\bar{M}_{12}\bar{M}_{22}^{i-2}\bar{N}_{21} = 0$ *for all* $i = 2, \ldots, \tilde{n}$.

COROLLARY 3.5. *Suppose that Proposition* 3.4 *holds; then* $\tilde{\bar{B}} = \begin{pmatrix} B & \bar{N}_{12} \\ \bar{N}_{21} & \bar{N}_{22} \end{pmatrix}$.

Proof. By definition, $\tilde{\bar{B}} = \bar{V}B\bar{Q} + \bar{N}$. Considering $\bar{Q} = \begin{pmatrix} I_m & 0 \end{pmatrix}$, where (0) is an $m \times (\tilde{m} - m)$ dimensional matrix and carrying out these operations, we obtain $\tilde{\bar{B}} = \begin{pmatrix} B & \bar{N}_{12} \\ \bar{N}_{21} & \bar{N}_{22} \end{pmatrix}$. ∎

PROPOSITION 3.6. *Consider the system* (3.8) *and the corresponding expanded system* (3.9) *verifying* (3.10)–(3.13)*; then* $\bar{L} = \begin{pmatrix} 0 & \bar{L}_{12} \\ \bar{L}_{21} & \bar{L}_{22} \end{pmatrix}$. (0) *denotes an* $l \times n$ *dimensional matrix in* $\bar{L}$ *and the other blocks satisfy* $\bar{L}_{12}\bar{M}_{22}^{i-2}\bar{M}_{21} = 0$ *for all* $i = 2, \ldots, \tilde{n}$.

COROLLARY 3.7. *Suppose that Proposition* 3.6 *holds; then* $\tilde{\bar{C}} = \begin{pmatrix} C & \bar{L}_{12} \\ \bar{L}_{21} & \bar{L}_{22} \end{pmatrix}$.

Proof. By definition, $\tilde{\bar{C}} = \bar{T}C\bar{U} + \bar{L}$. Consider $\bar{T} = \begin{pmatrix} I_l \\ 0 \end{pmatrix}$, where (0) is an $(\tilde{l} - l) \times l$ dimensional matrix. Carrying out these operations, we obtain the matrix $\tilde{\bar{C}} = \begin{pmatrix} C & \bar{L}_{12} \\ \bar{L}_{21} & \bar{L}_{22} \end{pmatrix}$. ∎

PROPOSITION 3.8. *Consider the system* (3.8) *and the corresponding expanded system* (3.9) *verifying* (3.10)–(3.13). *Let* $\bar{M}$, $\bar{N}$, $\bar{L}$ *be the matrices* $\bar{M} = \begin{pmatrix} 0 & \bar{M}_{12} \\ \bar{M}_{21} & \bar{M}_{22} \end{pmatrix}$, $\bar{N} = \begin{pmatrix} 0 & \bar{N}_{12} \\ \bar{N}_{21} & \bar{N}_{22} \end{pmatrix}$, $\bar{L} = \begin{pmatrix} 0 & \bar{L}_{12} \\ \bar{L}_{21} & \bar{L}_{22} \end{pmatrix}$, *where* (0) *denotes matrices of appropriate dimensions. The matrices* $\bar{M}$, $\bar{N}$, *and* $\bar{L}$ *verify Propositions* 3.2, 3.4, *and* 3.6*; then* $\bar{L}_{12}\bar{M}_{22}^{i-2}\bar{N}_{21} = 0$ *for all* $i = 2, \ldots, \tilde{n} + 1$.

Note. The matrix conditions imposed by the inclusion principle in (3.13) have been reduced to conditions on submatrices in the form

$$\left.\begin{aligned} \bar{M}_{12}\bar{M}_{22}^{i-2}\bar{M}_{21} &= 0 \\ \bar{M}_{12}\bar{M}_{22}^{i-2}\bar{N}_{21} &= 0 \\ \bar{L}_{12}\bar{M}_{22}^{i-2}\bar{M}_{21} &= 0 \end{aligned}\right\} \text{ for all } i = 2, \ldots, \tilde{n}, \tag{3.18}$$

$$\bar{L}_{12}\bar{M}_{22}^{i-2}\bar{N}_{21} = 0 \quad \text{for all } i = 2, \ldots, \tilde{n} + 1,$$

where $\bar{M} = \begin{pmatrix} 0 & \bar{M}_{12} \\ \bar{M}_{21} & \bar{M}_{22} \end{pmatrix}$, $\bar{N} = \begin{pmatrix} 0 & \bar{N}_{12} \\ \bar{N}_{21} & \bar{N}_{22} \end{pmatrix}$, and $\bar{L} = \begin{pmatrix} 0 & \bar{L}_{12} \\ \bar{L}_{21} & \bar{L}_{22} \end{pmatrix}$. These matrices are expressed in the new basis. The following theorem expresses the complementary matrices M, N, and L in the initial basis.

THEOREM 3.9. *Consider the system* (1.1) *and the corresponding expanded system* (1.2) *verifying* (1.9)*; then the complementary matrices* M, N, *and* L *of* (3.14) *have the following form:*

$$M = \begin{pmatrix} 0 & M_{12} & -M_{12} & 0 \\ M_{21} & M_{22} & M_{23} & M_{24} \\ -M_{21} & -(M_{22}+M_{23}+M_{33}) & M_{33} & -M_{24} \\ 0 & M_{42} & -M_{42} & 0 \end{pmatrix}, \tag{3.19}$$

$$N = \begin{pmatrix} 0 & N_{12} & -N_{12} & 0 \\ N_{21} & N_{22} & N_{23} & N_{24} \\ -N_{21} & -(N_{22}+N_{23}+N_{33}) & N_{33} & -N_{24} \\ 0 & N_{42} & -N_{42} & 0 \end{pmatrix}, \tag{3.20}$$

$$L = \begin{pmatrix} 0 & L_{12} & -L_{12} & 0 \\ L_{21} & L_{22} & L_{23} & L_{24} \\ -L_{21} & -(L_{22}+L_{23}+L_{33}) & L_{33} & -L_{24} \\ 0 & L_{42} & -L_{42} & 0 \end{pmatrix}. \tag{3.21}$$

The following conditions hold:

(3.22)
$$\left.\begin{array}{l}\begin{pmatrix} M_{12} \\ M_{23}+M_{33} \\ M_{42} \end{pmatrix} \left(M_{22}+M_{33} \right)^{i-2} \begin{pmatrix} M_{21} & M_{22}+M_{23} & M_{24} \end{pmatrix} = 0 \\ \begin{pmatrix} M_{12} \\ M_{23}+M_{33} \\ M_{42} \end{pmatrix} \left(M_{22}+M_{33} \right)^{i-2} \begin{pmatrix} N_{21} & N_{22}+N_{23} & N_{24} \end{pmatrix} = 0 \\ \begin{pmatrix} L_{12} \\ L_{23}+L_{33} \\ L_{42} \end{pmatrix} \left(M_{22}+M_{33} \right)^{i-2} \begin{pmatrix} M_{21} & M_{22}+M_{23} & M_{24} \end{pmatrix} = 0 \end{array}\right\} \text{for all } i = 2, \ldots, \tilde{n},$$

$$\begin{pmatrix} L_{12} \\ L_{23}+L_{33} \\ L_{42} \end{pmatrix} \left(M_{22}+M_{33} \right)^{i-2} \begin{pmatrix} N_{21} & N_{22}+N_{23} & N_{24} \end{pmatrix} = 0 \text{ for all } i = 2, \ldots, \tilde{n}+1.$$

Proof. Since $\bar{M}_{11} = 0$, the following relations are obtained from (3.16):

$$\begin{array}{lll} M_{11} = 0, & M_{13} = -M_{12}, & M_{14} = 0, \\ M_{31} = -M_{21}, & M_{32} = -(M_{22} + M_{23} + M_{33}), & M_{34} = -M_{24}, \\ M_{41} = 0, & M_{43} = -M_{42}, & M_{44} = 0. \end{array} \tag{3.23}$$

Substituting (3.23) into the general matrix M given by (3.14) leads to

$$M = \begin{pmatrix} 0 & M_{12} & -M_{12} & 0 \\ M_{21} & M_{22} & M_{23} & M_{24} \\ -M_{21} & -(M_{22}+M_{23}+M_{33}) & M_{33} & -M_{24} \\ 0 & M_{42} & -M_{42} & 0 \end{pmatrix}, \tag{3.24}$$

which proves (3.19). Equations (3.20) and (3.21) have a similar proof.

By using (3.23) in (3.16), we obtain

$$\bar{M}_{12} = \begin{pmatrix} 2M_{12} \\ -M_{23}-M_{33} \\ 2M_{42} \end{pmatrix}, \ \bar{M}_{21} = \begin{pmatrix} M_{21} & M_{22}+M_{23} & M_{24} \end{pmatrix}, \ \bar{M}_{22} = \begin{pmatrix} M_{22}+M_{33} \end{pmatrix}. \tag{3.25}$$

Similar expressions are obtained for submatrices $\bar{N}_{ij}$ and $\bar{L}_{ij}$ for all $i, j = 1, 2$. Substituting all of these expressions into (3.18) results in (3.22). ∎

COROLLARY 3.10. *Suppose that Theorem* 3.9 *holds; then the matrix* $\tilde{A}$ *is*

$$\tilde{A} = \begin{pmatrix} A_{11} & \frac{1}{2}A_{12}+M_{12} & \frac{1}{2}A_{12}-M_{12} & A_{13} \\ A_{21}+M_{21} & \frac{1}{2}A_{22}+M_{22} & \frac{1}{2}A_{22}+M_{23} & A_{23}+M_{24} \\ A_{21}-M_{21} & \frac{1}{2}A_{22}-(M_{22}+M_{23}+M_{33}) & \frac{1}{2}A_{22}+M_{33} & A_{23}-M_{24} \\ A_{31} & \frac{1}{2}A_{32}+M_{42} & \frac{1}{2}A_{32}-M_{42} & A_{33} \end{pmatrix}. \tag{3.26}$$

Proof. By definition, $\tilde{A} = VAU + M$ and by using the matrix A given in (3.2) and the matrix M of (3.19), we prove the corollary. ∎

COROLLARY 3.11. *Suppose that Theorem* 3.9 *holds; then the matrix* $\tilde{B}$ *is*

$$\tilde{B} = \begin{pmatrix} B_{11} & \frac{1}{2}B_{12}+N_{12} & \frac{1}{2}B_{12}-N_{12} & B_{13} \\ B_{21}+N_{21} & \frac{1}{2}B_{22}+N_{22} & \frac{1}{2}B_{22}+N_{23} & B_{23}+N_{24} \\ B_{21}-N_{21} & \frac{1}{2}B_{22}-(N_{22}+N_{23}+N_{33}) & \frac{1}{2}B_{22}+N_{33} & B_{23}-N_{24} \\ B_{31} & \frac{1}{2}B_{32}+N_{42} & \frac{1}{2}B_{32}-N_{42} & B_{33} \end{pmatrix}. \tag{3.27}$$

Proof. Since $\tilde{B} = VBQ + N$, by using the matrix N given in (3.20) and the matrix B defined in (3.2), we prove the corollary. ∎

COROLLARY 3.12. *Suppose that Theorem* 3.9 *holds; then the matrix* $\tilde{C}$ *is*

$$\tilde{C} = \begin{pmatrix} C_{11} & \frac{1}{2}C_{12}+L_{12} & \frac{1}{2}C_{12}-L_{12} & C_{13} \\ C_{21}+L_{21} & \frac{1}{2}C_{22}+L_{22} & \frac{1}{2}C_{22}+L_{23} & C_{23}+L_{24} \\ C_{21}-L_{21} & \frac{1}{2}C_{22}-(L_{22}+L_{23}+L_{33}) & \frac{1}{2}C_{22}+L_{33} & C_{23}-L_{24} \\ C_{31} & \frac{1}{2}C_{32}+L_{42} & \frac{1}{2}C_{32}-L_{42} & C_{33} \end{pmatrix}. \tag{3.28}$$

Proof. From $\tilde{C} = TCU + L$ and by using L as in (3.21) and C given in (3.2), we prove this corollary. □

3.2. Expansion forms. Theorem 3.9 gives us a block structure for the complementary matrices as well as the conditions to be satisfied by the blocks to guarantee the inclusion principle. In this section, we explore such conditions to identify possible choices of complementary matrices.

By using (3.22) for $i = 2$, we obtain

$$\begin{aligned} &\begin{pmatrix} M_{12} \\ M_{23}+M_{33} \\ M_{42} \end{pmatrix} \begin{pmatrix} M_{21} & M_{22}+M_{23} & M_{24} \end{pmatrix} = 0, \\ &\begin{pmatrix} M_{12} \\ M_{23}+M_{33} \\ M_{42} \end{pmatrix} \begin{pmatrix} N_{21} & N_{22}+N_{23} & N_{24} \end{pmatrix} = 0, \\ &\begin{pmatrix} L_{12} \\ L_{23}+L_{33} \\ L_{42} \end{pmatrix} \begin{pmatrix} M_{21} & M_{22}+M_{23} & M_{24} \end{pmatrix} = 0, \\ &\begin{pmatrix} L_{12} \\ L_{23}+L_{33} \\ L_{42} \end{pmatrix} \begin{pmatrix} N_{21} & N_{22}+N_{23} & N_{24} \end{pmatrix} = 0. \end{aligned} \tag{3.29}$$

We may identify two broad cases from (3.29), that is, when the matrix rows are zeros or when the matrix columns are zeros, respectively. The first case results in the following complementary matrices:

$$\begin{aligned} &M_{12} = 0, \quad M_{23} + M_{33} = 0, \quad M_{42} = 0, \\ &L_{12} = 0, \quad L_{23} + L_{33} = 0, \quad L_{42} = 0. \end{aligned} \tag{3.30}$$

The second case results in the following complementary matrices:

$$\begin{aligned} &M_{21} = 0, \quad M_{22} + M_{23} = 0, \quad M_{24} = 0, \\ &N_{21} = 0, \quad N_{22} + N_{23} = 0, \quad N_{24} = 0. \end{aligned} \tag{3.31}$$

3.3. Restrictions and aggregations. The two above-mentioned cases include restrictions and aggregations as important particular cases. Therefore, we characterize the structure of complementary matrices of these two cases in this subsection.

PROPOSITION 3.13. *The system* $\mathbf{S}$ *is a restriction of the system* $\tilde{\tilde{\mathbf{S}}}$ *if and only if*

$$\begin{aligned} &\text{(a)} \quad \bar{M} = \begin{pmatrix} 0 & \bar{M}_{12} \\ 0 & \bar{M}_{22} \end{pmatrix}, \\ &\text{(b)} \quad \bar{N} = \begin{pmatrix} 0 & \bar{N}_{12} \\ 0 & \bar{N}_{22} \end{pmatrix}, \\ &\text{(c)} \quad \bar{L} = \begin{pmatrix} 0 & \bar{L}_{12} \\ 0 & \bar{L}_{22} \end{pmatrix}. \end{aligned} \tag{3.32}$$

Proof. By using Theorem 1.7, we have

$$\begin{array}{ll} \text{(a)} & \bar{M}\bar{V} = 0, \\ \text{(b)} & \bar{N}\bar{R} = 0, \\ \text{(c)} & \bar{L}\bar{V} = 0. \end{array} \tag{3.33}$$

Taking into account that $\bar{M}$ is given by (3.15) and that $\bar{V}$ has the form $\bar{V} = \left(\begin{smallmatrix} I_n \\ 0 \end{smallmatrix}\right)$, from $\bar{M}\bar{V} = 0$ it results in $\bar{M}_{11} = 0$ and $\bar{M}_{21} = 0$ so that $\bar{M} = \left(\begin{smallmatrix} 0 & \bar{M}_{12} \\ 0 & \bar{M}_{22} \end{smallmatrix}\right)$. The proof of cases (b) and (c) is similar. ∎

COROLLARY 3.14. *Let* $\mathbf{S}$ *be a restriction of the system* $\tilde{\tilde{\mathbf{S}}}$*; then the matrices* $\tilde{\tilde{A}}$*,* $\tilde{\tilde{B}}$*, and* $\tilde{\tilde{C}}$ *have the following form:*

$$\tilde{\tilde{A}} = \begin{pmatrix} A & \bar{M}_{12} \\ 0 & \bar{M}_{22} \end{pmatrix}, \quad \tilde{\tilde{B}} = \begin{pmatrix} B & \bar{N}_{12} \\ 0 & \bar{N}_{22} \end{pmatrix}, \quad \tilde{\tilde{C}} = \begin{pmatrix} C & \bar{L}_{12} \\ 0 & \bar{L}_{22} \end{pmatrix}. \tag{3.34}$$

Proof. The substitution of the matrices $\bar{M}_{21} = 0$, $\bar{N}_{21} = 0$, and $\bar{L}_{21} = 0$ into the matrices $\tilde{\tilde{A}}$, $\tilde{\tilde{B}}$, and $\tilde{\tilde{C}}$ obtained in Corollaries 3.3, 3.5, and 3.7, respectively, results in (3.34). ∎

Now, we give the conditions for complementary matrices M, N, and L in the initial basis.

THEOREM 3.15. *The system* $\mathbf{S}$ *is a restriction of the system* $\tilde{\mathbf{S}}$ *if and only if the matrices* M*,* N*, and* L *have the following form:*

$$\begin{array}{ll} \text{(a)} & M = \begin{pmatrix} 0 & M_{12} & -M_{12} & 0 \\ 0 & M_{22} & -M_{22} & 0 \\ 0 & M_{32} & -M_{32} & 0 \\ 0 & M_{42} & -M_{42} & 0 \end{pmatrix}, \\ \text{(b)} & N = \begin{pmatrix} 0 & N_{12} & -N_{12} & 0 \\ 0 & N_{22} & -N_{22} & 0 \\ 0 & N_{32} & -N_{32} & 0 \\ 0 & N_{42} & -N_{42} & 0 \end{pmatrix}, \\ \text{(c)} & L = \begin{pmatrix} 0 & L_{12} & -L_{12} & 0 \\ 0 & L_{22} & -L_{22} & 0 \\ 0 & L_{32} & -L_{32} & 0 \\ 0 & L_{42} & -L_{42} & 0 \end{pmatrix}. \end{array} \tag{3.35}$$

Proof. Let M be the matrix given by (3.19). It follows from Proposition 3.13 that $\bar{M}_{11} = 0$, $\bar{M}_{21} = 0$. We get (3.23) and simultaneously

$$M_{21} = 0, \qquad M_{22} + M_{23} = 0, \qquad M_{24} = 0 \tag{3.36}$$

from (3.25). Substituting these results into the matrix M given in (3.19), we conclude case (a). The proofs of (b) and (c) are analogous. ∎

COROLLARY 3.16. *Let* $\mathbf{S}$ *be a restriction of the system* $\tilde{\mathbf{S}}$*. Then the matrices* $\tilde{A}$*,* $\tilde{B}$*, and* $\tilde{C}$ *have the following form:*

$$\tilde{A} = \begin{pmatrix} A_{11} & \frac{1}{2}A_{12}+M_{12} & \frac{1}{2}A_{12}-M_{12} & A_{13} \\ A_{21} & \frac{1}{2}A_{22}+M_{22} & \frac{1}{2}A_{22}-M_{22} & A_{23} \\ A_{21} & \frac{1}{2}A_{22}+M_{32} & \frac{1}{2}A_{22}-M_{32} & A_{23} \\ A_{31} & \frac{1}{2}A_{32}+M_{42} & \frac{1}{2}A_{32}-M_{42} & A_{33} \end{pmatrix}, \tag{3.37}$$

$$\tilde{B} = \begin{pmatrix} B_{11} & \frac{1}{2}B_{12}+N_{12} & \frac{1}{2}B_{12}-N_{12} & B_{13} \\ B_{21} & \frac{1}{2}B_{22}+N_{22} & \frac{1}{2}B_{22}-N_{22} & B_{23} \\ B_{21} & \frac{1}{2}B_{22}+N_{32} & \frac{1}{2}B_{22}-N_{32} & B_{23} \\ B_{31} & \frac{1}{2}B_{32}+N_{42} & \frac{1}{2}B_{32}-N_{42} & B_{33} \end{pmatrix}, \tag{3.38}$$

$$\tilde{C} = \begin{pmatrix} C_{11} & \frac{1}{2}C_{12}+L_{12} & \frac{1}{2}C_{12}-L_{12} & C_{13} \\ C_{21} & \frac{1}{2}C_{22}+L_{22} & \frac{1}{2}C_{22}-L_{22} & C_{23} \\ C_{21} & \frac{1}{2}C_{22}+L_{32} & \frac{1}{2}C_{22}-L_{32} & C_{23} \\ C_{31} & \frac{1}{2}C_{32}+L_{42} & \frac{1}{2}C_{32}-L_{42} & C_{33} \end{pmatrix}. \tag{3.39}$$

Proof. Let the matrices $\tilde{A}$, $\tilde{B}$, and $\tilde{C}$ be given by Corollaries 3.10, 3.11, and 3.12, respectively. Let M, N, and L be given by (3.35). By using these matrices in (3.26)–(3.28), the corollary is proved. ∎

The same way of reasoning holds for aggregations.

PROPOSITION 3.17. *The system* $\mathbf{S}$ *is an aggregation of the system* $\tilde{\tilde{\mathbf{S}}}$ *if and only if*

$$\begin{aligned} &\text{(a)} \quad \bar{M} = \begin{pmatrix} 0 & 0 \\ \bar{M}_{21} & \bar{M}_{22} \end{pmatrix}, \\ &\text{(b)} \quad \bar{N} = \begin{pmatrix} 0 & 0 \\ \bar{N}_{21} & \bar{N}_{22} \end{pmatrix}, \\ &\text{(c)} \quad \bar{L} = \begin{pmatrix} 0 & 0 \\ \bar{L}_{21} & \bar{L}_{22} \end{pmatrix}. \end{aligned} \tag{3.40}$$

Proof. By Theorem 1.8, we know that any aggregation must verify

$$\begin{aligned} &\text{(a)} \quad \bar{U}\bar{M} = 0, \\ &\text{(b)} \quad \bar{U}\bar{N} = 0, \\ &\text{(c)} \quad \bar{S}\bar{L} = 0. \end{aligned} \tag{3.41}$$

Taking into account that $\bar{M}$ is given by (3.15) and that $\bar{U}$ has the form $\bar{U} = (\, I_n \;\; 0\,)$, from $\bar{U}\bar{M} = 0$ it results in $\bar{M}_{11} = 0$ and $\bar{M}_{12} = 0$ so that $\bar{M} = \left(\begin{smallmatrix} 0 & 0 \\ \bar{M}_{21} & \bar{M}_{22} \end{smallmatrix}\right)$. The proofs of cases (b) and (c) are similar. ∎

COROLLARY 3.18. *Let* $\mathbf{S}$ *be an aggregation of the system* $\tilde{\tilde{\mathbf{S}}}$. *Then the matrices* $\tilde{\tilde{A}}$, $\tilde{\tilde{B}}$, $\tilde{\tilde{C}}$ *have the following form:*

$$\tilde{\tilde{A}} = \left(\begin{smallmatrix} A & 0 \\ \bar{M}_{21} & \bar{M}_{22} \end{smallmatrix}\right), \quad \tilde{\tilde{B}} = \left(\begin{smallmatrix} B & 0 \\ \bar{N}_{21} & \bar{N}_{22} \end{smallmatrix}\right), \quad \tilde{\tilde{C}} = \left(\begin{smallmatrix} C & 0 \\ \bar{L}_{21} & \bar{L}_{22} \end{smallmatrix}\right). \tag{3.42}$$

Proof. To prove the corollary, we substitute $\bar{M}_{12} = 0$, $\bar{N}_{12} = 0$, and $\bar{L}_{12} = 0$ into the matrices $\tilde{\tilde{A}}$, $\tilde{\tilde{B}}$, and $\tilde{\tilde{C}}$ given in Corollaries 3.3, 3.5, and 3.7, respectively. ∎

Since we want to know the structure of complementary matrices of aggregations in the expanded system $\tilde{\mathbf{S}}$, we come back to the initial basis. These structures are characterized by the following theorem.

THEOREM 3.19. *The system* $\mathbf{S}$ *is an aggregation of the system* $\tilde{\mathbf{S}}$ *if and only if*

the matrices M, N, and L are given as follows:

$$\text{(a)}\quad M = \begin{pmatrix} 0 & 0 & 0 & 0 \\ M_{21} & M_{22} & M_{23} & M_{24} \\ -M_{21} & -M_{22} & -M_{23} & -M_{24} \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\text{(b)}\quad N = \begin{pmatrix} 0 & 0 & 0 & 0 \\ N_{21} & N_{22} & N_{23} & N_{24} \\ -N_{21} & -N_{22} & -N_{23} & -N_{24} \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{3.43}$$

$$\text{(c)}\quad L = \begin{pmatrix} 0 & 0 & 0 & 0 \\ L_{21} & L_{22} & L_{23} & L_{24} \\ -L_{21} & -L_{22} & -L_{23} & -L_{24} \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Proof. Consider the matrix M given by Theorem 3.9. It follows from Proposition 3.17 that $\bar{M}_{11} = 0$ and $\bar{M}_{12} = 0$. We obtain (3.23) together with the equations

$$M_{12} - M_{13} = 0, \quad M_{22} - M_{23} + M_{32} - M_{33} = 0, \quad M_{42} - M_{43} = 0. \tag{3.44}$$

This leads to

$$M = \begin{pmatrix} 0 & 0 & 0 & 0 \\ M_{21} & M_{22} & M_{23} & M_{24} \\ -M_{21} & -M_{22} & -M_{23} & -M_{24} \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

in case (a) only by substituting the obtained relations into the matrix M given by (3.19). The proofs of cases (b) and (c) are similar. □

COROLLARY 3.20. *Let* $\mathbf{S}$ *be an aggregation of the system* $\tilde{\mathbf{S}}$. *Then the matrices* $\tilde{A}$, $\tilde{B}$, *and* $\tilde{C}$ *have the following form:*

$$\tilde{A} = \begin{pmatrix} A_{11} & \frac{1}{2}A_{12} & \frac{1}{2}A_{12} & A_{13} \\ A_{21}+M_{21} & \frac{1}{2}A_{22}+M_{22} & \frac{1}{2}A_{22}+M_{23} & A_{23}+M_{24} \\ A_{21}-M_{21} & \frac{1}{2}A_{22}-M_{22} & \frac{1}{2}A_{22}-M_{23} & A_{23}-M_{24} \\ A_{31} & \frac{1}{2}A_{32} & \frac{1}{2}A_{32} & A_{33} \end{pmatrix},$$

$$\tilde{B} = \begin{pmatrix} B_{11} & \frac{1}{2}B_{12} & \frac{1}{2}B_{12} & B_{13} \\ B_{21}+N_{21} & \frac{1}{2}B_{22}+N_{22} & \frac{1}{2}B_{22}+N_{23} & B_{23}+N_{24} \\ B_{21}-N_{21} & \frac{1}{2}B_{22}-N_{22} & \frac{1}{2}B_{22}-N_{23} & B_{23}-N_{24} \\ B_{31} & \frac{1}{2}B_{32} & \frac{1}{2}B_{32} & B_{33} \end{pmatrix}, \tag{3.45}$$

$$\tilde{C} = \begin{pmatrix} C_{11} & \frac{1}{2}C_{12} & \frac{1}{2}C_{12} & C_{13} \\ C_{21}+L_{21} & \frac{1}{2}C_{22}+L_{22} & \frac{1}{2}C_{22}+L_{23} & C_{23}+L_{24} \\ C_{21}-L_{21} & \frac{1}{2}C_{22}-L_{22} & \frac{1}{2}C_{22}-L_{23} & C_{23}-L_{24} \\ C_{31} & \frac{1}{2}C_{32} & \frac{1}{2}C_{32} & C_{33} \end{pmatrix}.$$

Proof. Since $\mathbf{S}$ is an aggregation of $\tilde{\mathbf{S}}$, then M, N, and L are given by (3.43). Substituting the structure of these matrices into (3.26)–(3.28), respectively, proves the corollary. □

3.4. Contractibility. We want to design the control laws in the expanded system $\tilde{\mathbf{S}}$ so that we can contract and implement them into the original system $\mathbf{S}$. Now, we determine the conditions when a control law designed in the expanded system $\tilde{\tilde{\mathbf{S}}}$ can be contracted into the initial system $\mathbf{S}$ in terms of the complementary matrices.

Notation. We denote the matrices that appear in the contractibility process as follows. The prototype complementary matrix has the following form:

$$F = \begin{array}{c} \\ m_1 \\ m_2 \\ m_2 \\ m_3 \end{array} \begin{array}{c} \begin{array}{cccc} n_1 & n_2 & n_2 & n_3 \end{array} \\ \begin{pmatrix} F_{11} & F_{12} & F_{13} & F_{14} \\ F_{21} & F_{22} & F_{23} & F_{24} \\ F_{31} & F_{32} & F_{33} & F_{34} \\ F_{41} & F_{42} & F_{43} & F_{44} \end{pmatrix} \end{array}. \tag{3.46}$$

Define

$$\bar{F} = \begin{pmatrix} \bar{F}_{11} & \bar{F}_{12} \\ \bar{F}_{21} & \bar{F}_{22} \end{pmatrix}, \tag{3.47}$$

where $\bar{F}_{11}$, $\bar{F}_{12}$, $\bar{F}_{21}$, $\bar{F}_{22}$ are $m \times n$, $m \times (\tilde{n} - n)$, $(\tilde{m} - m) \times n$, $(\tilde{m} - m) \times (\tilde{n} - n)$ dimensional matrices, respectively. Similarly, let us denote the prototype gain matrix

$$K = \begin{array}{c} \\ m_1 \\ m_2 \\ m_3 \end{array} \begin{array}{c} \begin{array}{ccc} n_1 & n_2 & n_3 \end{array} \\ \begin{pmatrix} K_{11} & K_{12} & K_{13} \\ K_{21} & K_{22} & K_{23} \\ K_{31} & K_{32} & K_{33} \end{pmatrix} \end{array}. \tag{3.48}$$

The gain matrix $\tilde{\bar{K}}$ for the system $\tilde{\tilde{\mathbf{S}}}$ has the form

$$\tilde{\bar{K}} = \bar{R} K \bar{U} + \bar{F}, \tag{3.49}$$

where

$$\begin{aligned} \tilde{\bar{K}} &= T_B^{-1} \tilde{K} T_A, \\ \bar{F} &= T_B^{-1} F T_A. \end{aligned} \tag{3.50}$$

PROPOSITION 3.21. *Let $\bar{F}$ be the matrix, where $\bar{F} = \left(\begin{smallmatrix} \bar{F}_{11} & \bar{F}_{12} \\ \bar{F}_{21} & \bar{F}_{22} \end{smallmatrix}\right)$. Then the submatrices $\bar{F}_{ij}$ have the form, for $i, j = 1, 2$,*

$$\begin{aligned} \bar{F}_{11} &= \begin{pmatrix} F_{11} & F_{12}+F_{13} & F_{14} \\ \frac{1}{2}(F_{21}+F_{31}) & \frac{1}{2}(F_{22}+F_{23}+F_{32}+F_{33}) & \frac{1}{2}(F_{24}+F_{34}) \\ F_{41} & F_{42}+F_{43} & F_{44} \end{pmatrix}, \\ \bar{F}_{12} &= \begin{pmatrix} F_{12}-F_{13} \\ \frac{1}{2}(F_{22}-F_{23}+F_{32}-F_{33}) \\ F_{42}-F_{43} \end{pmatrix}, \\ \bar{F}_{21} &= \frac{1}{2} \begin{pmatrix} F_{21}-F_{31} & F_{22}+F_{23}-F_{32}-F_{33} & F_{24}-F_{34} \end{pmatrix}, \\ \bar{F}_{22} &= \frac{1}{2} \begin{pmatrix} F_{22}-F_{23}-F_{32}+F_{33} \end{pmatrix}. \end{aligned} \tag{3.51}$$

Proof. Let $\bar{F} = T_B^{-1} F T_A$ be given by (3.50), where T_A and T_B^{-1} are given by (3.3). The proposition is proved by comparing the resulting product with the matrix defined by $\bar{F} = \left(\begin{smallmatrix} \bar{F}_{11} & \bar{F}_{12} \\ \bar{F}_{21} & \bar{F}_{22} \end{smallmatrix}\right)$. ∎

COROLLARY 3.22. *Suppose that Proposition* 3.21 *holds; then the matrix* $\tilde{\bar{K}}$ *of the expanded system* $\tilde{\bar{\mathbf{S}}}$ *is* $\tilde{\bar{K}} = \left(\begin{smallmatrix} K & \bar{F}_{12} \\ \bar{F}_{21} & \bar{F}_{22} \end{smallmatrix}\right)$.

Proof. Considering $\tilde{\bar{K}}$ as in (3.49), $\bar{R} = \left(\begin{smallmatrix} I_m \\ 0 \end{smallmatrix}\right)$ and $\bar{U} = \left(\begin{smallmatrix} I_n & 0 \end{smallmatrix}\right)$, where (0) denotes matrices of appropriate dimensions, we conclude the corollary. ∎

By Definition 1.3 of contractibility, we know that $\tilde{\bar{u}} = -\tilde{\bar{K}}\tilde{\bar{x}}$ is contractible to the control law $u = -Kx$ whenever $Kx(t; x_0, u) = \bar{Q}\tilde{\bar{K}}\tilde{\bar{x}}(t; \bar{V}x_0, \bar{R}u)$. Then we can present the following theorem for the expanded system $\tilde{\bar{\mathbf{S}}}$.

THEOREM 3.23. *A control law* $\tilde{\bar{u}} = -\tilde{\bar{K}}\tilde{\bar{x}}$ *given in the expanded system* $\tilde{\bar{\mathbf{S}}}$ *is contractible to the control law* $u = -Kx$ *of the system* $\mathbf{S}$ *if and only if*

$$\begin{array}{ll} \text{(a)} & \bar{Q}\bar{F}\bar{M}^{i-1}\bar{V} = 0, \\ \text{(b)} & \bar{Q}\bar{F}\bar{M}^{i-1}\bar{N}\bar{R} = 0 \end{array} \tag{3.52}$$

for all $i = 1, \ldots, \tilde{n}$.

Proof. It is known that

$$x(t) = e^{At}x_0 + \int_o^t e^{A(t-\tau)}Bu(\tau)d(\tau) \tag{3.53}$$

and

$$\tilde{\bar{x}}(t) = e^{\tilde{\bar{A}}t}\bar{V}x_0 + \int_o^t e^{\tilde{\bar{A}}(t-\tau)}\tilde{\bar{B}}\bar{R}u(\tau)d(\tau), \tag{3.54}$$

where

$$e^{At} = I + At + \frac{1}{2!}A^2t^2 + \frac{1}{3!}A^3t^3 + \cdots \tag{3.55}$$

and

$$e^{\tilde{\bar{A}}t} = I + \tilde{\bar{A}}t + \frac{1}{2!}\tilde{\bar{A}}^2t^2 + \frac{1}{3!}\tilde{\bar{A}}^3t^3 + \cdots . \tag{3.56}$$

Since we require $Kx(t; x_0, u) = \bar{Q}\tilde{\bar{K}}\tilde{\bar{x}}(t; \bar{V}x_0, \bar{R}u)$ when considering $\tilde{\bar{K}} = \bar{R}K\bar{U} + \bar{F}$, we get $Kx(t) = [\bar{Q}\bar{R}K\bar{U} + \bar{Q}\bar{F}]\tilde{\bar{x}}(t)$, and consequently $Kx(t) = K\bar{U}\tilde{\bar{x}}(t) + \bar{Q}\bar{F}\tilde{\bar{x}}(t)$. Then multiplying K by x in (3.53), as well as multiplying $K\bar{U}$ and $\bar{Q}\bar{F}$ by $\tilde{\bar{x}}$ in (3.54), the results coincide if and only if

$$\begin{array}{ll} \text{(a)} & \bar{Q}\bar{F}\bar{M}^{i-1}\bar{V} = 0, \\ \text{(b)} & \bar{Q}\bar{F}\bar{M}^{i-1}\bar{N}\bar{R} = 0 \end{array} \tag{3.57}$$

hold for all $i = 1, \ldots, \tilde{n}$. ∎

Another equivalent form of presenting Theorem 3.23 provides the following theorem without explicitly using the complementary matrix F.

THEOREM 3.24. *A control law* $\tilde{\bar{u}} = -\tilde{\bar{K}}\tilde{\bar{x}}$ *given in the expanded system* $\tilde{\bar{\mathbf{S}}}$ *is contractible to the control law* $u = -Kx$ *of the initial* $\mathbf{S}$ *if and only if*

$$\begin{array}{ll} \text{(a)} & \bar{Q}\tilde{\bar{K}}\bar{V} = K, \\ \text{(b)} & \bar{Q}\tilde{\bar{K}}\bar{M}^{i-1}\bar{V} = 0 \ \textit{ for all } i = 2, \ldots, \tilde{n}, \\ \text{(c)} & \bar{Q}\tilde{\bar{K}}\bar{M}^{i-1}\bar{N}\bar{R} = 0 \ \textit{ for all } i = 1, \ldots, \tilde{n}. \end{array} \tag{3.58}$$

Proof. To prove Theorem 3.24, we follow the equivalence with Theorem 3.23. Let $\tilde{\bar{K}} = \bar{R}K\bar{U} + \bar{F}$ be given by (3.49); then $\bar{F} = \tilde{\bar{K}} - \bar{R}K\bar{U}$. Substituting $\bar{F}$ into (a) and (b) in Theorem 3.23, we obtain

$$\begin{aligned}
&\text{(a)} \quad \bar{Q}[\tilde{\bar{K}} - \bar{R}K\bar{U}]\bar{M}^{i-1}\bar{V} = 0, \quad \text{and consequently } \bar{Q}\tilde{\bar{K}}\bar{M}^{i-1}\bar{V} = K\bar{U}\bar{M}^{i-1}\bar{V},\\
&\text{(b)} \quad \bar{Q}[\tilde{\bar{K}} - \bar{R}K\bar{U}]\bar{M}^{i-1}\bar{N}\bar{R} = 0, \quad \text{and consequently } \bar{Q}\tilde{\bar{K}}\bar{M}^{i-1}\bar{N}\bar{R} = K\bar{U}\bar{M}^{i-1}\bar{N}\bar{R},
\end{aligned}$$

for all $i = 1, \ldots, \tilde{n}$. From (3.13) it follows $\bar{U}\bar{M}^i\bar{V}{=}0$ for all $i = 1, \ldots, \tilde{n}$. Therefore, according to (a), we have

$$\begin{aligned}
\bar{Q}\tilde{\bar{K}}\bar{V} &= K,\\
\bar{Q}\tilde{\bar{K}}\bar{M}^{i-1}\bar{V} &= 0
\end{aligned} \tag{3.59}$$

for all $i = 2, \ldots, \tilde{n}$. By (3.13) we know that $\bar{U}\bar{M}^{i-1}\bar{N}\bar{R} = 0$, and consequently $\bar{Q}\tilde{\bar{K}}\bar{M}^{i-1}\bar{N}\bar{R} = 0$, for all $i = 1, \ldots, \tilde{n}$. ∎

Note. So far we do not know the form of the complementary matrix F and the conditions which must be satisfied in order to get a control law contractible to the initial system. The following theorem answers this problem giving the contractibility conditions for control laws between the systems $\tilde{\bar{\mathbf{S}}}$ and $\mathbf{S}$. Further, we introduce an equivalent theorem for the systems $\tilde{\mathbf{S}}$ and $\mathbf{S}$.

THEOREM 3.25. *A control law $\tilde{\bar{u}} = -\tilde{\bar{K}}\tilde{\bar{x}}$ given in the expanded system $\tilde{\bar{\mathbf{S}}}$ is contractible to the control law $u = -Kx$ in the system $\mathbf{S}$ if and only if $\bar{F} = \begin{pmatrix} 0 & \bar{F}_{12} \\ \bar{F}_{21} & \bar{F}_{22} \end{pmatrix}$, satisfying $\bar{F}_{12}\bar{M}_{22}^{i-2}\bar{M}_{21} = 0$ and $\bar{F}_{12}\bar{M}_{22}^{i-2}\bar{N}_{21} = 0$ for all $i = 2, \ldots, \tilde{n} + 1$.*

Proof. We consider the relations (3.52). Considering the set of conditions (a), we get $\bar{Q}\bar{F}\bar{V} = 0$ for $i = 1$, that is, $\bar{F}_{11} = 0$ and consequently $\bar{F} = \begin{pmatrix} 0 & \bar{F}_{12} \\ \bar{F}_{21} & \bar{F}_{22} \end{pmatrix}$. Consider the matrix $\bar{M}^{\mu}$ given by (3.17). We have $\bar{Q}\bar{F}\bar{M}^{k-1}\bar{V} = 0$ for $i = k \geq 2$. It implies $\bar{F}_{12}\bar{M}_{22}^{k-1}\bar{M}_{21} = 0$. Repeating this process, we have $\bar{Q}\bar{F}\bar{M}^{\tilde{n}-1}\bar{V} = 0$ for $i = \tilde{n}$, so that we obtain $\bar{F}_{12}\bar{M}_{22}^{\tilde{n}-1}\bar{M}_{21} = 0$. Thus $\bar{F}_{12}\bar{M}_{22}^{\tilde{n}-2}\bar{M}_{21} = 0$ holds for all $i = 2, \ldots, \tilde{n} + 1$.

Consider now the set of conditions (b). $\bar{Q}\bar{F}\bar{N}\bar{R} = 0$ holds for $i = 1$ and therefore $\bar{F}_{12}\bar{N}_{21} = 0$. Consider again the matrix $\bar{M}^{\mu}$ given by (3.17). Then $\bar{Q}\bar{F}\bar{M}^{k-1}\bar{N}\bar{R} = 0$ for $i = k \geq 2$ and consequently $\bar{F}_{12}\bar{M}_{22}^{k-1}\bar{N}_{21} = 0$ by using the results of Proposition 3.4. Finally, $\bar{Q}\bar{F}\bar{M}^{\tilde{n}-1}\bar{N}\bar{R} = 0$ holds for $i = \tilde{n}$, that is, $\bar{F}_{12}\bar{M}_{22}^{\tilde{n}-1}\bar{N}_{21} = 0$. Summarizing this process, we get $\bar{F}_{12}\bar{M}_{22}^{\tilde{n}-2}\bar{N}_{21} = 0$ for all $i = 2, \ldots, \tilde{n} + 1$. ∎

THEOREM 3.26. *A control law $\tilde{u} = -\tilde{K}\tilde{x}$ given in the expanded system $\tilde{\mathbf{S}}$ is contractible to the control law $u = -Kx$ of the system $\mathbf{S}$ if and only if the matrix F has the form*

$$F = \begin{pmatrix} 0 & F_{12} & -F_{12} & 0 \\ F_{21} & F_{22} & F_{23} & F_{24} \\ -F_{21} & -(F_{22}+F_{23}+F_{33}) & F_{33} & -F_{24} \\ 0 & F_{42} & -F_{42} & 0 \end{pmatrix} \tag{3.60}$$

and the conditions

$$\begin{aligned}
&\begin{pmatrix} F_{12} \\ F_{23}+F_{33} \\ F_{42} \end{pmatrix} \begin{pmatrix} M_{22}+M_{33} \end{pmatrix}^{i-2} \begin{pmatrix} M_{21} & M_{22}+M_{23} & M_{24} \end{pmatrix} = 0,\\
&\begin{pmatrix} F_{12} \\ F_{23}+F_{33} \\ F_{42} \end{pmatrix} \begin{pmatrix} M_{22}+M_{33} \end{pmatrix}^{i-2} \begin{pmatrix} N_{21} & N_{22}+N_{23} & N_{24} \end{pmatrix} = 0
\end{aligned} \tag{3.61}$$

hold for all $i = 2, \ldots, \tilde{n} + 1$.

Proof. Consider the matrix $\bar{F} = T_B^{-1} F T_A$ given by (3.50). Then, carrying out this product, we obtain

$$\bar{F} = \begin{pmatrix} F_{11} & F_{12}+F_{13} & F_{14} & F_{12}-F_{13} \\ \frac{1}{2}(F_{21}+F_{31}) & \frac{1}{2}(F_{22}+F_{23}+F_{32}+F_{33}) & \frac{1}{2}(F_{24}+F_{34}) & \frac{1}{2}(F_{22}-F_{23}+F_{32}-F_{33}) \\ F_{41} & F_{42}+F_{43} & F_{44} & F_{42}-F_{43} \\ \frac{1}{2}(F_{21}-F_{31}) & \frac{1}{2}(F_{22}+F_{23}-F_{32}-F_{33}) & \frac{1}{2}(F_{24}-F_{34}) & \frac{1}{2}(F_{22}-F_{23}-F_{32}+F_{33}) \end{pmatrix}. \tag{3.62}$$

Moreover, according to Theorem 3.25, $\bar{F} = \left(\begin{smallmatrix} 0 & \bar{F}_{12} \\ \bar{F}_{21} & \bar{F}_{22} \end{smallmatrix}\right)$. Since $\bar{F}_{11} = 0$ in (3.51), we obtain F given by (3.60). $\bar{F}_{12}$ has the form

$$\bar{F}_{12} = \begin{pmatrix} F_{12}-F_{13} \\ \frac{1}{2}(F_{22}-F_{23}+F_{32}-F_{33}) \\ F_{42}-F_{43} \end{pmatrix}$$

when considering (3.51). Now, considering $\bar{M}_{21}$, $\bar{M}_{22}$, and $\bar{N}_{21}$ given by Theorem 3.9 and imposing the relations given by Theorem 3.25, we conclude that

$$\begin{aligned} &\begin{pmatrix} F_{12} \\ F_{23}+F_{33} \\ F_{42} \end{pmatrix} \left(M_{22}+M_{33} \right)^{i-2} \begin{pmatrix} M_{21} & M_{22}+M_{23} & M_{24} \end{pmatrix} = 0, \\ &\begin{pmatrix} F_{12} \\ F_{23}+F_{33} \\ F_{42} \end{pmatrix} \left(M_{22}+M_{33} \right)^{i-2} \begin{pmatrix} N_{21} & N_{22}+N_{23} & N_{24} \end{pmatrix} = 0 \end{aligned} \tag{3.63}$$

for all $i = 2, \ldots, \tilde{n} + 1$. □

Now, we specialize the general contractibility results for the case of restrictions. This is motivated by the well-known experience when considering the control design for this case.

PROPOSITION 3.27. *Let* $\mathbf{S}$ *be a restriction of the system* $\tilde{\mathbf{S}}$. *Then any control law* $\tilde{u} = -\tilde{K}\tilde{x}$ *of* $\tilde{\mathbf{S}}$ *is contractible to the control law* $u = -Kx$ *of* $\mathbf{S}$.

Proof. Theorem 3.26 provides the necessary and sufficient condition for the control law $\tilde{u} = -\tilde{K}\tilde{x}$ of $\tilde{\mathbf{S}}$ to be contractible to the control law $u = -Kx$. Then the matrix F must satisfy the relations

$$\begin{aligned} &\begin{pmatrix} F_{12} \\ F_{23}+F_{33} \\ F_{42} \end{pmatrix} \left(M_{22}+M_{33} \right)^{i-2} \begin{pmatrix} M_{21} & M_{22}+M_{23} & M_{24} \end{pmatrix} = 0, \\ &\begin{pmatrix} F_{12} \\ F_{23}+F_{33} \\ F_{42} \end{pmatrix} \left(M_{22}+M_{33} \right)^{i-2} \begin{pmatrix} N_{21} & N_{22}+N_{23} & N_{24} \end{pmatrix} = 0 \end{aligned} \tag{3.64}$$

for all $i = 2, \ldots, \tilde{n} + 1$. Since we suppose that $\mathbf{S}$ is a restriction of $\tilde{\mathbf{S}}$, the following relations appear:

$$\begin{aligned} &M_{21} = 0, \quad M_{22} + M_{23} = 0, \quad M_{24} = 0, \\ &N_{21} = 0, \quad N_{22} + N_{23} = 0, \quad N_{24} = 0. \end{aligned} \tag{3.65}$$

We deduce easily that $\begin{pmatrix} M_{21} & M_{22}+M_{23} & M_{24} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}$ and $\begin{pmatrix} N_{21} & N_{22}+N_{23} & N_{24} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}$ from these relations, whereby the conditions (3.64) hold for any complementary matrix F of the form (3.60). □

Note. It is well known that any control law is contractible in the case of restrictions. Proposition 3.27 has revisited this result in the sense that it has proved that any control law is contractible independently of F. However, when considering the structure of F given by (3.60) we can take advantage of it to design different contractible control laws in the expanded space.

3.5. Decentralized control. Consider that we have an original system **S** exhibiting strong overlapping connections among their subsystems. The idea of decentralized control in the framework of the inclusion principle means to expand the above-mentioned original system to another one in the expanded space that is composed of weakly coupled subsystems. In this way, control design methods for weakly coupled subsystems can be applied. In other words, we attempt to get an expanded system with maximum of zero off-diagonal blocks. The possible maximum of off-diagonal zero blocks shows the following matrix:

$$\tilde{A} = \left(\begin{array}{cc:cc} A_{11} & A_{12} & * & * \\ A_{21} & A_{22} & * & * \\ \hdashline * & * & A_{22} & A_{23} \\ * & * & A_{32} & A_{33} \end{array}\right), \tag{3.66}$$

where * denotes the possible zero blocks.

Two well-known classes of complementary matrices verifying Theorem 1.4 have been extensively considered in the literature [3], [4], [5], [7], [6], [10], which lead to the following cases of matrix $\tilde{A}$:

(3.67)

$$\tilde{A} = \left(\begin{array}{cc:cc} A_{11} & A_{12} & 0 & A_{13} \\ A_{21} & A_{22} & 0 & A_{23} \\ \hdashline A_{21} & 0 & A_{22} & A_{23} \\ A_{31} & 0 & A_{32} & A_{33} \end{array}\right) \quad \text{or} \quad \tilde{A} = \left(\begin{array}{cc:cc} A_{11} & A_{12} & A_{12} & A_{13} \\ A_{21} & A_{22} & 0 & 0 \\ \hdashline 0 & 0 & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{32} & A_{33} \end{array}\right).$$

When considering the procedure proposed in this paper, we get more flexibility in reducing the off-diagonal nonzero blocks. For instance, by applying Theorem 3.9, in some situations we can select the complementary matrices in the form

$$\begin{aligned} &M_{12} = \tfrac{1}{2}A_{12}, && M_{21} = A_{21}, && M_{22} = \tfrac{1}{2}A_{22}, \\ &M_{23} = -\tfrac{1}{2}A_{22}, && M_{24} = -A_{23}, && M_{33} = \tfrac{1}{2}A_{22}, \\ &M_{42} = -\tfrac{1}{2}A_{32}, \end{aligned} \tag{3.68}$$

and we get the resulting expanded matrix $\tilde{A}$ in the form

$$\tilde{A} = \left(\begin{array}{cc:cc} A_{11} & A_{12} & 0 & A_{13} \\ 2A_{21} & A_{22} & 0 & 0 \\ \hdashline 0 & 0 & A_{22} & 2A_{23} \\ A_{31} & 0 & A_{32} & A_{33} \end{array}\right). \tag{3.69}$$

To simplify, we use the same notation for $\tilde{A}$ in (3.67) and (3.69). Only the blocks A_{13} and A_{31} cannot be eliminated. Let us note that the same zero reduction procedure holds for the matrices $\tilde{B}$ and $\tilde{C}$. This procedure is not related to any specific feedback control law. It is expected that when considering both control matrices in the original and expanded spaces, respectively, K and $\tilde{K}$, and the corresponding complementary matrix F as defined in (1.8), the blocks $(\cdot)_{13}$ and $(\cdot)_{31}$ in all above-mentioned matrices can be weakened or eliminated. In this sense we could think of choosing the appropriate matrix F such that

$$\min_{N,F} \{\|\tilde{A}_{\text{off}} - \tilde{B}_{\text{off}} F - NF\|\}, \tag{3.70}$$

where $\tilde{A}_{\text{off}} = \tilde{A} - \text{diag}\{\tilde{A}_{11}, \tilde{A}_{22}\}$, $\tilde{B}_{\text{off}} = \tilde{B} - \text{diag}\{\tilde{B}_{11}, \tilde{B}_{22}\}$. The generalization to more than two subsystems is straightforward. Let us leave the solution of (3.70) as an opportunity for future research.

4. Conclusion. The main result of this paper is that the complementary matrices of the transformations involved in the input-state-output inclusion principle for linear time invariant systems may be constructed employing a block matrix structure in a very flexible manner within the expansion-contraction process. Two general classes of this process have been identified in which aggregations and restrictions have been characterized. Within both classes, submatrices can be arbitrarily selected so that the designer can obtain a wide variety of expanded systems. Further, the contractibility of controllers has been systematically characterized within this framework. The decentralized control design has been discussed. The proposed structure of the expansion-contraction process can be useful for the analysis and design of overlapping decentralized controllers.

REFERENCES

[1] L. Bakule and J. Lunze, *Decentralized Design of Feedback Control for Large-Scale Systems*, Kybernetika (Prague) Suppl. 24, 1988.

[2] L. Bakule and J. Rodellar, *Decentralized control and overlapping decomposition of mechanical systems. Part* 1: *System decomposition. Part* 2: *Decentralized stabilization*, Internat. J. Control, 61 (1995), pp. 559–587.

[3] A. Iftar and U. Özgüner, *Contractible controller design and optimal control with state and input inclusion*, Automatica J. IFAC, 26 (1990), pp. 593–597.

[4] M. Ikeda and D.D. Šiljak, *Overlapping decompositions, expansions and contractions of dynamic systems*, Large Scale Systems, 1 (1980), pp. 29–38.

[5] M. Ikeda and D.D. Šiljak, *Generalized decompositions of dynamic systems and vector Lyapunov functions*, IEEE Trans. Automat. Control, 26 (1981), pp. 1118–1125.

[6] M. Ikeda and D.D. Šiljak, *Overlapping decentralized control with input, state and output inclusion*, Control Theory Adv. Tech., 2 (1986), pp. 155–172.

[7] M. Ikeda, D.D. Šiljak, and D.E. White, *An inclusion principle for dynamic systems*, IEEE Trans. Automat. Control, 29 (1984), pp. 244–249.

[8] J.M. Rossell, J. Rodellar, and P. Rubió, *The generalized inclusion principle: New expansion forms*, in Proceedings of the Meeting on Matrix Analysis and Applications, University of Sevilla, Sevilla, Spain, 1997, pp. 375–382 (in Spanish).

[9] J.M. Rossell, *Contribution to Decentralized Control of Large-Scale Systems via Overlapping Models*, Ph.D. thesis, Technical University of Catalunya, Barcelona, Spain, 1998 (in Spanish).

[10] D.D. Šiljak, *Decentralized Control of Complex Systems*, Academic Press, New York, 1991.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 4, pp. 1156–1162
© 2000 Society for Industrial and Applied Mathematics

ON SHARY'S ALGEBRAIC APPROACH FOR LINEAR INTERVAL EQUATIONS*

ARNOLD NEUMAIER†

Abstract. A recent method by Shary for enclosing the solution set of a system of linear interval equations is derived in a new way. It is shown that the method converges to the fixed-point inverse and that it has finite termination with probability 1.

Key words. linear interval equation, fixed-point inverse, inner subtraction, H-matrix, finite termination

AMS subject classification. 65G10

PII. S0895479898347472

1. Introduction.

There are a variety of stationary iterative methods for enclosing the *solution set*

$$\mathbf{A}^H\mathbf{b} = \square\{x \in \mathbb{R}^n \mid Ax = b \quad \text{for some } A \in \mathbf{A},\ b \in \mathbf{b}\} \tag{1}$$

of a system of linear equations with interval coefficient matrix $\mathbf{A} \in \mathbb{IR}^{n\times n}$ and interval right-hand side $\mathbf{b} \in \mathbb{IR}^n$. (Here $\square S$ denotes the interval hull, i.e., the smallest interval vector containing the bounded set S.) By preconditioning, the system (1) is usually reduced to another whose coefficient matrix is an H-matrix.

A detailed discussion of the enclosure methods and preconditioning techniques known in 1990 is in [3], together with an analysis of the approximation power of the methods. Other recent advances concern a method by Hansen [2], simplified and refined by Rohn [5] and Ning and Kearfott [4], giving optimal enclosures for problems involving an H-matrix with diagonal midpoint, arising by midpoint preconditioning.

Recently, Shary [6] introduced a new algorithm (which he called the "algebraic approach") for enclosing the solution set (1) when $\mathbf{A}$ is an H-matrix. While it is also an iterative method, he empirically observed that, in exact arithmetic, the limit is usually achieved in a finite number of iterations. In this respect, the method resembles the conjugate gradient method for linear (noninterval) equations.

In the following, we rederive Shary's method in a way that makes the finite termination property explicit. We also show that the limit interval vector of Shary's algebraic approach is the fixed-point inverse of $\mathbf{A}$ applied to $\mathbf{b}$, as defined in Neumaier [3].

In this paper, notation is as in [3], except that interval quantities are in bold face and intermediate computations may result in *improper interval vectors,* where some lower bound exceeds the corresponding upper bound. We write $\mathbb{IR}^n$ for the set of proper interval vectors and $\mathbb{I}'\mathbb{R}^n$ for the set of proper or improper interval vectors of size n.

2. A new derivation of Shary's method.

Shary's algebraic method is based on the fixed-point equation

$$\mathbf{z} = \mathbf{M}\mathbf{z} + G^{-1}\mathbf{b}, \tag{2}$$

*Received by the editors November 12, 1998; accepted for publication (in revised form) by I.C.F. Ipsen March 17, 1999; published electronically March 21, 2000.
http://www.siam.org/journals/simax/21-4/34747.html
†Institut für Mathematik, Universität Wien, Strudlhofgasse 4, A-1090 Wien, Austria (neum@cma.univie.ac.at, http://solon.cma.univie.ac.at/~neum).

where $\mathbf{M} \in \mathbb{IR}^{n\times n}$ is an interval matrix, $\mathbf{b} \in \mathbb{IR}^n$ is an interval vector, and G is a diagonal matrix with real, nonzero entries. To define Shary's algorithms in a way that makes the finite termination property apparent, we introduce some machinery.

Shary uses the concept of immersion that identifies an interval vector $\mathbf{x} = [\underline{x}, \bar{x}] \in \mathbb{I}'\mathbb{R}^n$ with the real vector $\left(\begin{smallmatrix} -\underline{x} \\ \bar{x} \end{smallmatrix}\right) \in \mathbb{R}^{2n}$; the minus sign is needed to preserve an order structure. While immersion respects addition, subtraction in $\mathbb{R}^{2n}$ does not correspond to subtraction in $\mathbb{I}'\mathbb{R}^n$, but rather to *inner subtraction* of two interval vectors $\mathbf{x}, \mathbf{y} \in \mathbb{I}'\mathbb{R}^n$, defined by

$$\mathbf{x} \stackrel{o}{-} \mathbf{y} := [\underline{x} - \underline{y}, \bar{x} - \bar{y}]. \tag{3}$$

Note that the result may be an improper interval vector even when $\mathbf{x}, \mathbf{y}$ are proper.

LEMMA 2.1.

$$\mathbf{p} = \mathbf{x} \stackrel{o}{-} \mathbf{z} \;\Leftrightarrow\; \mathbf{z} = \mathbf{x} \stackrel{o}{-} \mathbf{p} \;\Leftrightarrow\; \mathbf{x} = \mathbf{z} + \mathbf{p}, \tag{4}$$

$$(\mathbf{z} + \mathbf{q}) \stackrel{o}{-} (\mathbf{z} + \mathbf{p}) = \mathbf{q} - \mathbf{p}. \tag{5}$$

Proof. Insert the definitions. □

The relation with the product is in general quite complicated. However, it turns out that for the present application one needs only a special case that resembles block matrix multiplication: We define the *extended product*

$$\hat{B} * \mathbf{x} := [B^1\underline{x} + B^2\bar{x},\ B^3\underline{x} + B^4\bar{x}] \tag{6}$$

of a real $(2n \times 2n)$-matrix

$$\hat{B} = \begin{pmatrix} B^1 & B^2 \\ B^3 & B^4 \end{pmatrix}$$

with four $n \times n$ blocks $B^1, \ldots, B^4$ with an interval vector $\mathbf{x} = [\underline{x}, \bar{x}] \in \mathbb{I}'\mathbb{R}^n$. Again, the result may be improper even when $\mathbf{x}$ is proper.

LEMMA 2.2.

$$\hat{B} * (\hat{C} * \mathbf{x}) = (\hat{B}\hat{C}) * \mathbf{x}, \tag{7}$$

$$\hat{B}^{-1} * (\hat{B} * \mathbf{x}) = \mathbf{x} \qquad \textit{if } \hat{B} \textit{ is nonsingular,} \tag{8}$$

$$\hat{B} * \mathbf{x} \stackrel{o}{-} \hat{C} * \mathbf{x} = (\hat{B} - \hat{C}) * \mathbf{x}. \tag{9}$$

Proof. The proof follows immediately from corresponding matrix properties in dimension $2n$. □

The importance of the extended product stems from the fact that it can be used to express standard interval matrix-vector products $\mathbf{Mx}$.

PROPOSITION 2.3. *For any $\mathbf{M} \in \mathbb{IR}^{n\times n}$ and $\mathbf{x} \in \mathbb{IR}^n$, there is a $(2n \times 2n)$-matrix*

$$\hat{M} = \begin{pmatrix} M^1 & M^2 \\ M^3 & M^4 \end{pmatrix} \qquad \textit{with } M^l_{ik} \in \{\underline{M}_{ik}, \overline{M}_{ik}, 0\} \textit{ for all } i,k,l$$

such that $\mathbf{Mx} = \hat{M} * \mathbf{x}$.

Proof. The key observation is that for two intervals $\mathbf{a}, \mathbf{b} \in \mathbb{IR}$,

$$\underline{\mathbf{ab}} = \min\{\underline{a}\underline{b}, \underline{a}\overline{b}, \overline{a}\underline{b}, \overline{a}\overline{b}\} = a^1\underline{b} + a^2\overline{b},$$

where (a^1, a^2) is one of $(\underline{a}, 0)$, $(0, \underline{a})$, $(\overline{a}, 0)$, and $(0, \overline{a})$, depending on which term in the min expression is the smallest. Therefore,

$$\underline{\mathbf{M}_{ik}\mathbf{x}_k} = M^1_{ik}\underline{x}_k + M^2_{ik}\bar{x}_k,$$

where $M^1_{ik}, M^2_{ik} \in \{\underline{M}_{ik}, \overline{M}_{ik}, 0\}$. Since

$$\underline{(\mathbf{Mx})_i} = \sum_{k=1}^{n} \underline{\mathbf{M}_{ik}\mathbf{x}_k} = \sum_{n=1}^{n} (M^1_{ik}\underline{x}_k + M^2_{ik}\bar{x}_k) = (M^1\underline{x} + M^2\bar{x})_i$$

for all i, we find $\underline{\mathbf{Mx}} = M^1\underline{x} + M^2\bar{x}$. By a similar argument, $\overline{\mathbf{Mx}} = M^3\underline{x} + M^4\bar{x}$, where $(M^3_{ik}, M^4_{ik}) \in \{\underline{M}_{ik}, \overline{M}_{ik}, 0\}$. ∎

Note that $\hat{M}$ depends on x and is not always unique; however, it is not difficult to extract from the proof an explicit algorithm for computing some $\hat{M}$ given $\mathbf{M}$ and $\mathbf{x}$. The fact that, independent of $\mathbf{x}$, there are only finitely many possible choices for $\hat{M}$ implies that the end points of an interval matrix-vector product are piecewise linear functions of the end points of the factors. In particular, *unless* $\mathbf{x}$ *happens to lie on one of the hypersurfaces where the linear pieces match,* $\hat{M}$ *is constant in a neighborhood of* $\mathbf{x}$. As we shall see, these observations explain the behavior of Shary's algorithm in practice.

We now use the extended product to give a new derivation of Shary's algorithm from which the finite termination property is apparent. Let $\mathbf{x}$ be an approximation to a solution $\mathbf{z}$ of the fixed-point equation (2) and suppose that the matrix $\hat{M}$ of Proposition 2.3 satisfies both

$$\mathbf{Mx} = \hat{M} * \mathbf{x} \quad \text{and} \quad \mathbf{Mz} = \hat{M} * \mathbf{z}. \tag{10}$$

As mentioned above, this is the generic case when $\mathbf{x}$ and $\mathbf{z}$ are sufficiently close and on the same linear piece of the multiplication operator.

THEOREM 2.4. *If* (10) *holds and* $\hat{M} - I$ *is invertible, then*

$$\mathbf{z} = \mathbf{x} \stackrel{o}{-} (\hat{M} - I)^{-1} * (\mathbf{Mx} + G^{-1}\mathbf{b} \stackrel{o}{-} \mathbf{x}). \tag{11}$$

Proof. Write $\mathbf{p} := \mathbf{x} \stackrel{o}{-} \mathbf{z}$, so that $\mathbf{x} = \mathbf{z} + \mathbf{p}$. Then

$$\begin{aligned}
\mathbf{Mx} + G^{-1}\mathbf{b} \stackrel{o}{-} \mathbf{x} &= \hat{M} * (\mathbf{z} + \mathbf{p}) + G^{-1}\mathbf{b} \stackrel{o}{-} (\mathbf{z} + \mathbf{p}) && \text{by (10)} \\
&= \hat{M} * \mathbf{z} + \hat{M} * \mathbf{p} + G^{-1}\mathbf{b} \stackrel{o}{-} (\mathbf{z} + \mathbf{p}) \\
&= \mathbf{Mz} + G^{-1}\mathbf{b} + \hat{M} * \mathbf{p} \stackrel{o}{-} (\mathbf{z} + \mathbf{p}) && \text{by (10)} \\
&= (\mathbf{z} + \hat{M} * \mathbf{p}) \stackrel{o}{-} (\mathbf{z} + \mathbf{p}) && \text{by (2)} \\
&= \hat{M} * \mathbf{p} \stackrel{o}{-} \mathbf{p} && \text{by (5)} \\
&= (\hat{M} - I) * \mathbf{p} && \text{by (9).}
\end{aligned}$$

Solving for $\mathbf{p}$ using (8) gives

$$\mathbf{p} = (\hat{M} - I)^{-1} * (\mathbf{Mx} + G^{-1}\mathbf{b} \stackrel{o}{-} \mathbf{x}),$$

and since $\mathbf{x} = \mathbf{z} + \mathbf{p}$ implies $\mathbf{z} = \mathbf{x} \stackrel{o}{-} \mathbf{p}$, the assertion (11) follows. ∎

Theorem 2.4 suggests the iteration

$$\begin{aligned} \mathbf{x}^{k+1} &= \mathbf{x}^k \stackrel{o}{-} (\hat{M}_k - I)^{-1} * (\mathbf{M}\mathbf{x}^k + G^{-1}\mathbf{b} \stackrel{o}{-} \mathbf{x}^k), \\ &\quad \text{with } \hat{M}_k * \mathbf{x}_k = \mathbf{M}\mathbf{x}_k \text{ from Proposition 2.3.} \end{aligned} \tag{12}$$

With the initialization

$$\mathbf{x}^0 = (\operatorname{mid} \mathbf{A})^{-1}\mathbf{b}, \tag{13}$$

this is just the algebraic method (with damping factor $\tau = 1$), as defined on p. 129 of Shary [6]. In particular, Theorem 2.4 implies that as soon as $\mathbf{x}_k$ reaches the neighborhood of $\mathbf{z}$ that ensures (10), the method produces $\mathbf{x}^{k+1} = \mathbf{z}$ in the next step.

Shary proves convergence of his method under a technical assumption (stated in the proof of Theorem 6.1 in [6]) that is satisfied if the entries of $\mathbf{M}$ are sufficiently narrow. Together with Theorem 2.4, this gives finite termination with probability 1 (namely, unless $\mathbf{z}$ lies on two linear pieces of the multiplication operator $\mathbf{M}$). (One cannot expect finite termination when a damping factor $\tau < 1$ is used, since then only part of the discrepancy from the exact result is corrected.)

3. *H*-matrices and the fixed-point inverse. A matrix $\mathbf{A} \in \mathbb{IR}^{n\times n}$ is called an *H-matrix* iff the comparison matrix $\langle \mathbf{A} \rangle$ defined by

$$\langle \mathbf{A} \rangle_{ii} = \langle \mathbf{A}_{ii} \rangle = \min\{|\alpha| \mid \alpha \in \mathbf{A}_{ii}\},$$

$$\langle \mathbf{A} \rangle_{ik} = -|\mathbf{A}_{ik}| = -\max\{|\alpha| \mid \alpha \in \mathbf{A}_{ik}\} \quad \text{for } i \neq k$$

is nonsingular and its inverse is nonnegative, $\langle \mathbf{A} \rangle^{-1} \geq 0$. As a consequence, $0 \notin \mathbf{A}_{ii}$ for all i.

For H-matrices, the theory in Neumaier [3, Chapter 4] shows that the best enclosure that can be achieved with stationary iterations based on triangular splitting is the *fixed-point solution set* $\mathbf{A}^F\mathbf{b}$, defined as the unique solution z of the interval equations

$$\mathbf{z}_i = \left(\mathbf{b}_i - \sum_{k\neq i} \mathbf{A}_{ik}\mathbf{z}_k \right) / \mathbf{A}_{ii} \qquad (i = 1, \ldots, n) \tag{14}$$

(see Theorem 4.4.4 in [3]). $\mathbf{A}^F\mathbf{b}$ exists for all H-matrices $\mathbf{A}$ and is computable by means of the (then converging) interval Gauss–Seidel iteration, but when $\langle \mathbf{A} \rangle$ is ill-conditioned, this iteration converges very slowly. It turns out that for a suitable choice of $\mathbf{M}$ and G in (2), Shary's algorithm also produces the fixed-point solution set $\mathbf{z} = \mathbf{A}^F\mathbf{b}$, but usually much faster.

Shary defines the *deviation* $\operatorname{dev}(\mathbf{a})$ of an interval $\mathbf{a} = [\underline{a}, \bar{a}] \in \mathbb{IR}$ from 0 to be the number

$$\operatorname{dev}(\mathbf{a}) := \begin{cases} \underline{a} & \text{if } |\underline{a}| \geq |\bar{a}|, \\ \bar{a} & \text{otherwise.} \end{cases} \tag{15}$$

Using the deviation, the quotient of two intervals can be represented as the solution of a univariate fixed-point equation.

LEMMA 3.1. *Let* $\mathbf{r}, \mathbf{a} \in \mathbb{IR}, 0 \notin \mathbf{a}$. *Then* $\mathbf{z} = \mathbf{r}/\mathbf{a}$ *satisfies the equation*

$$\mathbf{z} = \operatorname{dev}(\mathbf{a})^{-1} \left((\operatorname{dev}(\mathbf{a}) - \mathbf{a})\mathbf{z} + \mathbf{r} \right). \tag{16}$$

Proof. Case 1. If $\mathbf{a} > 0,\ \mathbf{r} > 0$, then

$$\operatorname{dev}(\mathbf{a}) = \bar{a}, \quad \mathbf{z} = [\underline{r}/\bar{a}, \bar{r}/\underline{a}],$$

and the right-hand side equals

$$\begin{aligned} \bar{a}^{-1}\left([0, \bar{a} - \underline{a}][\underline{r}/\bar{a}, \bar{r}/\underline{a}] + [\underline{r}, \bar{r}]\right) &= \bar{a}^{-1}\left([0, (\bar{a} - \underline{a})\bar{r}/\underline{a}] + [\underline{r}, \bar{r}]\right) \\ &= \bar{a}^{-1}[\underline{r}, \bar{a}\bar{r}/\underline{a}] = [\underline{r}/\bar{a}, \bar{r}/\underline{a}] = \mathbf{z}. \end{aligned}$$

Case 2. If $\mathbf{a} > 0,\ \mathbf{r} \ni 0$, then

$$\operatorname{dev}(\mathbf{a}) = \bar{a}, \quad \mathbf{z} = [\underline{r}/\underline{a}, \bar{r}/\underline{a}],$$

and the right-hand side equals

$$\begin{aligned} \bar{a}^{-1}\left([0, \bar{a} - \underline{a}][\underline{r}/\underline{a}, \bar{r}/\underline{a}] + [\underline{r}, \bar{r}]\right) &= \bar{a}^{-1}\left([(\bar{a} - \underline{a})\underline{r}/\underline{a}, (\bar{a} - \underline{a})\bar{r}/\underline{a}] + [\underline{r}, \bar{r}]\right) \\ &= \bar{a}^{-1}[\bar{a}\underline{r}/\underline{a}, \bar{a}\bar{r}/\underline{a}] = [\underline{r}/\underline{a}, \bar{r}/\underline{a}] = \mathbf{z}. \end{aligned}$$

The other cases can be reduced to one of these two by changing the signs of $\mathbf{r}$ and/or $\mathbf{a}$. □

As explained in Shary [6, pp. 129–130], Shary's algorithm for enclosing the solution set (1) is based on the fixed-point equation (2), where

$$G := \operatorname{Diag}\left(\operatorname{dev}(\mathbf{A}_{ii})\right), \quad \mathbf{M} = G^{-1}(G - \mathbf{A}), \tag{17}$$

and the spectral radius $\rho(|\mathbf{M}|)$ of $|\mathbf{M}|$ is assumed to be less than one. (Absolute values are taken componentwise.) The condition $\mathbf{A}_{ii} \neq 0$ is also needed to ensure that G is invertible. (Some of Shary's theory is more general, but the only situation worked out in the algorithmic stage is the one stated here.)

THEOREM 3.2. *The matrix $\mathbf{A} \in \mathbb{IR}^{n\times n}$ is an H-matrix iff the matrix G in* (17) *is invertible and the spectral radius of $|\mathbf{M}|$ is less than one. In this case, $\mathbf{z} = \mathbf{A}^F\mathbf{b}$ is the unique solution of the fixed-point equation* (2).

Proof. Suppose first that G is invertible and $\rho(|\mathbf{M}|) < 1$. By the Perron–Frobenius theory, $I - |\mathbf{M}|$ is invertible and its inverse is nonnegative. Since G is diagonal with positive diagonal entries and $\mathbf{A} = G(I - \mathbf{M})$, we have $\langle\mathbf{A}\rangle = \langle G\rangle(I - |\mathbf{M}|)$ and $\langle\mathbf{A}\rangle^{-1} = (I - |\mathbf{M}|)^{-1}\langle G\rangle^{-1} \geq 0$. Hence $\mathbf{A}$ is an H-matrix.

Conversely, suppose that $\mathbf{A}$ is an H-matrix. Then $0 \neq \mathbf{A}_{ii}$ whence $\operatorname{dev}(\mathbf{A}_{ii}) \neq 0$ and G is invertible. Moreover, $I - |\mathbf{M}| = \langle G\rangle^{-1}\langle\mathbf{A}\rangle$ is invertible, with inverse $(I - |\mathbf{M}|)^{-1} = \langle\mathbf{A}\rangle^{-1}\langle G\rangle \geq 0$. Again by the Perron–Frobenius theory, this implies that $\rho(|\mathbf{M}|) < 1$. Now let $\mathbf{z} = \mathbf{A}^F\mathbf{b}$. By (14) we have $\mathbf{z}_i = \mathbf{r}_i/\mathbf{A}_{ii}$, where

$$\mathbf{r}_i = \mathbf{b}_i - \sum_{k\neq i} \mathbf{A}_{ik}\mathbf{z}_k. \tag{18}$$

Using $\operatorname{dev}(\mathbf{A}_{ii}) = G_{ii}$ we find

$$\begin{aligned} \mathbf{z}_i &= G_{ii}^{-1}\left((G_{ii} - \mathbf{A}_{ii})\mathbf{z}_i + \mathbf{r}_i\right) && \text{by Lemma 3.1} \\ &= G_{ii}^{-1}\left((G_{ii} - \mathbf{A}_{ii})\mathbf{z}_i - \sum_{k\neq i}\mathbf{A}_{ik}\mathbf{z}_k + \mathbf{b}_i\right) && \text{by (18)} \\ &= G_{ii}^{-1}\left((G - \mathbf{A})\mathbf{z} + \mathbf{b}\right)_i && \text{since } G \text{ is diagonal} \\ &= \left(\mathbf{M}\mathbf{z} + G^{-1}\mathbf{b}\right)_i && \text{by (17).} \end{aligned}$$

Hence $\mathbf{z}$ is a solution of the fixed-point equation (2). Since $|\mathbf{M}|$ has spectral radius < 1, this equation has a unique solution, so $\mathbf{z} = \mathbf{A}^F\mathbf{b}$ is the only solution of (2). ∎

Together with the good overestimation properties of $\mathbf{A}^F\mathbf{b}$ (derived in Neumaier [3]) when $\mathbf{A}$ is strictly diagonally dominant, this result explains the good numerical properties of the enclosures computed in Shary [6].

The theory in [3] suggests, as an alternative to (13), the initialization

$$\mathbf{x}^0 = [-u, u], \quad \text{where } u = \langle \mathbf{A} \rangle^{-1} |\mathbf{b}|, \tag{19}$$

since this is always an enclosure for $\mathbf{A}^F\mathbf{b}$, and we already have $\mathbf{x}^0 = \mathbf{A}^F\mathbf{b}$ when the midpoint of $\mathbf{b}$ is zero.

Consider, for example, the interval system

$$\mathbf{A} = \begin{pmatrix} [2,4] & [-2,0] \\ [-1,0] & [2,4] \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} [1,2] \\ [1,2] \end{pmatrix}$$

by Barth and Nuding [1]. We find

$$u = \langle \mathbf{A} \rangle^{-1} |\mathbf{b}| = \begin{pmatrix} 2 & -2 \\ -1 & 2 \end{pmatrix}^{-1} \begin{pmatrix} 2 \\ 2 \end{pmatrix} = \begin{pmatrix} 4 \\ 3 \end{pmatrix},$$

and Shary's method with starting point (19) converges in two iterations:

$$\mathbf{x}^0 = \begin{pmatrix} [-4,4] \\ [-3,3] \end{pmatrix}, \quad \mathbf{x}^1 = \begin{pmatrix} [2,4] \\ [1.5,3] \end{pmatrix}, \quad \mathbf{x}^2 = \begin{pmatrix} [0.25,4] \\ [0.25,3] \end{pmatrix} = \mathbf{A}^F\mathbf{b}.$$

One sees that, unlike for Gauss–Seidel-like methods, convergence in Shary's method need not be inclusion monotone. A weaker order relation under which the method is monotone is given in Shary [6].

The following questions remain before the method is fully understood:

1. Is the H-matrix property, or at least $\rho(|\mathbf{M}|) < \frac{1}{2}$, sufficient for the convergence of (12)?

2. Are there examples where the iteration (12) converges to $\mathbf{z} = \mathbf{A}^F\mathbf{b}$ but $\mathbf{x}^k \neq \mathbf{z}$ for all k?

Finally we remark that the finite termination property of the algebraic method proved here for the computation of enclosures of the solution set is applicable (with hardly any change in derivation and proof) also for the algebraic method applied in Shary [7] to the computation of tolerance sets (inner solutions) for linear interval equations. It is likely that the above ideas will also be useful for the analysis of the algebraic method applied to other generalized solution sets.

Acknowledgment. I want to thank Sergey Shary for useful comments on an earlier version of this paper.

REFERENCES

[1] W. Barth and E. Nuding, *Optimale Lösung von Intervallgleichungssystemen*, Computing, 12 (1974), pp. 117–125.

[2] E. Hansen, *Bounding the solution of interval linear equations*, SIAM J. Numer. Anal., 29 (1992), pp. 1493–1503.

[3] A. Neumaier, *Interval Methods for Systems of Equations*, Cambridge University Press, Cambridge, UK, 1990.

[4] S. Ning and R.B. Kearfott, *A comparison of some methods for solving linear interval equations*, SIAM J. Numer. Anal., 34 (1997), pp. 1289–1305.

[5] J. Rohn, *Cheap and tight bounds: The recent result by E. Hansen can be made more efficient*, Interval Comput., 4 (1993), pp. 13–21.
[6] S.P. Shary, *Algebraic approach in the "outer problem" for interval linear equations*, Reliab. Comput., 3 (1997), pp. 103–135.
[7] S.P. Shary, *Algebraic approach to the interval linear static identification, tolerance, and control problems, or one more application of Kaucher arithmetic*, Reliab. Comput. 2 (1996), pp. 3–33.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 4, pp. 1163–1184
© 2000 Society for Industrial and Applied Mathematics

A CONTINUATION METHOD FOR A RIGHT DEFINITE TWO-PARAMETER EIGENVALUE PROBLEM*

BOR PLESTENJAK†

Abstract. The continuation method has been successfully applied to the classical $Ax = \lambda x$ and to the generalized $Ax = \lambda Bx$ eigenvalue problems. Shimasaki applied the continuation method to the right definite two-parameter problem, which resulted in a discretization of a two-parameter Sturm–Liouville problem. We show that the continuation method can be used for a general right definite two-parameter problem and we give a sketch of the algorithm. For a local convergent method we use the tensor Rayleigh quotient iteration (TRQI), which is a generalization of the Rayleigh iterative method to two-parameter problems. We show its convergence and compare it with Newton's method and with the generalized Rayleigh quotient iteration (GRQI), studied by Ji, Jiang, and Lee.

Key words. right definite two-parameter problem, continuation method, Newton's method, Rayleigh quotient iteration

AMS subject classifications. 65F15, 65H20, 15A69, 15A18

PII. S0895479898346193

1. Introduction.

We consider a two-parameter eigenvalue problem

$$\begin{aligned} A_1x &= \lambda B_1x + \mu C_1x, \\ A_2y &= \lambda B_2y + \mu C_2y, \end{aligned} \tag{1.1}$$

where A_i, B_i, C_i are symmetric $n_i \times n_i$ matrices over $\mathbb{R}$, $i = 1, 2$. We also require a definiteness condition

$$D(x,y) := \begin{vmatrix} x^TB_1x & x^TC_1x \\ y^TB_2y & y^TC_2y \end{vmatrix} \geq \delta_1 > 0 \tag{1.2}$$

for all vectors $\|x\| = \|y\| = 1$. We call problem (1.1) *right definite* [27] if the condition (1.2) holds. If in addition the matrices B_1 and C_2 are positive definite, then we call the problem *diagonal right definite*.

We say that (λ, μ) is an *eigenvalue* of the problem (1.1) if

$$\ker(A_i - \lambda B_i - \mu C_i) \neq \{0\}, \quad i = 1, 2.$$

If $\dim \ker(A_i - \lambda B_i - \mu C_i) = 1$ for $i = 1, 2$, then (λ, μ) is a *simple eigenvalue* of (1.1).

On the tensor product space $V := \mathbb{R}^{n_1} \otimes \mathbb{R}^{n_2}$ of the dimension $N := n_1n_2$ we define operator determinants

$$\Delta_0 = \begin{vmatrix} B_1^\dagger & C_1^\dagger \\ B_2^\dagger & C_2^\dagger \end{vmatrix}, \quad \Delta_1 = \begin{vmatrix} A_1^\dagger & C_1^\dagger \\ A_2^\dagger & C_2^\dagger \end{vmatrix}, \quad \Delta_2 = \begin{vmatrix} B_1^\dagger & A_1^\dagger \\ B_2^\dagger & A_2^\dagger \end{vmatrix},$$

where $A_1^\dagger, A_2^\dagger, B_1^\dagger, B_2^\dagger, C_1^\dagger, C_2^\dagger$ are the induced linear transformations on V. For instance, $A_1^\dagger$ is defined on a decomposable tensor $x \otimes y$ by $A_1^\dagger(x \otimes y) = A_1x \otimes y$,

*Received by the editors October 19, 1998; accepted for publication (in revised form) by M. Chu August 18, 1999; published electronically March 30, 2000.
http://www.siam.org/journals/simax/21-4/34619.html

†IMFM/TCS, University of Ljubljana, Jadranska 19, SI-1000 Ljubljana, Slovenia (bor.plestenjak@fmf.uni-lj.si).

and this definition is extended to all of V by linearity and continuity. Similarly, $A_2^{\dagger}(x \otimes y) = x \otimes A_2 y$.

A two-parameter system (1.1) is called nonsingular if the corresponding operator determinant Δ_0 is invertible. In the case of a nonsingular two-parameter system, the problem (1.1) is equivalent to the simultaneous problem in V:

$$\begin{aligned} \Delta_1 z &= \lambda \Delta_0 z, \\ \Delta_2 z &= \mu \Delta_0 z \end{aligned} \tag{1.3}$$

for decomposable tensors z (see Atkinson [3] for details). Atkinson [3, Theorem 7.8.2] also showed that right definiteness is equivalent to the condition that the operator determinant Δ_0 is positive definite.

Several numerical algorithms can be applied to the right definite two-parameter eigenvalue problem (1.1). Bohte [9] used Newton's method to find an eigenvalue pair. The simultaneous problem (1.3) of the right definite problem (1.1) can be treated by standard numerical methods [25, 16]. Müller [21] used the continuation method to compute eigenvalue curves starting from a given eigenvalue, and Browne and Sleeman [11] used the gradient method [7, 8]. Blum and Chang [6] derived the minimum residual quotient iteration (MRQI) for the problem $Ax = \lambda Bx + \mu Cx$ subject to $\|x\| = 1$ and $f(x) = 0$, where f is a real functional. Ji, Jiang, and Lee [18] generalized their approach to the right definite two-parameter problem (1.1) and derived the generalized Rayleigh quotient iteration (GRQI).

Multiparameter eigenvalue problems arise in a variety of applications [2], particularly in mathematical physics when the method of separation of variables is used to solve boundary value problems [27]. When the separation constants cannot be decoupled, two-parameter Sturm–Liouville problems of the form

$$-\Big(p_i(x_i)y_i'(x_i)\Big)' + q_i(x_i)y_i(x_i) = \Big(\lambda a_{i1}(x_i) + \mu a_{i2}(x_i)\Big)y_i(x_i), \tag{1.4}$$

where $x_i \in [a_i, b_i]$, with boundary conditions

$$y_i(a_i)\cos\alpha_i - y_i'(a_i)\sin\alpha_i = 0, \quad 0 \le \alpha_i < \pi,$$

$$y_i(b_i)\cos\beta_i - y_i'(b_i)\sin\beta_i = 0, \quad 0 < \beta_i \le \pi,$$

will arise, where $p_i', q_i, a_{i1}, a_{i2}$ are real valued and continuous for $i = 1, 2$. Many numerical methods for two-parameter Sturm–Liouville problems have been proposed; see, for example, [4, 5, 14, 15].

By using finite differences, problem (1.4) can be converted into a right definite problem (1.1), where matrices A_i, B_i, C_i satisfy the following conditions, which we will denote by TBC as in [17]: A_i is an irreducible tridiagonal matrix and B_i, C_i are diagonal matrices, such that diagonal elements are all strictly positive or all strictly negative for $i = 1, 2$. Some numerical methods have been developed especially for problems which satisfy TBC: Ji [17] uses the two-dimensional bisection method and Shimasaki [23] uses the continuation method with quadratic convergence rate [24] (cf. [1] for details about the continuation method).

The continuation method has been successfully applied to the one-parameter eigenvalue problems; see, for example, [12, 13, 19, 20]. We show that the continuation method with a homotopy similar to the one used in [23] can be applied to a general two-parameter problem (1.1), which does not necessarily satisfy TBC.

In section 2 we show that every right definite problem can be transformed to a diagonal right definite problem by only a simple linear substitution of parameters λ and μ. In section 3 we construct a homotopy, derive bounds for the minimum distance between eigenvectors, and give a sketch of the algorithm based on the continuation method. For a local convergent method in the continuation method we use tensor Rayleigh quotient iteration (TRQI). We derive TRQI in section 4, show its convergence, and compare it with another generalization of the Rayleigh quotient, suggested by Ji, Jiang, and Lee [18]. In section 5 we give a numerical example which reflects the behavior of our continuation method.

2. Transformation to a diagonal right definite problem.

LEMMA 2.1. *If* (1.1) *is a right definite problem, then*

(i) *at least one of the matrices B_1 and B_2 is definite (positive or negative);*

(ii) *at least one of the matrices C_1 and C_2 is definite (positive or negative).*

Proof. First we show that at least one of the matrices B_1, B_2 is definite.

If not, there exist vectors $x_0, y_0 \neq 0$ such that $x_0^T B_1 x_0 = 0$ and $y_0^T B_2 y_0 = 0$, which gives the counterexample $D(x_0, y_0) = 0$. It follows that at least one of the matrices B_1, B_2 is definite. In the same manner we show that at least one of the matrices C_1, C_2 is definite. □

Since problem (1.1) is right definite, there exist $i, j \in \{1, 2\}$ such that B_i and C_j are definite matrices (positive or negative). This is enough to transform the right definite problem (1.1) to a diagonal right definite problem.

LEMMA 2.2. *If problem* (1.1) *is right definite, then it is possible to transform it to a diagonal right definite problem by a linear substitution of parameters λ and μ.*

Proof. Let us first consider all possibilities with the positive definite matrix B_1.

(i) The matrix C_2 is positive definite. This is already the situation we are looking for, so no transformation is needed.

(ii) The matrix C_2 is negative definite. Since for all vectors $x, y \neq 0$ we have $(x^T B_1 x)(y^T C_2 y) < 0$ and $D(x, y) > 0$, it follows that both matrices C_1 and B_2 are definite. If not, there would exist a pair $x_0, y_0 \neq 0$ such that $(x_0^T C_1 x_0)(y_0^T B_2 y_0) = 0$ and $D(x_0, y_0) < 0$, but this is not possible. Since B_2 is a definite matrix, there exists $k \in \mathbb{R}$, such that $C_2 + kB_2$ is a positive definite matrix.

It is easy to see that the substitution $\lambda \mapsto \lambda + k\mu$ gives

$$A_1 x = \lambda B_1 x + \mu(C_1 + kB_1)x,$$
$$A_2 y = \lambda B_2 y + \mu(C_2 + kB_2)y,$$

which is a desired situation.

(iii) The matrix C_1 is definite (positive or negative). From $D(x, y) > 0$ and $x^T B_1 x > 0$ for all vectors $x, y \neq 0$, it follows that

$$y^T C_2 y - \frac{x^T C_1 x}{x^T B_1 x} y^T B_2 y > 0.$$

Consequently, for an arbitrarily chosen vector $x_0 \neq 0$, the matrix

$$C_2 - \frac{x_0^T C_1 x_0}{x_0^T B_1 x_0} B_2$$

is positive definite and the substitution $\lambda \mapsto \lambda - k\mu$ gives

$$A_1 x = \lambda B_1 x + \mu(C_1 - kB_1)x,$$
$$A_2 y = \lambda B_2 y + \mu(C_2 - kB_2)y,$$

where

$$k = \frac{x_0^T C_1 x_0}{x_0^T B_1 x_0}.$$

If B_1 is a negative definite matrix, we multiply both equations in (1.1) by -1. This turns B_1 into a positive definite matrix while the determinant $D(x,y)$ remains the same. Once B_1 is a positive definite matrix, we can use (i), (ii), or (iii). Similarly, if B_2 is definite, we exchange equations in (1.1) and multiply one of the equations by -1 so that $D(x,y)$ remains unchanged. This shows that it is enough to prove the lemma only for the case when the matrix B_1 is definite. ∎

3. A continuation method. As shown in the previous section, we can assume that the matrices B_1 and C_2 in (1.1) are positive definite. In this case we construct the following homotopy:

$$H : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \mathbb{R} \times \mathbb{R} \times [0,1] \longrightarrow \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \mathbb{R} \times \mathbb{R},$$

$$H(x,y,\lambda,\mu,t) := \begin{bmatrix} (1-t)W_1 x + tA_1 x - \lambda B_1 x - t\mu C_1 x \\ (1-t)W_2 y + tA_2 y - t\lambda B_2 y - \mu C_2 y \\ \frac{1}{2}(1 - x^T x) \\ \frac{1}{2}(1 - y^T y) \end{bmatrix}, \tag{3.1}$$

where W_i are symmetric $n_i \times n_i$ matrices, such that eigenproblems

$$W_1 x = \lambda B_1 x \tag{3.2}$$

and

$$W_2 y = \mu C_2 y \tag{3.3}$$

have n_1 and n_2 distinct eigenvalues, respectively.

It is obvious that a solution of $H(x,y,\lambda,\mu,t) = 0$ is a solution of the two-parameter problem

$$\begin{aligned} (1-t)W_1 x + tA_1 x &= \lambda B_1 x + \mu t C_1 x, \\ (1-t)W_2 y + tA_2 y &= \lambda t B_2 y + \mu C_2 y, \end{aligned} \tag{3.4}$$

which is equal to (1.1) for $t = 1$ and equal to (3.2), (3.3) for $t = 0$. Associated with the problem (3.4) are operator determinants

$$\Delta_0(t) = \begin{vmatrix} B_1^\dagger & tC_1^\dagger \\ tB_2^\dagger & C_2^\dagger \end{vmatrix},$$

$$\Delta_1(t) = \begin{vmatrix} (1-t)W_1^\dagger + tA_1^\dagger & tC_1^\dagger \\ (1-t)W_2^\dagger + tA_2^\dagger & C_2^\dagger \end{vmatrix},$$

$$\Delta_2(t) = \begin{vmatrix} B_1^\dagger & (1-t)W_1^\dagger + tA_1^\dagger \\ tB_2^\dagger & (1-t)W_2^\dagger + tA_2^\dagger \end{vmatrix},$$

where † denotes the induced linear transformation on V.

For a start we will show that the problem (3.4) is right definite for all $t \in [0,1]$. We denote the determinant

$$\begin{vmatrix} x^T B_1 x & t x^T C_1 x \\ t y^T B_2 y & y^T C_2 y \end{vmatrix}$$

by $D(x,y,t)$. It is easy to see that for arbitrary vectors x, y and $t \in [0,1]$ we have

$$\min\{D(x,y,0), D(x,y,1)\} \le D(x,y,t). \tag{3.5}$$

LEMMA 3.1. *The problem* (3.4) *is right definite for all* $t \in [0,1]$.

Proof. As B_1 and C_2 are positive definite matrices and problem (1.1) is right definite, it follows that

$$D(x,y,0), D(x,y,1) > 0$$

for all vectors $x, y \neq 0$. Now (3.5) yields

$$D(x,y,t) > 0$$

for all $x, y \neq 0$. ∎

A simple consequence of Lemma 3.1 is that there exists a constant $\delta_2 > 0$ such that

$$D(x,y,t) \ge \delta_2 \tag{3.6}$$

for all vectors $\|x\| = \|y\| = 1$ and $t \in [0,1]$. It is easy to see from (3.5) that

$$\delta_2 := \min\left\{ \min_{\|x\|=\|y\|=1} D(x,y,0), \min_{\|x\|=\|y\|=1} D(x,y,1) \right\} \tag{3.7}$$

satisfies (3.6).

Let

$$\Gamma := \{(x,y,\lambda,\mu,t) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \mathbb{R} \times \mathbb{R} \times [0,1] \mid H(x,y,\lambda,\mu,t) = 0\}$$

denote a solution set for the homotopy (3.1).

If we denote

$$\begin{aligned} S_1(t) &= (1-t)W_1 + tA_1 - \lambda B_1 - t\mu C_1, \\ S_2(t) &= (1-t)W_2 + tA_2 - t\lambda B_2 - \mu C_2, \end{aligned}$$

then we can write the Jacobian of the homotopy (3.1) as

$$J(x,y,\lambda,\mu,t) = \begin{bmatrix} S_1(t) & 0 & -B_1 x & -tC_1 x \\ 0 & S_2(t) & -tB_2 y & -C_2 y \\ -x^T & 0 & 0 & 0 \\ 0 & -y^T & 0 & 0 \end{bmatrix}. \tag{3.8}$$

DEFINITION 3.2. *We say that* $(x,y,\lambda,\mu,t) \in \Gamma$ *is a singular point of the homotopy* (3.1) *if the Jacobian* $J(x,y,\lambda,\mu,t)$ *is singular.*

LEMMA 3.3. *Singular points of the homotopy* (3.1) *are precisely those points* $(x, y, \lambda, \mu, t) \in \Gamma$ *where* (λ, μ) *is a multiple eigenvalue of* (3.4).

Proof. First we prove that if $(x, y, \lambda, \mu, t) \in \Gamma$ and (λ, μ) is a simple eigenvalue of (3.4), then the Jacobian (3.8) is nonsingular. Let $(z, w, a, b) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \mathbb{R} \times \mathbb{R}$ be such that

$$J(x, y, \lambda, \mu, t) \begin{bmatrix} z \\ w \\ a \\ b \end{bmatrix} = 0.$$

It follows that

$$S_1(t)z - aB_1x - btC_1x = 0, \tag{3.9}$$

$$S_2(t)w - atB_2y - bC_2y = 0, \tag{3.10}$$

$$x^T z = 0, \tag{3.11}$$

$$y^T w = 0. \tag{3.12}$$

Since $H(x, y, \lambda, \mu, t) = 0$,

$$S_1(t)x = 0, \tag{3.13}$$

$$x^T x = 1, \tag{3.14}$$

$$S_2(t)y = 0, \tag{3.15}$$

$$y^T y = 1. \tag{3.16}$$

From (3.13), (3.9), and $S_1(t) = S_1^T(t)$ it follows that

$$0 = z^T S_1(t)x = x^T S_1^T(t)z = x^T S_1(t)z = ax^T B_1 x + btx^T C_1 x. \tag{3.17}$$

Similarly it follows from (3.15) and (3.10) that

$$0 = w^T S_2(t)y = y^T S_2^T(t)w = y^T S_2(t)w = aty^T B_2 y + by^T C_2 y. \tag{3.18}$$

Equations (3.17) and (3.18) form a 2×2 homogeneous linear system for a and b. Lemma 3.1 implies that this linear system is nonsingular; thus the only solution is

$$a = b = 0.$$

From (3.9) and (3.10) it now follows $S_1(t)z = 0$ and $S_2(t)w = 0$. Since (λ, μ) is a simple eigenvalue for (3.4) we have $\operatorname{rank}(S_i(t)) = n_i - 1$ and there exist $\alpha, \beta \in \mathbb{R}$ such that

$$z = \alpha x$$

and

$$w = \beta y.$$

It follows from (3.11), (3.12), (3.14), and (3.16) that $\alpha = \beta = 0$.

For the second part, let (λ, μ) be a multiple eigenvalue of (3.4). Without loss of generality we can assume that $\operatorname{rank}(S_1(t)) < n_1 - 1$. There exists a vector $z \in \mathbb{R}^{n_1}$ such that $z \neq 0$, $S_1(t)z = 0$ and $x^T z = 0$. It follows that

$$J(x, y, \lambda, \mu, t) \begin{bmatrix} z \\ 0 \\ 0 \\ 0 \end{bmatrix} = 0$$

and the Jacobian $J(x, y, \lambda, \mu, t)$ is singular. □

If problem (1.1) satisfies the TBC condition, then it is easy to see that the homotopy (3.1) has no singular points. The reason for this is the well-known fact that irreducible tridiagonal matrices cannot have multiple eigenvalues [22]. Shimasaki [23] makes use of this in his algorithm.

Since problem (3.4) is right definite, it has N real eigenvalues for every $t \in [0, 1]$. Therefore, there exist N homotopy curves

$$\vartheta_i(t) := (x_i(t), y_i(t), \lambda_i(t), \mu_i(t), t), \quad i = 1, \ldots, N,$$

parametrizable with respect to t, such that $\vartheta_i(t) \in \Gamma$. The following theorem, which is a result of Browne and Sleeman, is of major importance, as it shows that $\vartheta_i(t)$ is a holomorphic function of t.

THEOREM 3.4 (see Browne and Sleeman [10]). *Eigenvalues $\lambda_i(t), \mu_i(t)$ and vectors $x_i(t), y_i(t)$, for $i = 1, \ldots, N$, are holomorphic functions of t on the interval $[0, 1]$ and $x_i(t) \otimes y_i(t)$ form a $\Delta_0(t)$-orthogonal basis of eigenvectors for (3.4), i.e., if $i \neq j$, then*

$$(x_i(t) \otimes y_i(t))^T \Delta_0(t)(x_j(t) \otimes y_j(t)) = 0.$$

Theorem 3.4 implies that we have N disjoint homotopy curves $\vartheta_i(t)$. This allows easy tracking of each curve from $t = 0$ to $t = 1$.

Eigenvalues of problem (3.4) are not necessarily distinct (except for $t = 0$), so the eigenvalue curves $\gamma_i(t) := (\lambda_i(t), \mu_i(t))$, $i = 1, \ldots, N$, can intersect. It follows from Lemma 3.3 that whenever at least two eigenvalue curves $\gamma_i(t)$ intersect, we have a singular point. When we trace an eigenvalue curve from $t = 0$ to $t = 1$ numerically, we must be aware that there is a danger of switching to another curve whenever we hit a singular point. A numerical method finds one eigenvector from the eigensubspace and since the eigensubspace at a singular point is at least two-dimensional, the method can accidentally pick an eigenvector from another homotopy curve.

Since we choose the matrices W_1 and W_2 in such a way that the homotopy has no singular points at $t = 0$, the homotopy has a singular point only at finitely many values $t \in (0, 1]$. The continuity of the eigenvalue curves ensures that we can jump over the singular point at t using $\vartheta_i(t - h)$ as an initial approximation for $\vartheta_i(t + h)$ if h is small enough.

In order to prevent switching from one homotopy curve to another, we will derive a bound for the constant $\eta > 0$, independent of t, such that if $x_i(t) \otimes y_i(t)$, and $x_j(t) \otimes y_j(t)$ are eigenvectors for (3.4) and unit vectors $x_i(t), y_i(t)$, and $x_j(t), y_j(t)$ are parts of homotopy curves $\vartheta_i(t)$ and $\vartheta_j(t)$, respectively, then

$$\max(\|x_i(t) - x_j(t)\|^2, \|y_i(t) - y_j(t)\|^2) \geq \eta$$

for all $i \neq j$ and $t \in [0, 1]$. Then we will write the sketch of an algorithm for calculating all eigentuples of problem (1.1) using the continuation method.

Before we can state the theorem about the existence of such an η, we have to prove some auxiliary results.

LEMMA 3.5. *Let $A \in \mathbb{R}^{n\times n}$ be a positive definite symmetric matrix and let x and y be unit vectors, such that $x^T Ay = 0$. It follows that*

$$\|x - y\|^2 \geq \frac{4\lambda_n}{\lambda_1 + \lambda_n},$$

where λ_1 and λ_n are the greatest and the smallest eigenvalue of A, respectively.

Proof. Since matrix A is symmetric, there exists an orthonormal basis of eigenvectors x_i, such that $Ax_i = \lambda_i x_i$, $i = 1, \ldots, n$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$, and we can write vectors x and y as

$$x = \sum_{i=1}^{n} \alpha_i x_i,$$
$$y = \sum_{i=1}^{n} \beta_i x_i.$$

The condition $x^T Ay = 0$ is now equivalent to

$$\sum_{i=1}^{n} \lambda_i \alpha_i \beta_i = 0. \tag{3.19}$$

We are searching for the minimum of $\|x - y\|^2$. Under conditions $\sum_{i=1}^{n} \alpha_i^2 = 1$ and $\sum_{i=1}^{n} \beta_i^2 = 1$ we have

$$\|x - y\|^2 = \sum_{i=1}^{n} (\alpha_i - \beta_i)^2 = 2 - 2\sum_{i=1}^{n} \alpha_i \beta_i. \tag{3.20}$$

It follows from (3.19) and from the fact that matrix A is positive definite that in (3.20) not all terms in the sum on the right-hand side can be of equal sign. It is easy to see that candidates for which the minimum is attained have the form

$$x = \alpha_1 x_1 + \alpha_n x_n,$$
$$y = -\alpha_1 x_1 + \alpha_n y_n.$$

Equation (3.19), together with $\|x\| = \|y\| = 1$, gives

$$\alpha_1^2 = \frac{\lambda_n}{\lambda_1 + \lambda_n}, \quad \alpha_n^2 = \frac{\lambda_1}{\lambda_1 + \lambda_n},$$

and

$$\|x - y\|^2 = \frac{4\lambda_n}{\lambda_1 + \lambda_n}. \qquad \square$$

LEMMA 3.6. *If x_1, x_2, y_1, y_2 are unit vectors such that*

$$\|(x_1 \otimes y_1) - (x_2 \otimes y_2)\|^2 \geq \epsilon^2,$$

then

$$\max(\|x_1 - x_2\|^2, \|y_1 - y_2\|^2) \geq 2\left(1 - \sqrt{1 - \frac{\epsilon^2}{2}}\right).$$

Proof. Since vectors x_1, x_2, y_1, y_2 are all normalized, we have

$$\|(x_1 \otimes y_1) - (x_2 \otimes y_2)\|^2 = 2 - 2\langle x_1, x_2\rangle\langle y_1, y_2\rangle \geq \epsilon^2.$$

It follows that

$$\langle x_1, x_2\rangle\langle y_1, y_2\rangle \leq 1 - \frac{\epsilon^2}{2}$$

and

$$\min(\langle x_1, x_2\rangle, \langle y_1, y_2\rangle) \leq \sqrt{1 - \frac{\epsilon^2}{2}}.$$

Finally, from

$$\|x_1 - x_2\|^2 = 2 - 2\langle x_1, x_2\rangle,$$

$$\|y_1 - y_2\|^2 = 2 - 2\langle y_1, y_2\rangle,$$

we get

$$\max(\|x_1 - x_2\|^2, \|y_1 - y_2\|^2) \geq 2\left(1 - \sqrt{1 - \frac{\epsilon^2}{2}}\right). \qquad \square$$

Now we are able to prove a theorem on the minimum distance between eigenvectors of the homotopy curves $\vartheta_i(t)$.

THEOREM 3.7. *Let $x_i(t) \otimes y_i(t)$ and $x_j(t) \otimes y_j(t)$ be eigenvectors of* (3.4)*, $i \neq j$ and $t \in [0, 1]$, where unit vectors $x_i(t), y_i(t)$ and $x_j(t), y_j(t)$ are parts of homotopy curves $\vartheta_i(t)$ and $\vartheta_j(t)$, respectively. Then there exists a constant $\eta > 0$, independent of t, such that*

$$\max(\|x_i(t) - x_j(t)\|^2, \|y_i(t) - y_j(t)\|^2) \geq \eta$$

for all $i \neq j$ and $t \in [0, 1]$.

Proof. By Lemma 3.1, the matrix $\Delta_0(t)$ is symmetric and positive definite for all $t \in [0, 1]$. From Theorem 3.4 we have

$$(x_i(t) \otimes y_i(t))^T \Delta_0(t)(x_j(t) \otimes y_j(t)) = 0$$

and we can apply Lemma 3.5. It follows that for every $t \in [0, 1]$ there exists a constant $\epsilon(t) > 0$, such that

$$\|x_i(t) \otimes y_i(t) - x_j(t) \otimes y_j(t)\|^2 \geq \epsilon^2(t).$$

By Lemma 3.6 we have

$$\max(\|x_i(t) - x_j(t)\|^2, \|y_i(t) - y_j(t)\|^2) \geq 2\left(1 - \sqrt{1 - \frac{\epsilon^2}{2}}\right),$$

where

$$\epsilon := \min_{t \in [0,1]} \epsilon(t)$$

and the constant we are looking for is

$$\eta := 2\left(1 - \sqrt{1 - \frac{\epsilon^2}{2}}\right). \qquad \square$$

In order to use Theorem 3.7 to obtain a bound η on the minimum distance between the eigenvectors for different eigenvalues of (3.4) we need information about the extreme eigenvalues of $\Delta_0(t)$. Since $\Delta_0(t)$ is a matrix of order N, the exact computation would demand too much computational work. Therefore, it is better to obtain an upper and a lower bound for the maximum and the minimum eigenvalues of $\Delta_0(t)$, respectively.

The upper bound is easily obtained from the fact that

$$\|\Delta_0(t)\| = \|B_1 \otimes C_2 - t^2 C_1 \otimes B_2\| \leq \|B_1\|\|C_2\| + \|C_1\|\|B_2\|.$$

It follows that $\lambda \leq \|B_1\|\|C_2\| + \|C_1\|\|B_2\|$ for an arbitrary eigenvalue λ of $\Delta_0(t)$, $t \in [0,1]$.

It is not so trivial to obtain a lower bound for the minimum eigenvalue of $\Delta_0(t)$. Let

$$\delta(t) := \min_{\|x\|=\|y\|=1} D(x,y,t).$$

Volkmer [26, Theorem 4.5] showed that

$$\delta(t) \cdot \min(\{n_1, n_2\})^{-1} \leq \lambda$$

for an arbitrary eigenvalue λ of $\Delta_0(t)$.

If follows from (3.6) that $\delta_2 \leq \delta(t)$ for every $t \in [0,1]$. We summarize these results in the following lemma.

LEMMA 3.8. *We consider the right definite problem* (3.4)*, where the assumption* (3.6) *is valid. Let* λ *be an eigenvalue of* $\Delta_0(t)$*,* $t \in [0,1]$*. Then*

$$\delta_2 \cdot \min(\{n_1, n_2\})^{-1} \leq \lambda \leq \|B_1\|\|C_2\| + \|C_1\|\|B_2\|. \tag{3.21}$$

Since $x_i(t)$ and $y_i(t)$ are holomorphic functions of t, for each $\eta > 0$ there exists $h > 0$, such that

$$\|x_i(t) - x_i(t')\|, \|y_i(t) - y_i(t')\| \leq \eta$$

when $|t - t'| \leq h$ and $t, t' \in [0,1]$. If we take h small enough, then $\vartheta_i(t)$ will be a good approximation for $\vartheta_i(t+h)$. If we use a numerical method which converges from the initial approximation $\vartheta_i(t)$ to the nearest solution of (3.4) at $t+h$, then switching from one curve to another is possible only when passing through a singular point.

We propose the following continuation method for tracing a homotopy curve $\vartheta(t) = \vartheta_i(t)$ from $t = 0$ to $t = 1$.

ALGORITHM 1.

Initialization.

(i) Solve (3.2) and (3.3) to obtain an initial eigenvalue $(\lambda(0), \mu(0))$ and initial eigenvectors $x(0)$, $y(0)$.

(ii) Set $t = 0$ and set the initial stepsize h.

Main loop.
Repeat:

1. If $t + h > 1$, then set $h = 1 - t$.
2. Use a numerical method to calculate $\vartheta(t + h)$ using $\vartheta(t)$ as an initial approximation.
3. Calculate $M := \max(\|x(t + h) - x(t)\|^2, \|y(t + h) - y(t)\|^2)$.
4. If
   (a) $M \geq c_1\eta$, where $c_1 \in (0, 1]$ is a prescribed constant and η is the constant from Theorem 3.7, or
   (b) $t + h = 1$ and the Jacobian $J(x(1), y(1), \lambda(1), \mu(1), 1)$ is singular,
   then set t to the last good value, set $h = 2h/3$, and go to step 2.
5. If $M \leq c_2\eta$, where $0 < c_2 < c_1$ is a prescribed constant, then set $h = 3h/2$.
6. Set $t = t + h$.

Until $t \geq 1 - \epsilon_1$, where $\epsilon_1 > 0$ is a prescribed constant.

The criterion in step 4(a) detects if the point $\vartheta(t + h)$ still lies on the same homotopy curve. If $M \geq c_1\eta$, then it is possible that we have switched to another curve. The reason is that either the stepsize h is too large or we have passed a singular point. The solution is to decrease the stepsize h and start again from the last good point t. Instead of halving the stepsize we choose to multiply it by 2/3 in order to avoid another calculation at the point where the problem was detected in step 4. By using this approach we jump over all singular points.

When checking for singular points, we could use Lemma 3.3 and check the singularity of the Jacobian $J(x(t), y(t), \lambda(t), \mu(t), t)$. These checks are necessary only when $t = 1$, since we cannot jump over $t = 1$. The jump is not possible since problem (3.4) is not necessarily positive definite for $t = 1 + h$. The solution at $t = 1$ is to always check the singularity of the Jacobian in addition to the criterion $M \geq c_1\eta$. We do this in step 4(b). In the case of a singular point at $t = 1$, one strategy is to successfully decrease the stepsize h and take $\vartheta(t)$ for the solution at $t = 1$ when t is close enough to 1.

In step 1 we adjust the stepsize if $t + h$ is greater than 1 and in step 5 we increase the stepsize when the homotopy curve is flat. The constants c_1 and c_2 have to be suitably chosen to balance the computational time and the precision of the final results.

4. Numerical methods. The main principle of the continuation method is that as the parameter t changes from 0 to 1, $\vartheta(t)$ is always a good initial approximation for $\vartheta(t+h)$. In order to follow a homotopy curve from $t = 0$ to $t = 1$ numerically, we need a local convergent method, which will converge to the eigentuple of a right definite two-parameter eigenvalue problem (1.1) providing a good initial approximation is given. For such a method we can take Newton's method.

We view $u = (x, y, \lambda, \mu)$ as a $(K+2)$-dimensional vector, where $K := n_1 + n_2$ and define

$$F(u) := \begin{bmatrix} A_1x - \lambda B_1x - \mu C_1x \\ A_2y - \lambda B_2y - \mu C_2y \\ \frac{1}{2}(x^Tx - 1) \\ \frac{1}{2}(y^Ty - 1) \end{bmatrix}.$$

Newton's method applied to the equation $F(u) = 0$ gives

$$F'(u_k)(u_{k+1} - u_k) = -F(u_k), \tag{4.1}$$

where

$$F'(u) = \begin{bmatrix} A_1 - \lambda B_1 - \mu C_1 & 0 & -B_1 x & -C_1 x \\ 0 & A_2 - \lambda B_2 - \mu C_2 & -B_2 y & -C_2 y \\ x^T & 0 & 0 & 0 \\ 0 & y^T & 0 & 0 \end{bmatrix}.$$

To perform one step (4.1) of Newton's method, we have to solve the linear system

$$\begin{bmatrix} A_1 - \lambda_k B_1 - \mu_k C_1 & 0 & -B_1 x_k & -C_1 x_k \\ 0 & A_2 - \lambda_k B_2 - \mu_k C_2 & -B_2 y_k & -C_2 y_k \\ x_k^T & 0 & 0 & 0 \\ 0 & y_k^T & 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta x_k \\ \Delta y_k \\ \Delta \lambda_k \\ \Delta \mu_k \end{bmatrix} = \begin{bmatrix} -(A_1 - \lambda_k B_1 - \mu_k C_1) x_k \\ -(A_2 - \lambda_k B_2 - \mu_k C_2) y_k \\ \frac{1}{2}(1 - x_k^T x_k) \\ \frac{1}{2}(1 - y_k^T y_k) \end{bmatrix} \tag{4.2}$$

and calculate new approximations:

$$\begin{aligned} x_{k+1} &= x_k + \Delta x_k, \\ y_{k+1} &= y_k + \Delta y_k, \\ \lambda_{k+1} &= \lambda_k + \Delta \lambda_k, \\ \mu_{k+1} &= \mu_k + \Delta \mu_k. \end{aligned}$$

The following equations can be derived from (4.2):

$$(A_1 - \lambda_k B_1 - \mu_k C_1) x_{k+1} = \Delta\lambda_k B_1 x_k + \Delta\mu_k C_1 x_k, \tag{4.3}$$

$$(A_2 - \lambda_k B_2 - \mu_k C_2) y_{k+1} = \Delta\lambda_k B_2 y_k + \Delta\mu_k C_2 y_k, \tag{4.4}$$

$$\begin{aligned} x_k^T x_{k+1} &= \tfrac{1}{2}(x_k^T x_k + 1), \\ y_k^T y_{k+1} &= \tfrac{1}{2}(y_k^T y_k + 1). \end{aligned}$$

To solve for $\Delta\lambda_k, \Delta\mu_k$ we denote

$$\begin{aligned} v_k &= (A_1 - \lambda_k B_1 - \mu_k C_1)^{-1} B_1 x_k, \\ w_k &= (A_1 - \lambda_k B_1 - \mu_k C_1)^{-1} C_1 x_k, \\ p_k &= (A_2 - \lambda_k B_2 - \mu_k C_2)^{-1} B_2 y_k, \\ q_k &= (A_2 - \lambda_k B_2 - \mu_k C_2)^{-1} C_2 y_k. \end{aligned} \tag{4.5}$$

We multiply the above equalities with x_k^T and y_k^T and obtain the following linear system:

$$\begin{bmatrix} x_k^T v_k & x_k^T w_k \\ y_k^T p_k & y_k^T q_k \end{bmatrix} \begin{bmatrix} \Delta\lambda_k \\ \Delta\mu_k \end{bmatrix} = \begin{bmatrix} \frac{1}{2}(x_k^T x_k + 1) \\ \frac{1}{2}(y_k^T y_k + 1) \end{bmatrix} \tag{4.6}$$

for $\Delta\lambda_k$ and $\Delta\mu_k$. After we solve system (4.6) for $\Delta\lambda_k, \Delta\mu_k$ we calculate x_{k+1} and y_{k+1} from (4.3) and (4.4).

To improve the convergence we use the following generalization of Rayleigh quotient to two-parameter problem (1.1). Remember that in the classical eigenvalue problem $Ax = \lambda x$ the formula for the Rayleigh quotient is

$$\rho(x) = \frac{x^T A x}{x^T x}.$$

If instead of problem (1.1) we consider the related system (1.3), then we define the Rayleigh quotient $\rho(x_k, y_k, A_1, A_2, B_1, B_2, C_1, C_2)$ as a pair (ρ_1, ρ_2), where

$$\begin{aligned} \rho_1 &= \frac{\langle \Delta_1 z, z\rangle}{\langle \Delta_0 z, z\rangle}, \\ \rho_2 &= \frac{\langle \Delta_2 z, z\rangle}{\langle \Delta_0 z, z\rangle}. \end{aligned} \tag{4.7}$$

We call (4.7) a *tensor Rayleigh quotient*. Since the two-parameter problem (1.1) is right definite, all tensor Rayleigh quotients are well defined.

Although it looks at first sight that a computation with $\Delta_0, \Delta_1, \Delta_2$, which are all of order N, will increase the computational work and make this method inefficient, it is possible to compute the tensor Rayleigh quotient quite efficiently. The reason is that we are always dealing only with the decomposable tensors $z = x \otimes y$. Then it is easy to see from the structure of Δ_i that

$$\begin{aligned} \rho_1 &= \frac{(x^T C_1 x)(y^T A_2 y) - (x^T A_1 x)(y^T C_2 y)}{(x^T B_1 x)(y^T C_2 y) - (x^T C_1 x)(y^T B_2 y)}, \\ \rho_2 &= \frac{(x^T A_1 x)(y^T B_2 y) - (x^T B_1 x)(y^T A_2 y)}{(x^T B_1 x)(y^T C_2 y) - (x^T C_1 x)(y^T B_2 y)}. \end{aligned} \tag{4.8}$$

This shows that we do not have to work explicitly with matrices Δ_i.

In a modification of Newton's method with the tensor Rayleigh quotient we respectively replace λ_k and μ_k with $\rho_{k,1}$ and $\rho_{k,2}$ in (4.3), (4.4), and (4.5). Let x_0 and y_0 be two initial approximations for the eigenvector of (1.1). An initial approximation for the eigenvalue is not needed, since we can calculate one using the tensor Rayleigh quotient. The following algorithm is called the TRQI algorithm.

ALGORITHM 2. THE TRQI ALGORITHM.

Repeat for $k = 0, 1, \ldots$:

1. Calculate the tensor Rayleigh quotient
$$(\rho_{k,1}, \rho_{k,2}) = \rho(x_k, y_k, A_1, A_2, B_1, B_2, C_1, C_2).$$
2. If any of the matrices $A_1 - \rho_{k,1} B_1 - \rho_{k,2} C_1$ or $A_2 - \rho_{k,1} B_2 - \rho_{k,2} C_2$ is singular, give $(\rho_{k,1}, \rho_{k,2})$ a small perturbation.
3. Calculate the vectors
$$\begin{aligned} v_k &= (A_1 - \rho_{k,1} B_1 - \rho_{k,2} C_1)^{-1} B_1 x_k, \\ w_k &= (A_1 - \rho_{k,1} B_1 - \rho_{k,2} C_1)^{-1} C_1 x_k, \\ p_k &= (A_2 - \rho_{k,1} B_2 - \rho_{k,2} C_2)^{-1} B_2 y_k, \\ q_k &= (A_2 - \rho_{k,1} B_2 - \rho_{k,2} C_2)^{-1} C_2 y_k. \end{aligned}$$
4. Calculate $\Delta\lambda_k$ and $\Delta\mu_k$ from the linear system
$$\begin{bmatrix} x_k^T v_k & x_k^T w_k \\ y_k^T p_k & y_k^T q_k \end{bmatrix} \begin{bmatrix} \Delta\lambda_k \\ \Delta\mu_k \end{bmatrix} = \begin{bmatrix} \frac{1}{2}(x_k^T x_k + 1) \\ \frac{1}{2}(y_k^T y_k + 1) \end{bmatrix}.$$

5. Calculate the vectors

$$\begin{aligned}\widetilde{x}_{k+1} &= \Delta\lambda_k v_k + \Delta\mu_k w_k,\\ \widetilde{y}_{k+1} &= \Delta\lambda_k p_k + \Delta\mu_k q_k,\\ x_{k+1} &= \widetilde{x}_{k+1}/\|\widetilde{x}_{k+1}\|,\\ y_{k+1} &= \widetilde{y}_{k+1}/\|\widetilde{y}_{k+1}\|.\end{aligned}$$

6. Set $k = k + 1$.

Until $\left(\|(A_1 - \rho_{k,1}B_1 - \rho_{k,2}C_1)x_{k+1}\|^2 + \|(A_2 - \rho_{k,1}B_2 - \rho_{k,2}C_2)y_{k+1}\|^2\right)^{\frac{1}{2}} \le \epsilon_2$.

To justify the terminating criterion we have to prove some results. Let us define the residual

$$r(x, y, \lambda, \mu) := \begin{bmatrix}(A_1 - \lambda B_1 - \mu C_1)x\\ (A_2 - \lambda B_2 - \mu C_2)y\end{bmatrix}. \tag{4.9}$$

LEMMA 4.1 (see Ji, Jiang, and Lee [18]). *Let λ, μ be arbitrary scalars and x, y arbitrary unit vectors. Let $(x^*, y^*, \lambda^*, \mu^*)$ be an eigentuple of* (1.1). *If the matrix*

$$\begin{bmatrix}x^T B_1 x^* & x^T C_1 x^*\\ y^T B_2 y^* & y^T C_2 y^*\end{bmatrix}$$

is nonsingular, then

$$\sqrt{(\lambda^* - \lambda)^2 + (\mu^* - \mu)^2} \le \left\| \begin{bmatrix}x^T B_1 x^* & x^T C_1 x^*\\ y^T B_2 y^* & y^T C_2 y^*\end{bmatrix}^{-1} \right\| \|r\|_2, \tag{4.10}$$

where $r = r(x, y, \lambda, \mu)$ is the residual (4.9).

Proof. Since $(x^*, y^*, \lambda^*, \mu^*)$ is an eigentuple, we have

$$\begin{aligned}A_1 x^* - \lambda^* B_1 x^* - \mu^* C_1 x^* = 0,\\ A_2 y^* - \lambda^* B_2 y^* - \mu^* C_2 y^* = 0.\end{aligned} \tag{4.11}$$

From (4.9) it follows that

$$\begin{aligned}x^T A_1 x^* - x^T \lambda B_1 x^* - x^T \mu C_1 x^* = x^{*T} r_1,\\ y^T A_2 y^* - y^T \lambda B_2 y^* - y^T \mu C_2 y^* = y^{*T} r_2,\end{aligned} \tag{4.12}$$

where $r = [r_1\ r_2]^T$. If we multiply relations (4.11) by x^T and y^T and subtract them from (4.12), we obtain the following linear system:

$$\begin{bmatrix}x^T B_1 x^* & x^T C_1 x^*\\ y^T B_2 y^* & y^T C_2 y^*\end{bmatrix} \begin{bmatrix}\lambda^* - \lambda\\ \mu^* - \mu\end{bmatrix} = \begin{bmatrix}x^{*T} r_1\\ y^{*T} r_2\end{bmatrix}.$$

The bound (4.10) follows readily. ∎

COROLLARY 4.2 (see Ji, Jiang, and Lee [18]). *If the sequence $(x_k, y_k, \rho_{k,1}, \rho_{k,2})$ converges to the eigentuple $(x^*, y^*, \lambda^*, \mu^*)$ of* (1.1), *then for a sufficiently large k we have*

$$\sqrt{(\lambda^* - \rho_{k,1})^2 + (\mu^* - \rho_{k,2})^2} \le c\, \delta_1^{-1} \|r_k\|_2,$$

where $r_k = r(x_k, y_k, \rho_{k,1}, \rho_{k,2})$ *is the residual* (4.9), c *is a constant independent of* r, k, *and* δ_1 *is the constant from* (1.2).

Proof. For a sufficiently large k we have $x_k = x^* + o(1)$ and $y_k = y^* + o(1)$. It follows that

$$\begin{vmatrix} x_k^T B_1 x^* & x_k^T C_1 x^* \\ y_k^T B_2 y^* & y_k^T C_2 y^* \end{vmatrix} = D(x^*, y^*) + o(1).$$

Since $D(x^*, y^*) \geq \delta_1$ we can apply Lemma 4.1. It is easy to see that there exists a constant c such that

$$\left\| \begin{bmatrix} x_k^T B_1 x^* & x_k^T C_1 x^* \\ y_k^T B_2 y^* & y_k^T C_2 y^* \end{bmatrix}^{-1} \right\| \leq \delta_1^{-1} c. \qquad \square$$

We will show that when TRQI converges, its convergence rate is quadratic. We denote by $u_k = (x_k, y_k, \rho_{k,1}, \rho_{k,2})$ the sequence generated by TRQI with the initial vectors x_0 and y_0. Here $\|x_k\| = \|y_k\| = 1$, $k = 0, 1, \ldots$. Let $u^* = (x^*, y^*, \rho_1^*, \rho_2^*)$ stand for a true solution of the problem, ρ_k for $(\rho_{k,1}, \rho_{k,2})$ and ρ^* for (ρ_1^*, ρ_2^*).

We need the following lemma on the error bound for the Rayleigh quotient.

LEMMA 4.3. *Let* A *be a symmetric matrix with eigenpairs* $Ax_i = \lambda_i x_i$ *for* $i = 1, \ldots, n$. *Let a unit vector* x *be an approximation for the eigenvector* x_1 *and let* $\mu = x^T Ax$ *be the Rayleigh quotient for* x. *Then*

$$|\mu - \lambda_1| \leq 2\|A\|\|x - x_1\|^2. \tag{4.13}$$

Proof. Since matrix A is symmetric, its eigenvectors form an orthogonal basis. In this basis we can write $x = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$, where $\alpha_1^2 + \cdots + \alpha_n^2 = 1$. It is easy to see that

$$\mu = x^T Ax = \lambda_1 \alpha_1^2 + \lambda_2 \alpha_2^2 + \cdots + \lambda_n \alpha_n^2$$

and

$$\|x - x_1\|^2 = (1 - \alpha_1)^2 + \alpha_2^2 + \cdots + \alpha_n^2 = 2(1 - \alpha_1).$$

It follows that

$$\mu - \lambda_1 = \sum_{i=2}^{n} (\lambda_i - \lambda_1)\alpha_i^2.$$

Finally,

$$\|\lambda_i - \lambda_1\| \leq 2\|A\|$$

yields

$$|\mu - \lambda_1| \leq 2\|A\|(1 - \alpha_1^2) \leq 4\|A\|(1 - \alpha_1) = 2\|A\|\|x - x_1\|^2. \qquad \square$$

THEOREM 4.4. *Suppose that* $F'(u^*)$ *is nonsingular. If the sequence* $\{u_k\}$ *converges to* u^* *as* $k \to \infty$, *then for* k *large enough there exists a constant* c *such that*

$$\|u_{k+1} - u^*\| \leq c\|u_k - u^*\|^2.$$

Proof. We notice that we obtain $u'_{k+1} := (\widetilde{x}_{k+1}, \widetilde{y}_{k+1}, \rho_{k,1} + \Delta\lambda_k, \rho_{k,2} + \Delta\mu_k)$ by Newton's method using the initial approximation u_k. Since $F'(u^*)$ is nonsingular, Newton's method has a quadratic convergence rate and for k large enough there exists a constant c_1 such that

$$\|\widetilde{x}_{k+1} - x^*\|, \|\widetilde{y}_{k+1} - y^*\| \le c_1 \|u_k - u^*\|^2.$$

It is easy to see that there exists a constant c_2 such that

$$\|x_{k+1} - x^*\|, \|y_{k+1} - y^*\| \le c_2 \|u_k - u^*\|^2. \tag{4.14}$$

In order to complete the proof we have to find a suitable bound for $\|\rho_{k+1} - \rho^*\|$. We can write the system (1.3) in the form

$$\begin{aligned} G_1 w &= \lambda w, \\ G_2 w &= \mu w, \end{aligned}$$

where $G_i = \Delta_0^{-1/2} \Delta_i \Delta_0^{-1/2}$ for $i = 1, 2$ and $w = \Delta_0^{1/2} z$. Let us write

$$\begin{aligned} z_{k+1} &= x_{k+1} \otimes y_{k+1}, \\ w_{k+1} &= \Delta_0^{1/2} z_{k+1}, \\ z^* &= x^* \otimes y^*, \\ w^* &= \Delta_0^{1/2} z^*. \end{aligned}$$

The tensor Rayleigh quotient $(\rho_{k+1,1}, \rho_{k+1,2}) = \rho(x_{k+1}, y_{k+1}, A_1, A_2, B_1, B_2, C_1, C_2)$ is equal to

$$\begin{aligned} \rho_{k+1,1} &= \frac{\langle G_1 w_{k+1}, w_{k+1} \rangle}{\langle w_{k+1}, w_{k+1} \rangle}, \\ \rho_{k+1,2} &= \frac{\langle G_2 w_{k+1}, w_{k+1} \rangle}{\langle w_{k+1}, w_{k+1} \rangle}. \end{aligned}$$

The matrices G_1, G_2 are symmetric and we can apply Lemma 4.3. It follows that

$$\|\rho_{k+1} - \rho^*\| \le c_3 \|w_{k+1} - w^*\|^2 \le c_3 \|\Delta_0^{1/2}\|^2 \|z_{k+1} - z^*\|^2, \tag{4.15}$$

where $c_3 := 2 \max(\|G_1\|, \|G_2\|)$. Together with (4.14) this gives a constant c_4, such that

$$\|\rho_{k+1} - \rho^*\| \le c_4 \|u_k - u^*\|^4. \tag{4.16}$$

The proof follows from the bounds (4.14) and (4.16). ∎

Let us remark that the bound (4.16) shows that in TRQI the eigenvalue part converges faster than in Newton's method. This justifies the numerical results in the next section.

Another generalization of Rayleigh quotient was studied by Blum and Chang in [6] and by Ji, Jiang, and Lee in [18]. In the classical eigenvalue problem $Ax = \lambda x$ the Rayleigh quotient

$$\rho(x) = \frac{x^T A x}{x^T x}$$

minimizes the residual $\|Ax - \mu x\|$ for a given $x \neq 0$ over all $\mu \in \mathbb{R}$. Based on this property we define the *generalized Rayleigh quotient* $\widetilde{\rho}(x, y, A_1, A_2, B_1, B_2, C_1, C_2)$ for the two-parameter problem (1.1) to be a pair $(\widetilde{\rho}_1, \widetilde{\rho}_2)$ which minimizes the norm of the residual

$$\left(\|(A_1 - s_1 B_1 - s_2 C_1)x\|^2 + \|(A_2 - s_1 B_2 - s_2 C_2)y\|^2\right)^{1/2} \tag{4.17}$$

for a given $x, y \neq 0$ over all $s_1, s_2 \in \mathbb{R}$.

LEMMA 4.5. *The generalized Rayleigh quotient* $\widetilde{\rho}(x, y, A_1, A_2, B_1, B_2, C_1, C_2)$ *is equal to*

$$\begin{aligned} \widetilde{\rho}_1 &= \frac{\langle Az, Bz\rangle\langle Cz, Cz\rangle - \langle Az, Cz\rangle\langle Bz, Cz\rangle}{\langle Bz, Bz\rangle\langle Cz, Cz\rangle - \langle Bz, Cz\rangle^2}, \\ \widetilde{\rho}_2 &= \frac{\langle Az, Cz\rangle\langle Bz, Bz\rangle - \langle Az, Bz\rangle\langle Bz, Cz\rangle}{\langle Bz, Bz\rangle\langle Cz, Cz\rangle - \langle Bz, Cz\rangle^2}, \end{aligned} \tag{4.18}$$

where

$$z = \begin{bmatrix} x \\ y \end{bmatrix}, \quad A = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}, \quad B = \begin{bmatrix} B_1 & 0 \\ 0 & B_2 \end{bmatrix}, \quad C = \begin{bmatrix} C_1 & 0 \\ 0 & C_2 \end{bmatrix}.$$

Proof. We consider the least squares problem

$$\min_{(\widetilde{\rho}_1, \widetilde{\rho}_2) \in \mathbb{R}^2} \|Az - \widetilde{\rho}_1 Bz - \widetilde{\rho}_2 Cz\| \tag{4.19}$$

and the corresponding overdetermined linear system

$$\begin{bmatrix} B_1 x & C_1 x \\ B_2 y & C_2 y \end{bmatrix} \begin{bmatrix} \widetilde{\rho}_1 \\ \widetilde{\rho}_2 \end{bmatrix} = \begin{bmatrix} A_1 x \\ A_2 y \end{bmatrix}. \tag{4.20}$$

It is obvious that the norm (4.17) is equal to the norm of the residual (4.19). Let us show that the overdetermined system (4.20) has rank 2. Namely, if the $K \times 2$ matrix

$$\begin{bmatrix} B_1 x & C_1 x \\ B_2 y & C_2 y \end{bmatrix} \tag{4.21}$$

has rank smaller than 2, then if follows that

$$\begin{vmatrix} x^T B_1 x & x^T C_1 x \\ y^T B_2 y & y^T C_2 y \end{vmatrix} = D(x, y) = 0,$$

which contradicts the fact that problem (1.1) is right definite.

Since the matrix (4.21) has rank 2, the least squares solution of (4.20) can be expressed as the solution of the normal equations:

$$\begin{bmatrix} (B_1 x)^T & (B_2 y)^T \\ (C_1 y)^T & (C_2 y)^T \end{bmatrix} \begin{bmatrix} B_1 x & C_1 x \\ B_2 y & C_2 y \end{bmatrix} \begin{bmatrix} \widetilde{\rho}_1 \\ \widetilde{\rho}_2 \end{bmatrix} = \begin{bmatrix} (B_1 x)^T & (B_2 y)^T \\ (C_1 y)^T & (C_2 y)^T \end{bmatrix} \begin{bmatrix} A_1 x \\ A_2 y \end{bmatrix}. \tag{4.22}$$

Using Cramer's rule it is easy to see that the solution of system (4.22) is identical to (4.18). ∎

If we compare the formulas (4.8) and (4.18), then we see that the tensor Rayleigh quotient requires fewer operations to compute than the generalized Rayleigh quotient. If we use the generalized Rayleigh quotient instead of the tensor Rayleigh quotient in step 1 of the TRQI algorithm, then we call this method the GRQI algorithm.

5. Numerical example. For a numerical example we take the following two-parameter system

$$\begin{bmatrix}1&0&0\\0&0&0\\0&0&0\end{bmatrix}x-\lambda\begin{bmatrix}2&1&1\\1&3&-0.5\\1&-0.5&2\end{bmatrix}x-\mu\begin{bmatrix}0&1&1\\1&0&1\\1&1&0\end{bmatrix}x=0,$$

$$\begin{bmatrix}0&0&0\\0&1&0\\0&0&0\end{bmatrix}y-\lambda\begin{bmatrix}1&1&1\\1&0&1\\1&1&1\end{bmatrix}y-\mu\begin{bmatrix}3&1&-1\\1&3&1\\-1&1&4\end{bmatrix}y=0.$$

Matrices are chosen in such a way that $(0,0)$ is a quadruple eigenvalue. For the construction of the homotopy (3.1) we take matrices

$$W_1=\begin{bmatrix}0&1&1\\1&1&1\\1&1&1\end{bmatrix}\quad\text{and}\quad W_2=\begin{bmatrix}1&1&1\\1&0&1\\1&1&1\end{bmatrix}.$$

It is easy to see that $(0,0)$ is a quadruple eigenvalue for the homotopy (3.1) at $t=1/2$. Thus the homotopy has a singular point at $t=1/2$ which some of the homotopy curves $\vartheta_i(t)$ have to pass.

Using the gradient method we calculate the value

$$\delta_2=\min\left\{\min_{\|x\|=\|y\|=1}D(x,y,0),\min_{\|x\|=\|y\|=1}D(x,y,1)\right\}$$

and obtain $\delta_2=0.2188$. It follows from Lemma 3.8 that the smallest eigenvalue of Δ_0 is greater than or equal to $\delta_2/3=0.0729$. (For a comparison, the exact smallest eigenvalue of Δ_0 is 0.2112.) The upper bound $\|B_1\|\|C_2\|+\|C_1\|\|B_2\|$ from (3.21) is equal to 22.6292. (For a comparison, the exact largest eigenvalue of Δ_0 is 18.8374.) Now we use Theorem 3.7 to obtain $\eta=0.0064$. This is the value we use in the continuation method.

For the initial eigentuple $\vartheta(0)$ we take

$$\begin{aligned}\lambda(0)&=-1.0445,\\ \mu(0)&=-0.4529,\\ x(0)&=(0.8067,-0.3435,-0.4808)^T,\\ y(0)&=(0.4326,-0.8326,0.3460)^T.\end{aligned}$$

We choose $h=0.02$ as the initial stepsize and set $c_1=0.5$ and $c_2=0.1$. We set $\epsilon_1=\epsilon_2=10^{-8}$ for the constants in the continuation method and the TRQI (respectively, GRQI) algorithm. As a result, the continuation method using these initial values takes 94 steps to reach $t=1$ from $t=0$. In each step a maximum of 3 steps of the TRQI algorithm and a maximum of 4 steps of the GRQI algorithm are needed to reach the prescribed precision.

The following figures present the behavior of the continuation method for these initial values. In Figure 5.1 we see the graphs of $\lambda(t)$ and $\mu(t)$. We see that the homotopy curve passes the singular point $(0,0)$ at $t=1/2$. As can be seen from Figures 5.2 and 5.3, where we see graphs of $x(t)$ and $y(t)$, no curve switching occurs in the singular point. Figure 5.4 depicts how the stepsize h changes from $t=0$ to $t=1$. If we compare Figure 5.4 with the previous ones, then it is clearly seen how the

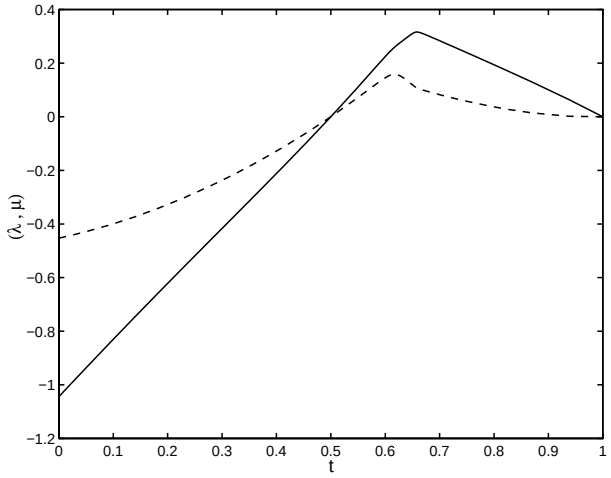


FIG. 5.1. *Eigenvalue pair* (λ, μ) *versus* t. *Solid line:* $\lambda(t)$*; dashed line:* $\mu(t)$.

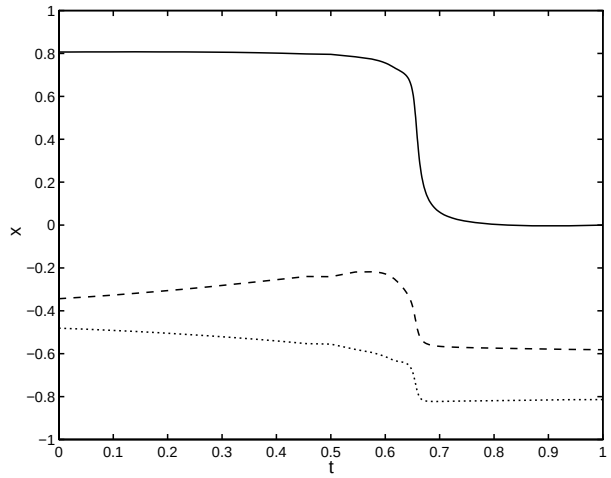


FIG. 5.2. *Eigenvector* x *versus* t. *Solid line:* $x_1(t)$*; dashed line:* $x_2(t)$*; dotted line:* $x_3(t)$.

TABLE 5.1

Statistics for all 9 *eigenvalues: starting* $\lambda(0)$, $\mu(0)$ *and obtained* $\lambda(1)$, $\mu(1)$*; number of steps of the continuation method; number of evaluations of Newton's method, TRQI, and GRQI.*

			No. of	Number of evaluations					Passes
i	$\lambda(0)$	$\mu(0)$	steps	Newton	TRQI	GRQI	$\lambda(1)$	$\mu(1)$	$(0,0)$
1	−1.0445	0.8279	159	360	323	493	−0.2154	0.3659	no
2	1.0445	0.8279	174	399	346	534	1.8297	1.2842	no
3	0	0.8279	216	501	430	654	0.0090	0.4623	no
4	−1.0445	0	129	289	254	395	0	0	yes
5	1.0445	0	54	145	106	179	1.0445	0	no
6	0	0	10	21	9	9	0	0	yes
7	−1.0445	−0.4529	94	270	189	316	0	0	yes
8	1.0445	−0.4529	161	405	321	538	4.8629	−4.0229	no
9	0	−0.4529	104	253	208	311	0	0	yes

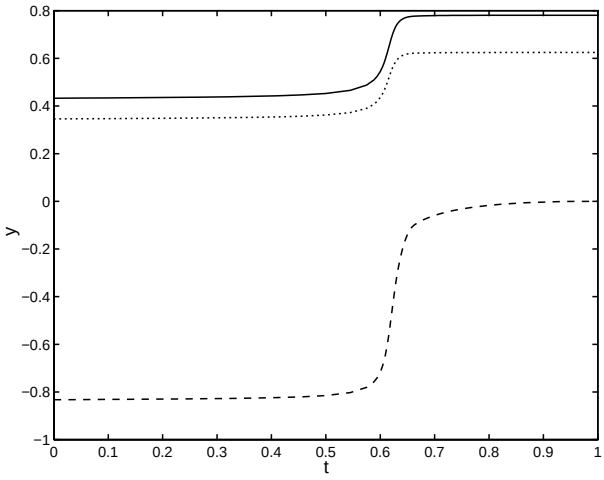


FIG. 5.3. *Eigenvector y versus t. Solid line: $y_1(t)$; dashed line: $y_2(t)$; dotted line: $y_3(t)$.*

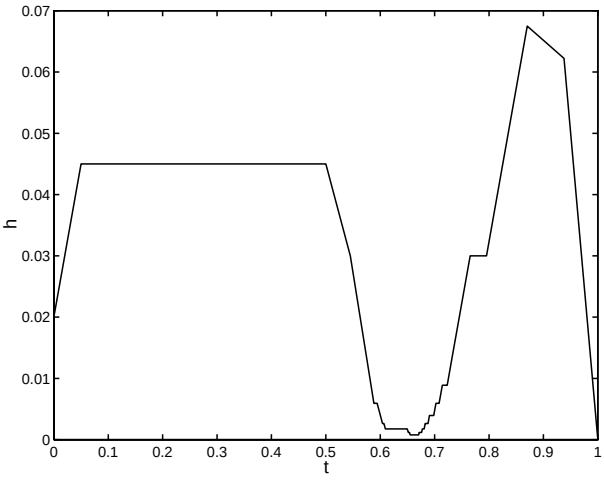


FIG. 5.4. *Stepsize h versus t.*

adaptive algorithm appropriately reduces or increases the stepsize h. We also notice that the stepsize h reduces near $t = 1$, where we have another singular point. It is clearly seen that the algorithm slows down in sensitive areas and speeds up when the homotopy curve is flat.

Similar results can be shown for the remaining eight homotopy curves. Table 5.1 presents the number of steps it takes the continuation method to calculate the eigentuples from $t = 0$ to $t = 1$. In all homotopy curves the majority of steps is calculated on the interval $[0.6, 0.7]$. The only exception is $i = 6$, where the number of steps is so small because $\vartheta_6(t)$ is a constant. As it can be seen from Table 5.1 and also from Figure 5.4, the large number of steps is a result of a steep homotopy curve and not a result of passing a singular point.

We also tested Newton's method without tensor or generalized Rayleigh quotient.

The numbers in the fourth, the fifth, and the sixth columns of Table 5.1 represent the number of calculated Newton iterations, tensor Rayleigh quotients, and generalized Rayleigh quotients, respectively. As predicted in Theorem 4.4 TRQI converges faster than Newton's method. On the other hand, GRQI is most expensive to compute and has the slowest convergence among these three methods. Usually 2 steps of TRQI and 3 steps of GRQI are needed for 1 step of the continuation method. Based on these results and based on the theory from the previous section, we recommend the use of TRQI instead of Newton's method or GRQI.

The number of steps could be further reduced by using better bounds for the norm of the difference between two distinct eigenvectors of the two-parameter right definite problem (1.1).

Acknowledgment. The author would like to thank Dr. Tomaž Košir for his helpful comments and suggestions.

REFERENCES

[1] E. L. Allgower and K. Georg, *Numerical Continuation Methods*, Springer-Verlag, Berlin, 1990.
[2] F. V. Atkinson, *Multiparameter spectral theory*, Bull. Amer. Math. Soc., 74 (1968), pp. 1–27.
[3] F. V. Atkinson, *Multiparameter Eigenvalue Problems*, Academic Press, New York, 1972.
[4] P. B. Bailey, *The automatic solution of two-parameter Sturm–Liouville eigenvalue problems in ordinary differential equations*, Appl. Math. Comput., 8 (1981), pp. 251–259.
[5] P. A. Binding, P. J. Browne, and X. Ji, *A numerical method using the Prüfer transformation for the calculation of eigenpairs of two-parameter Sturm–Liouville problems*, IMA J. Numer. Anal., 13 (1993), pp. 559–569.
[6] E. K. Blum and A. F. Chang, *A numerical method for the solution of the double eigenvalue problem*, J. Inst. Math. Appl., 22 (1978), pp. 29–41.
[7] E. K. Blum and A. R. Curtis, *A convergent gradient method for matrix eigenvector-eigentuple problems*, Numer. Math., 31 (1978), pp. 247–263.
[8] E. K. Blum and P. B. Geltner, *Numerical solution of eigentuple-eigenvector problems in Hilbert spaces by a gradient method*, Numer. Math., 31 (1978), pp. 231–246.
[9] Z. Bohte, *Numerical solution of some two-parameter eigenvalue problems*, Slov. Acad. Sci. Art., (1982), pp. 17–28.
[10] P. J. Browne and B. D. Sleeman, *Analytic perturbations of multiparameter eigenvalue problems*, Quart. J. Math. Oxford Ser. (2), 30 (1979), pp. 257–263.
[11] P. J. Browne and B. D. Sleeman, *A numerical technique for multiparameter eigenvalue problems*, IMA J. Numer. Anal., 2 (1982), pp. 451–457.
[12] M. T. Chu, *A simple application of the homotopy method to symmetric eigenvalue problems*, Linear Algebra Appl., 59 (1984), pp. 85–90.
[13] M. T. Chu, T. Y. Li, and T. Sauer, *Homotopy method for general λ-matrix problems*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 528–536.
[14] L. Fox, L. Hayes, and D. F. Mayers, *The double eigenvalue problem*, in Topics in Numerical Analysis, Proc. Roy. Irish Acad. Conf., J. Miller, ed., Academic Press, New York, 1973, pp. 93–112.
[15] B. A. Hargrave and B. D. Sleeman, *The numerical solution of two-parameter eigenvalue problems in ordinary differential equations with an application to the problem of diffraction by a plane angular sector*, J. Inst. Math. Appl., 14 (1974), pp. 9–22.
[16] X. Ji, *Numerical solution of joint eigenpairs of a family of commutative matrices*, Appl. Math. Lett., 4 (1991), pp. 57–60.
[17] X. Ji, *A two-dimensional bisection method for solving two-parameter eigenvalue problems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1085–1093.
[18] X. R. Ji, H. Jiang, and B. H. K. Lee, *A Generalized Rayleigh Quotient Iteration for Coupled Eigenvalue Problems*, manuscript.
[19] T. Y. Li and N. H. Rhee, *Homotopy algorithm for symmetric eigenvalue problems*, Numer. Math., 55 (1989), pp. 265–280.
[20] T. Y. Li, Z. Zeng, and L. Cong, *Solving eigenvalue problems of real nonsymmetric matrices with real homotopies*, SIAM J. Numer. Anal., 29 (1992), pp. 229–248.

[21] R. E. Müller, *Numerical solution of multiparameter eigenvalue problems*, Z. Angew. Math. Mech., 62 (1982), pp. 681–686.
[22] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.
[23] M. Shimasaki, *Homotopy algorithm for two-parameter eigenvalue problems*, Trans. Japan SIAM, 5 (1995), pp. 121–129 (in Japanese).
[24] M. Shimasaki, *Numerical method based on homotopy algorithm for two-parameter problems*, Trans. Japan SIAM, 6 (1996), pp. 205-218 (in Japanese).
[25] T. Slivnik and G. Tomšič, *A numerical method for the solution of two-parameter eigenvalue problem*, J. Comput. Appl. Math., 15 (1986), pp. 109–115.
[26] H. Volkmer, *On the minimal eigenvalue of a positive definite operator determinant*, Proc. Roy. Soc. Edinburgh Sect. A, 103 (1986), pp. 201–208.
[27] H. Volkmer, *Multiparameter Problems and Expansion Theorems*, Lecture Notes in Math. 1356, Springer-Verlag, Berlin, New York, 1988.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 4, pp. 1185–1201
© 2000 Society for Industrial and Applied Mathematics

A BLOCK ALGORITHM FOR MATRIX 1-NORM ESTIMATION, WITH AN APPLICATION TO 1-NORM PSEUDOSPECTRA*

NICHOLAS J. HIGHAM† AND FRANÇOISE TISSEUR†

Abstract. The matrix 1-norm estimation algorithm used in LAPACK and various other software libraries and packages has proved to be a valuable tool. However, it has the limitations that it offers the user no control over the accuracy and reliability of the estimate and that it is based on level 2 BLAS operations. A block generalization of the 1-norm power method underlying the estimator is derived here and developed into a practical algorithm applicable to both real and complex matrices. The algorithm works with $n \times t$ matrices, where t is a parameter. For $t = 1$ the original algorithm is recovered, but with two improvements (one for real matrices and one for complex matrices). The accuracy and reliability of the estimates generally increase with t and the computational kernels are level 3 BLAS operations for $t > 1$. The last $t-1$ columns of the starting matrix are randomly chosen, giving the algorithm a statistical flavor. As a by-product of our investigations we identify a matrix for which the 1-norm power method takes the maximum number of iterations. As an application of the new estimator we show how it can be used to efficiently approximate 1-norm pseudospectra.

Key words. matrix 1-norm, matrix norm estimation, matrix condition number, condition number estimation, p-norm power method, 1-norm pseudospectrum, LAPACK, level 3 BLAS

AMS subject classification. 65F35

PII. S0895479899356080

1. Introduction.

Research in matrix condition number estimation began in the 1970s with the problem of cheaply estimating the condition number $\kappa(A) = \|A\|\|A^{-1}\|$ and an approximate null vector of a square matrix A given some factorization of it. The earliest algorithm is one of Gragg and Stewart [10]. It was improved by Cline, Moler, Stewart, and Wilkinson [4], leading to the 1-norm condition estimation algorithm used in LINPACK [8] and later included in MATLAB (function `rcond`).

During the 1980s, attention was drawn to various componentwise condition numbers and it was recognized that most condition estimation problems can be reduced to the estimation of $\|A\|$ when matrix-vector products Ax and A^Tx can be cheaply computed [2], [16, Sec. 14.1]. Hager [12] derived an algorithm for the 1-norm that is a special case of the more general p-norm power method proposed by Boyd [3] and later investigated by Tao [18]. Hager's algorithm was modified by Higham [14] and incorporated in LAPACK (routine `xLACON`) [1] and MATLAB (function `condest`).

The LINPACK and LAPACK estimators both produce estimates that in practice are almost always within a factor 10 and 3, respectively, of the quantities they are estimating [13], [14], [15]. This has been entirely adequate for applications where only an order of magnitude estimate is required, such as the evaluation of error bounds. However, in some applications an estimate with one or more correct digits is required (see, for example, the pseudospectra application described in section 4), and for these the LINPACK and LAPACK estimators have the drawback that they offer the user no way to control or improve the accuracy of the estimate. Here, "accuracy" refers to

*Received by the editors May 10, 1999; accepted for publication by A. Edelman July 21, 1999; published electronically March 30, 2000. This work was supported by Engineering and Physical Sciences Research Council grants GR/L76532 and GR/L94314.

http://www.siam.org/journals/simax/21-4/35608.html

†Department of Mathematics, University of Manchester, Manchester, M13 9PL, England (higham@ma.man.ac.uk, ftisseur@ma.man.ac.uk).

TABLE 1
Empirical probabilities that $\min\{\widetilde{\phi}_s/\|A\|_1, \|A\|_1/\widetilde{\phi}_s\} \geq \alpha$ *for one* A *of the form* `inv(randn(100))` *and* $N(0,1)$ *vectors* x_j.

α	$s=1$	4	8	12	16	20
0.99	0.01	0.02	0.04	0.03	0.03	0.04
0.9	0.10	0.23	0.31	0.31	0.35	0.35
0.5	0.62	0.92	0.98	0.99	1.00	1.00
0.3	0.88	0.99	1.00	1.00	1.00	1.00
0.1	1.00	1.00	1.00	1.00	1.00	1.00

average case behavior. Also of interest for a norm estimator is its worst case behavior, that is, its "reliability."

Since estimating $\|A\|$ cheaply appears inevitably to admit the possibility of arbitrarily poor estimates (although proving so is an open problem [6]), one might look for an approach for which probabilistic statements can be made about the accuracy of the estimate. The definition $\|A\| = \max_{x\neq 0} \|Ax\|/\|x\|$ of a subordinate matrix norm suggests the estimate

$$\phi_s = \max\left\{ \frac{\|Ax_j\|}{\|x_j\|} : j = 1\colon s \right\} \leq \|A\|, \tag{1.1}$$

where s is a parameter and the x_j are independently chosen random vectors. In the case of the 2-norm ($\|x\|_2 = (x^T x)^{1/2}$) and an appropriate distribution of the x_j, explicit bounds are available on the probability of such estimates being within a given factor of $\|A\|$ [7]. As is done in [11] for certain estimates of the Frobenius norm we can scale our estimates $\widetilde{\phi}_s \leftarrow \theta_s \phi_s$, where the constant θ_s is chosen so that the expected value of $\widetilde{\phi}_s$ is $\|A\|$ (note that $\widetilde{\phi}_s$ can therefore be greater or less than $\|A\|$). For the 1-norm ($\|x\|_1 = \sum_i |x_i|$), which is our interest here, we investigate this approach empirically. For a fixed matrix A of the form, in MATLAB notation, `inv(randn(100))`, Table 1 shows the observed probabilities that $\min\{\widetilde{\phi}_s/\|A\|_1, \|A\|_1/\widetilde{\phi}_s\} \geq \alpha$ for various α and s, based on 1000 separate evaluations of the ϕ_s with vectors x_j from the normal $N(0,1)$ distribution, and where θ_s is determined empirically so that the mean of the $\widetilde{\phi}_s$ is $\|A\|_1$.

The table shows that even with $s = n/5 = 20$ samples only 35% of the estimates were within a factor 0.9 of the true norm. The statistical sampling technique is clearly too crude to be useful for obtaining estimates with correct digits. One way to exploit the information contained in the vectors Ax_j is to regard them as first iterates from the 1-norm power method with starting vectors x_j and to continue to iterate. These considerations motivate the block generalization of the 1-norm power method that we present in this paper. Our block power method works with a matrix with t columns instead of a vector. To give a feel for how our new estimator compares with the sampling technique, we applied the estimator (Algorithm 2.4) to 1000 random matrices of the form `inv(randn(100))`. The results are shown in Table 2; the estimates for $t = 1\colon 5$ are obtained at approximately the same cost as the estimates $\widetilde{\phi}_s$ for $s = 4, 8, 12, 16, 20$, respectively. The superiority of the new estimator is clear.

In section 2 we derive the block 1-norm power method and develop it into a practical algorithm for both real and complex matrices. In section 3 we present numerical experiments that give insight into the behavior of the algorithm. An application involving complex matrices is given in section 4, where we describe how the algorithm

TABLE 2
Empirical probabilities that est $\geq \alpha\|A\|_1$, *for* est *from Algorithm* 2.4 *and* A *of the form* `inv(randn(100))`.

α	$t=1$	2	3	4	5
0.99	0.86	0.93	0.98	0.98	0.99
0.9	0.94	0.98	1.00	1.00	1.00
0.5	1.00	1.00	1.00	1.00	1.00

can be used to approximate 1-norm pseudospectra. Conclusions are presented in section 5.

Finally, we note that although our work is specific to the 1-norm, the ∞-norm can be estimated by applying our algorithm to A^*, since $\|A\|_\infty = \|A^*\|_1$.

2. Block 1-norm power method. The 1-norm power method is a special case of Boyd's p-norm power method [3] and was derived independently by Hager [12]. For a real matrix A we denote by $\mathrm{sign}(A)$ the matrix with (i,j) element 1 or -1 according as $a_{ij} \geq 0$ or $a_{ij} < 0$. The jth column of the identity matrix is denoted by e_j.

ALGORITHM 2.1 (1-norm power method). *Given* $A \in \mathbb{R}^{n\times n}$ *this algorithm computes* γ *and* x *such that* $\gamma \leq \|A\|_1$ *and* $\|Ax\|_1 = \gamma\|x\|_1$.

$x = \mathrm{ones}(n,1)/n$
repeat
  $y = Ax$
  $\xi = \mathrm{sign}(y)$
  $z = A^T\xi$
  if $\|z\|_\infty \leq z^T x$
    $\gamma = \|y\|_1$
    quit
  end
  $x = e_j$, where $|z_j| = \|z\|_\infty$ (smallest such j)
end

Algorithm 2.1 was modified by Higham [14, Alg. 4.1] (see also [15], [16, Alg. 14.4]) to improve its reliability and efficiency. The modifications that improve the reliability are, first, to force at least two iterations and, second, to take as the final estimate the maximum of that produced by the algorithm and

$$\|Ab\|_1/\|b\|_1, \qquad b_i = (-1)^{i+1}\left(1 + \frac{i-1}{n-1}\right), \qquad i = 1\colon n. \tag{2.1}$$

The vector b is a heuristic choice intended to "pick out" any large elements of A in those cases where such elements fail to be revealed during the course of the algorithm. Efficiency is improved by terminating the algorithm after computing ξ if it is the same as the previous ξ, since it can be shown that convergence would otherwise be declared after the subsequent computation of z.

To obtain a more accurate and reliable estimate than that provided by Algorithm 2.1 we could run the algorithm t times in succession on t different starting vectors. This idea was suggested in [12], with each starting vector being the mean of the unit vectors e_j not already visited and with the algorithm being prohibited from visiting unit vectors previously visited. Note that the estimates obtained this way are nondecreasing in t. This approach has two weaknesses: it allows limited communication of information between the t different iterations and the highest level computational kernel remains matrix-vector multiplication. We therefore develop a

block algorithm that works with an $n \times t$ matrix as a whole instead of t separate n-vectors. The block approach offers the potential of better estimates, through providing more information on which to base decisions, and it allows the use of level 3 BLAS operations, thus promising greater efficiency. The following algorithm estimates not just the 1-norm of A, but, as a by-product, the 1-norms of the t columns of A having largest 1-norms.

ALGORITHM 2.2 (block 1-norm power method). *Given $A \in \mathbb{R}^{n\times n}$ and a positive integer t, this algorithm computes vectors g and* ind *with $g_j = \|A(:, \mathrm{ind}_j)\|_1$ and $g_1 \geq \cdots \geq g_t$ such that g_j is a lower bound for the 1-norm of the column of A of jth largest 1-norm.*

Choose starting matrix $X \in \mathbb{R}^{n\times t}$ with columns of unit 1-norm.
repeat
  $Y = AX$
  $g_j = \|Y(:, j)\|_1$, $j = 1\colon t$
  Sort g so that $g_1 \geq \cdots \geq g_t$.
  ind_best $= \mathrm{ind}_j$ where $g_1 = \|Y(:, j)\|_1$
  $S = \mathrm{sign}(Y)$
  $Z = A^T S$
  $h_i = \|Z(i, :)\|_\infty$, $\mathrm{ind}_i = i$, $i = 1\colon n$
  if $\max(h_i) \leq Z(:, \mathrm{ind_best})^T X(:, \mathrm{ind_best})$, quit, end
  Sort h so that $h_1 \geq \cdots \geq h_t$ and reorder ind correspondingly.
  $X(:, j) = e_{\mathrm{ind}_j}$, $j = 1\colon t$
end

Like the basic 1-norm power method ($t = 1$) [14], Algorithm 2.2 has the attractive property that it generates increasing sequences of estimates. Denote with a superscript "(k)" quantities from the kth iteration of the loop in Algorithm 2.2 and let a_j denote the jth column of A.

LEMMA 2.3. *The sorted vectors $g^{(k)}$ and $h^{(k)}$ satisfy*

$$g_1^{(k)} \leq h_1^{(k)} \leq g_1^{(k+1)}, \qquad k \geq 1 \tag{2.2}$$

and

$$g_j^{(k)} \leq g_j^{(k+1)}, \qquad j = 1\colon t, \quad k \geq 2. \tag{2.3}$$

Proof. First, we have

$$g_1^{(k)} = \max_{1\leq j\leq t} \|y_j^{(k)}\|_1 =: \|y_r^{(k)}\|_1.$$

But

$$\|y_r^{(k)}\|_1 = {s_r^{(k)}}^T y_r^{(k)} = {s_r^{(k)}}^T A x_r^{(k)} = {z_r^{(k)}}^T x_r^{(k)}.$$

Thus

$$g_1^{(k)} = {z_r^{(k)}}^T x_r^{(k)} \leq \|z_r^{(k)}\|_\infty \|x_r^{(k)}\|_1 = \|z_r^{(k)}\|_\infty \leq \max_{i,j} |z_{ij}^{(k)}| = h_1^{(k)}.$$

Furthermore, if $h_1^{(k)} = \|Z^{(k)}(r, :)\|_\infty$, then

$$h_1^{(k)} = \max_j |a_r^T s_j^{(k)}| \leq \|a_r\|_1 = \|y_1^{(k+1)}\|_1 \leq g_1^{(k+1)}.$$

To prove (2.3), assume without loss of generality that $x_j^{(k-1)} = e_j$, $j = 1\colon t$. Then $Y^{(k)} = [a_1 \ \dots \ a_t]$ and $Z^{(k)}$ has the form

$$Z^{(k)} = \begin{bmatrix} \|a_1\|_1 & \alpha_1 & \dots & \alpha_1 \\ \alpha_2 & \|a_2\|_1 & \dots & \alpha_2 \\ \vdots & & \ddots & \vdots \\ \alpha_t & \dots & \alpha_t & \|a_t\|_1 \\ \alpha_{t+1} & & & \alpha_{t+1} \\ \vdots & & & \vdots \\ \alpha_n & \dots & \dots & \alpha_n \end{bmatrix},$$

where each α_i in row i is, in general, different, and $\alpha_i \le \|a_i\|_1$, $i = 1\colon n$. The algorithm chooses the ind values corresponding to the t rows of $Z^{(k)}$ with largest ∞-norm, and the $g_j^{(k+1)}$ on the next stage are at least as large as these ∞-norms. Since row j of $Z^{(k)}$ has ∞-norm $\|a_j\|_1$ for $j = 1\colon t$ and $\{g_j^{(k)}\}_{j=1}^t = \{\|a_j\|_1\}_{j=1}^t$, inequality (2.3) follows. ∎

Algorithm 2.2 has three possible sources of inefficiency. First, the columns of S are vectors of ± 1s and so a pair of columns s_i and s_j may be parallel ($s_i = \pm s_j$), in which case $z_i = \pm z_j$ and in computing Z a matrix-vector multiplication is redundant. There can be up to $t-1$ redundant matrix-vector products per iteration, this maximum being achieved on the second iteration with S the matrix of ones when A has nonnegative elements. The second possible inefficiency is that a column of S may be parallel to one from the previous iteration, in which case the formation of $Z = A^T S$ again involves redundant computation. We choose to detect parallel columns and replace them by random vectors rand$\{-1, 1\}$ not already in S or the previous S; here, rand$\{-1, 1\}$ denotes a vector with elements from the uniform distribution on the set $\{-1, 1\}$. The detection is done by forming inner products between columns and looking for elements of magnitude n. The total cost of these computations is $O(nt^2)$ flops, which is negligible compared with the $2n^2t$ flops required for each matrix product in the algorithm, since $t \ll n$ in practice.

Our strategy could be extended to check for parallel columns between the current S and all previous S; we return to this possibility in section 3. If all the columns of S are parallel to columns from the previous iteration, then it is easy to see that Algorithm 2.2 is about to converge; we therefore immediately terminate the iteration without computing Z.

Finally, on step k we can have $x_j^{(k)} = e_s = x_i^{(r)}$ for some i and some $r < k$, so that the computation of $Ax_j^{(k)}$ then repeats an earlier computation. Repeated vectors e_j are easily avoided by keeping track of the indices of all the previously used e_j and selecting $\mathrm{ind}_1, \dots, \mathrm{ind}_t$ from among the indices not previously used. If all the indices are repeats, then again we prematurely terminate the iteration, saving a matrix product.

Note that the first and third inefficiencies are possible only for $t > 1$, while the second can happen for the original 1-norm power method.

Our strategy of detecting redundant computations and replacing them by computations that provide new information has three benefits. First, it can reduce the amount of computation, through premature detection of convergence. Second, it can lead to better estimates. For the columns of S this depends on the random

replacement vectors generated, but for the e_j the improvement is deterministic up to future replacements of columns of S. The third benefit is that the dimensions of the matrix multiplications remain constant at each iteration, as opposed to varying if we simply skip redundant computations; this helps us to make efficient use of the computing resources.

The next algorithm incorporates these modifications. The algorithm is forced to take at least 2 (and at most itmax) iterations so that it computes at least t columns of A; it also explicitly identifies the approximate maximizing vector that achieves the norm estimate.

ALGORITHM 2.4 (practical block 1-norm estimator). *Given $A \in \mathbb{R}^{n\times n}$ and positive integers t and* itmax ≥ 2*, this algorithm computes a scalar* est *and vectors v and w such that* est $\leq \|A\|_1$*, $w = Av$, and $\|w\|_1 = \text{est}\|v\|_1$.*

Choose starting matrix $X \in \mathbb{R}^{n\times t}$ with columns of unit 1-norm.
ind_hist = [ ] % Integer vector recording indices of used unit vectors e_j.
$\text{est}_{\text{old}} = 0$, ind = zeros$(n, 1)$, S = zeros(n, t)
for $k = 1, 2, \ldots$
  $Y = AX$
  est = max$\{\, \|Y(:, j)\|_1 : j = 1\colon t \,\}$
  if est $> \text{est}_{\text{old}}$ or $k = 2$
    ind_best $= \text{ind}_j$ where est $= \|Y(:, j)\|_1$, $w = Y(:, \text{ind_best})$
  end
(1) if $k \geq 2$ and est $\leq \text{est}_{\text{old}}$, est $= \text{est}_{\text{old}}$, goto (6), end
  $\text{est}_{\text{old}} = \text{est}$, $S_{\text{old}} = S$
  if $k >$ itmax, goto (6), end
  $S = \text{sign}(Y)$
(2) If every column of S is parallel to a column of S_{old}, goto (6), end
  if $t > 1$
    Ensure that no column of S is parallel to another column of S
    or to a column of S_{old} by replacing columns of S by rand$\{-1, 1\}$.
  end
(3) $Z = A^T S$
  $h_i = \|Z(i, :)\|_\infty$, $\text{ind}_i = i$, $i = 1\colon n$
(4) if $k \geq 2$ and $\max(h_i) = h_{\text{ind_best}}$, goto (6), end
  Sort h so that $h_1 \geq \cdots \geq h_n$ and re-order ind correspondingly.
  if $t > 1$
(5) If ind$(1\colon t)$ is contained in ind_hist, goto (6), end
    Replace ind$(1\colon t)$ by the first t indices in ind$(1\colon n)$ that are
    not in ind_hist.
  end
  $X(:, j) = e_{\text{ind}_j}$, $j = 1\colon t$
  ind_hist = [ind_hist ind$(1\colon t)$]
end
(6) $v = e_{\text{ind_best}}$

Note that this algorithm does not explicitly compute lower bounds for the 1-norms of all t largest columns of A. If this information is required (and we are not aware of any applications in which it is needed), then it can be obtained by keeping track of the largest $\|y_j^{(k)}\|_1$ values encountered.

Inequality (2.2), now expressed as $\text{est}^{(k)} \leq h_1^{(k)} \leq \text{est}^{(k+1)}$, is still valid, except that $h_1^{(k)} > \text{est}^{(k+1)}$ is possible on the last iteration if the original $\text{ind}_1^{(k)}$ is a repeat

(this event is handled by the test (1)). However, (2.3) is no longer true, because of the avoidance of repeated indices.

How do Algorithms 2.4 and 2.2 compare? Ignoring the itmax test, Algorithm 2.4 terminates in at most $n/t+1$ iterations, since t vertices e_j are visited on each iteration after the first and no vertex can be visited more than once. The same cannot be said of Algorithm 2.2 because of the possibility of repeated vertices. Algorithm 2.4 can produce a smaller estimate than Algorithm 2.2: when a redundant computation is avoided, the new information computed can lead to an apparently more promising vertex (based on the relative sizes of the h_i) replacing one that actually corresponds to a larger column of A. However, it is more likely, when Algorithm 2.4 and Algorithm 2.2 produce different results, that Algorithm 2.4 produces a better estimate, as in the following example.

Example 2.5. For a certain 10×10 matrix A, with $t = 2$ and a random starting matrix X_1 we obtained the following results. For Algorithm 2.2,

```
1:  (0, 8.10e-001) (0, 6.13e-001)
2:  (10, 1.77e+000) (4, 1.76e+000)
Underestimation ratio: 1.99e-001
```

The first column denotes the iteration number. The kth row gives the sorted $g_i^{(k)}$, $j = 1{:}t$, each one preceded by the corresponding index ind_i. (Since X_1 does not have columns e_j, the ind_i for the first iteration are shown as zero.) For Algorithm 2.4,

```
1:  (0, 8.10e-001) (0, 6.13e-001)
2:  (10, 1.77e+000) (4, 1.76e+000)
    1 parallel column between S and S_old
3:  (2, 8.87e+000) (5, 4.59e+000)
Exact estimate!
```

Algorithm 2.2 converges after two iterations and produces an estimate too small by a factor 5. However, on the second iteration Algorithm 2.4 detects a column of S parallel to one of S_{old} and replaces it. The new column produces a different Z matrix and causes the convergence test (4) to be failed. The extra iteration visits the (unique) column of maximum 1-norm, so an exact estimate is obtained.

When $t = 1$, Algorithm 2.4 differs from the modified version of Algorithm 2.1 used in LAPACK 3.0 [14, Alg. 4.1], [16, Alg. 14.4] in two ways. First, Algorithm 2.4 does not use the "extra estimate" (2.1). Second, the LAPACK algorithm checks for $\xi^{(k)} = \xi^{(k-1)}$ but not for $\xi^{(k)} = -\xi^{(k-1)}$ as does Algorithm 2.4. This is an oversight. We recommend that LAPACK's `xLACON` be modified to include the extra test. This change will not affect the estimates produced but will sometimes reduce the number of iterations.

We now explain our choice of starting matrix. We take the first column of X to be the vector of 1s, which is the starting vector used in Algorithm 2.1. This has the advantage that for a matrix with nonnegative elements the algorithm converges with an exact estimate on the second iteration, and such matrices arise in applications, for example as a stochastic matrix or as the inverse of an M-matrix. The remaining columns are chosen as $\mathrm{rand}\{-1, 1\}$, with a check for and correction of parallel columns, exactly as for S in the body of the algorithm. We choose random vectors because it is difficult to argue for any particular fixed vectors and because randomness lessens the importance of counterexamples (see the comments in the next section).

Next, we consider complex matrices, which arise in the pseudospectrum application of section 4. Everything in this section remains valid for complex matrices provided that $\mathrm{sign}(A)$ is redefined as the matrix $(a_{ij}/|a_{ij}|)$ (and $\mathrm{sign}(0) = 1$) and

transposes are replaced by conjugate transposes. The matrix S is now complex with elements of unit modulus and we are much less likely to find parallel columns of S from one iteration to the next or within the current S. Therefore for complex matrices we omit the tests for parallel columns. However, we take the same, real, starting matrix. There is one further question in the complex case. In the analogue of Algorithm 2.1 for complex matrices in [14, Alg. 5.1] z is defined as $z = \mathrm{Re}(A^*\xi)$, based on subgradient considerations. In our block algorithm should we take $Z = A^*S$ or $Z = \mathrm{Re}(A^*S)$? The former can be justified from heuristic considerations and preserves more information about A. We return to this question in the next section.

The motivation for Algorithm 2.4 is to enable more accurate and reliable estimates to be obtained than are provided by the 1-norm power method. The question arises of how the accuracy and reliability of the estimates varies with t. Little can be said theoretically because, unlike the approach of [12] mentioned at the start of this section, the estimates are not monotonic in t. If we run Algorithm 2.4 for t_1 and for $t_2 > t_1$, using a common set of t_1 starting vectors, we can obtain a smaller estimate for t_2 than for t_1, because a less promising choice of unit vector e_j can turn out to be better than a more promising choice made with more available information. Nonmonotonicity is unlikely, however, and we argue that it is a price worth paying for the other advantages that accrue. In the next section we investigate the behavior of Algorithm 2.4 empirically.

3. Numerical experiments. Our aim in this section is to answer the following questions about Algorithm 2.4, bearing in mind that for $t = 1$ the algorithm is an implementation of the well-understood 1-norm power method.

1. How does the accuracy and reliability of the norm estimates vary with t?
2. How good are the norm estimates in general?
3. How does the number of iterations behave for $t > 1$?

Note that we are not searching for counterexamples, as was done for previous condition estimators [4], [5], [14]. We know that for a fixed starting matrix and any $t \ll n$ there must be families of matrices whose norm is underestimated by an arbitrarily large factor, since the algorithm samples the behavior of the $n \times n$ matrix A on fewer than n vectors. But since the algorithm uses a random starting matrix for $t > 1$, each counterexample will be valid only for particular starting matrices.

All our tests have been performed with MATLAB.

Our first group of tests deals with random real matrices. Amongst the matrices A we used were the following:

1. A from the normal $N(0,1)$ distribution (denoted `randn`) and its inverse, orthogonal QR factor, upper triangular part, and inverse of the upper triangular part.
2. A from the uniform distribution on the set $\{-1, 0, 1\}$ (denoted `rand({-1,0,1})`), A^{-1} from the uniform distribution on the interval $[0, 1]$.
3. A and A^{-1} of the form $U\Sigma V^T$ where U and V are random orthogonal matrices and $\Sigma = \mathrm{diag}(\sigma_i)$, with the singular values σ_i distributed exponentially, arithmetically, or with all except the smallest equal to unity, and with 2-norm condition number ranging from 1 to 10^{16}.

Note that we omit, for example, matrices from the uniform $[0, 1]$ and uniform $\{-1, 1\}$ distributions, because for such matrices Algorithm 2.4 is easily seen to produce the exact norm for all t.

We chose n and t in the range $25 \le n \le 250$, $1 \le t \le 10$.

For each test matrix we recorded a variety of statistics including the underestimation ratio $\text{est}/\|A\|_1 \le 1$, averaged and minimized over each type of A for fixed n and t, the relative error $|\text{est} - \|A\|_1|/\|A\|_1$, and the number of iterations. We declared an estimate exact if the relative error was no larger than 10^{-14} (the unit roundoff is of order 10^{-16}). For a given matrix A we first generated a starting matrix X_1 with $t_{\max}$ columns, where $t_{\max}$ is the largest value of t to be used, and then ran Algorithm 2.4 for $t = 1\colon t_{\max}$ using starting matrix $X_1(:, 1\colon t)$. In this way we could see the effect of increasing t. In particular, we checked what percentage of the estimates for a given t were at least as large as the estimates for all smaller t; we denote this "improve%." We set itmax = 5 in Algorithm 2.4.

First, we give some general comments on the results.

1. Increasing t usually gave larger average and minimum underestimation ratios, though there were exceptions. The quantity improve% was 100 about half the time and never less than 78.
2. The number of iterations averaged between 2 and 3 throughout, with maxima ranging from 2 to 5 depending on the type of matrix. Thus increasing t from 1 has little effect on the number of iterations—an important fact that could not be predicted from the theory. Although there are specially constructed examples for which the 1-norm power method requires many iterations (one is described below), it is rare for the limit of 5 iterations to come into effect.
3. Throughout the tests we also computed the extra estimate (2.1) used by the LAPACK norm estimator [14]. As expected, in none of our tests (with random matrices) was the extra estimate larger than the estimate provided by Algorithm 2.4 with $t = 1$.

In Tables 3 and 4 we show detailed results for two particular types of random matrix from among those described above, with $n = 100$: `inv(randn)` and `rand({-1,0,1})`. The columns headed "Products" show the average and maximum total number of matrix products AX and A^TS. In each case 5000 matrices were used. For the matrices `inv(randn)` taking $t = 2$ significantly improves the worst-case and average estimates and the proportion of exact estimates over $t = 1$, while for $t = 4$ the estimate is exact almost 98% of the time. For the matrices `rand({-1,0,1})` the improvements as t increases are less dramatic but still useful; notice that exactly four matrix products were required in every case. As well as recording the number of products, we checked how convergence was achieved. For the matrices `rand({-1,0,1})` convergence was always achieved at the test (4) in Algorithm 2.4, while for `inv(randn)` convergence was declared at tests (4) and (2) in approximately 96% and 4% of the cases, respectively (with just a few instances of convergence at (5) for $t \ge 2$). The last two columns of Table 3 show the average and maximum number of parallel columns of S that were detected. A small number of repeated e_j vectors were detected and replaced (the largest average was 0.03, occurring for $t = 2$, and the maximum number of 7 occurred for $t = 8$). No parallel columns or repeated e_j were detected for the matrices in Table 4.

The strategy of replacing parallel columns of S and repeated e_j vectors has little effect on the overall performance of Algorithm 2.4 in our tests with random matrices. Since particular examples can be found where it is beneficial (see Example 2.5) and the cost is negligible, we feel its use is worthwhile. However, we see no advantage to extending the strategy to compare the columns of S with those of all previous S matrices.

TABLE 3
Results for 5000 *matrices* `inv(randn)` *of dimension* 100.

	Underest. ratio			Products			Parallel cols.	
t	min	average	% exact	average	max	improve%	average	max
1	0.176	0.979	83.40	4.3	10		0.00	0
2	0.507	0.993	92.64	4.0	8	98.42	0.09	2
3	0.628	0.997	96.40	4.0	6	98.88	0.22	4
4	0.702	0.999	97.98	4.0	6	99.46	0.37	6
5	0.780	0.999	98.98	4.0	6	99.80	0.52	8
6	0.798	1.000	99.40	4.0	6	99.92	0.68	9
7	0.885	1.000	99.62	4.0	6	99.94	0.86	11
8	0.889	1.000	99.78	4.0	6	99.96	1.04	13
9	0.893	1.000	99.88	4.0	4	100.00	1.22	15
10	0.893	1.000	99.92	4.0	4	100.00	1.39	17

TABLE 4
Results for 5000 *matrices* `rand({-1,0,1})` *of dimension* 100.

	Underest. ratio			Products		
t	min	average	% exact	average	max	improve%
1	0.530	0.836	3.42	4	4	
2	0.588	0.883	6.80	4	4	88.62
3	0.676	0.904	10.12	4	4	87.68
4	0.708	0.917	13.00	4	4	88.08
5	0.733	0.928	16.48	4	4	88.86
6	0.743	0.935	19.24	4	4	88.70
7	0.757	0.941	22.24	4	4	89.20
8	0.761	0.946	25.60	4	4	89.88
9	0.775	0.951	28.56	4	4	90.74
10	0.775	0.956	31.64	4	4	91.44

Higham [15] gives a tridiagonal matrix for which the 1-norm power method (Algorithm 2.1) requires n iterations to converge. We have constructed a matrix for which the maximum $n+1$ iterations is required. It is the inverse of a bidiagonal matrix:

$$A_n(\alpha) = -\begin{bmatrix} 1 & \alpha & & & \\ & 1 & \alpha & & \\ & & 1 & \ddots & \\ & & & \ddots & \alpha \\ & & & & 1 \end{bmatrix}^{-1}$$
$$= -\begin{bmatrix} 1 & -\alpha & \alpha^2 & \dots & (-\alpha)^{n-1} \\ & 1 & -\alpha & & \vdots \\ & & 1 & \ddots & \vdots \\ & & & \ddots & -\alpha \\ & & & & 1 \end{bmatrix} \in \mathbb{R}^{n\times n}, \qquad 0 < \alpha < 1.$$

(The minus sign in front of the matrix is necessary!) It is straightforward to show that when Algorithm 2.1 is applied to $A_n(\alpha)$ it produces, for $k = 1\colon n$,

$$x^{(k)} = e_k, \qquad y^{(k+1)} = Ae_k, \qquad \|y^{(k+1)}\|_1 = \frac{1-\alpha^k}{1-\alpha},$$
$$\xi^{(k+1)} = [\,(-1)^k \quad (-1)^{k-1} \quad \dots \quad -1 \quad 1 \quad \dots \quad 1\,], \qquad \|z^{(k+1)}\|_\infty = |z_{k+1}^{(k+1)}|.$$

TABLE 5
Results for matrix $A_{100}(1-10^{-6})$, *with* 1000 *repetitions.*

	Underest. ratio			Products		
t	min	average	% exact	average	max	improve%
1	0.050	0.050	0.00	11.0	11*	
2	0.290	0.901	60.80	7.8	11*	100.00
3	0.510	0.975	84.90	6.5	11*	96.40
4	0.650	0.997	97.60	5.4	11*	99.60
5	0.840	0.999	99.30	4.9	11*	99.70
6	1.000	1.000	100.00	4.6	11*	100.00
7	1.000	1.000	100.00	4.3	11*	100.00
8	1.000	1.000	100.00	4.2	8	100.00
9	1.000	1.000	100.00	4.1	8	100.00
10	1.000	1.000	100.00	4.1	8	100.00

Thus every column of $A_n(\alpha)$ is computed, in order from first to last, and the exact norm is obtained. If the algorithm is terminated after p iterations, then it produces the estimate $(1-\alpha^{p-1})/(1-\alpha)$, with underestimation ratio $(1-\alpha^{p-1})/(1-\alpha^n)$. For $t = 1$, Algorithm 2.4 behaves in exactly the same way.

We applied Algorithm 2.4 1000 times to $A_{100}(1-10^{-6})$ (with itmax = 5, as in all our tests). The results are shown in Table 5; in the sixth column, "11*" denotes that convergence was declared because the iteration limit was reached (the percentage of such occurrences varied from 100% for $t = 1$ to 0.1% for $t = 7$). The underestimation ratio for $t = 1$ agrees with the theory and is unacceptably small (note that there is no randomness and hence only one estimate for $t = 1$). But for all $t \geq 2$ the average norm estimates are satisfactory. The extra estimate (2.1) has the value 0.561; thus it significantly improves the estimate for $t = 1$ but is worse than the average estimates for all greater t.

For complex matrices we have tried both $Z = A^*S$ and $Z = \mathrm{Re}(A^*S)$ at (3) in Algorithm 2.4. Tables 6 and 7 compare the two choices for 5000 100×100 random complex matrices of the form `inv(rand+i*rand)`, where `rand` is a matrix from the uniform distribution on the interval $[0, 1]$. In this test, larger underestimation ratios are obtained for $Z = A^*S$, the percentage of exact estimates is higher, and the statistics on the number of matrix products are slightly better. In other tests we have found the complex choice of Z always to perform at least as well, overall, as the real choice (see, for example, the riffle shuffle example in section 4). We therefore keep Z complex in Algorithm 2.4. In version 3.0 of LAPACK the norm estimator has been modified from that in version 2.0 to keep Z (here a vector) complex.

Finally, how does Algorithm 2.4 compare with the suggestion of Hager mentioned at the beginning of section 2 of running the 1-norm power method t times in succession? In tests with random matrices we have found Hager's approach to produce surprisingly good norm estimates, but they are generally inferior to those from Algorithm 2.4. Since Hager's approach is based entirely on level 2 BLAS operations, Algorithm 2.4 is clearly preferred.

4. Computing 1-norm pseudospectra. In this section we apply the complex version of Algorithm 2.4 to the computation of 1-norm pseudospectra. For $\epsilon \geq 0$ and any subordinate matrix norm the ϵ-pseudospectrum of $A \in \mathbb{C}^{n\times n}$ is defined by [22]

$$\Lambda_\epsilon(A) = \{\, z \in \mathbb{C} : z \text{ is an eigenvalue of } A + E \text{ for some } E \text{ with} \|E\| \leq \epsilon \,\},$$

TABLE 6
Results for 5000 *matrices* `inv(rand+i*rand)` *of dimension* 100*, with* $Z = A^*S$.

	Underest. ratio			Products		
t	min	average	% exact	average	max	improve%
1	0.456	0.980	76.04	4.2	8	
2	0.688	0.994	89.92	4.0	8	98.46
3	0.746	0.997	95.30	4.0	6	99.14
4	0.763	0.999	97.46	4.0	6	99.34
5	0.763	0.999	98.68	4.0	6	99.62
6	0.859	1.000	99.18	4.0	6	99.82
7	0.859	1.000	99.50	4.0	4	99.88
8	0.859	1.000	99.70	4.0	4	99.96
9	0.859	1.000	99.86	4.0	4	100.00
10	0.954	1.000	99.90	4.0	4	100.00

TABLE 7
Results for 5000 *matrices* `inv(rand+i*rand)` *of dimension* 100*, with* $Z = \mathrm{Re}(A^*S)$.

	Underest. ratio			Products		
t	min	average	% exact	average	max	improve%
1	0.403	0.908	36.82	4.2	8	
2	0.503	0.955	56.94	4.1	8	93.98
3	0.633	0.975	71.24	4.0	8	94.92
4	0.641	0.986	81.10	4.0	6	96.22
5	0.641	0.992	87.52	4.0	6	97.48
6	0.690	0.995	92.36	4.0	6	98.54
7	0.781	0.997	95.00	4.0	6	99.20
8	0.801	0.998	96.70	4.0	6	99.44
9	0.819	0.999	98.12	4.0	6	99.74
10	0.822	0.999	98.68	4.0	6	99.84

or, equivalently, in terms of the resolvent $(zI - A)^{-1}$, as

$$\Lambda_\epsilon(A) = \{\, z \in \mathbb{C} : \|(zI - A)^{-1}\| \geq \epsilon^{-1} \,\}.$$

Most published work on pseudospectra has dealt with the 2-norm, and the utility of 2-norm pseudospectra in revealing the effects of nonnormality is well appreciated [19], [20], [22].

The 2-norm and any other p-norm of an $n \times n$ matrix differ by a factor at most $\sqrt{n}$. For small n, pseudospectra therefore do not vary much between different p-norms. However, Jónsson and Trefethen have shown [17] that in Markov chain applications the choice of norm for pseudospectra can be crucial. In Markov chains representing a random walk on an m-dimensional hypercube and a riffle shuffle of m cards, the transition matrices are of dimension exponential in m and factorial in m, respectively. It is known that when measured in an appropriate way, these random processes converge to a steady state not gradually but suddenly after a certain number of steps. As these processes involve powers of matrices of possibly huge dimension, and as the 1-norm is the natural norm for probability,[1] 1-norm pseudospectra are an important tool for explaining the transient behavior [17].

One of the most useful graphical representations of pseudospectra is a plot of level curves of the resolvent. We therefore consider the standard approach of evaluating $\|(zI - A)^{-1}\|_1$ on an equally spaced grid of points z in some region of interest in the

[1] Because of the use of row vectors in the Markov chain literature, it is actually the ∞-norm that is relevant. By using $\|A\|_1 = \|A^T\|_\infty$ we can continue to work with the 1-norm.

complex plane and sending the results to a contour plotter. A variety of methods for carrying out these computations for the 2-norm are surveyed by Trefethen [21], but most of the ideas employed are not directly applicable to the 1-norm.

Explicitly forming $(zI - A)^{-1}$ at each grid point is computationally expensive. A more efficient approach is to factorize $P(zI - A) = LU$ at each point z by LU factorization with partial pivoting and to use Algorithm 2.4 to *estimate* $\|(zI-A)^{-1}\|_1$; the matrix multiplications in the algorithm become triangular solves with multiple right-hand sides. This approach can take advantage of sparsity in A. However, the method still requires $O(n^3)$ operations per grid point when A is full.

We consider instead a more efficient approach that is applicable when we can compute a Schur factorization [9, Chap. 7]

$$A = QTQ^*, \tag{4.1}$$

where Q is unitary and T is upper triangular. Given this factorization we have

$$zI - A = Q(zI - T)Q^*, \tag{4.2}$$

so that forming a matrix product with $(zI - A)^{-1}$ or its conjugate transpose reduces to solving a multiple right-hand side triangular system and multiplying by Q and Q^*. Given this initial decomposition the cost per grid point of estimating the resolvent norm using Algorithm 2.4 is just $O(n^2t)$ flops—a substantial saving over the first approach, and of the same order of magnitude as the cost of standard methods for computing 2-norm pseudospectra (provided t is small). In place of the Schur decomposition we could use a Hessenberg decomposition, computed either by Gauss transformations or Householder transformations [9, Sec. 7.4]; these decompositions are less expensive, but the cost per grid point is a larger multiple of n^2 flops because of the need to factorize a Hessenberg matrix.

In more detail our approach is as follows.

ALGORITHM 4.1 (1-norm pseudospectra estimation). *Given $A \in \mathbb{C}^{n\times n}$ and a positive integer t, this algorithm estimates $\|(zI - A)^{-1}\|_1$ on a specified grid of points in the complex plane, using Algorithm* 2.4 *with parameter t.*

```
Compute the Schur factorization A = QTQ*.
for each grid point z
    Apply the complex version of Algorithm 2.4 to (zI - A)^{-1}
    with parameter t, using the representation (4.2).
end
```

Our experience is that Algorithm 4.1 frequently leads to visually acceptable contour plots even for $t = 1$. As the following example shows, however, a larger value of t may be needed. We take an example from Markov chains: the Gilbert–Shannon–Reeds model of a riffle shuffle on a deck of n cards. The transition matrix P is of dimension $n!$. Remarkably, as Jónsson and Trefethen explain [17], the dimension of the matrix can be reduced to n by certain transformations that preserve the 1-norms of powers and of the resolvent. For this experiment we took $n = 52$ and, as in [17], we computed pseudospectra of the "decay matrix" $A = P - \lim_{k\to\infty} P^k$, working with the reduced form of A.

Figure 1 shows approximations to the 1-norm pseudospectra computed on a 100×100 grid, with $t = 1, 2, 3$ in Algorithm 2.4. Contours are plotted for $\epsilon = 10^{-1}, 10^{-1.5}, \ldots, 10^{-4}$, the dashed line marks the unit circle, and the eigenvalues are plotted as dots. We did not exploit the fact that, since A is real, the pseudospectra are symmetric about the real axis. The contour plot for $t = 1$ is clearly incorrect in

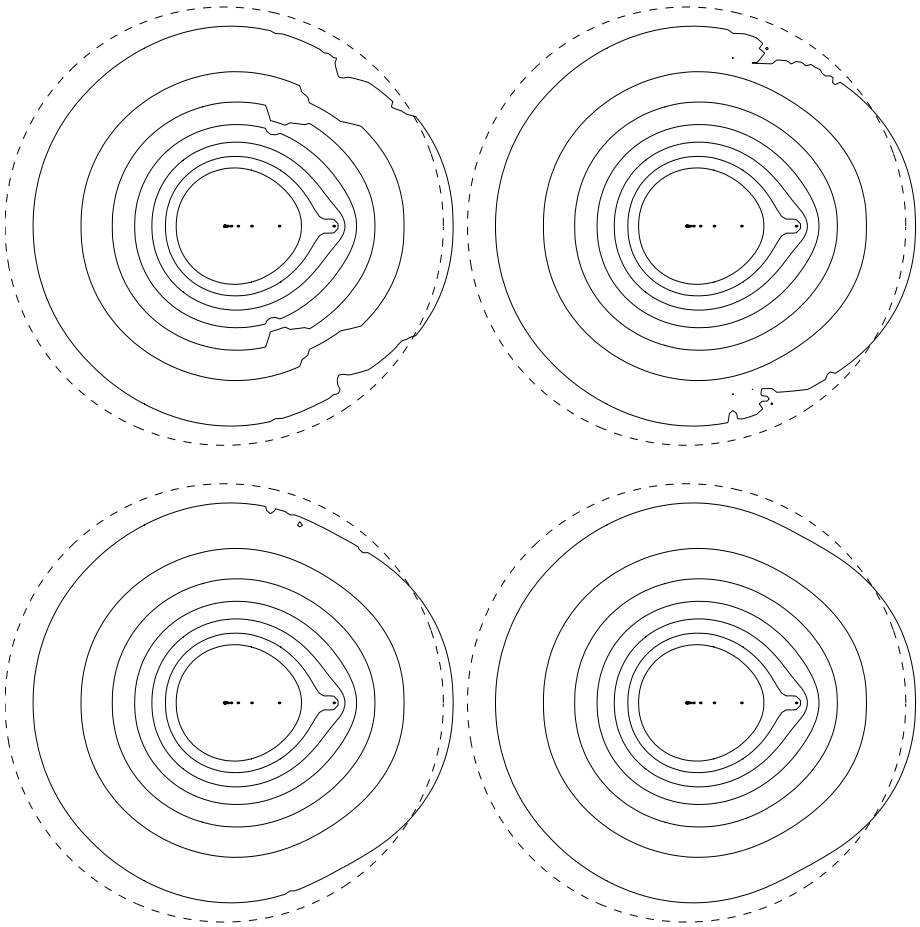


FIG. 1. 1-*norm pseudospectra for the riffle shuffle. Clockwise from top left:* $t = 1$ *and* $Z = \mathrm{Re}(A^*S)$*,* $t = 1$*,* $t = 2$*,* $t = 3$*.*

the outer contour, while $t = 2$ yields an improvement and the plot for $t = 3$ is correct to visual accuracy. Table 8 summarizes the key statistics from these computations, showing that on average the norm estimates had about t correct significant digits for $t = 1, 2, 3$. The figure and the table confirm that it is better to keep Z complex in Algorithm 2.4.

We give a further example, in which A is a spectral discretization of an integral operator of Landau [21, Sec. 21]; like the operator, A is complex and symmetric. We took dimension $n = 250$ and Fresnel number 8. As noted in [21] for the 2-norm, a fine grid is needed to resolve the details for this example; we used a 200×200 grid. Figure 2 shows the computed pseudospectra for $t = 1, 4, 8, 16$, and Table 9 summarizes the statistics for these values and for $t = 2$. Contours are plotted for $\epsilon = 10^{-1}, 10^{-1.5}, \ldots, 10^{-10}$, the dashed line marks the unit circle, and the eigenvalues are plotted as dots. In Table 9, "11^*" denotes that convergence was declared because the iteration limit was reached (the percentage of such occurrences was 0.12% for $t = 1$ and 0.025% for $t = 2$). For $t = 1$ one of the contour lines misses an eigenvalue

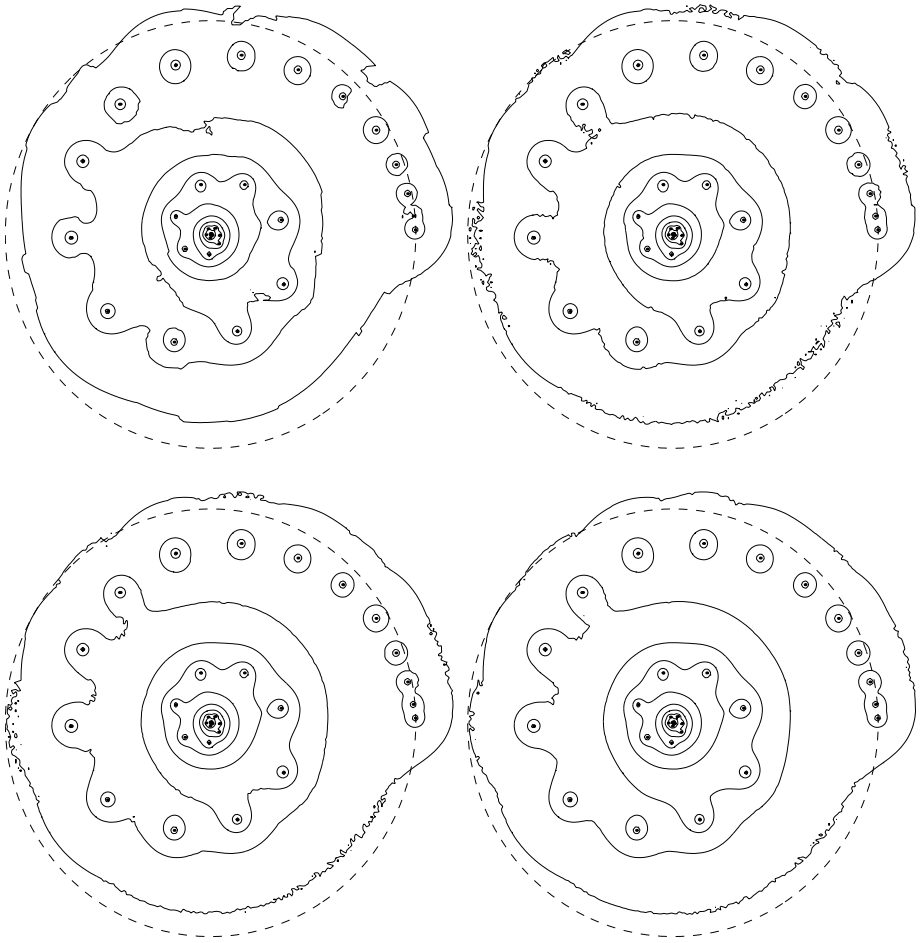


FIG. 2. 1-*norm pseudospectra for the Landau matrix. Clockwise from top left:* $t = 1$, $t = 4$, $t = 8$, $t = 16$.

in the northwest corner of the plot. For $t = 16$ the plot differs from the exact one only by some tiny oscillations in two outer contours. While Algorithm 2.4 performs well even for small t, as measured by the underestimation ratio, quite accurate values of the 1-norm of the resolvent are needed in this example in order to produce smooth contours.

5. Conclusions. We have derived a new matrix 1-norm estimation algorithm, Algorithm 2.4, with a number of key features. Most importantly, the algorithm has a parameter t that can be used to control the accuracy and reliability of the estimate (which is actually a lower bound). While there is no guarantee that increasing t increases the estimate (leaving aside the fact that the starting matrix is partly random), the estimate typically does increase with t, leading quickly to one or more correct significant digits in the estimate. A crucial property of the algorithm is that the number of iterations and matrix products required for convergence is essentially independent of t (for random matrices about two iterations are required on average,

TABLE 8
Results for riffle shuffle example. $t = 1^*$ *denotes* $t = 1$ *with* $Z = \mathrm{Re}(A^*S)$.

	Underest. ratio			Products	
t	min	average	% exact	average	max
1^*	0.464	0.963	80.06	4.5	10
1	0.486	0.978	85.90	4.1	8
2	0.621	0.989	89.88	4.1	8
3	0.686	0.996	94.30	4.0	6

TABLE 9
Results for Landau matrix example.

	Underest. ratio			Products	
t	min	average	% exact	average	max
1	0.096	0.920	13.87	4.2	11^*
2	0.511	0.954	17.95	4.1	11^*
4	0.666	0.971	29.00	4.0	10
8	0.786	0.984	43.22	4.0	8
16	0.846	0.993	61.02	4.0	6

corresponding to four products of $n \times n$ and $n \times t$ matrices). The algorithm avoids redundant computations and keeps constant the size of the matrix multiplications. In future work we intend to investigate how the choice of t affects the efficiency of the algorithm in a high-performance computing environment.

Unlike the statistically based norm estimation techniques in [7], [11] which currently apply only to real matrices, our algorithm handles both real and complex matrices.

Since our algorithm uses a partly random starting matrix for $t \geq 2$, it is natural to ask whether bounds, valid for all A, can be obtained on the probability of the estimate being within a certain factor of $\|A\|_1$. We feel that the very features of the algorithm that make it so effective make it difficult or impossible to derive useful bounds of this type.

For $t = 1$ our algorithm is very similar to the estimator in LAPACK, the differences being that our algorithm omits the extra estimate (2.1) and for real matrices we test for parallel rather than simply repeated sign vectors (which improves the efficiency). Unlike in the estimator in LAPACK 2.0, for complex matrices we do not take the real part of the z vector (which improves both the quality of the estimates and the efficiency), and this change has been incorporated into LAPACK 3.0.

The new algorithm makes an attractive replacement for the existing LAPACK estimator. The value $t = 1$ would be the natural default choice (with the extra estimate (2.1) included for extra reliability and backward compatibility) and a user willing to pay more for more accurate and reliable 1-norm estimates would have the option of choosing a larger t.

Acknowledgment. We thank Nick Trefethen for providing M-files that compute the riffle shuffle and Landau matrices used in section 4.

REFERENCES

[1] E. ANDERSON, Z. BAI, C. H. BISCHOF, S. BLACKFORD, J. W. DEMMEL, J. J. DONGARRA, J. J. DU CROZ, A. GREENBAUM, S. J. HAMMARLING, A. MCKENNEY, AND D. C. SORENSEN, *LAPACK Users' Guide*, 3rd ed., SIAM, Philadelphia, PA, 1999.

[2] M. Arioli, J. W. Demmel, and I. S. Duff, *Solving sparse linear systems with sparse backward error*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 165–190.
[3] D. W. Boyd, *The power method for ℓ^p norms*, Linear Algebra Appl., 9 (1974), pp. 95–101.
[4] A. K. Cline, C. B. Moler, G. W. Stewart, and J. H. Wilkinson, *An estimate for the condition number of a matrix*, SIAM J. Numer. Anal., 16 (1979), pp. 368–375.
[5] A. K. Cline and R. K. Rew, *A set of counter-examples to three condition number estimators*, SIAM J. Sci. Stat. Comput., 4 (1983), pp. 602–611.
[6] J. W. Demmel, *Open Problems in Numerical Linear Algebra*, IMA Preprint Series #961, Institute for Mathematics and its Applications, University of Minnesota, Minneapolis, MN, 1992.
[7] J. D. Dixon, *Estimating extremal eigenvalues and condition numbers of matrices*, SIAM J. Numer. Anal., 20 (1983), pp. 812–814.
[8] J. J. Dongarra, J. R. Bunch, C. B. Moler, and G. W. Stewart, *LINPACK Users' Guide*, SIAM, Philadelphia, PA, 1979.
[9] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[10] W. B. Gragg and G. W. Stewart, *A stable variant of the secant method for solving nonlinear equations*, SIAM J. Numer. Anal., 13 (1976), pp. 889–903.
[11] T. Gudmundsson, C. S. Kenney, and A. J. Laub, *Small-sample statistical estimates for matrix norms*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 776–792.
[12] W. W. Hager, *Condition estimates*, SIAM J. Sci. Stat. Comput., 5 (1984), pp. 311–316.
[13] N. J. Higham, *A survey of condition number estimation for triangular matrices*, SIAM Rev., 29 (1987), pp. 575–596.
[14] N. J. Higham, *FORTRAN codes for estimating the one-norm of a real or complex matrix, with applications to condition estimation (Algorithm* 674*)*, ACM Trans. Math. Software, 14 (1988), pp. 381–396.
[15] N. J. Higham, *Experience with a matrix norm estimator*, SIAM J. Sci. Stat. Comput., 11 (1990), pp. 804–809.
[16] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, PA, 1996.
[17] G. F. Jónsson and L. N. Trefethen, *A numerical analyst looks at the "cutoff phenomenon" in card shuffling and other Markov chains*, in Numerical Analysis 1997, Proceedings of the 17th Dundee Biennial Conference, D. F. Griffiths, D. J. Higham, and G. A. Watson, eds., Pitman Res. Notes Math. Ser. 380, Addison Wesley Longman, Harlow, Essex, UK, 1998, pp. 150–178.
[18] P. D. Tao, *Convergence of a subgradient method for computing the bound norm of matrices*, Linear Algebra Appl., 62 (1984), pp. 163–182 (in French).
[19] L. N. Trefethen, *Pseudospectra of matrices*, in Numerical Analysis 1991, Proceedings of the 14th Dundee Conference, D. F. Griffiths and G. A. Watson, eds., Pitman Res. Notes Math. 260, Longman Scientific and Technical, Essex, UK, 1992, pp. 234–266.
[20] L. N. Trefethen, *Pseudospectra of linear operators*, SIAM Rev., 39 (1997), pp. 383–406.
[21] L. N. Trefethen, *Computation of pseudospectra*, in Acta Numerica 8, A. Iseries, ed., Cambridge University Press, Cambridge, UK, 1999, pp. 247–295.
[22] L. N. Trefethen, *Spectra and Pseudospectra: The Behavior of Non-Normal Matrices and Operators*, 1999, in preparation.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 4, pp. 1202–1228
© 2000 Society for Industrial and Applied Mathematics

AN EFFICIENT AND STABLE ALGORITHM FOR THE SYMMETRIC-DEFINITE GENERALIZED EIGENVALUE PROBLEM*

S. CHANDRASEKARAN†

Abstract. A new, efficient, and stable algorithm for computing all the eigenvalues and eigenvectors of the problem $Ax = \lambda Bx$, where A is symmetric indefinite and B is symmetric positive definite, is proposed.

Key words. generalized eigenvalues, eigenvalues, eigenvectors, error analysis, deflation, perturbation bounds, pencils, stable algorithms

AMS subject classifications. 15A18, 15A22, 15A23, 15A42, 47A75, 65F15

PII. S0895479897316308

1. Introduction.

In this paper we consider the problem of computing the eigenvalues and eigenvectors of the pencil $Ax = \lambda Bx$, where A is a real symmetric-indefinite matrix and B is a real symmetric positive-definite matrix. Mathematically, this problem is equivalent to computing the eigendecomposition of the symmetric matrix $G^{-1}AG^{-T}$, where $B = GG^T$. Unfortunately, the approach is not numerically stable, but it does reveal some important properties about the eigenvalues and eigenvectors. First, all the eigenvalues must be real. Second, the eigenvector matrix diagonalizes both A and B simultaneously. In finite precision, the transformation $G^{-1}AG^{-T}$ leads to violation of the second property, while the QZ algorithm violates the first property.

In this paper we propose a new algorithm which satisfies both properties and is numerically stable and efficient.

Previous work on this problem, when the matrices are dense, has involved either trying to implement the transformation $G^{-1}AG^{-T}$ accurately or extending Jacobi, QZ, or other iterative type methods. See section 8.7 in [2], section 5.68 in [8], and chapter 15 in [6] for a summary of earlier work. Iterative methods, which can be used for both dense and sparse problems, have been studied more extensively. See [7] for a more extensive guide to the literature.

The outline of this paper is as follows. To convey the basic ideas we first outline the algorithm assuming that the symmetric eigenvalue problem can be solved exactly. We then point out the difficulties introduced by inexact calculations and the methods we propose for overcoming them. This is followed by an error analysis to prove the stability of the algorithm. We then discuss implementation issues and describe the experimental results which validate our claims. As part of the error analysis, we also establish a perturbation bound for the smallest eigenvalues in magnitude, which we believe to be new.

1.1. The key idea.

The problem can be viewed as the simultaneous LDL^T factorization of the matrices A and B, where L is now no longer constrained to be a triangular matrix. As is well known, while the LDL^T factorization of B (which is symmetric positive definite) is stable even without pivoting (Cholesky factorization),

*Received by the editors February 7, 1997; accepted for publication (in revised form) by F. T. Luk July 14, 1999; published electronically March 30, 2000. This research was supported in part by NSF grant CCR-9734290.

http://www.siam.org/journals/simax/21-4/31630.html

†Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 (shiv@ece.ucsb.edu).

the LDL^T factorization of A, which is symmetric indefinite, can be severely unstable without pivoting. When viewed from this framework, the problem is to discover a "permutation" (actually, any orthogonal transform) such that both A and B can be simultaneously factored into LDL^T, while assuring at the same time that stability is maintained in the factorization of A. This is a difficult task. The key idea in this paper is that the correct pivoting order is obtained when the eigenvalues are deflated in *decreasing order of magnitude*. This is an oversimplification, but it helps in understanding the rest of the paper.

2. Notation. We will follow Householder's convention [3] and denote matrices by capital Roman letters and vectors by small Roman letters. Scalars will be denoted by small Greek letters. Due to the necessity of keeping track of the various variables of the algorithm in the error analysis, we will avoid statements such as $A := A + I$. Instead the same statement will be denoted by $A_{;i+1} \leftarrow A_{;i} + I$. That is, the letters after the semicolon in the suffix help keep track of the position in the algorithm where the particular variable was created. The letters after the semicolon also give cues to where the particular variable was created. For example, $A_{;i;q}$ is used to indicate the matrix $A_{;i}$ after a similarity transformation by a matrix labeled by Q. Similarly the letters w, t, and s also appear in subscripts after semicolons and their cues can be inferred from their definition contexts. Also the notation $A_{2,3;i}$ denotes the element in the $(2,3)$ position of the matrix $A_{;i}$, and $A_{j;i}$ denotes the jth column of the matrix $A_{;i}$. We will use Matlab-type notation [5] for submatrices. Therefore, $A_{[i:n,p:q];i}$ denotes the submatrix of $A_{;i}$ formed from those elements in rows i to n which are also in columns p to q. When there is no necessity to refer to the elements of a matrix we will drop the semicolon in the suffix. This will be clear from the context. A bar over a variable indicates that it is a submatrix.

We use $\equiv$ when variables are being defined for the first time. In the description of algorithms we use left and right arrows to denote assignment. Therefore, $X_{;i} \leftarrow A_{;i}B_{;i}$ means that the product of the matrices $A_{;i}$ and $B_{;i}$ is assigned to the variable $X_{;i}$. Whereas,

$$A_{;i}B_{;i} \rightarrow U_{;i}\Sigma_{;i}V_{;i}^T, \qquad \text{compute SVD},$$

means that the SVD of the product $A_{;i}B_{;i}$ is computed, and $U_{;i}$ is assigned the left singular vector matrix, $\Sigma_{;i}$ is assigned the matrix of singular values, and $V_{;i}$ is assigned the right singular vector matrix.

3. In infinite precision. In this section we first present the algorithm assuming that all calculations (including some eigendecompositions!) can be done exactly. This is to enable us to present the basic ideas in an uncluttered manner.

We assume that A is nonsingular. If it is not, then the zero eigenvalues can be easily deflated from the problem, as detailed in section 5.

We note that the main difficulty in finding the eigenvalues and eigenvectors is due to finite precision effects. So the objective is to try to design an algorithm which will work for problems, where the symmetric positive-definite matrix B is almost numerically singular. For otherwise, if the matrix B is well-conditioned, we can compute the transformation $G^{-1}AG^{-T}$ with little loss of accuracy.

One key observation we make now is that even when both A and B are highly ill-conditioned we can compute the matrix $G^TA^{-1}G$ to sufficient accuracy so as to enable us to compute its largest eigenvalues in magnitude to backward accuracy. One way to proceed now is to deflate these computed eigenvalues from A and B and to

work recursively on the smaller pencil. Unfortunately, the deflation requires transformations whose cumulative condition number is as high as the condition number of B. So we seem to fare no better.

This is where our next key observation comes in: if we deflate the eigenvalues of $G^T A^{-1} G$ in the order of decreasing size in magnitude, then we can ensure that the resulting sequence of deflating transformations *can* be implemented in a numerically stable manner. The rest of this paper is devoted to showing why and how this can be done.

We first begin by showing why it is necessary to deflate the eigenvalues in the order of decreasing size in magnitude. For that purpose we present a version of the algorithm here which assumes that all computations can be done exactly, including some eigenvalue decompositions.

ALGORITHM I. USING EXACT EIGENDECOMPOSITIONS.

Begin

1. $A_{;1} \equiv A; \qquad B \equiv B_{;1} \rightarrow U \Sigma U^T$; compute eigendecomposition of B.
2. $\sqrt{\Sigma} U^T A_{;1}^{-1} U \sqrt{\Sigma} \rightarrow V \Lambda V^T$; compute the eigendecomposition of equivalent symmetric matrix such that the eigenvalues are ordered from largest to smallest in magnitude.
3. $X_{;1} \equiv X \leftarrow A_{;1}^{-1} U \sqrt{\Sigma} V$; compute the generalized eigenvectors of the pencil.
4. We now deflate the eigenvectors from the pencil.
   For $i = 1$ **to** n **do**
   (a) Compute Householder transform $Q_{;i}$ such that $Q_{;i} X_{i;i}$ is parallel to e_i.
   (b) $A_{;i;q} \leftarrow Q_{;i} A_{;i} Q_{;i}^T, \quad B_{;i;q} \leftarrow Q_{;i} B_{;i} Q_{;i}^T, \quad X_{;i;q} \leftarrow Q_{;i} X_{;i}$.
   (c) Compute Householder transform W_i such that $W_i A_{i;i;q}$ has zeros below the $(i+1)$st component. It follows that $W_i B_{i;i;q}$ also has zeros below the $(i+1)$st component.
   (d) $A_{;i;w} \leftarrow W_i A_{;i;q} W_i^T, \quad B_{;i;w} \leftarrow W_i B_{;i;q} W_i^T, \quad X_{;i;w} \leftarrow W_i X_{;i;q}$.
   (e) Note that $X_{i;i;w}$ is still parallel to e_i. Therefore, $A_{i;i;w}$ and $B_{i;i;w}$ are parallel and in the span of e_i and e_{i+1}. Therefore, we can find one elementary Gauss transform L_i^{-1} such that $L_i^{-1} A_{i;i;w}$ and $L_i^{-1} B_{i;i;w}$ are both parallel to e_i.
   (f) $A_{;i+1} \leftarrow L_i^{-1} A_{;i;w} L_i^{-T}, \quad B_{;i+1} \leftarrow L_i^{-1} B_{;i;w} L_i^{-T}, \quad X_{;i+1} \leftarrow L_i^T X_{;i;w}$.
   endfor
5. We now have

$$(L_n^{-1} W_n Q_{;n}) \cdots (L_1^{-1} W_1 Q_{;1}) A ((L_n^{-1} W_n Q_{;n}) \cdots (L_1^{-1} W_1 Q_{;1}))^T \equiv D_A,$$
$$(L_n^{-1} W_n Q_{;n}) \cdots (L_1^{-1} W_1 Q_{;1}) B ((L_n^{-1} W_n Q_{;n}) \cdots (L_1^{-1} W_1 Q_{;1}))^T \equiv D_B,$$

where D_A and D_B are diagonal matrices. Now define

$$C \equiv Q_{;1}^T W_1^T L_1 \cdots Q_{;n}^T W_n^T L_n. \tag{1}$$

Then we have $C D_A C^T = A$ and $C D_B C^T = B$, where D_A and D_B are diagonal matrices. Therefore, the generalized eigenvalues can be obtained as the ratios of the diagonal elements of D_A and D_B, and the generalized eigenvectors can be obtained from the inverse of C by using the factored form (1).

End

We now look in more detail at the transforms $Q_{;i}$, W_i, and L_i used during the deflation process. We claim that

$$X_{;i} \equiv \begin{array}{c} \\ i-1 \\ n-i+1 \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} D_{;X;i} & 0 & 0 \\ 0 & x_i & \bar{X}_{;i} \end{pmatrix} \end{array}, \tag{2}$$

$$A_{;i} \equiv \begin{array}{c} \\ i-1 \\ n-i+1 \end{array} \begin{array}{c} \begin{array}{cc} i-1 & n-i+1 \end{array} \\ \begin{pmatrix} \bar{D}_{;A;i} & 0 \\ 0 & \bar{A}_{;i} \end{pmatrix} \end{array}, \tag{3}$$

$$B_{;i} \equiv \begin{array}{c} \\ i-1 \\ n-i+1 \end{array} \begin{array}{c} \begin{array}{cc} i-1 & n-i+1 \end{array} \\ \begin{pmatrix} \bar{D}_{;B;i} & 0 \\ 0 & \bar{B}_{;i} \end{pmatrix} \end{array}, \tag{4}$$

where D's denote diagonal matrices. These facts will be proved by induction. It is obviously true for $i = 1$. Invoking the induction hypothesis, we see that

$$Q_{;i} \equiv \begin{array}{c} \\ i-1 \\ n-i+1 \end{array} \begin{array}{c} \begin{array}{cc} i-1 & n-i+1 \end{array} \\ \begin{pmatrix} I & 0 \\ 0 & \bar{Q}_i \end{pmatrix} \end{array}, \qquad \bar{Q}_i x_i = \pm \|x_i\| e_i,$$

$$X_{;i;q} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} D_{;X;i} & 0 & 0 \\ 0 & \pm\|x_i\| & h_i^T \\ 0 & 0 & \bar{X}_{;i;q} \end{pmatrix} \end{array},$$

$$A_{;i;q} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} \bar{D}_{;A;i} & 0 & 0 \\ 0 & \alpha_i & a_{;i;q}^T \\ 0 & a_{;i;q} & \bar{A}_{;i;q} \end{pmatrix} \end{array},$$

$$B_{;i;q} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} \bar{D}_{;B;i} & 0 & 0 \\ 0 & \beta_i & b_{;i;q}^T \\ 0 & b_{;i;q} & \bar{B}_{;i;q} \end{pmatrix} \end{array}, \qquad \lambda_i a_{;i;q} = b_{;i;q},$$

where $\Lambda \equiv \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Therefore,

$$W_i \equiv \begin{array}{c} \\ i \\ n-i \end{array} \begin{array}{c} \begin{array}{cc} i & n-i \end{array} \\ \begin{pmatrix} I & 0 \\ 0 & \bar{W}_i \end{pmatrix} \end{array}, \qquad \bar{W}_i a_{;i;q} = \pm \|a_{;i;q}\| e_i,$$

$$X_{;i;w} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} D_{;X;i} & 0 & 0 \\ 0 & \pm\|x_i\| & h_i^T \\ 0 & 0 & \bar{X}_{;i;w} \end{pmatrix} \end{array},$$

$$A_{;i;w} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} \bar{D}_{;A;i} & 0 & 0 \\ 0 & \alpha_i & \pm\|a_{;i;q}\| e_1^T \\ 0 & \pm\|a_{;i;q}\| e_1 & \bar{A}_{;i;w} \end{pmatrix} \end{array},$$

$$B_{;i;w} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} \bar{D}_{;B;i} & 0 & 0 \\ 0 & \beta_i & \pm\|b_{;i;q}\|e_1^T \\ 0 & \pm\|b_{;i;q}\|e_1 & \bar{B}_{;i;w} \end{pmatrix} \end{array},$$

$$\begin{pmatrix} \beta_i \\ \pm\|b_{;i;q}\| \end{pmatrix} = \lambda_i \begin{pmatrix} \alpha_i \\ \pm\|a_{;i;q}\| \end{pmatrix}.$$

Therefore, we have that

$$L_i^{-1} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} I & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \rho_i e_1 & I \end{pmatrix} \end{array}, \qquad \rho_i = -\frac{\pm\|b_{;i;q}\|}{\beta_i} = -\frac{\pm\|a_{;i;q}\|}{\alpha_i},$$

$$X_{;i+1} = \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} D_{;X;i} & 0 & 0 \\ 0 & \pm\|x_i\| & 0 \\ 0 & 0 & \bar{X}_{;i;w} \end{pmatrix} \end{array},$$

$$A_{;i+1} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} \bar{D}_{;A;i} & 0 & 0 \\ 0 & \alpha_i & 0 \\ 0 & 0 & \bar{A}_{;i;w} - \frac{\|a_{;i;q}\|^2}{\alpha_i} e_1 e_1^T \end{pmatrix} \end{array},$$

$$B_{;i+1} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} \bar{D}_{;B;i} & 0 & 0 \\ 0 & \beta_i & 0 \\ 0 & 0 & \bar{B}_{;i;w} - \frac{\|b_{;i;q}\|^2}{\beta_i} e_1 e_1^T \end{pmatrix} \end{array},$$

where the structure of the ith row of $X_{;i+1}$ is obtained by looking at the form of $A_{;i+1}$ and $B_{;i+1}$. This completes our induction and proves the structures assumed in (2), (3), and (4).

We now show that the norms of the Schur complements of A generated by the transforms L_i grow no faster than those which occur in Gaussian elimination with partial pivoting. This is one of the reasons why our approach leads to a numerically stable algorithm.

Define the following two submatrices of $A_{;i;w}$ and $B_{;i;w}$:

$$\bar{A}_{;i;l} \equiv \begin{array}{c} \\ 1 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 1 & 1 \end{array} \\ \begin{pmatrix} \alpha_i & \pm\|a_{;i;q}\| \\ \pm\|a_{;i;q}\| & \bar{A}_{1,1;i;w} \end{pmatrix} \end{array},$$

$$\bar{B}_{;i;l} \equiv \begin{array}{c} \\ 1 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 1 & 1 \end{array} \\ \begin{pmatrix} \beta_i & \pm\|b_{;i;q}\| \\ \pm\|b_{;i;q}\| & \bar{B}_{1,1;i;w} \end{pmatrix} \end{array} = \begin{pmatrix} \lambda_i\alpha_i & \pm\|a_{;i;q}\|\lambda_i \\ \pm\|a_{;i;q}\|\lambda_i & \bar{B}_{1,1;i;w} \end{pmatrix}.$$

Note that $\bar{B}_{;i;l}$ is symmetric positive definite. From this we get, using determinants, that

$$\bar{B}_{1,1;i;w}\lambda_i\alpha_i > \|a_{;i;q}\|^2\lambda_i^2,$$

which implies that

$$|\lambda_i| < \bar{B}_{1,1;i;w} \frac{|\alpha_i|}{\|a_{;i;q}\|^2}. \tag{5}$$

Since we required the following ordering of the eigenvalues

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|,$$

it follows that λ_i is also the largest eigenvalue in magnitude of the pencil $\lambda \bar{A}_{;i;l} x = \bar{B}_{;i;l}$. Using the variational characterization of eigenvalues, we get

$$|\lambda_i| > \frac{\bar{B}_{1,1;i;w}}{|\bar{A}_{1,1;i;w}|}. \tag{6}$$

From inequalities (5) and (6) we get

$$|\bar{A}_{1,1;i;w}| > \frac{\|a_{;i;q}\|^2}{|\alpha_i|}.$$

Therefore,

$$\left| \bar{A}_{1,1;i;w} - \frac{\|a_{;i;q}\|^2}{\alpha_i} \right| < 2|\bar{A}_{1,1;i;w}|. \tag{7}$$

This indicates that the elements of D_A can grow at most like 2^n. Also,

$$\sqrt{|\alpha_i|}|\rho_i| = \sqrt{\frac{\|a_{;i;q}\|^2}{|\alpha_i|}} < \sqrt{|\bar{A}_{1,1;i;w}|}.$$

These facts by themselves are not sufficient to establish the numerical stability of the algorithm. We now proceed to look at the effects of errors in the eigendecomposition computation.

4. Error propagation. In this section we assume that $\|B\| = \|A\| = 1$. We consider the effects of the truncation error

$$G \equiv U\sqrt{\Sigma}, \qquad G^T A^{-1} G \to \hat{V}\hat{\Lambda}\hat{V}^T + E,$$

where $\|E\| \leq \epsilon\|\hat{\Lambda}\|$. Recovering the generalized eigenvector, we have that

$$GG^T(A^{-1}G\hat{v}_i) + GE\hat{v}_i = \hat{\lambda}_i A(A^{-1}G\hat{v}_i).$$

Define $\hat{x}_i \equiv A^{-1}G\hat{v}_i$ and rearrange the above expression to get it in normalized backward error form:

$$\begin{aligned} GG^T \frac{\hat{x}_i}{\|\hat{x}_i\|} + \frac{GE\hat{v}_i}{\|\hat{x}_i\|} &= \hat{\lambda}_i A \frac{\hat{x}_i}{\|\hat{x}_i\|} \qquad \text{if } |\hat{\lambda}_i| < 1, \\ \frac{1}{\hat{\lambda}_i} GG^T \frac{\hat{x}_i}{\|\hat{x}_i\|} + \frac{GE\hat{v}_i}{\hat{\lambda}_i\|\hat{x}_i\|} &= A \frac{\hat{x}_i}{\|\hat{x}_i\|} \qquad \text{if } |\hat{\lambda}_i| \geq 1. \end{aligned}$$

From the above two equations it is clear that not all computed eigenpairs, $(\hat{\lambda}_i, \hat{x}_i)$, will be sufficiently accurate, and possibly no eigenpair is exact.

We modify the algorithm to take care of these possibilities. The new variables in the modified algorithm will have a "$;t$" in their suffix to distinguish them from similar variables occurring in Algorithm I.

ALGORITHM II. USING INEXACT EIGENDECOMPOSITIONS.

Begin

1. Assume $\|A\| = \|B\|$; else rescale A and B.
2. Assume A is nonsingular.
3. $A \equiv A_{;1;t}$; $\quad B \equiv B_{;1;t} \rightarrow U_{;1;t}\Sigma_{;1;t}U_{;1;t}^T$; compute eigendecomposition of B.
4. $\sqrt{\Sigma_{;1;t}}U_{;1;t}^T A_{;1;t}^{-1}U_{;1;t}\sqrt{\Sigma_{;1;t}} \rightarrow V_{;1;t}\Lambda_{;1;t}V_{;1;t}^T + E_{;1;t}$; compute the eigendecomposition of the equivalent symmetric matrix such that the eigenvalues are ordered from largest to smallest in magnitude.
5. $X_{;1;t} \leftarrow A_{;1;t}^{-1}U_{;1;t}\sqrt{\Sigma_{;1;t}}V_{;1;t}$; compute the generalized eigenvectors of the pencil.
6. $i \leftarrow 1$; we now deflate the eigenvectors from the pencil.
   While $(i < n)$ **do**
   (a) **While** $(i < n)$ **and**
   $(\|(\lambda_{i;i;t}A_{;i;t} - B_{;i;t})X_{i;i;t}\| \leq \epsilon\|X_{i;i;t}\|(|\lambda_{i;i;t}|\ \|A_{[i:n],[i:n];i;t}\| + \|B_{[i:n],[i:n];i;t}\|)$
   do
   (i) Compute Householder transform $Q_{;i;t}$ such that $Q_{;i;t}X_{i;i;t}$ is parallel to e_i.
   (ii) $A_{;i;q;t} \leftarrow Q_{;i;t}A_{;i;t}Q_{;i;t}^T$, $\quad B_{;i;q;t} \leftarrow Q_{;i;t}B_{;i;t}Q_{;i;t}^T$, $\quad X_{;i;q;t} \leftarrow Q_{;i;t}X_{;i;t}$
   (iii) **If** $|\lambda_{i;i;t}| \geq 1$, **then**
   A. $(A_{i;i;q;t} \leftarrow B_{i;i;q;t}/\lambda_{i;i;t};\ A_{i,[1:n];i;q;t} \leftarrow A_{i;i;q;t}^T) \rightarrow A_{;i;p;t}$; make ith column and row of $A_{;i;q;t}$ exactly parallel to that of $B_{;i;q;t}$.
   B. Compute Householder transform $W_{i;t}$ such that $W_{i;t}B_{i;i;q;t}$ has zeros below the $(i+1)$st component.
   C. $A_{;i;w;t} \leftarrow W_{i;t}A_{;i;p;t}W_i^T$, $\quad B_{;i;w;t} \leftarrow W_{i;t}B_{;i;q;t}W_{i;t}^T$, $\quad X_{;i;w;t} \leftarrow W_{i;t}X_{;i;q;t}$.
   D. Find an elementary Gauss transform $L_{;i;t}^{-1}$ such that $L_{;i;t}^{-1}B_{i;i;w;t}$ is parallel to e_i.
   E. $A_{;i+1;t} \leftarrow L_{;i;t}^{-1}A_{;i;w;t}L_{;i;t}^{-T}$, $\ B_{;i+1;t} \leftarrow L_{;i;t}^{-1}B_{;i;w;t}L_{;i;t}^{-T}$, $\ X_{;i+1;t} \leftarrow L_{;i;t}^T X_{;i;w;t}$.
   (iv) **else**; $(|\lambda_{i;i;t}| < 1)$
   A. $(B_{i;i;q;t} \leftarrow \lambda_{i;i;t}A_{i;i;q;t};\ B_{i,[1:n];i;q;t} \leftarrow B_{i;i;q;t}^T) \rightarrow B_{;i;p;t}$; make ith column and row of $B_{;i;q;t}$ exactly parallel to that of $A_{;i;q;t}$.
   B. Compute Householder transform $W_{i;t}$ such that $W_{i;t}A_{i;i;q;t}$ has zeros below the $(i+1)$st component.
   C. $B_{;i;w;t} \leftarrow W_{i;t}B_{;i;p;t}W_i^T$, $\quad A_{;i;w;t} \leftarrow W_{i;t}A_{;i;q;t}W_{i;t}^T$, $\quad X_{;i;w;t} \leftarrow W_{i;t}X_{;i;q;t}$.
   D. Find an elementary Gauss transform $L_{;i;t}^{-1}$ such that $L_{;i;t}^{-1}A_{i;i;w;t}$ is parallel to e_i.
   E. $B_{;i+1;t} \leftarrow L_{;i;t}^{-1}B_{;i;w;t}L_{;i;t}^{-T}$, $\ A_{;i+1;t} \leftarrow L_{;i;t}^{-1}A_{;i;w;t}L_{;i;t}^{-T}$, $\ X_{;i+1;t} \leftarrow L_{;i;t}^T X_{;i;w;t}$.
   endif
   (v) $\Lambda_{;i+1;t} \leftarrow \Lambda_{;i;t}$.
   (vi) $i \leftarrow i + 1$.
   endwhile
   (b) **If** $(i < n)$, **then**
   (i) $B_{;i;t} \rightarrow U_{;i;t}\Sigma_{;i;t}U_{;i;t}^T$; recompute eigendecomposition of $B_{;i;t}$.
   (ii) $\sqrt{\Sigma_{;i;t}}U_{;i;t}^T A_{;i;t}^{-1}U_{;i;t}\sqrt{\Sigma_{;i;t}} \rightarrow V_{;i;t}\Lambda_{;i;t}V_{;i;t}^T + E_{;i;t}$; recompute the eigendecomposition of equivalent symmetric matrix corresponding to the lower-right $(n-i+1) \times (n-i+1)$ block, such that the eigenvalues

of the submatrix are ordered from largest to smallest in magnitude.

(iii) $X_{:i:t} \leftarrow A_{:i:t}^{-1} U_{:i:t} \sqrt{\Sigma_{:i:t}} V_{:i:t}$; recompute the generalized eigenvectors of the pencil.

endif

endwhile

7. We now have the two diagonal matrices

$$D_{A:t} \equiv A_{:n:t}, \qquad D_{B:t} \equiv B_{:n:t},$$

and the accumulated transform

$$C_{:t} \equiv Q_{:1:t}^T W_{:1:t}^T L_{:1:t} \cdots Q_{:n:t}^T W_{:n:t}^T L_{:n:t},$$

such that

$$C_{:t} D_{B:t} C_{:t}^T = B + E_{:B:t},$$

$$\|E_{:B:t}\| \leq 2n\epsilon(\|A\| + n\|B\|) + O(\epsilon^2), \tag{8}$$

$$\frac{\|(D_{i,i;B;t} A - D_{i,i;A;t} B) C_{:t}^{-T} e_i\|}{\|C_{:t}^{-T} e_i\| \, (|D_{i,i;B;t}| \, \|A\| + |D_{i,i;A;t}| \, \|B\|)} \leq 5n^2\epsilon + O(\epsilon^2). \tag{9}$$

End

The important difference between Algorithm I and Algorithm II is that the latter deflates eigenvectors only after checking that they are sufficiently accurate, and if they are not it recomputes them.

Inequality (9) establishes that each computed eigenpair is computed to backward accuracy. Inequality (8) establishes that all the eigenpairs have been computed, in the sense that the eigenvector matrix diagonalizes B to backward accuracy. We now prove the claims in step 7 of Algorithm II.

We first claim that

$$A_{:i:t} \equiv \begin{array}{c} \\ i-1 \\ n-i+1 \end{array} \begin{array}{c} \begin{array}{cc} i-1 & n-i+1 \end{array} \\ \begin{pmatrix} \bar{D}_{:A:i:t} & 0 \\ 0 & \bar{A}_{:i:t} \end{pmatrix} \end{array}, \tag{10}$$

$$B_{:i:t} \equiv \begin{array}{c} \\ i-1 \\ n-i+1 \end{array} \begin{array}{c} \begin{array}{cc} i-1 & n-i+1 \end{array} \\ \begin{pmatrix} \bar{D}_{:B:i:t} & 0 \\ 0 & \bar{B}_{:i:t} \end{pmatrix} \end{array}. \tag{11}$$

These facts will be proved by induction. It is obviously true for $i = 1$. For general i, two cases are possible: either $X_{:i:t}$ has been computed in step 6(b)(iii) (or step 5) of Algorithm II or we have successfully passed the test of the while loop in step 6(a) and are entering step 6(a)(i).

We first consider the case when $X_{:i:t}$ is being computed in step 6(b) (or step 5). Since this is obviously true for $i = 1$, we can invoke the induction hypothesis. We now establish that we will pass the test

$$\|(\lambda_{i:i:t} A_{:i:t} - B_{:i:t}) X_{i:i:t}\| \leq \epsilon \|X_{i:i:t}\| (|\lambda_{i:i:t}| \, \|A_{[i:n],[i:n];i:t}\| + \|B_{[i:n],[i:n];i:t}\|) \tag{12}$$

in step 6(a). Using the induction hypothesis we can conclude that

$$U_{:i:t} \equiv \begin{array}{c} \\ i-1 \\ n-i+1 \end{array} \begin{array}{c} \begin{array}{cc} i-1 & n-i+1 \end{array} \\ \begin{pmatrix} I & 0 \\ 0 & \bar{U}_{:i:t} \end{pmatrix} \end{array},$$

$$\Sigma_{;i;t} \equiv \begin{array}{c} \\ i-1 \\ n-i+1 \end{array} \begin{array}{c} \begin{array}{cc} i-1 & n-i+1 \end{array} \\ \begin{pmatrix} \bar{\Sigma}_{;i;u;t} & 0 \\ 0 & \bar{\Sigma}_{;i;l;t} \end{pmatrix} \end{array},$$

such that $\bar{B}_{;i;t} = \bar{U}_{;i;t}\bar{\Sigma}_{;i;l;t}\bar{U}^T_{;i;t}$. Therefore, we have

$$V_{;i;t} \equiv \begin{array}{c} \\ i-1 \\ n-i+1 \end{array} \begin{array}{c} \begin{array}{cc} i-1 & n-i+1 \end{array} \\ \begin{pmatrix} I & 0 \\ 0 & \bar{V}_{;i;t} \end{pmatrix} \end{array},$$

$$\Lambda_{;i;t} \equiv \begin{array}{c} \\ i-1 \\ n-i+1 \end{array} \begin{array}{c} \begin{array}{cc} i-1 & n-i+1 \end{array} \\ \begin{pmatrix} \bar{\Lambda}_{;i;u;t} & 0 \\ 0 & \bar{\Lambda}_{;i;l;t} \end{pmatrix} \end{array},$$

where we assume that

$$\left\| \sqrt{\bar{\Sigma}_{;i;l;t}}\bar{U}^T_{;i;t}\bar{A}^{-1}_{;i;t}\bar{U}_{;i;t}\sqrt{\bar{\Sigma}_{;i;l;t}} - \bar{V}_{;i;t}\bar{\Lambda}_{;i;l;t}\bar{V}^T_{;i;t} \right\| = \|\bar{E}_{;i;t}\| \leq \epsilon \|\bar{\Lambda}_{;i;l;t}\|. \tag{13}$$

Therefore, this establishes that in step 6(b)(iii) we will have

$$X_{;i;t} \equiv \begin{array}{c} \\ i-1 \\ n-i+1 \end{array} \begin{array}{c} \begin{array}{cc} i-1 & n-i+1 \end{array} \\ \begin{pmatrix} D_{;X;i;t} & 0 \\ 0 & \bar{X}_{;i;t} \end{pmatrix} \end{array}.$$

Therefore, the test (12) can be rewritten as

$$\|(\lambda_{i;i;t}\bar{A}_{;i;t} - \bar{B}_{;i;t})\bar{X}_{i;i;t}\| \leq \epsilon \|\bar{X}_{;i;t}\|(|\lambda_{i;i;t}|\ \|\bar{A}_{;i;t}\| + \|\bar{B}_{;i;t}\|). \tag{14}$$

From (13) we have that

$$\begin{aligned} \bar{B}_{;i;t}\left(\bar{A}^{-1}_{;i;t}\bar{U}_{;i;t}\sqrt{\bar{\Sigma}_{;i;l;t}}\bar{V}_{j;i;t}\right) + \bar{U}_{;i;t}\sqrt{\bar{\Sigma}_{;i;l;t}}\bar{E}_{;i;t}\bar{V}_{j;i;t} \\ = \lambda_{j;i;l;t}\bar{A}_{;i;t}\left(\bar{A}^{-1}_{;i;t}\bar{U}_{;i;t}\sqrt{\bar{\Sigma}_{;i;l;t}}\bar{V}_{j;i;t}\right), \qquad j = 1, \ldots, n-i+1, \end{aligned}$$

or

$$\bar{B}_{;i;t}\bar{X}_{j;i;t} + \bar{U}_{;i;t}\sqrt{\bar{\Sigma}_{;i;l;t}}\bar{E}_{;i;t}\bar{V}_{j;i;t} = \lambda_{j;i;l;t}\bar{A}_{;i;t}\bar{X}_{j;i;t}, \quad j = 1, \ldots, n-i+1. \tag{15}$$

Since $\|\bar{\Lambda}_{;i;l;t}\| = |\lambda_{1;i;l;t}|$, we have that

$$\begin{aligned} |\lambda_{1;i;l;t}|(1-\epsilon) &\leq \left\|\bar{U}_{;i;t}\sqrt{\bar{\Sigma}_{;i;l;t}}\right\| \ \left\|\bar{A}^{-1}_{;i;t}\bar{U}_{;i;t}\sqrt{\bar{\Sigma}_{;i;l;t}}\bar{V}_{1;i;t}\right\| \\ &= \left\|\bar{U}_{;i;t}\sqrt{\bar{\Sigma}_{;i;l;t}}\right\| \ \left\|\bar{X}_{1;i;t}\right\|. \end{aligned}$$

Therefore,

$$\frac{\|\bar{U}_{;i;t}\sqrt{\bar{\Sigma}_{;i;l;t}}\bar{E}_{;i;t}\bar{V}_{1;i;t}\|}{\|\bar{X}_{1;i;t}\|} \leq \frac{\epsilon}{1-\epsilon}\|\bar{\Sigma}_{;i;l;t}\|. \tag{16}$$

Using arguments similar to the derivation of (6), we can show that

$$|\lambda_{1;i;l;t}|(1+\epsilon) \geq \frac{\|\bar{B}_{;i;t}\|}{\|\bar{A}_{;i;t}\|}.$$

Hence it follows that

$$|\lambda_{1;i;l;t}|\|\bar{A}_{;i;t}\| + \|\bar{B}_{;i;t}\| \geq \|\bar{B}_{;i;t}\|\frac{2+\epsilon}{1+\epsilon}. \tag{17}$$

Since

$$\frac{2+\epsilon}{1+\epsilon} \geq \frac{1}{1-\epsilon},$$

for sufficiently small ϵ, it follows from (15), (16), and (17) that the test (14) will be passed by at least $\bar{X}_{1;i;t}$.

Now we go ahead and analyze step 6(a) of Algorithm II. By the induction hypothesis and the preceding discussion, it is clear that

$$Q_{;i;t} \equiv \begin{array}{c} \\ i-1 \\ n-i+1 \end{array} \begin{array}{c} \begin{array}{cc} i-1 & n-i+1 \end{array} \\ \begin{pmatrix} I & 0 \\ 0 & \bar{Q}_{i;t} \end{pmatrix} \end{array}, \qquad \bar{Q}_i X_{i;i;t} = \pm\|X_{i;i;t}\|e_i, \tag{18}$$

$$X_{;i;q;t} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} D_{;X;i;t} & 0 & 0 \\ 0 & \pm\|X_{i;i;t}\| & h_{i;t}^T \\ 0 & 0 & \bar{X}_{;i;q;t} \end{pmatrix} \end{array},$$

$$A_{;i;q;t} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} \bar{D}_{;A;i;t} & 0 & 0 \\ 0 & \alpha_{i;t} & a_{;i;q;t}^T \\ 0 & a_{;i;q;t} & \bar{A}_{;i;q;t} \end{pmatrix} \end{array}, \tag{19}$$

$$B_{;i;q;t} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} \bar{D}_{;B;i;t} & 0 & 0 \\ 0 & \beta_{i;t} & b_{;i;q;t}^T \\ 0 & b_{;i;q;t} & \bar{B}_{;i;q;t} \end{pmatrix} \end{array}, \tag{20}$$

where we now have that

$$\left\| \lambda_{i;i;t} \begin{pmatrix} \alpha_{i;t} \\ a_{;i;q;t} \end{pmatrix} - \begin{pmatrix} \beta_{i;t} \\ b_{;i;q;t} \end{pmatrix} \right\| \leq \epsilon(|\lambda_{i;i;t}|\ \|\bar{A}_{;i;t}\| + \|\bar{B}_{;i;t}\|). \tag{21}$$

That is, $(\,\alpha_{i;t} \quad a_{;i;q;t}^T\,)^T$ and $(\,\beta_{i;t} \quad b_{;i;q;t}^T\,)^T$ are no longer perfectly parallel.

At this stage the algorithm can follow two paths, depending on $\lambda_{i;i;t}$ (step 6(a)(iii)). We will first follow the path taken when $|\lambda_{i;i;t}| \geq 1$.

In step 6(a)(iii)(A) we perturb $A_{;i;q;t}$ such that its ith column is parallel to that of $B_{;i;q;t}$. Hence we have that

$$A_{;i;p;t} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} \bar{D}_{;A;i;t} & 0 & 0 \\ 0 & \beta_{i;t}/\lambda_{i;i;t} & b_{;i;q;t}^T/\lambda_{i;i;t} \\ 0 & b_{;i;q;t}/\lambda_{i;i;t} & \bar{A}_{;i;q;t} \end{pmatrix} \end{array}. \tag{22}$$

Since $|\lambda_{i;i;t}| \geq 1$, it follows from (21) that

$$\|A_{;i;q;t} - A_{;i;p;t}\| \leq \epsilon(\|\bar{A}_{;i;t}\| + \|\bar{B}_{;i;t}\|). \tag{23}$$

The remaining steps in 6(a)(iii) are similar to Algorithm I and we have accordingly that

$$W_{;i;t} \equiv \begin{array}{c} \\ i \\ n-i \end{array} \begin{array}{c} \begin{array}{cc} i & n-i \end{array} \\ \begin{pmatrix} I & 0 \\ 0 & \bar{W}_{;i;t} \end{pmatrix} \end{array}, \qquad \bar{W}_i b_{;i;q;t} = \pm\|b_{;i;q;t}\| e_i, \tag{24}$$

$$X_{;i;w;t} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} D_{;X;i;t} & 0 & 0 \\ 0 & \pm\|X_{i;i;t}\| & h_{;i;t}^T \\ 0 & 0 & \bar{X}_{;i;w;t} \end{pmatrix} \end{array},$$

$$A_{;i;w;1;t} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} \bar{D}_{;A;i;t} & 0 & 0 \\ 0 & \beta_{i;t}/\lambda_{i;i;t} & \pm\|b_{;i;q;t}\| e_1^T/\lambda_{i;i;t} \\ 0 & \pm\|b_{;i;q;t}\| e_1/\lambda_{i;i;t} & \bar{A}_{;i;w;t} \end{pmatrix} \end{array},$$

$$B_{;i;w;t} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} \bar{D}_{;B;i;t} & 0 & 0 \\ 0 & \beta_{i;t} & \pm\|b_{;i;q;t}\| e_1^T \\ 0 & \pm\|b_{;i;q;t}\| e_1 & \bar{B}_{;i;w;t} \end{pmatrix} \end{array}.$$

Next we apply a suitable elementary Gauss transform to complete the deflation of the eigenvector:

$$L_{;i;t}^{-1} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} I & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \rho_{i;t} e_1 & I \end{pmatrix} \end{array}, \qquad \rho_{i;t} = -\frac{\pm\|b_{;i;q}\|}{\beta_i}, \tag{25}$$

$$X_{;i+1;t} = \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} D_{;X;i;t} & 0 & 0 \\ 0 & \pm\|X_{i;i;t}\| & * \\ 0 & 0 & \bar{X}_{;i;w;t} \end{pmatrix} \end{array},$$

$$A_{;i+1;t} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} \bar{D}_{;A;i;t} & 0 & 0 \\ 0 & \beta_{i;t}/\lambda_{i;i;t} & 0 \\ 0 & 0 & \bar{A}_{;i;w;t} - \frac{\|b_{;i;q;t}\|^2}{\lambda_{i;i;t}\beta_{i;t}} e_1 e_1^T \end{pmatrix} \end{array}, \tag{26}$$

$$B_{;i+1;t} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} \bar{D}_{;B;i;t} & 0 & 0 \\ 0 & \beta_{i;t} & 0 \\ 0 & 0 & \bar{B}_{;i;w;t} - \frac{\|b_{;i;q;t}\|^2}{\beta_{i;t}} e_1 e_1^T \end{pmatrix} \end{array},$$

where

$$\left\| \frac{\|b_{;i;q;t}\|^2}{\lambda_{i;i;t}\beta_{i;t}} e_1 e_1^T \right\| \leq \left\| \frac{\|b_{;i;q;t}\|^2}{\beta_{i;t}} e_1 e_1^T \right\|, \tag{27}$$

since $|\lambda_{i;i;t}| \geq 1$.

Therefore, as long as $|\lambda_{i;i;t}| \geq 1$, the application of $L_{i;t}^{-1}$ to $A_{;i;w;t}$ is numerically stable (actually, backward stable with respect to A). This will follow from the more general analysis to be carried out below. We now analyze the case when $|\lambda_{i;i;t}| < 1$.

In step 6(a)(iv)(A) we perturb $B_{;i;q;t}$ such that its ith column is parallel to that of $A_{;i;q;t}$. Hence we have that

$$B_{;i;p;t} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} \bar{D}_{;B;i;t} & 0 & 0 \\ 0 & \alpha_{i;t}\lambda_{i;i;t} & \lambda_{i;i;t}a_{;i;q;t}^T \\ 0 & \lambda_{i;i;t}a_{;i;q;t} & \bar{B}_{;i;q;t} \end{pmatrix} \end{array}. \tag{28}$$

It follows from (21) that

$$\|B_{;i;q;t} - B_{;i;p;t}\| \leq \epsilon(|\lambda_{i;i;t}|\|\bar{A}_{;i;t}\| + \|\bar{B}_{;i;t}\|). \tag{29}$$

The remaining steps in 6(a)(iv) are similar to the steps in 6(a)(iii) and we have accordingly that

$$W_{;i;t} \equiv \begin{array}{c} \\ i \\ n-i \end{array} \begin{array}{c} \begin{array}{cc} i & n-i \end{array} \\ \begin{pmatrix} I & 0 \\ 0 & \bar{W}_{;i;t} \end{pmatrix} \end{array}, \qquad \bar{W}_i a_{;i;q;t} = \pm\|a_{;i;q;t}\|e_i, \tag{30}$$

$$X_{;i;w;t} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} D_{;X;i;t} & 0 & 0 \\ 0 & \pm\|X_{i;i;t}\| & h_{;i;t}^T \\ 0 & 0 & \bar{X}_{;i;w;t} \end{pmatrix} \end{array},$$

$$B_{;i;w;1;t} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} \bar{D}_{;B;i;t} & 0 & 0 \\ 0 & \lambda_{i;i;t}\alpha_{i;t} & \pm\|a_{;i;q;t}\|\lambda_{i;i;t}e_1^T \\ 0 & \pm\|a_{;i;q;t}\|\lambda_{i;i;t}e_1 & \bar{B}_{;i;w;t} \end{pmatrix} \end{array},$$

$$A_{;i;w;t} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} \bar{D}_{;A;i;t} & 0 & 0 \\ 0 & \alpha_{i;t} & \pm\|a_{;i;q;t}\|e_1^T \\ 0 & \pm\|a_{;i;q;t}\|e_1 & \bar{A}_{;i;w;t} \end{pmatrix} \end{array}.$$

Next we apply a suitable elementary Gauss transform to complete the deflation of the eigenvector:

$$L_{;i;t}^{-1} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} I & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \rho_{i;t}e_1 & I \end{pmatrix} \end{array}, \qquad \rho_{i;t} = -\frac{\pm\|a_{;i;q}\|}{\alpha_i}, \tag{31}$$

$$X_{;i+1;t} = \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} D_{;X;i;t} & 0 & 0 \\ 0 & \pm\|X_{i;i;t}\| & * \\ 0 & 0 & \bar{X}_{;i;w;t} \end{pmatrix} \end{array},$$

$$B_{;i+1;t} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} \bar{D}_{;B;i;t} & 0 & 0 \\ 0 & \lambda_{i;i;t}\alpha_{i;t} & 0 \\ 0 & 0 & \bar{B}_{;i;w;t} - \frac{\lambda_{i;i;t}\|a_{;i;q;t}\|^2}{\alpha_{i;t}}e_1e_1^T \end{pmatrix} \end{array},$$

$$A_{;i+1;t} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} \bar{D}_{;A;i;t} & 0 & 0 \\ 0 & \alpha_{i;t} & 0 \\ 0 & 0 & \bar{A}_{;i;w;t} - \frac{\|a_{;i;q;t}\|^2}{\alpha_{i;t}} e_1 e_1^T \end{pmatrix} \end{array}. \tag{32}$$

This completes the proof by induction of (10) and (11). We are now ready to determine the global error of Algorithm II.

We first consider the case when the eigenvalues are larger than one in magnitude. Then the only error which affects the accuracy of the final answer is when we perturb $A_{;i;q;t}$ to get $A_{;i;p;t}$. Let

$$E_{;i;A;t} \equiv A_{;i;p;t} - A_{;i;q;t}, \tag{33}$$

where, by (23),

$$\|E_{;i;A;t}\| \leq \epsilon(\|\bar{A}_{;i;t}\| + \|\bar{B}_{;i;t}\|). \tag{34}$$

We will show that this error can be written as a small backward error in $A_{;i-1;p;t}$ or $A_{;i-1;t}$, depending on whether $|\lambda_{i-1;i-1;t}|$ was strictly smaller than one or not. Let

$$A_{;i-1;s;t} \equiv \begin{cases} A_{;i-1;t} & \text{if } |\lambda_{i-1;i-1;t}| < 1, \\ A_{;i-1;p;t} & \text{otherwise,} \end{cases}$$

and extend the definition of $E_{;i;A;t}$ by $E_{;i;A;t} \equiv A_{;i;s;t} - A_{;i;q;t}$. From an examination of Algorithm II it follows that

$$A_{;i;q;t} = L_{;i-1;t}^{-1} W_{;i-1;t} Q_{;i-1;t} A_{;i-1;s;t} Q_{;i-1;t}^T W_{;i-1;t}^T L_{;i-1;t}^{-T}.$$

Substituting in (33) we get

$$A_{;i;p;t} = L_{;i-1;t}^{-1} W_{;i-1;t} Q_{;i-1;t} (A_{;i-1;s;t} + E_{;i;A;b;t}) Q_{;i-1;t}^T W_{;i-1;t}^T L_{;i-1;t}^{-T}, \tag{35}$$

where

$$E_{;i;A;b;t} = Q_{;i-1;t}^T W_{;i-1;t}^T L_{;i-1;t} E_{;i;A;t} L_{;i-1;t}^T W_{;i-1;t} Q_{;i-1;t}.$$

From the structure equations (19), (22), and (25), we can conclude that

$$L_{;i-1;t} E_{;i;A;t} L_{;i-1;t}^T = E_{;i;A;t}.$$

Therefore,

$$\|E_{;i;A;b;t}\| = \|E_{;i;A;t}\|. \tag{36}$$

Furthermore, observe that $E_{;i;A;b;t}$, like $E_{;i;A;t}$, is nonzero only in the lower-right $(n-i+1) \times (n-i+1)$ block. Therefore, it follows from the structure equations (18), (24), and (25) that the matrix

$$Q_{;i-2;t}^T W_{;i-2;t}^T L_{;i-2;t} E_{;i;A;b;t} L_{;i-2;t}^T W_{;i-2;t} Q_{;i-2;t}$$

is nonzero only in the lower-right $(n-i+1) \times (n-i+1)$ block, and that

$$L_{;i-2;t} E_{;i;A;b;t} L_{;i-2;t}^T = E_{;i;A;b;t}.$$

Since this can be extended by induction, we can conclude using (35) and (36) that

$$D_{;A;t} = A_{;n;p;t} = C_{;t}^{-1}(A + E_{;A;t})C_{;t}^{-T},$$

where

$$\|E_{;A;t}\| \leq \sum_{i=1}^{n} \mu_i \|E_{;i;A;t}\| \leq \epsilon \left(\sum_{i=1}^{n} \mu_i \|\bar{A}_{;i;t}\| + \sum_{i=1}^{n} \mu_i \|\bar{B}_{;i;t}\| \right).$$

The last inequality follows from (34), and μ_i is defined as follows:

$$\mu_i = \begin{cases} 1 & \text{if } |\lambda_{i;i;t}| \geq 1, \\ 0 & \text{otherwise.} \end{cases}$$

We can carry out a similar analysis when $|\lambda_{i;i;t}| < 1$. Now the only error which affects the accuracy of the final answer is

$$E_{;i;B;t} \equiv B_{;i;p;t} - B_{;i;q;t}, \tag{37}$$

where, by (29),

$$\|E_{;i;B;t}\| \leq \epsilon(|\lambda_{i;i;t}| \ \|\bar{A}_{;i;t}\| + \|\bar{B}_{;i;t}\|). \tag{38}$$

We will show that this error can be written as a small backward error in $B_{;i-1;p;t}$ or $B_{;i-1;t}$, depending on whether $|\lambda_{i-1;i-1;t}|$ was strictly smaller than one or not. Let

$$B_{;i-1;s;t} \equiv \begin{cases} B_{;i-1;t} & \text{if } |\lambda_{i-1;i-1;t}| \geq 1, \\ B_{;i-1;p;t} & \text{otherwise,} \end{cases}$$

and extend the definition of $E_{;i;B;t}$ by $E_{;i;B;t} \equiv B_{;i;s;t} - B_{;i;q;t}$. From an examination of Algorithm II it follows that

$$B_{;i;q;t} = L_{;i-1;t}^{-1} W_{;i-1;t} Q_{;i-1;t} B_{;i-1;s;t} Q_{;i-1;t}^{T} W_{;i-1;t}^{T} L_{;i-1;t}^{-T}.$$

Substituting in (37) we get

$$B_{;i;p;t} = L_{;i-1;t}^{-1} W_{;i-1;t} Q_{;i-1;t} (B_{;i-1;s;t} + E_{;i;B;b;t}) Q_{;i-1;t}^{T} W_{;i-1;t}^{T} L_{;i-1;t}^{-T}, \tag{39}$$

where

$$E_{;i;B;b;t} = Q_{;i-1;t}^{T} W_{;i-1;t}^{T} L_{;i-1;t} E_{;i;B;t} L_{;i-1;t}^{T} W_{;i-1;t} Q_{;i-1;t}.$$

From the structure equations (20), (28), and (31), we can conclude that

$$L_{;i-1;t} E_{;i;B;t} L_{;i-1;t}^{T} = E_{;i;B;t}.$$

Therefore,

$$\|E_{;i;B;b;t}\| = \|E_{;i;B;t}\|. \tag{40}$$

Furthermore, observe that $E_{;i;B;b;t}$, like $E_{;i;B;t}$, is nonzero only in the lower-right $(n-i+1) \times (n-i+1)$ block. Therefore it follows from the structure equations (18), (30), and (31) that the matrix

$$Q_{;i-2;t}^{T} W_{;i-2;t}^{T} L_{;i-2;t} E_{;i;B;b;t} L_{;i-2;t}^{T} W_{;i-2;t} Q_{;i-2;t}$$

is nonzero only in the lower-right $(n-i+1)\times(n-i+1)$ block, and that

$$L_{;i-2;t}E_{;i;B;b;t}L_{;i-2;t}^T = E_{;i;B;b;t}.$$

Since this can be extended by induction, we can conclude using (39) and (40) that

$$D_{;B;t} = B_{;n;p;t} = C_{;t}^{-1}(B + E_{;B;t})C_{;t}^{-T},$$

where

$$\begin{aligned}\|E_{;B;t}\| &\leq \sum_{i=1}^{n}(1-\mu_i)\|E_{;i;B;t}\| \\ &\leq \epsilon\left(\sum_{i=1}^{n}(1-\mu_i)|\lambda_{i;i;t}|\ \|\bar{A}_{;i;t}\| + \sum_{i=1}^{n}(1-\mu_i)\|\bar{B}_{;i;t}\|\right). \qquad (41)\end{aligned}$$

The last inequality follows from (38).

Therefore, the error analysis boils down to obtaining good bounds on $\|\bar{A}_{;i;t}\|$ and $\|\bar{B}_{;i;t}\|$. We proceed to do that now.

We shall assume that

$$\|B^{-1}\|^{-1} > \|E_{;B;t}\|, \qquad (42)$$

so that $B_{;i;t}$ is always symmetric positive definite. (We shall give more practical conditions later.) Therefore, $\bar{B}_{;i;t}$ is the Schur complement of a symmetric positive-definite matrix and it follows that

$$\sum_{i=1}^{n}\|\bar{B}_{;i;t}\| \leq n\|B\| + O(\epsilon). \qquad (43)$$

Now consider

$$\sum_{i=1}^{n}(1-\mu_i)|\lambda_{i;i;t}|\ \|\bar{A}_{;i;t}\|.$$

From (26) and (32) we obtain

$$\bar{A}_{;i+1;t} = \bar{A}_{;i;w;t} - (1-\mu_i)\frac{\|a_{;i;q;t}\|^2}{\alpha_{i;t}}e_1e_1^T - \mu_i\frac{\|b_{;i;q;t}\|^2}{\lambda_{i;i;t}\beta_{i;t}}e_1e_1^T. \qquad (44)$$

By the assumption (42), $B_{;i;w;t}$ and $B_{;i;w;1;t}$ are symmetric positive definite. Hence we have that

$$\mu_i\left\|\frac{\|b_{;i;q;t}\|^2}{\lambda_{i;i;t}\beta_{i;t}}e_1e_1^T\right\| \leq \mu_i\|B_{;i;w;t}\| = \mu_i\|B\| + O(\epsilon) \qquad (45)$$

and

$$(1-\mu_i)\left\|\lambda_{i;i;t}\frac{\|a_{;i;q;t}\|^2}{\alpha_{i;t}}e_1e_1^T\right\| \leq (1-\mu_i)\|B_{;i;w;1;t}\| = (1-\mu_i)\|B\| + O(\epsilon). \qquad (46)$$

From (44), (45), and (46), we get

$$\begin{aligned}|\lambda_{i+1;i+1;t}|\ \|\bar{A}_{;i+1;t}\| &\leq \left|\frac{\lambda_{i+1;i+1;t}}{\lambda_{i;i;t}}\right|\ |\lambda_{i;i;t}|\ \|\bar{A}_{;i;t}\| \\ &\quad + \left((1-\mu_i)\left|\frac{\lambda_{i+1;i+1;t}}{\lambda_{i;i;t}}\right| + \mu_i\right)\|B\| + O(\epsilon). \qquad (47)\end{aligned}$$

Let k be the smallest integer such that $|\lambda_{k;k;t}| < 1$. Then it follows from (44) and (45) that

$$\|\bar{A}_{;i;t}\| \leq \|A\| + i\|B\|, \qquad i < k, \tag{48}$$

$$|\lambda_{k;k;t}|\ \|\bar{A}_{;k;t}\| \leq \|A\| + k\|B\|. \tag{49}$$

Now let

$$\tau_i \equiv \max\left(1, \left|\frac{\lambda_{i+1;i+1;t}}{\lambda_{i;i;t}}\right|\right).$$

Then we get from (47) and (49) that

$$|\lambda_{i+1;i+1;t}|\ \|\bar{A}_{;i+1;t}\| \leq (\|A\| + 2i\ \|B\|)\ \Pi_{j=k}^{i}\tau_j + O(\epsilon), \quad i \geq k. \tag{50}$$

This inequality is very loose.

Note that we cannot expect $\tau_i = 1$, because Algorithm II does *not* compute the exact eigenvalues of the pencil. From (58) it follows that

$$\tau_i \leq 1 + c_1\epsilon\kappa(\bar{B}_{;i;t}) + c_2\epsilon|\lambda_{i;i;t}|\ \|\bar{A}_{;i;t}\|\ \|\bar{B}_{;i;t}^{-1}\| + O(\epsilon^2),$$

where c_1 and c_2 are constants. Using (47), (49), and the fact that $\|A\| = \|B\|$, we obtain

$$\tau_i \leq 1 + c_3 i\epsilon\kappa(B) + O(\epsilon^2) \qquad \text{for } i \geq k,$$

where c_3 is a constant. Henceforth we make the assumption

$$c_3 n^2 \kappa(B)\epsilon < 1, \tag{51}$$

which is stronger than the previous assumption (42). Under this assumption it follows that

$$\Pi_{j=k}^{i}\tau_j \leq 2 + O(\epsilon^2).$$

Therefore, we can conclude from (50) that

$$\sum_{i=1}^{n}(1-\mu_i)|\lambda_{i;i;t}|\ \|\bar{A}_{;i;t}\| \leq 2n(\|A\| + n\|B\|) + O(\epsilon). \tag{52}$$

Substituting (52) and (43) in (41) we obtain

$$\|E_{;B;t}\| = \|C_{;t}D_{;B;t}C_{;t}^T - B\| \leq 2n\epsilon(\|A\| + n\|B\|) + O(\epsilon^2). \tag{53}$$

In other words the factorization $C_{;t}D_{;B;t}C_{;t}^T$ is backward stable (to first order) with respect to B, which is the first claim (8) of step 7 of Algorithm II. The main difficulty in proving this claim is that the perturbation in B depends upon the norms of the Schur complements of A, which can theoretically grow exponentially. The surprise is that the decomposition of B is backward stable *in spite of* this possibility.

We now turn our attention to proving the second claim (9), namely, that each eigenpair has a small residual with respect to A and B.

Let $C_{;i;t}$ denote the partially accumulated transform

$$C_{;i;t} \equiv Q_{;1;t}^T W_{;1;t}^T L_{;1;t} \cdots Q_{;i;t}^T W_{;i;t}^T L_{;i;t}^T.$$

Then note that the ith eigenvector is also given by

$$C_{;t}^{-1} e_i = C_{;i;t}^{-1} e_i.$$

This follows from (18), (24), and (25). Therefore, it is enough to prove that

$$\|(\lambda_{i;i;t} A - B) C_{;i;t}^{-T} e_i\| \leq 5n^2 \epsilon \, \|C_{;i;t}^{-T} e_i\| \, (|\lambda_{i;i;t}| \, \|A\| + \|B\|) + O(\epsilon^2)$$

in order to establish our claim. From (10) and (11), and the arguments leading to (36) and (40), it follows that

$$\|A - C_{;i;t} A_{;i+1;t} C_{;i;t}^T\| \leq \sum_{l=1}^{i} \|E_{;l;A;t}\|,$$

$$\|B - C_{;i;t} B_{;i+1;t} C_{;i;t}^T\| \leq \sum_{l=1}^{i} \|E_{;l;B;t}\|.$$

Therefore,

$$\begin{aligned} \|(\lambda_{i;i;t} A - B) C_{;i;t}^{-T} e_i\| \leq{}& \|(\lambda_{i;i;t} D_{i,i;A;t} - D_{i,i;B;t}) C_{;i;t} e_i\| \\ &+ \|C_{;i;t}^{-T} e_i\| \sum_{l=1}^{i} (|\lambda_{i;i;t}| \, \|E_{;l;A;t}\| + \|E_{;l;B;t}\|). \end{aligned}$$

Since

$$\lambda_{i;i;t} D_{i,i;A;t} = D_{i,i;B;t},$$

we have

$$\|(\lambda_{i;i;t} A - B) C_{;i;t}^{-T} e_i\| \leq \|C_{;i;t}^{-T} e_i\| \sum_{l=1}^{i} (|\lambda_{i;i;t}| \, \|E_{;l;A;t}\| + \|E_{;l;B;t}\|).$$

Note that $\|E_{;l;A;t}\|$ is multiplied by $|\lambda_{i;i;t}|$. This is crucial in the analysis which follows. From inequalities (41) and (53) it follows that

$$\sum_{l=1}^{i} \|E_{;l;B;t}\| \leq 2n\epsilon(\|A\| + n\|B\|) + O(\epsilon^2).$$

Now using our previous definition that k is the smallest integer such that $|\lambda_{k;k;t}| < 1$, we observe that if $i < k$, then from (34), (43), and (48) it follows that

$$|\lambda_{i;i;t}| \sum_{l=1}^{i} \|E_{;l;A;t}\| \leq |\lambda_{i;i;t}| \, 2n\epsilon(\|A\| + n\|B\|) + O(\epsilon^2), \qquad i < k.$$

When $i \geq k$, it follows from (49) that

$$|\lambda_{i;i;t}| \, \|\bar{A}_{;l;t}\| = \frac{|\lambda_{i;i;t}|}{|\lambda_{l;l;t}|} |\lambda_{l;l;t}| \, \|\bar{A}_{;l;t}\| \leq 2(\|A\| + 2l\|B\|) + O(\epsilon^2), \qquad l \leq i.$$

Combining this with (34) we obtain

$$\sum_{l=1}^{i} |\lambda_{i;i;t}| \, \|E_{;l;A;t}\| \leq 2n\epsilon(\|A\| + n\|B\|) + O(\epsilon^2), \qquad i \geq k.$$

Therefore, we can conclude that

$$\|(\lambda_{i;i;t}A - B)C_{;i;t}^{-T}e_i\| \leq \|C_{;i;t}^{-T}e_i\|(2 + |\lambda_{i;i;t}|)2n\epsilon(\|A\| + n\|B\|) + O(\epsilon^2).$$

Since $\|A\| = \|B\|$, this can be rewritten as

$$\|(\lambda_{i;i;t}A - B)C_{;i;t}^{-T}e_i\| \leq 5n^2\epsilon \, \|C_{;i;t}^{-T}e_i\| \, (|\lambda_{i;i;t}| \, \|A\| + \|B\|) + O(\epsilon^2),$$

which establishes our claim that each eigenpair has a small residual with respect to A and B. This is true *in spite of* the fact that the factorization of A is potentially unstable theoretically!

In fact using an analysis similar to that used to establish (53) we can show that

$$\|E_{;A;t}\| \leq p(n)\gamma(\|A\| + \|B\|)\epsilon,$$

where $p(n)$ is a low-order polynomial in n, and γ is the "growth factor" defined to be

$$\gamma \equiv \max_i \frac{\|\bar{A}_{;i;t}\|}{\|A\|}.$$

From (7) it follows that γ can grow at most exponentially in n, but in numerous numerical experiments (see section 6) we have never observed it. This is similar to the situation in Gaussian elimination with partial pivoting where the growth factor can potentially grow exponentially but is rarely observed [4]. Nevertheless, our algorithm will compute all eigenpairs reliably no matter how large or small γ happens to be.

We have now established that Algorithm II computes *each* eigenpair to backward accuracy (see (9)), and that it also computes *all* the eigenvectors, in the sense that they diagonalize B to backward accuracy (see (8)). Furthermore, by establishing inequality (see (12)), we have proved that Algorithm II is an $O(mn^3)$ algorithm, where m is the number of times step 6(b) is executed ($m < n$).

In extensive numerical experiments with Algorithm II the largest m we have observed is 7 for a pencil of size 500 by 500. Usually m is smaller than 4. Therefore, we conjecture that m is $O(\log(1/\epsilon))$. The reasoning for the conjecture is as follows.

We examine how many of the eigenvectors computed in step 6(b) will pass the residual test in step 6(a).

The analysis is an extension of the one used to prove inequality (14). We will continue to use the same notation. Immediately after step 6(b) we have (15), which we repeat here for convenience:

$$\bar{B}_{;i;t}\bar{X}_{j;i;t} + \bar{U}_{;i;t}\sqrt{\bar{\Sigma}_{;i;l;t}}\bar{E}_{;i;t}\bar{V}_{j;i;t} = \lambda_{j;i;l;t}\bar{A}_{;i;t}\bar{X}_{j;i;t}, \quad j = 1, \ldots, n-i+1. \tag{15}$$

Furthermore, from (13) we have that

$$\sqrt{\bar{\Sigma}_{;i;l;t}}\bar{U}_{;i;t}^T\bar{A}_{;i;t}^{-1}\bar{U}_{;i;t}\sqrt{\bar{\Sigma}_{;i;l;t}} - \bar{V}_{;i;t}\bar{\Lambda}_{;i;l;t}\bar{V}_{;i;t}^T = \bar{E}_{;i;t},$$

where

$$\|\bar{E}_{;i;t}\| \leq \epsilon\|\bar{\Lambda}_{;i;l;t}\|.$$

Therefore,

$$\begin{aligned}|\lambda_{j;i;l;t}| - \epsilon\|\bar{\Lambda}_{;i;l;t}\| &\leq \left\|\bar{U}_{;i;t}\sqrt{\bar{\Sigma}_{;i;l;t}}\right\| \, \left\|\bar{A}_{;i;t}^{-1}\bar{U}_{;i;t}\sqrt{\bar{\Sigma}_{;i;l;t}}\bar{V}_{j;i;t}\right\| \\ &= \left\|\bar{U}_{;i;t}\sqrt{\bar{\Sigma}_{;i;l;t}}\right\| \, \left\|\bar{X}_{j;i;t}\right\|.\end{aligned}$$

Hence

$$\frac{\|\bar{U}_{;i;t}\sqrt{\bar{\Sigma}_{;i;l;t}}\bar{E}_{;i;t}\bar{V}_{j;i;t}\|}{\|\bar{X}_{j;i;t}\|} \leq \frac{\epsilon}{\frac{|\lambda_{j;i;l;t}|}{\|\bar{\Lambda}_{;i;l;t}\|} - \epsilon}\|\bar{\Sigma}_{;i;l;t}\|. \tag{54}$$

From the variational characterization of eigenvalues we have

$$\|\bar{\Lambda}_{;i;l;t}\|(1+\epsilon) \geq \frac{\|\bar{B}_{;i;t}\|}{\|\bar{A}_{;i;t}\|}.$$

Therefore,

$$\begin{aligned}|\lambda_{j;i;l;t}|\|\bar{A}_{;i;t}\| + \|\bar{B}_{;i;t}\| &= \frac{|\lambda_{j;i;l;t}|}{|\lambda_{1;i;l;t}|}|\lambda_{1;i;l;t}|\|\bar{A}_{;i;t}\| + \|\bar{B}_{;i;t}\| \\ &\geq \frac{|\lambda_{j;i;l;t}|}{|\lambda_{1;i;l;t}|}\frac{\|\bar{B}_{;i;t}\|}{1+\epsilon} + \|\bar{B}_{;i;t}\| \\ &= \|\bar{B}_{;i;t}\|\frac{\frac{|\lambda_{j;i;l;t}|}{|\lambda_{1;i;l;t}|} + 1 + \epsilon}{1+\epsilon}. \end{aligned} \tag{55}$$

Now, if

$$\frac{|\lambda_{j;i;l;t}|}{|\lambda_{1;i;l;t}|} > \frac{\sqrt{5}-1}{2} + 3\epsilon, \tag{56}$$

then

$$\frac{1}{\frac{|\lambda_{j;i;l;t}|}{\|\bar{\Lambda}_{;i;l;t}\|} - \epsilon} < \frac{\frac{|\lambda_{j;i;l;t}|}{|\lambda_{1;i;l;t}|} + 1 + \epsilon}{1+\epsilon}.$$

Therefore, from (15), (54), and (55), we can conclude that

$$\|(\lambda_{j;i;t}\bar{A}_{;i;t} - \bar{B}_{;i;t})\bar{X}_{j;i;t}\| \leq \epsilon\|\bar{X}_{;i;t}\|(|\lambda_{j;i;t}|\ \|\bar{A}_{;i;t}\| + \|\bar{B}_{;i;t}\|)$$

for all j between 1 and $n-i+1$ such that inequality (56) is satisfied.

In other words, the eigenvectors corresponding to the eigenvalues in a *significant* interval around the largest eigenvalue will be sufficiently accurate to pass the residual test. The algorithm proceeds by deflating the eigenvector corresponding to the largest eigenvalue in magnitude, $\lambda_{1;i;t}$. The important question now is how many of the originally accurate eigenvectors will continue to pass the residual test after the deflation. In numerical experiments most of these eigenvectors continue to pass the residual test after the deflation. Our best explanation of this phenomenon is that the eigenvectors seem to be ordered in rank-revealed form and that the *total* loss of accuracy due to the nonorthogonal transforms is $O(\kappa(B)\epsilon)$. This explains our conjecture. Doubtless, more extensive investigations are needed and they are being carried out.

5. Implementation issues. Algorithm II was structured to make the error analysis easy. Here we present some necessary details for a practical implementation.

Singular A. So far, throughout our analysis, we had assumed that $\bar{A}_{;i;t}$ was nonsingular. This is hard to maintain, since we have no real control on the growth of the condition number of $\bar{A}_{;i;t}$. Furthermore, the additive perturbations can make some $\bar{A}_{;i;t}$ singular. Therefore, we need to explicitly deal with singular $\bar{A}_{;i;t}$.

In step 6(b) we explicitly check if the condition number of $\bar{A}_{;i;t}$ is bigger than a small constant times the reciprocal of the machine precision. If it is, we compute all the eigenvectors of $\bar{A}_{;i;t}$ corresponding to its small eigenvalues (of the size of the machine precision). Compute the Householder transform which maps the eigenvector corresponding to the smallest eigenvalue in magnitude to e_1. Apply this to $\bar{A}_{;i;t}$ and $\bar{B}_{;i;t}$. The first column and row of $\bar{A}_{;i;t}$ will be numerically zero. They should be explicitly set to zero. Now by elementary Gauss transforms the first column and row of $\bar{B}_{;i;t}$ are eliminated. This will have no effect on $\bar{A}_{;i;t}$. This process is continued until all the eigenvectors of $\bar{A}_{;i;t}$ corresponding to the small eigenvalues (in magnitude) have been deflated. Note that this procedure has no adverse effect on the accuracy and speed of Algorithm II.

Residual test. The residual test in step 6(a) is too strong and should be weakened as follows:

$$\|(\lambda_{i;i;t}A_{;i;t} - B_{;i;t})X_{i;i;t}\| \leq \epsilon\|X_{i;i;t}\|(|\lambda_{i;i;t}| \max_{j\leq i}\|\bar{A}_{;j;t}\| + \max_{j\leq i}\|\bar{B}_{;j;t}\|).$$

This is to account for the fact that the Schur complements of A and B might decrease in norm during the deflation process.

Well-conditioned submatrices. If either $\bar{A}_{;i;t}$ or $\bar{B}_{;i;t}$ happen to be well-conditioned, then it improves performance to directly compute the eigenvalues and eigenvectors of either $\bar{B}_{;i;t}\bar{A}^{-1}_{;i;t}$ or $\bar{A}_{;i;t}\bar{B}^{-1}_{;i;t}$, respectively (in symmetric form, of course), and avoid deflating them.

QR factorization of $C_{;t}$. The matrix $C_{;t}$ can be stored in factorized form as in step 7 of Algorithm II, or it can be stored in QR-factored form, which might be more convenient for the user. To do this we observe that

$$C_{;t} = Q^T_{;1;t}W^T_{;1;t}\cdots Q^T_{;n;t}W^T_{;n;t}\tilde{L}_{;1;t}\cdots\tilde{L}_{;n;t},$$

where $\tilde{L}_{;i;t}$ is still an elementary Gauss transform of the type

$$\tilde{L}_{;i;t} \equiv \begin{array}{c} \\ i-1 \\ 1 \\ n-i \end{array} \begin{array}{c} \begin{array}{ccc} i-1 & 1 & n-i \end{array} \\ \begin{pmatrix} I & 0 & 0 \\ 0 & 1 & 0 \\ 0 & y_{;i;t} & I \end{pmatrix} \end{array},$$

and $y_{;i;t}$ can be obtained easily from $\rho_{i;t}$ and the transforms $Q_{;j;t}$ and $W_{;j;t}$ for $j > i$. Due to this, the QR factorization of $C_{;t}$ can be computed efficiently in $O(n^3)$ flops.

Numerical issues. So far we have tacitly assumed that we can compute the large eigenvalues and corresponding eigenvectors of $\sqrt{\Sigma_{;i;t}}U^T_{;i;t}A^{-1}_{;i;t}U_{;i;t}\sqrt{\Sigma_{;i;t}}$ in step 6(b)(ii) to sufficient accuracy. That task is nontrivial numerically, and in this section we detail the mechanism for carrying it out.

We first compute the eigendecomposition of $A_{;i;t} = Z_{;i}\Delta_{;i}Z^T_{;i}$. Then we compute the required product as

$$\sqrt{\Sigma_{;i;t}}\left((U^T_{;i;t}Z_{;i})\Delta_{;i}(Z^T_{;i}U_{;i;t})\right)\sqrt{\Sigma_{;i;t}}$$

in the order suggested by the parentheses. It is important to follow the suggested order of operations. When the expression $(U^T_{;i;t}Z_{;i})\Delta_{;i}(Z^T_{;i}U_{;i;t})$ is being evaluated it should be done in outer-product form, exploiting the fact that $\Delta_{;i}$ is diagonal. Then a round-off error analysis can be used to show that if $Y_{;i}$ denotes the computed product, then

$$Y_{;i} + E_{;Y;i} = \sqrt{\Sigma_{;i;t}}U^T_{;i;t}(A_{;i;t} + E_{;A;m;i})^{-1}U_{;i;t}\sqrt{\Sigma_{;i;t}},$$

where

$$\|E_{;A;m;i}\| \leq n^2\epsilon\|A_{;i;t}\|,$$
$$\|E_{;Y;i}\| \leq n^2\epsilon\|\sqrt{\Sigma_{;i;t}}\|\ \|A_{;i;t}^{-1}U_{;i;t}\sqrt{\Sigma_{;i;t}}\|.$$

Since $\|Y_{;i}\|$ can be as small as $\sigma_{\min}(\Sigma_{;i;t})/\|A_{;i;t}\|$, it follows that the computed $Y_{;i}$ can fail to be stable by approximately $\epsilon\sqrt{\kappa(\Sigma_{;i;t})}$ digits. Since we have assumed that $\kappa(B)\epsilon < 1$, it follows that $Y_{;i}$ is accurate to at least half the digits. Therefore, if the computed eigenvectors of $Y_{;i}$ corresponding to the largest eigenvalues turn out to be too inaccurate to pass the residual test, we can efficiently correct them by inverse iteration (see chapter 5 in [1]).

Round-off errors. The analysis of the round-off errors incurred in the algorithm is similar to the analysis of the eigenvalue decomposition truncation errors carried out for Algorithm II. We just combine the standard error analysis techniques of Gaussian elimination with the truncation error analysis of section 4. The final error bounds have larger polynomials in n.

6. Numerical experiments. We now describe some numerical experiments that were carried out to test the accuracy and efficiency of Algorithm II. The algorithm was implemented in Matlab [5] and run on a Sun SPARCstation 20. The machine precision was approximately 10^{-16} (denoted by ϵ_{mach}). The algorithm was tested on three classes of randomly generated pencils with different characteristics.

The results for the first class of test matrices are shown in Figure 1, those for the second class of test matrices in Figure 2, and those for the third class of test matrices in Figure 3. In each class the matrices of all sizes from 10 to 100 were tested. In all the figures the horizontal axis represents the matrix size.

Figures 1(a), 2(a), and 3(a) show $n\kappa(A)$ for the experimental run being reported, where n denotes the matrix size.

Figures 1(b), 2(b), and 3(b) similarly show $n\kappa(B)$.

Figures 1(c), 2(c), and 3(c) show the error in the factorization of A, which is defined as follows:

$$\frac{\|A - C_{;t}D_{A;t}C_{;t}^T\|}{\|A\|}\frac{\epsilon_{mach}}{n\epsilon}.$$

The reason for the normalization factor is that from inequality (8) we expect the error to be bounded by a quadratic polynomial in n times ϵ. Since the bounds in error analysis tend to be conservative we chose to normalize by a linear polynomial instead.

Figures 1(d), 2(d), and 3(d) show the error in the factorization of B:

$$\frac{\|B - C_{;t}D_{B;t}C_{;t}^T\|}{\|B\|}\frac{\epsilon_{mach}}{n\epsilon}.$$

The plots in Figures 1(e), 2(e), and 3(e) display the accuracy of the computed eigenvalues measured by the expression

$$\max_i \frac{\sigma_{\min}(D_{i,i;B;t}A - D_{i,i;A;t}B)}{n\sigma_{\max}(D_{i,i;B;t}A - D_{i,i;A;t}B)}.$$

This expression measures the accuracy of the computed eigenvalues independently of the computed eigenvectors.

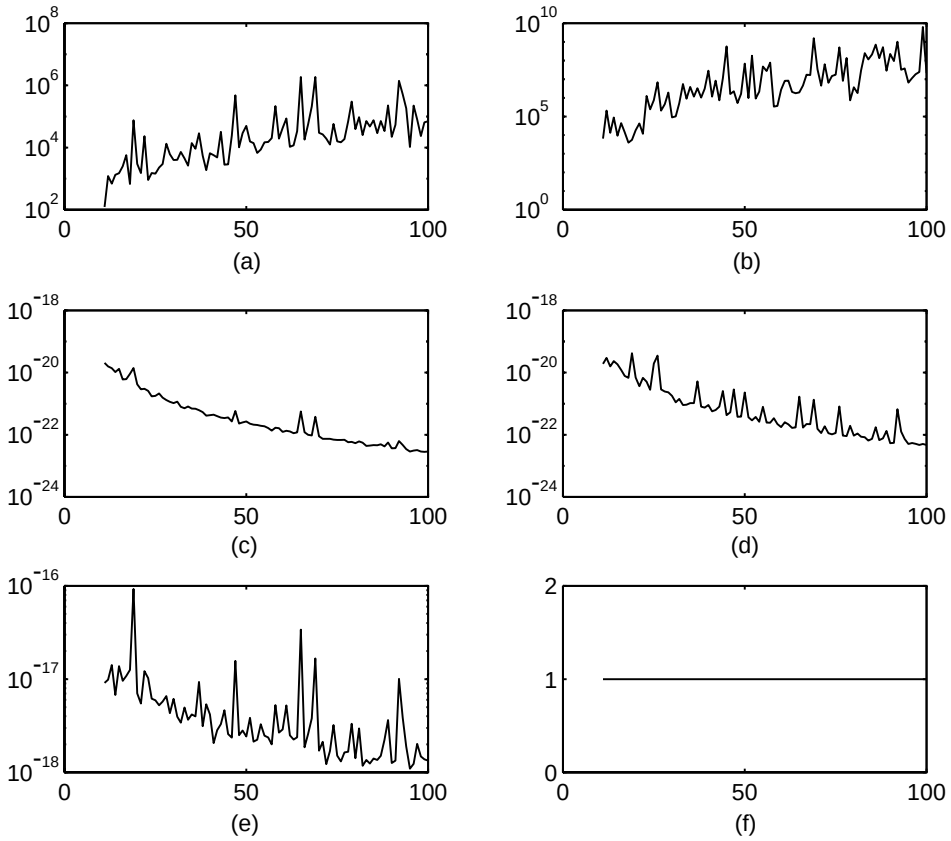


FIG. 1. *First test class.*

Finally, Figures 1(f), 2(f), and 3(f) display the number of times step 6(b) is executed. This gives an indication of the efficiency of the algorithm since the cost of Algorithm II is $O(mn^3)$, where m is the number of times step 6(b) is executed. We remind the reader that in step 6(b) we compute eigenvalue decompositions of dense matrices, and it can cost $O(n^3)$ flops. Hence it essentially determines how efficient the algorithm will be. Of course, we want to ensure that step 6(b) is executed as few times as possible.

For the experiments presented here we chose ϵ to be equal to $20n^{1.5}\epsilon_{mach}$. This was based on an estimate of the average accuracy of the solver for the symmetric eigenvalue problem. Choosing a smaller ϵ can make the algorithm inefficient with little gain in accuracy.

The first class of test matrices is generated as follows. To create the matrix A of order n, we generate a random (normal distribution) $n \times n$ matrix, take its symmetric part, and normalize it by its 1-norm. The matrix tends to have a moderately high condition number, as can be seen from Figure 1(a). To generate the matrix B of order n, we take a random (normal distribution) matrix of order n, multiply it by its transpose (to make it symmetric positive definite), and then normalize it by its 1-norm. The matrix tends to be more ill-conditioned than A, as shown in Figure 1(b). As can be seen from Figure 1(c) the computed eigenvectors diagonalize A in each instance to full backward accuracy. That is, there is no evidence of a large growth factor.

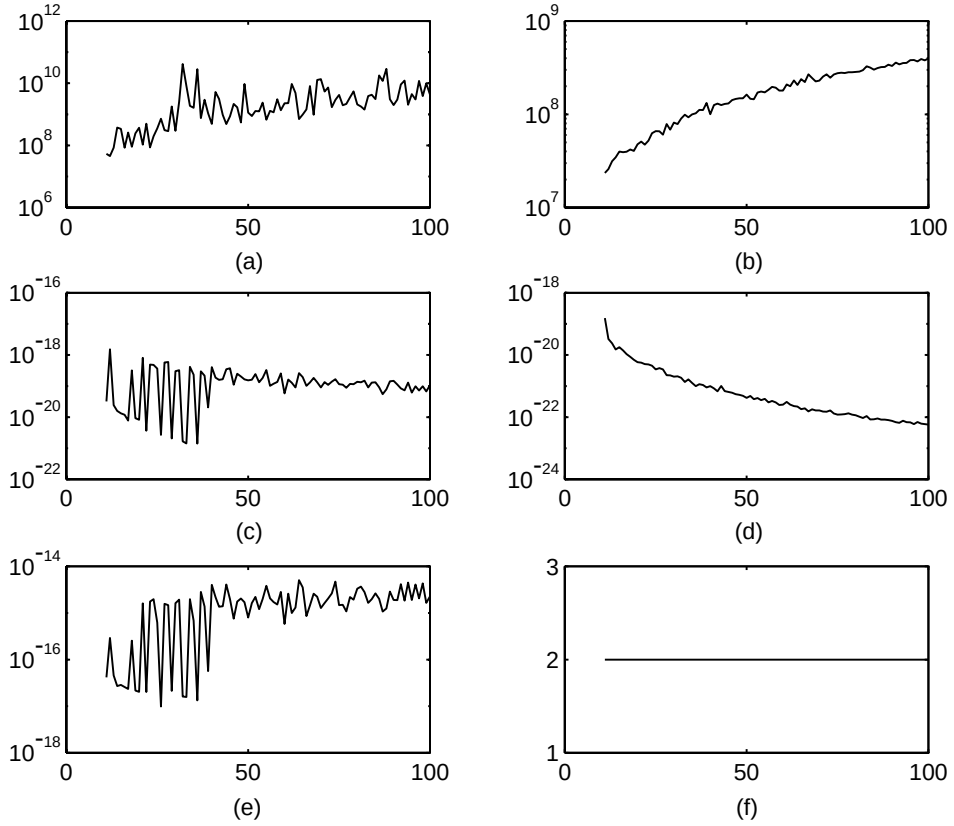


FIG. 2. *Second test class.*

As expected from the error analysis, Figure 1(d) demonstrates that the computed eigenvectors diagonalize B to full backward accuracy. From Figure 1(e) we see that the eigenvalues are computed with better accuracy than the error analysis indicates. Finally, Figure 1(f) shows that the algorithm is very efficient on this test class, as we never require more than one execution of step 6(b).

The second class of test matrices was generated as follows. Let $c_b \equiv 10^{-8}n$, where n denotes the order of the matrix. To create A, we first generated a random (normal distribution) matrix, took its symmetric part, and computed its eigenvalues, $\lambda_1 \leq \cdots \leq \lambda_n$. We then added $-\lambda_{[n/2]} + c_b\lambda_n$ times the identity matrix to the symmetric part, normalized the result by its 1-norm, and obtained A. The general result is a symmetric indefinite matrix with a rather large condition number, as seen in Figure 2(a). To create B, we generate a random (normal distribution) matrix of order n, extract its symmetric part, compute its eigenvalues, $\mu_1 \leq \cdots \leq \mu_n$, and add $|\mu_1| + c_b \max(\mu_1, \mu_n)$ times the identity to the symmetric part. This almost always results in a symmetric positive-definite matrix, with a moderately large condition number, as can be seen in Figure 2(b). Note that in the run shown in Figure 2, the matrix B is generally better conditioned than A, though both in general have rather large condition numbers. Again Figure 2(c) indicates that the computed eigenvectors diagonalize A to backward stability in each case, which is much better than what the error analysis predicts. Figure 2(d) shows that the eigenvectors diagonalize B to

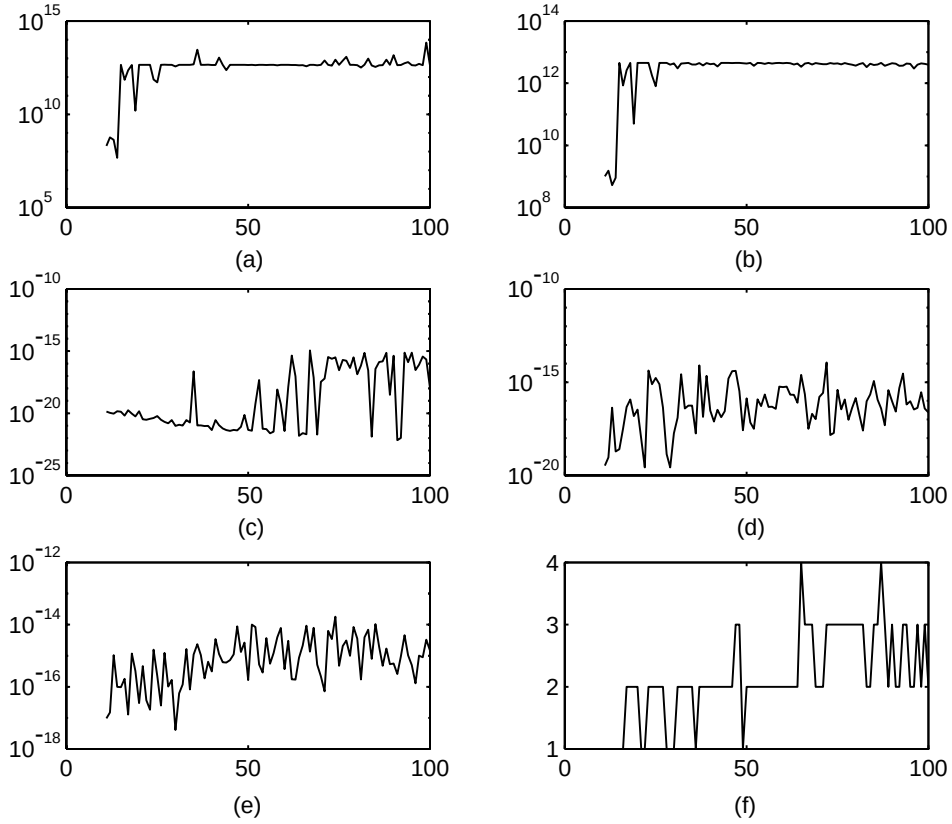


FIG. 3. *Third test class.*

full backward accuracy (as the error analysis predicts). Figure 2(e) shows that the individual eigenvalues are highly accurate. Figure 2(f) again shows the high efficiency of the algorithm. In this run we never execute step 6(b) more than two times. The main differences between the first test class and this one (the second) is that A is generally more ill-conditioned in the latter class and that B is made ill-conditioned by shifting (second class) rather than squaring (first class).

The third test class of matrices is deliberately chosen to try to make the algorithm perform badly. From the error analysis we identify the following features which such a test class must satisfy: First, the eigenvalues must not be too small or too large in magnitude; second, the eigenvalues must be highly clustered, preferably with eigenvalue gaps in a wide range; third, the eigenvalues must be highly ill-conditioned. We generate such a test class of matrices as follows. We first generate a random (uniform distribution) lower-triangular matrix, L. We then generate a diagonal matrix, D, whose ith element is a positive random number times $(-1.25)^i$. Then D is normalized by its 2-norm. The matrix A is taken to be equal to LDL^T. We now multiply the ith diagonal element of D by $1 + c_i$, where c_i is a random number uniformly distributed between -10^{-6} and 10^{-6}. The matrix B is taken to be equal to LDL^T, where we use the new D now. From the way we created A and B we can see that eigenvalues will be tightly clustered around $+1$ and -1. From Figures 3(a) and 3(b) we see that the matrices A and B are highly ill-conditioned, making the eigenvalues also highly

ill-conditioned. But from Figures 3(c) and 3(d) we see that the computed eigenvectors diagonalize A and B to full backward accuracy. From Figure 3(e) we see that eigenvalues are all accurate. From Figure 3(f) we see that now the algorithm executes step 6(b) more frequently than for the previous two test classes, but nevertheless the number of iterations never exceeds 4. This gives further evidence of the efficiency and robustness of the algorithm.

6.1. Some numerical examples. In this section we give some examples where conventional algorithms for the symmetric-definite generalized eigenvalue problem can give erroneous answers.

We first provide an example where Matlab's QZ algorithm computes eigenvalues with large imaginary parts. The algorithm described in this paper has no such problems and returns all the eigenvalues and eigenvectors to full backward accuracy. To proceed, we define the matrices A and B as follows:

$$A = \begin{pmatrix} -0.00116607575822 & -0.00039530283170 & -0.00076250811040 & 0.00112106013448 & -0.00065381674531 \\ -0.00039530283170 & -0.00013392875708 & -0.00025887606360 & 0.00038006880479 & -0.00022125054686 \\ -0.00076250811040 & -0.00025887606360 & -0.00049676987879 & 0.00073294540527 & -0.00042943322558 \\ 0.00112106013448 & 0.00038006880479 & 0.00073294540527 & -0.00107777377871 & 0.00062870745935 \\ -0.00065381674531 & -0.00022125054686 & -0.00042943322558 & 0.00062870745935 & -0.00036464347683 \end{pmatrix}$$

and

$$B = \begin{pmatrix} 0.00116607575822 & 0.00039530283170 & 0.00076250811040 & -0.00112106013448 & 0.00065381674531 \\ 0.00039530283170 & 0.00013408869830 & 0.00025810855444 & -0.00038001602253 & 0.00022204073848 \\ 0.00076250811040 & 0.00025810855444 & 0.00050045292133 & -0.00073319869124 & 0.00042564133688 \\ -0.00112106013448 & -0.00038001602253 & -0.00073319869124 & 0.00107779119829 & -0.00062844668852 \\ 0.00065381674531 & 0.00022204073848 & 0.00042564133688 & -0.00062844668852 & 0.00036854742882 \end{pmatrix}.$$

When we computed the eigenvalues using "eig(A,B)" in Matlab, one of the eigenvalues it returned was $-0.96956258140331 + 0.00009155375727i$, which as one can see has a large imaginary part for a computation carried out with approximately 14 decimal digits of accuracy.

The next example is one where the conventional idea of converting $Ax = \lambda Bx$ into $G^{-1}AG^{-T}x = \lambda x$, where $B = GG^T$, fails to do well. Let

$$A = \begin{pmatrix} 0.21472417430628 & 0.17400616567923 & -0.02769100324675 & -0.13782812711118 & 0.34789671111569 \\ 0.17400616567923 & -0.79254771778534 & -0.06583322926616 & -0.21873063212490 & 0.52253219013903 \\ -0.02769100324675 & -0.06583322926616 & 0.30681582819807 & -0.85955850791110 & -0.39898863450352 \\ -0.13782812711118 & -0.21873063212490 & -0.85955850791110 & -0.09425601545807 & -0.38773515091776 \\ 0.34789671111569 & 0.52253219013903 & -0.39898863450352 & -0.38773515091776 & 0.36526373073906 \end{pmatrix}$$

and

$$B = \operatorname{diag} \begin{pmatrix} 1.00000000000000 \\ 0.00050000000000 \\ 0.00000025000000 \\ 0.00000000012500 \\ 0.00000000000006 \end{pmatrix}.$$

When this example is solved in Matlab using "eig($G^{-1}AG^{-T}$)," the largest normalized residual is approximately 10^{-9}, which is too big for a computation carried out in double-precision arithmetic for a 5×5 problem.

7. Perturbation of generalized eigenvalues. In this section we develop the necessary perturbation theory of generalized eigenvalues of symmetric-definite pencils.

Let

$$\begin{aligned} B &= GG^T, \\ G^T A^{-1} G &= V_1 \Lambda_1 V_1^T + E_1, \\ B + E_B &= GFFG^T, \quad F = F^T, \\ FG^T(A + E_A)^{-1} GF &= V_2 \Lambda_2 V_2^T + E_2, \end{aligned}$$

where

$$\begin{aligned}\|E_1\| &\le \epsilon\|\Lambda_1\|,\\ \|E_2\| &\le \epsilon\|\Lambda_2\|,\\ \|E_B\| &\le \epsilon\|B\|,\\ \|E_A\| &\le \epsilon\|A\|,\\ \epsilon\kappa(B) &< 1.\end{aligned}$$

Since $FF = I + G^{-1}E_BG^{-T}$, we have

$$G - GF = G(I - F) = -E_BG^{-T}(I + F)^{-1}.$$

Using the fact that F is symmetric positive definite, we have

$$\|GF - G\| \le \epsilon\frac{1}{2}\|B\|\,\|G^{-1}\| + O(\epsilon^2).$$

Therefore, we can expand to first order in ϵ to obtain

$$\begin{aligned}V_2\Lambda_2V_2^T + E_2 = {} & V_1\Lambda_1V_1^T + E_1 + (GF - G)^TA^{-1}G + G^TA^{-1}(GF - G)\\ & + G^TA^{-1}E_AA^{-1}G + O(\epsilon^2).\end{aligned} \tag{57}$$

Also, since

$$A^{-1}G = G^{-T}V_1\Lambda_1V_1^T + O(\epsilon),$$

we obtain from (57) that

$$\|\Lambda_2\|(1 - \epsilon) \le \|\Lambda_1\|(1 + \epsilon + \epsilon\kappa(B) + \epsilon\|\Lambda_1\|\,\|A\|\,\|B^{-1}\|) + O(\epsilon^2),$$

where $\kappa(B) = \|B\|\,\|B^{-1}\|$ is the condition number of B. Therefore,

$$\frac{\|\Lambda_2\|}{\|\Lambda_1\|} \le 1 + 3\epsilon\kappa(B) + \epsilon\|\Lambda_1\|\,\|A\|\,\|B^{-1}\| + O(\epsilon^2). \tag{58}$$

The important thing to observe here is that the first-order perturbation bound is not directly dependent on the condition number of A.

8. Conclusion. We have presented a new technique for the stable deflation of eigenpairs from the pencil $Ax = \lambda Bx$, where A is symmetric indefinite and B is symmetric positive definite. We have given an error analysis to prove the stability of the algorithm, as well as numerical evidence for its efficiency and robustness even for almost singular problems. Hence we have shown that numerically reliable and efficient software for the symmetric-definite generalized eigenvalue problem can be written without unnecessarily restricting the class of problems on which it can work. We also see that we do not need extra precision beyond what is available intrinsically in the hardware. In addition we presented a new eigenvalue perturbation bound.

Acknowledgments. We would like to thank Ming Gu for useful discussions and the referees for many constructive suggestions.

REFERENCES

[1] S. Chandrasekaran, *When is a Linear System Ill-Conditioned?*, Ph.D. thesis, Yale University, New Haven, CT, 1994.
[2] G. Golub and C. van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1996.
[3] A. Householder, *The Theory of Matrices in Numerical Analysis*, Dover, New York, 1964.
[4] W. Kahan, *Numerical linear algebra*, Canad. Math. Bull., 9 (1966), pp. 757–801.
[5] The MathWorks, Inc., *Matlab Reference Guide*, Natick, MA, 1992.
[6] B. Parlett, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.
[7] G. Stewart, *A bibliographical tour of the large, sparse generalized eigenvalue problem*, in Sparse Matrix Computations, J. Bunch and D. Rose, eds., Academic Press, New York, 1976, pp. 113–130.
[8] J. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 4, pp. 1229–1235
© 2000 Society for Industrial and Applied Mathematics

ON THE SOLUTION SETS OF LINEAR COMPLEMENTARITY PROBLEMS*

G. S. R. MURTHY†, T. PARTHASARATHY‡, AND B. SRIPARNA‡

Abstract. In this article we consider two problems related to the solution sets of linear complementarity problems (LCPs)—one on the connectedness and the other on the convexity. In Jones and Gowda [*Linear Algebra Appl.*, 246 (1996), pp. 299–312], it was shown that the solution sets of LCPs arising out of $P_0 \cap Q$-matrices are connected, and they conjectured that this is true even in the case of $P_0 \cap Q_0$-matrices. We verify this, at least in the case of nonnegative matrices. Our second problem is related to the class of fully copositive (C_0^f)-matrices introduced in Murthy and Parthasarathy [*Math. Programming*, 82 (1998), pp. 401–411]. The class $C_0^f \cap Q_0$, which contains the class of positive semidefinite matrices, has several properties that positive semidefinite matrices have. This article further supplements this by showing that the solution sets arising from LCPs with $C_0^f \cap Q_0$-matrices and their transposes are convex. This means that $C_0^f \cap Q_0$-matrices are sufficient matrices, another well known class in the theory of linear complementarity problem introduced by Cottle, Pang, and Venkateswaran [*Linear Algebra Appl.*, 114/115 (1989), pp. 231–249].

Key words. linear complementarity problem, matrix classes, connectedness, convex solutions sets

AMS subject classification. 90C33

PII. S0895479898336975

1. Introduction.

Given a matrix $A \in \boldsymbol{R}^{n\times n}$ and $q \in \boldsymbol{R}^n$ the linear complementarity problem (LCP) (q, A) is to find a vector $z \in \boldsymbol{R}^n$ such that

$$Az + q \geq 0, \qquad z \geq 0 \quad \text{and} \quad z^t(Az + q) = 0. \tag{1.1}$$

In order to present our main results, we shall briefly introduce the notation and background. The reader may refer to [3] for all details as well as notation. Let $F(q, A) = \{z \in \boldsymbol{R}^n_+ \ : \ Az + q \geq 0 \ \}$ and $S(q, A) = \{z \in F(q, A) \ : \ (Az + q)^t z = 0\}$. The class $\boldsymbol{Q}$ consists of all real square matrices A such that $S(q, A) \neq \emptyset \ \forall q$, and the class $\boldsymbol{Q}_0$ consists of all real square matrices A such that $S(q, A) \neq \emptyset$ whenever $F(q, A) \neq \emptyset$; A is said to be completely $\boldsymbol{Q}_0$ if all principal submatrices of A are in $\boldsymbol{Q}_0$. Completely $\boldsymbol{Q}_0$-matrices are denoted by $\bar{\boldsymbol{Q}}_0$. A matrix A is said to be copositive if the quadratic form $x^t Ax \geq 0$ holds for all nonnegative x; A is said to be fully copositive if every principal pivotal transform (PPT) (see below for a definition of PPT) of A is copositive; A is said to be a $\boldsymbol{P}_0$-matrix if all its principal minors are nonnegative; A is called an $\boldsymbol{R}_0$-matrix if $(0, A)$ has a unique solution; A is called a connected matrix if $S(q, A)$ is connected for every q; A is said to be column sufficient (CS) if $\forall x \in \mathbf{R}^n$ the implication

$$x_i(Ax)_i \leq 0 \ \forall \ i \text{ implies } x_i(Ax)_i = 0 \ \forall \ i;$$

A is said to be row sufficient if A^t is CS, and A is said to be sufficient if A is both row and column sufficient.

*Received by the editors March 16, 1998; accepted for publication (in revised form) by R. Cottle September 16, 1998; published electronically March 30, 2000.
http://www.siam.org/journals/simax/21-4/33697.html

†Indian Statistical Institute, Street No. 8, Habshiguda, Hyderabad 500 007, India (murthygsr@hotmail.com).

‡Indian Statistical Institute, 7, SJS Sansanwal Marg, New Delhi 110 016, India.

Cao and Ferris [2] introduced the class of connected matrices and showed that LCPs arising from this class can be processed by Lemke's algorithm. Later Jones and Gowda [5] showed that $\boldsymbol{P}_0 \cap \boldsymbol{Q}$-matrices are connected matrices and conjectured that $\boldsymbol{P}_o \cap \boldsymbol{Q}_o$-matrices are also connected matrices. The class of connected matrices is denoted by $\boldsymbol{E}_c$ (this notation was introduced by Jones and Gowda). See [2, 5] for more results on this class. In section 2, we verify that the conjecture is true at least in the case of nonnegative matrices.

The class of sufficient matrices was introduced by Cottle, Pang, and Venkateswaran [4]. An equivalent characterization of CS matrices is that A is CS iff $S(q, A)$ is convex $\forall q$ (see [3, Theorem 3.5.8]).

The classes of copositive and fully copositive matrices are denoted by $\boldsymbol{C}_0$ and $\boldsymbol{C}_0^f$, respectively. Murthy and Parthasarathy introduced $\boldsymbol{C}_0^f$ and derived a number of interesting results related to this class (see [6, 7, 8]). In [6], it was conjectured that if $A \in \boldsymbol{C}_0^f \cap \boldsymbol{Q}_0$, then it is positive semidefinite (A is said to be positive semidefinite if $x^t Ax \geq 0\ \forall x$). It is known that positive semidefinite matrices are sufficient. In section 3, we will show that $\boldsymbol{C}_0^f \cap \boldsymbol{Q}_0$-matrices are sufficient.

For any positive integer n, $\bar{n}$ stands for the set $\{1, 2, \ldots, n\}$ and for any subset α of $\bar{n}$, $\bar{\alpha}$ denotes its complement with respect to $\bar{n}$. For any $A \in \boldsymbol{R}^{n\times n}$ and index sets α, β, $A_{\alpha\beta}$ is obtained by dropping rows and columns corresponding to $\bar{\alpha}$ and $\bar{\beta}$, respectively, from A. When β consists of a single index, say $\beta = \{j\}$, we write $A_{\alpha j}$ for $A_{\alpha\beta}$. For any $x \in \boldsymbol{R}^n$, x_α is obtained from x by dropping coordinates corresponding to $\bar{\alpha}$; and x_i denotes the ith coordinate of x.

Let $A \in \boldsymbol{R}^{n\times n}$ and $\alpha \subseteq \bar{n}$ be such that det $A_{\alpha\alpha} \neq 0$.

DEFINITION 1.1. *The matrix M defined by*

$$M_{\alpha\alpha} = (A_{\alpha\alpha})^{-1},\ M_{\alpha\bar{\alpha}} = -M_{\alpha\alpha}A_{\alpha\bar{\alpha}},\ M_{\bar{\alpha}\alpha} = A_{\bar{\alpha}\alpha}M_{\alpha\alpha},\ M_{\bar{\alpha}\bar{\alpha}} = A_{\bar{\alpha}\bar{\alpha}} - M_{\bar{\alpha}\alpha}A_{\alpha\bar{\alpha}}$$

is called the PPT of A with respect to α. For any $q \in \boldsymbol{R}^n$, its PPT with respect to A and α is defined by p, where $p_\alpha = -A_{\alpha\alpha}q_\alpha$ and $p_{\bar{\alpha}} = q_{\bar{\alpha}} - A_{\bar{\alpha}\alpha}(A_{\alpha\alpha})^{-1}q_\alpha$. Together, (p, M) is called the PPT of (q, A).

Note that a PPT is defined only with respect to those α for which det $A_{\alpha\alpha} \neq 0$. By convention, when $\alpha = \emptyset$, det $A_{\alpha\alpha} = 1$ and $M = A$ (see [3]).

Remark 1.2. PPTs preserve many interesting properties. If A is in $\boldsymbol{Q}_0$, then every PPT of A is also in $\boldsymbol{Q}_0$. This is so even for $\boldsymbol{Q}$, $\boldsymbol{P}_0$, and $\boldsymbol{C}_0^f$-matrices.

There is a one-to-one correspondence between solutions of (q, A) and those of its PPTs. This is stated below without proof. The details can be seen in [3].

Let (q, A) and (p, M) be as defined in Definition 1.1.

THEOREM 1.3. *The following statements hold:*

(i) $z \in S(q, A)$ *iff* $z' \in S(p, M)$ *where* $z'_\alpha = (Az + q)_\alpha$ *and* $z'_{\bar{\alpha}} = z_{\bar{\alpha}}$.

(ii) *For any* $\lambda \in [0, 1]$ *any* $x, z \in S(q, A), \lambda x + (1-\lambda)z \in S(q, A)$ *iff* $\lambda x' + (1-\lambda)z'$ *in* $S(p, M)$, *where* x' *and* z' *are as defined in* (i).

A diagonal matrix S with the diagonal entries as 1 or -1 is called a *signature matrix.* It is easy to verify that if A is CS, then SAS is CS for all signature matrices S.

2. On connectedness. Our main concern in this section is to show that $S(q, A)$ is connected $\forall q$ when A is a nonnegative $\boldsymbol{P}_0 \cap \boldsymbol{Q}_0$-matrix. We do this by first obtaining a slightly more general result and then applying it to the case of nonnegative matrices.

It is known that $A \in \boldsymbol{R}_0$ iff $S(q, A)$ is bounded for every q (see [3, Proposition 3.9.23]). We need a slightly more general version of this result, stated below. For the sake of completeness, we repeat the proof given in [3, Proposition 3.9.23].

PROPOSITION 2.1. *A matrix $A \in \boldsymbol{R}_0$ iff the set $T = \cup_{q \in H} S(q, A)$ is bounded for every bounded set H.*

Proof. We shall prove the "only if" part. If $A \in \boldsymbol{R}_0$, then $S(q, A)$ is bounded for every q (see [3, Proposition 3.9.23]). Assume, to the contrary, that T is unbounded. Then there exists a sequence $z^k \in T$ such that $\|z^k\| \to \infty$ as $k \to \infty$. Since $z^k \in T$, there exists, for each $k, q^k \in H$ such that $z^k \in S(q^k, A)$. Without loss of generality, we may assume $z^k/\|z^k\|$ converges to some x^0. It follows that $\|x^0\| = 1$. As H is bounded, $q/\|z^k\|$ converges to 0. It follows that x^0 is a nontrivial solution of $(0, A)$. This contradicts that $A \in \boldsymbol{R}_0$. □

LEMMA 2.2. *Suppose $S_1, S_2, \ldots, S_k$ are closed convex sets. If $\cup S_j$ is connected, then it is path connected.*

We use the following results repeatedly in this paper. The first part of the theorem is due to Aganagic and Cottle [1] and the second part is due to Jones and Gowda [5].

THEOREM 2.3. *Let $A \in \boldsymbol{R}^{n \times n}$.*

(i) *If $A \in \boldsymbol{P}_0$, then $A \in \boldsymbol{Q}$ iff $A \in \boldsymbol{R}_0$.*

(ii) *If $A \in \boldsymbol{R}_0$, then $A \in \boldsymbol{P}_0$ iff $A \in \boldsymbol{E}_c$.*

THEOREM 2.4. *Let $A \in \boldsymbol{R}^{n \times n}$ and $q \in \boldsymbol{R}^n$. The following statements hold:*

(i) *$S(q, A)$ is a finite union of closed convex sets.*

(ii) *$S(q, A)$ is connected iff $S(q, A)$ is path connected.*

Proof. (i) For each $\beta \subseteq \bar{n}$, define

$$S(\beta) = \{z \in S(q, A) : A_\beta.\ z + q_\beta = 0, z_{\bar{\alpha}} = 0\}.$$

Clearly, $S(\beta)$ is a closed convex set and $S(q, A) = \cup_\beta S(\beta)$. Thus, $S(q, A)$ is a finite union of closed convex sets.

(ii) This follows from Lemma 2.2. □

THEOREM 2.5. *Let $A \in \boldsymbol{R}^{n \times n}$ be such that for some index set α (possibly empty), $A_{\bar{\alpha}} = 0$. If $A_{\alpha\alpha} \in \boldsymbol{P}_0 \cap \boldsymbol{Q}$, then $S(q, A)$ is connected for every q.*

Proof. We shall prove this induction on n. It is easy to verify the result for $n = 1$. Assume the theorem to be true $\forall (n-1) \times (n-1)$ matrices, $n > 1$. Let $A \in \boldsymbol{R}^{n \times n}$ satisfy the hypothesis of the theorem. Let q be such that $S(q, A) \neq \emptyset$ (this implies $q_{\bar{\alpha}} \geq 0$). If $\bar{\alpha} = \emptyset$, then by Theorem 2.3(i), A is an $\boldsymbol{R}_0 \cap \boldsymbol{P}_0$-matrix, and from Theorem 2.3(ii), $S(q, A)$ is connected. On the other hand, if $\alpha = \emptyset$, then $A = 0$, and it is easy to verify that $S(q, A)$ is connected in this case. So assume α and $\bar{\alpha}$ are nonempty. Without loss of generality, assume $\alpha = \{1, 2, \ldots, k\}, k < n$.

If $q_i > 0$ for some $i \in \bar{\alpha}$, say $i = n$, then $S(q, A) = \{(x^t, 0)^t : x \in S(q_\beta, A_{\beta\beta})\}$, where $\beta = \{1, 2, \ldots, n-1\}$. By induction hypothesis, $S(q_\beta, A_{\beta\beta})$ is connected and hence $S(q, A)$ is connected. So assume $q_{\bar{\alpha}} = 0$.

To complete the proof, we will show that $S(q, A)$ is path connected which will, in turn, imply connectedness. From Theorem 2.4, write

$$S(q, A) = S_1 \cup S_2 \cup \cdots \cup S_h, \tag{2.1}$$

where each S_j is a closed convex set. Define

$$S_0 = \{z \in S(q, A) : z_{\bar{\alpha}} = 0\}.$$

Note that $S_0 = \{z \in \boldsymbol{R}^n : z_\alpha \in S(q_\alpha, A_{\alpha\alpha})$ and $z_{\bar{\alpha}} = 0\}$, and as $A_{\alpha\alpha} \in \boldsymbol{P}_0 \cap \boldsymbol{R}_0, S(q_\alpha, A_{\alpha\alpha})$ is a nonempty connected set. By Theorem 2.4, $S(q_\alpha, A_{\alpha\alpha})$ is path connected. It follows that S_0 is path connected.

Fix a $z^0 \in S_0$. We will show that every element $\bar{z}$ of $S(q, A)$ has a path connecting it to z^0. That is, we will exhibit a continuous function

$$F(t) : [0, 1] \to S(q, A), \qquad \text{such that} \quad F(0) = z^0 \quad \text{and} \quad F(1) = \bar{z}.$$

This will imply that $S(q, A)$ is path connected. So fix any $\bar{z} \in S(q, A)$. Let V be the path component of $S(q, A)$ containing $\bar{z}$. From (2.1), since $S_j s$ are convex, there exists a $J \subseteq \{1, 2, \ldots, h\}$ such that $V = \cup_{j \in \mathbf{J}} S_j$. Consequently, V is a closed set. If $V \cap S_0 \neq \emptyset$, then there is a path connecting z^0 and $\bar{z}$. So assume

$$V \cap S_0 = \emptyset. \tag{2.2}$$

Define

$$\lambda_0 = \inf_{x \in V} p^t z,$$

where p is the vector whose first k coordinates are equal to 0 and whose last $(n - k)$ coordinates are equal to 1 ($p^t z$ is the sum of last $(n - k)$ coordinates of z).

Claim 2.6. There exists a $z^* \in V$ such that $p^t z^* = \lambda_0$. From the definition of λ_0, there exists a sequence $\{z^m\}$ in V such that $p^t z^m$ converges to λ_0. Note that $\{z^m_{\bar{\alpha}}\}$ is bounded. Also, for each $m, z^m_\alpha \in S(q^m_\alpha, A_{\alpha\alpha})$, where $q^m_\alpha = q_\alpha + A_{\alpha\bar{\alpha}} z^m_{\bar{\alpha}}$. As $\{z^m_{\bar{\alpha}}\}$ is bounded, $\{q^m_\alpha\}$ is bounded. Since $A_{\alpha\alpha} \in \boldsymbol{P}_0 \cap \boldsymbol{Q}$, from Proposition 2.1, $\{z^m_\alpha\}$ is bounded. Thus, $\{z^m\}$ is a bounded sequence. Without loss of generality, assume that $z^m \to z^*$. As V is closed, $z^* \in V$. This proves the claim.

Since $V \cap S_0 = \emptyset, \lambda_0 > 0$ and $z^*_{\bar{\alpha}} \neq 0$. Define $\beta = \{i \in \bar{\alpha} : z^*_i > 0\}$. Let m_0 be such that $z^*_i - \frac{1}{m_0} > 0 \ \forall i \in \beta$. For each m, define z^m as

$$z^m_\alpha = x^m, \qquad z^m_i = z^*_i - \frac{1}{m} \quad \text{for} \quad i \in \beta, \qquad \text{and} \quad z^m_i = 0 \quad \text{for} \quad i \in \bar{\alpha} \setminus \beta,$$

where x^m is any particular solution of $S(q^m_\alpha, A_{\alpha\alpha})$ and $q^m_\alpha = q_\alpha + \sum_{i \in \beta}(z^*_i - \frac{1}{m}) A_{\alpha i}$. Then for each $m \geq m_0, z^m \in S(q, A)$. Without loss of generality, we may assume $z^m \in S_1 \ \forall m$. Note that $p^t z^{m_o} < \lambda_0$. From the definitions of λ_0 and V, it follows that

$$V \cap S_1 = \emptyset. \tag{2.3}$$

Since $\{q^m_\alpha\}$ is bounded, and since $A_{\alpha\alpha} \in \boldsymbol{R}_0$, of $A_{\alpha\alpha}, \{x^m\}$ is bounded. Hence, $\{z^m\}$ is bounded. Without loss of generality, assume $z^m \to u$. Since S_1 is closed, $u \in S_1$. Observe that $u_{\bar{\alpha}} = z^*_{\bar{\alpha}}$.

Define $q^0_\alpha = q_\alpha + \sum_{i \in \beta} z^*_i A_{\alpha i}$. Observe that u_α and z^*_α belong to $S(q^0_\alpha, A_{\alpha\alpha})$. Since $A_{\alpha\alpha} \in \boldsymbol{P}_0 \cap \boldsymbol{R}_0, S(q^0_\alpha, A_{\alpha\alpha})$ is path connected. Note that $x \in S(q^0_\alpha, A_{\alpha\alpha})$ iff $z \in S(q, A)$, where $z_\alpha = x$ and $z_{\bar{\alpha}} = z^*_{\bar{\alpha}}$. Let

$$p(t) : [0, 1] \to S(q^0_\alpha, A_{\alpha\alpha}) \quad \text{with} \quad p(0) = u_\alpha \quad \text{and} \quad p(1) = z^*_\alpha$$

be a path connecting u_α and z^*_α. Then, $F(t) : [0, 1] \to S(q, A)$ defined by

$$F(t)_\alpha = p(t) \quad \text{and} \quad F(t)_{\bar{\alpha}} = z^*_{\bar{\alpha}}$$

is a path between u and z^*. From the definition of $V, u \in V$. But this contradicts that $V \cap S_1 = \emptyset$. It follows that $\lambda_0 = 0$ and $S_0 \subseteq V$. Therefore, there is a path connecting z^0 and $\bar{z}$. This completes the proof of the theorem. □

COROLLARY 2.7. *Suppose $A \in \boldsymbol{R}^{n\times n}$ is a nonnegative $\boldsymbol{P}_0 \cap \boldsymbol{Q}_0$-matrix. Then $S(q, A)$ is connected.*

Proof. Let $\alpha = \{i : a_{ii} > 0\}$. From Theorem 2.5 of [6], it follows that $A_{\alpha\alpha}$ is a $\boldsymbol{Q}$-matrix and that $A_{\bar{\alpha}} = 0$. □

We shall need the next two results for our results in the next section.

LEMMA 2.8. *Let $A \in \boldsymbol{R}^{n\times n}$ and $q \in \boldsymbol{R}^n$. Suppose $(\bar{q}, M)$ is a PPT of (q, A). Then $S(q, A)$ is connected iff $S(\bar{q}, M)$ is connected.*

Proof. This follows from the fact that the affine mapping $L : S(q, A) \to S(\bar{q}, M)$ given by

$$L\begin{pmatrix} u_\alpha \\ u_{\bar{\alpha}} \end{pmatrix} = \begin{pmatrix} (Au+q)_\alpha \\ u_{\bar{\alpha}} \end{pmatrix}$$

is invertible (and continuous) and thus a homeomorphism. □

PROPOSITION 2.9. *Let $A \in \boldsymbol{R}^{n\times n}$. The set $S(0, A)$ is path connected.*

Proof. Follows from the fact that if $z \in S(0, A)$, then $\lambda z \in S(0, A)$ $\forall \lambda \geq 0$. □

3. On convexity. Our main result in this section is to show that $\boldsymbol{C}_0^f \cap \boldsymbol{Q}_0$-matrices are sufficient. See [3, 4] for details on sufficient matrices. We do this by showing that $S(q, A)$ as well as $S(q, A^t)$ are convex $\forall q$. But before doing this, we shall present the following result, which is interesting in its own right.

In [6], Murthy and Parthasarathy considered the problem: If (q, A) has a solution, when can we say that (q, A) has solution with a complementary basis? The authors showed that if A has property (D) (A is said to have property (D) if the columns of $A_{\bar{\alpha}\alpha}$ are linearly dependent $\forall \alpha$ such that $\det A_{\alpha\alpha} = 0$), then whenever $S(q, A) \neq \emptyset, (q, A)$ has a solution with a complementary basis. It is interesting to observe that the existence of solutions with complementary bases can be asserted even for the $\boldsymbol{Q}_0$-matrices. Consequently, Theorem 2.19 and Corollary 2.20 of [6] are also valid even without the assumption that A has property (D).

THEOREM 3.1. *Let $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{Q}_0$. Suppose $q \in \boldsymbol{R}^n$ is such that $S(q, A) \neq \emptyset$. Then there exists an index set α such that $\det A_{\alpha\alpha} \neq 0$ and a $z \in S(q, A)$ such that $z_{\bar{\alpha}} = 0$.*

Proof. Let $K(A) = \{q \in \boldsymbol{R}^n : S(q, A) \neq \emptyset\}$. It is well known that $A \in \boldsymbol{Q}_0$ iff $K(A)$ is convex. Let $q \in K(A)$. Note that e, the vector of ones, is in the interior of $K(A)$. Since K(A) is convex, all the points on the line segment joining e and q, except possibly q, are in the interior of $K(A)$. Therefore, there exists a sequence q^k in the interior of $K(A)$ which converges to q. Since there are only finitely many complementary cones, we may assume, without loss of generality, that all the q^k's belong to one full complementary cone. Since complementary cones are closed, it follows that q belongs to the full complementary cone which implies the conclusion of the theorem. □

As an application of the above results, we will first show that solution sets of LCPs arising from $\boldsymbol{C}_0^f \cap \boldsymbol{Q}_0$-matrices are connected.

THEOREM 3.2. *Suppose $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{C}_0^f \cap \boldsymbol{Q}_0$. Then $S(q, A)$ is connected for every q.*

Proof. We shall prove this by induction on n. The result is already known for $n = 1, 2$. So assume that the result is true $\forall (n-1)\times(n-1)$ matrices. Let $A \in \boldsymbol{R}^{n\times n}$ satisfy the hypothesis of the theorem and let q be such that $S(q, A) \neq \emptyset$. From Lemma 2.8, Proposition 2.9, and Theorem 3.1 we may assume, without loss of generality, that

$0 \neq q \geq 0$. Let $\alpha = \{i : q_i = 0\}$. Then $\bar{\alpha} \neq \emptyset$. Note that $S(q, A) = \{z : z_\alpha \in S(q_\alpha, A_{\alpha\alpha}),\ z_{\bar{\alpha}} = 0\}$. To see this, let $z \in S(q, A)$. Then $z^t Az + z^t q = 0$. As $A \in \boldsymbol{C}_0$ and $q \geq 0,\ z^t q = 0$. Since $q_{\bar{\alpha}} > 0,\ z_{\bar{\alpha}} = 0$. By induction hypothesis, $S(q_\alpha, A_{\alpha\alpha}$ is connected. It follows that $S(q, A)$ is connected. □

LEMMA 3.3. *Let $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{P}_0$ and let $q \in \boldsymbol{R}^n$. If x and z are two distinct solutions of (q, A), then there exists an index i such that either $x_i = z_i = 0$ or $y_i = w_i = 0$ where $y = Ax + q$ and $w = Az + q$.*

Proof. Let $\alpha = \{j : x_j > 0\}$. To the contrary, assume that $x+z > 0$ and $y+w > 0$. Then, we have $x_\alpha > 0,\ x_{\bar{\alpha}} = 0,\ y_\alpha = 0,\ y_{\bar{\alpha}} > 0,\ z_\alpha = 0,\ z_{\bar{\alpha}} > 0,\ w_\alpha > 0$ and $w_{\bar{\alpha}} = 0$. Note that, $q_\alpha = -A_{\alpha\alpha}x_\alpha$ and $q_{\bar{\alpha}} = -A_{\bar{\alpha}\bar{\alpha}}z_{\bar{\alpha}}$. Let $u = z - x$. Let us examine $u_k(Au)_k$. If $k \in \alpha$, then $(Au)_k = -A_{k\alpha}x_\alpha + A_{k\bar{\alpha}}z_{\bar{\alpha}} = q_k + A_{k\bar{\alpha}}z_{\bar{\alpha}} = w_k > 0$. Similarly, we can verify that for $k \in \bar{\alpha}, (Au)_k < 0$. Thus, we have

$$u_k(Au)_k < 0 \quad \text{for each } k.$$

From [3, Theorem 3.4.2], it follows that A is not a $\boldsymbol{P}_0$-matrix. From this contradiction, we conclude the proof of the lemma. □

We now present our main result.

THEOREM 3.4. *Let $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{C}_0^f \cap \boldsymbol{Q}_0$. Then A is sufficient.*

Proof. We first show that A is column sufficient. We shall show this by induction on the order of A. From [7, Theorem 4.9], the result is true for $n = 1, 2$. So assume the result $\forall (n-1)\times(n-1)$ matrices, $n > 1$. Let $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{C}_0^f \cap \boldsymbol{Q}_0$. From [3, Theorem 3.1.8], it suffices to show that $S(q, A)$ is convex $\forall q$. Let $q \in \boldsymbol{R}^n$ be arbitrary. If $S(q, A) = \emptyset$, then there is nothing to show. Suppose $S(q, A) \neq \emptyset$. Let $x, z \in S(q, A)$ and $\lambda \in (0, 1)$. From [7, Theorem 4.5] $A \in \boldsymbol{P}_0$. From Lemma 3.3, there exists an i such that either $x_i = z_i = 0$ or $y_i = w_i = 0$.

Case (i): $x_i = z_i = 0$. In this case $x_\alpha, z_\alpha \in S(q_\alpha, A_{\alpha\alpha})$, where $\alpha = \overline{\{i\}}$. From induction hypothesis, $\lambda x_\alpha + (1-\lambda)z_\alpha \in S(q_\alpha, A_{\alpha\alpha})$. This in turn implies that $\lambda x + (1-\lambda)z \in S(q, A)$.

Case (ii): $y_i = w_i = 0$. For simplicity, assume $i = 1$. Here we consider three subcases: (a) $a_{11} > 0$, (b) $a_{11} = 0$ and $a_{1j} \neq 0$, and (c) $A_{1\bar{n}} = 0$.

Consider case (a). Let (q, M) be PPT of (q, A) with respect to $\{1\}$. Then $\bar{x}$ and $\bar{z}$ defined by $\bar{x}_1 = 0,\ \bar{x}_\alpha = x_\alpha,\ \bar{z}_1 = 0$, and $\bar{z}_\alpha = z_\alpha$ are solutions of $(\bar{q}, M)$. Here $\alpha = \overline{\{1\}}$. From induction hypothesis, $\lambda x_\alpha + (1-\lambda)z_\alpha \in (\bar{q}_\alpha, M_{\alpha\alpha})$. This in turn implies $\lambda x + (1-\lambda)z \in S(q, A)$.

Next, consider case (b). Again we may assume, for simplicity, $j = 2$. From [7, Lemma 3.3], $\det A_{\beta\beta} > 0$ where $\beta = \{1, 2\}$. Let $(\bar{q}, M)$ be PPT of (q, A). We can argue exactly as in case (a) to conclude that $\lambda x + (1-\lambda)z \in S(q, A)$.

Last, consider case (c). From [7, Lemma 3.3], $A_{\bar{n}1} = 0$. Here $x_\alpha,\ z_\alpha \in S(q_\alpha, A_{\alpha\alpha})$. It can be verified that $\lambda x + (1-\lambda)z \in S(q, A)$. With this we conclude that $\boldsymbol{C}_0^f \cap \boldsymbol{Q}_0$-matrices are column sufficient.

We will now show that A is row sufficient by showing that $S(q, A^t)$ is convex $\forall q$. Once again we use induction. The result is true for $n \leq 2$. So assume the theorem for $(n-1)\times(n-1)$ matrices, $n > 1$. Let $A \in \boldsymbol{R}^{n\times n}$ satisfy the hypothesis of the theorem. Let $B = A^t$. Let $x, z \in S(q, B)$ and $\lambda \in (0, 1)$. Since $B \in \boldsymbol{P}_0$, it suffices to consider the two cases (i) and (ii) as above. The proof under case (i) is the same as above, so we leave this for the reader to verify.

Consider case (ii). Without loss of generality assume $i = 1$. As before we consider the three subcases: (a) $b_{11} > 0$, (b) $b_{11} = 0$ and $b_{1j} \neq 0$, and (c) $B_{1\bar{n}} = 0$. Consider case (a). Let M and D be the PPTs of A and B with respect to $\alpha = \{1\}$, respectively.

Note that $D_{\bar{\alpha}\bar{\alpha}} = (M_{\bar{\alpha}\bar{\alpha}})^t$. From this and induction it follow that $D_{\bar{\alpha}\bar{\alpha}}$ is column sufficient. Since $y_1 = w_1 = 0$, it follows that $\lambda x + (1-\lambda)z \in S(q, B)$. Next, consider case (b). We may assume $j = 2$ for simplicity. Then we can write

$$A = \begin{bmatrix} 0 & a & b^t \\ -a & a_{22} & A_{2\beta} \\ -b & A_{\beta 2} & A_{\beta\beta} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 0 & -a & -b^t \\ a & a_{22} & A^t_{\beta 2} \\ b & A^t_{2\beta} & A^t_{\beta\beta} \end{bmatrix},$$

where $\beta = \overline{\{1,2\}}$. Let M and D be the PPTs of A and B with respect to $\bar{\beta}$, respectively. We will show that $D_{\bar{\alpha}\bar{\alpha}}$ is column sufficient, where $\alpha = \{1\}$. Note that

$$M_{\bar{\alpha}\bar{\alpha}} = \begin{bmatrix} 0 & \frac{-1}{a} b^t \\ \frac{1}{a} b & M_{\beta\beta} \end{bmatrix} \quad \text{and} \quad D_{\bar{\alpha}\bar{\alpha}} = \begin{bmatrix} 0 & \frac{-1}{a} b^t \\ \frac{1}{a} b & M^t_{\beta\beta} \end{bmatrix}.$$

By induction, $M^t_{\bar{\alpha}\bar{\alpha}}$ is column sufficient and hence $SM^t_{\bar{\alpha}\bar{\alpha}}S$, where S is the $(n-1) \times (n-1)$ signature matrix with its first diagonal entry equal to -1 and the rest of the diagonal entries equal to 1. But $D_{\bar{\alpha}\bar{\alpha}} = SM^t_{\bar{\alpha}\bar{\alpha}}S$. Since $y_1 = w_1 = 0$, we can show as before $\lambda x + (1-\lambda)z \in S(q, B)$. □

4. Summary. In this article, we have considered two open problems. The first problem is to decide whether every solution set of LCP formed by any $\boldsymbol{P}_0 \cap \boldsymbol{Q}_0$-matrix is connected. The second problem is to decide whether $\boldsymbol{C}^f_0 \cap \boldsymbol{Q}_0$-matrices are positive semidefinite. As of this writing, the problems are still open. Our attempt towards these problems is only partial. For the first problem, we verify that nonnegative $\boldsymbol{P}_0 \cap \boldsymbol{Q}_0$-matrices are connected matrices. As for the second problem, we show that $\boldsymbol{C}^f_0 \cap \boldsymbol{Q}_0$-matrices are sufficient. This belongs to a series of results showing that $\boldsymbol{C}^f_0 \cap \boldsymbol{Q}_0$-matrices and positive semidefinite matrices have many properties in common. In [7, 8], it was shown that if $A \in \boldsymbol{C}^f_0 \cap \boldsymbol{Q}_0$, then A, A^t and all their PPTs are completely $\boldsymbol{Q}_0$-matrices, a strong property that positive semidefinite matrices have.

REFERENCES

[1] M. Aganagic and R. W. Cottle, *A note on $\boldsymbol{Q}$-matrices*, Math. Programming, 16 (1979), pp. 374–377.

[2] M. Cao and M. C. Ferris, *$\boldsymbol{P}_c$-matrices and the linear complementarity problem*, Linear Algebra Appl., 246 (1996), pp. 299–312.

[3] R. W. Cottle, J. S. Pang, and R. E. Stone, *The Linear Complementarity Problem*, Academic Press, Inc., Boston, 1992.

[4] R. W. Cottle, J. S. Pang, and V. Venkateswaran, *Sufficient matrices and the linear complementarity problem*, Linear Algebra Appl., 114/115 (1989), pp. 231–249.

[5] C. Jones and M. S. Gowda, *On the connectedness of solution sets of linear complementarity problem*, Linear Algebra Appl., 272 (1998), pp. 33–44.

[6] G. S. R. Murthy and T. Parthasarathy, *Some properties of fully semimonotone $\boldsymbol{Q}_0$-matrices,* SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1268–1286.

[7] G. S. R. Murthy and T. Parthasarathy, *Fully copositive matrices*, Math. Programming, 82 (1998), pp. 401–411.

[8] G. S. R. Murthy, T. Parthasarathy, and B. Sriparna, *Some recent results on the linear complementarity problem*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 898–905.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 4, pp. 1236–1252
© 2000 Society for Industrial and Applied Mathematics

SPARSE APPROXIMATE INVERSE SMOOTHER FOR MULTIGRID*

WEI-PAI TANG† AND WING LOK WAN‡

Abstract. Various forms of sparse approximate inverses (SAI) have been shown to be useful for preconditioning. Their potential usefulness in a parallel environment has motivated much interest in recent years. However, the capability of an approximate inverse in eliminating the local error has not yet been fully exploited in multigrid algorithms. A careful examination of the iteration matrices of these approximate inverses indicates their superiority in smoothing the high-frequency error in addition to their inherent parallelism. We propose a new class of SAI smoothers in this paper and present their analytic smoothing factors for constant coefficient PDEs. The following are several distinctive features that make this technique special:

- By adjusting the quality of the approximate inverse, the smoothing factor can be improved accordingly. For hard problems, this is useful.
- In contrast to the ordering sensitivity of other smoothing techniques, this technique is ordering independent.
- In general, the sequential performance of many superior parallel algorithms is not very competitive. This technique is useful in both parallel and sequential computations.

Our theoretical and numerical results have demonstrated the effectiveness of this new technique.

Key words. sparse approximate inverse, smoother, multigrid

AMS subject classifications. 15A09, 15A23, 65F10, 65F50, 65Y05

PII. S0895479899339342

1. Introduction.

The effectiveness of multigrid is based on two main components: *smoothing* and *coarse grid correction*. The smoothing process, usually carried out by a relaxation method, dampens the high-frequency error components. The coarse grid correction process, carried out by an interpolation, approximates the low-frequency components on the coarser grids. The idea is that the combination of the two will result in a significant error reduction independent of the problem size [22] and hence lead to an efficient solution procedure.

Relaxation methods such as Richardson, Jacobi, and Gauss–Seidel are often used as smoothers for multigrid, although other iterative methods, for instance, incomplete LU factorization (ILU) [23, 41] and (preconditioned) conjugate gradient [1, 7] methods, or even ODE solvers [26] such as Runge–Kutta methods, have also been used for specific problems. Relaxation methods are particularly useful for multigrid since they are simple, and they all have the common property of removing the high-frequency error components [5, 22].

*Received by the editors January 18, 1999; accepted for publication (in revised form) by G. Golub October 6, 1999; published electronically April 18, 2000.

http://www.siam.org/journals/simax/21-4/33934.html

†Department of Computer Science, University of Waterloo, Waterloo, ON, Canada N2L 3G1 (wptang@bryce1.math.uwaterloo.ca). This author was supported by the Natural Sciences and Engineering Research Council of Canada, by the Information Technology Research Centre, which is funded by the Province of Ontario, and by NASA under contract NAS 2-96027 between NASA and the Universities Space Research Association (USRA).

‡SCCM Program, Gates Building 2B, Stanford University, Stanford, CA 94305-9025 (wan@sccm.stanford.edu). Former address: Department of Mathematics, University of California, Los Angeles, CA 90095-1555. This author was partially supported by the NSF under contract ASC-9720257, NASA Ames under contract NAG2-1238, Sandia National Laboratory under contract LG-4440, Lawrence Livermore National Laboratory under contract W07405-Eng-48, Department of Energy under the Accelerated Strategic Computing Initiative (ASCI), and the Alfred P. Sloan Foundation as a Doctoral Dissertation Fellow.

In practice, Gauss–Seidel is usually the most effective smoother among relaxation methods. A drawback, however, is that Gauss–Seidel is a very sequential algorithm. A parallel version may be obtained by a special ordering of the unknowns, for example, red-black ordering for the five-point stencil operator on a square grid. The Jacobi method, on the other hand, is a very parallel method, but its smoothing efficiency is not as good as that of the Gauss–Seidel method. It is also well known that Gauss–Seidel smoothers do not work well for anisotropic problems and discontinuous coefficient problems. In this paper, we propose a new class of smoothers derived from sparse approximate inverse (SAI) preconditioners. It has a smoothing efficiency similar to Gauss–Seidel and it is independent of ordering. Moreover, for hard problems, we can improve the smoothing efficiency by adaptively adjusting the quality of the approximate inverse, for instance, by adding more nonzeros. Consequently, this new technique is more robust than using Gauss–Seidel smoothers. Our numerical tests verify this statement.

We remark that Huckle [25] independently experimented with an SAI smoother for multigrid by modifying the standard Gauss–Seidel iteration. Specifically, instead of using the exact inverse of the lower triangular matrix, he used an SAI computed by the techniques described in [20]. In our approach, we do not restrict ourselves only to Gauss–Seidel. Indeed, we replace the Gauss–Seidel smoother entirely by an SAI smoother. The resulting multigrid is efficient, and we have more flexibility in improving the smoothing quality for hard problems.

In section 2, we describe the construction of the SAI smoother. In section 5, we analyze the smoothing property of the proposed smoother both analytically and numerically for constant coefficient PDEs. Comparison to the Gauss–Seidel smoother is also given. Some techniques to improve the smoothing character for anisotropic problems are presented. Finally, in section 6, we show the effectiveness of the SAI as a smoother for multigrid on a variety of problems, including anisotropic problems, discontinuous coefficient problems, and unstructured grid problems.

2. SAI smoothers. Various techniques have been proposed for an effective SAI preconditioner [2, 4, 9, 11, 18, 20, 29, 30, 37]. However, the goal of constructing an effective smoother is very different from finding a good preconditioner. For a powerful preconditioner, the capability for removing both the high- and low-frequency errors is essential. In contrast, a good smoother may just dampen the high-frequency errors effectively. In this respect, much of the weaknesses [37] of SAI preconditioners become the strengths for SAI smoothers. Our proposal is to explore this to construct a powerful smoother.

Most SAI approaches seek a sparse matrix M so that the error of the residual

$$\mathcal{E} = MA - I$$

is minimized in some measure. The sparsity constraint often limits the effectiveness of M as a preconditioner due to the locality of the sparse approximation. This lack of global approximation has created many difficulties for a powerful preconditioner. Various additional techniques are required to improve the quality of an SAI preconditioner [8, 37]. The requirement for a good smoother, however, can take advantage of the locality of the sparse approximation. In the following sections, we propose a systematic approach for constructing sparsity patterns and a modified least squares formulation for SAI smoothers. The cost of this simpler form of *approximate inverse–local least squares approximation* is much less expensive than other typical SAI approaches, and yet it possesses a very effective smoothing property which we will show

in section 5.

2.1. Modified Frobenius norm construction. The construction of our SAI smoothers is based on the Frobenius norm approach described by Benson [2] and Benson and Frederickson [3]. Consider the left preconditioned linear system

$$MAy = Mb,$$

where M is a left preconditioner. An SAI preconditioner M is defined as a solution to the following minimization problem:

$$\min_M \|MA - I\|_F^2$$

subject to some constraints on the number and position of the nonzero entries of M. The minimization problem is equivalent to n independent least squares problems which can be solved in parallel:

$$\min_{m_j} \|A^T m_j - e_j\|_2, \qquad j = 1, \ldots, n, \tag{2.1}$$

where m_j is the jth row of M and e_j is the jth column of I.

In practice, the size of the matrix A may be too large for an effective solution of (2.1). However, since we are looking for a sparse solution M, the number of nonzeros of m_j is small and can be bounded. Thus, the matrix A in (2.1) can be replaced by a submatrix of itself, A_j, whose columns correspond to the nonzero entries of m_j. Let the compressed dense vector of m_j be m'_j, which contains only the nonzeros of m_j. The least squares problem (2.1) becomes

$$\min_{m_j} \|A_j^T m'_j - e_j\|_2, \qquad j = 1, \ldots, n. \tag{2.2}$$

Our modified approach simplifies (2.2) further by selecting a subset of nonzero rows of A_j. In the following, we describe a systematic procedure of constructing submatrices of A for defining the m_j's. In contrast to many other SAI approaches, this modified Frobenius norm approach has direct control of the number and location of the nonzeros of the m_j's as well as the complexity of the least squares problems a priori.

2.2. Algorithm. The submatrix construction algorithm is described by using graph theory notation. A sparse matrix A can be represented by a digraph $\mathcal{G} = (\mathcal{O}, \mathcal{E})$ [14]. For multigrid methods, the graph often is the mesh of the PDE solution domain. Define $\mathcal{L}_k(o_i)$, the k-level neighbor set of node o_i, the nodes of which are a distance $k+1$ or less from o_i. The use of the level concept to define the sparsity pattern for incomplete factorizations is widely present in ILU preconditioning [21, 12]. The 0-level neighbor set $\mathcal{L}_0(o_i)$ contains all the nodes which directly connect to node o_i. For PDE problems with a second-order finite difference/finite element discretization, $\mathcal{L}_0(o_i)$ is just the set of stencil points. Similarly, define $\mathcal{W}_k(o_i)$, the kth wavefront of node o_i, as the set of nodes which are a distance $k+1$ from node o_i.

Remark. It is known that the inverse of a matrix arising from a PDE problem is an approximation of the Green's function for that PDE. A row of this inverse for the grid node o_i would be an approximation of the Green's function when one variable is fixed at the given location. For many two-dimensional elliptic PDEs, this row can be portrayed as a decay surface with a singularity at node o_i. This leads to the wavefront decay behavior where the wavefront center is the corresponding node o_i. Let A_{ij}^{-1} be

the element of the inverse A^{-1} at location (i, j). The size of A_{ij}^{-1} is then associated with the unit contribution from node o_j to node o_i. With the wavefront decay, the set $\mathcal{L}_k(o_i)$ includes the k most influential wavefronts for o_i. This motivates us to choose $\mathcal{L}_k(o_i)$ for the sparsity pattern to approximate the discrete Green's function at node o_i. The computation of these locations is also inexpensive. An effective choice of the sparsity pattern can have a significant impact on the quality and the cost of the SAI.

In section 2.1, we mentioned that further saving is required to yield an effective approximate inverse. The least squares problems (2.2) have to be replaced by some much smaller tasks. Different algorithms for constructing an approximate inverse lead to different approaches for choosing the sparsity pattern of m_j and a submatrix of A_j in (2.2). After the submatrix $\widetilde{A}_j$ is chosen, (2.2) is replaced by a smaller least squares problem,

$$\min_{m_j} \|\widetilde{A_j^T} m'_j - e_j\|_2.$$

In constructing an effective SAI smoother, two important issues arise:

- The sparsity pattern of m_j. Based on the wavefront decay observation, we will choose $\mathcal{L}_k(o_j)$. A larger k is related to the higher cost for constructing the SAI smoother. For a regular problem, $k = 0$ is often good enough.
- What kind of local error should be removed effectively with this smoother? If we choose the submatrix of A_j^T corresponding to the rows for the nearest neighbors $\mathcal{L}_l(o_j)$, the resulting smoother will remove the error in this neighborhood more effectively, since the errors of this particular SAI on these locations are optimized by the least squares problem. For anisotropic problems, the error is not distributed evenly. This neighborhood measuring by a simple distance is not physically correct, and a more sophisticated technique should be introduced (see section 4). It is clear that a higher level l provides an approximation which is good for more neighbors. Notice that l is usually larger than k. Our experience indicates that $k = l$ is not a good choice.

Now, we define the rectangular submatrix[1]

$$A_{k,l} \equiv A(\mathcal{L}_k(o_j), \mathcal{L}_l(o_j)) \tag{2.3}$$

as the (k, l)-level local matrix of node o_j. This matrix takes rows corresponding to nodes $\mathcal{L}_k(o_j)$ and columns corresponding to nodes $\mathcal{L}_l(o_j)$ from A.

We introduce the (k, l)-*level local least squares approximation* of an inverse as follows. For each node o_j (row) of the approximation, the least squares solution m_{o_j} of

$$A_{k,l}^T m'_{o_j} = e_{o_j} \tag{2.4}$$

is taken for the nonzero values at the corresponding location $\mathcal{L}_k(o_j)$ of the sparse approximation of the discrete Green's function corresponding to node o_j. Here, e_{o_j} is a unit basis vector of size $|\mathcal{L}_l(o_j)|$, with one in the position corresponding to o_j and zeros elsewhere. In other words, we inject each element of the least squares solution m'_{o_j} back into a zero vector m_{o_j} at the corresponding locations in $\mathcal{L}_k(o_j)$ and use this sparse row to approximate the row o_j of A^{-1}.

[1]The MATLAB notation is adopted for extracting a submatrix from a given matrix A.

Remark. The first SAI approach proposed by Benson [2] is precisely the $(0,0)$-level local least squares approximation. However, it was not a good choice for either preconditioner or smoother.

The relaxation method

$$x^{\mu+1} = x^{\mu} + M(b - Ax^{\mu})$$

resulting from the SAI M constructed by the (k,l)-level local least squares approximation is called the (k,l)-level SAI smoothing.

2.3. Example: Constant coefficient PDEs. We illustrate the (0,1)-level local least squares approximation method by a model constant coefficient second-order elliptic PDE:

$$w_1 u_{xx} + w_2 u_{yy} + s u_x + t u_y = f(x,y), \qquad (x,y) \in \Omega,$$
$$u|_{\Gamma} = g(x,y)$$

on a square domain Ω. Suppose we use a $p \times q$ regular grid. The discretization matrix A is penta-diagonal and of size $(p \times q)^2$, using the conventional 5-point stencil finite difference method. Moreover, the resulting difference equation can be generally written as

$$b u_{i-1,j} + d u_{i,j-1} + a u_{i,j} + c u_{i+1,j} + e u_{i,j+1} = h^2 f_{i,j}, \qquad 1 \le i \le p, \quad 1 \le j \le q,$$

where a, b, c, d, and e are constants. Here we present the picture of the stencil and the construction of a $(0,1)$-level submatrix corresponding to the center node on a 5×5 grid in Figure 2.1. The vertical (horizontal) dashed lines indicate the columns (rows) to be extracted from the original matrix to form the $(0,1)$-level submatrix. The arrows indicate the position of the center node. The circulated nonzeros on the cross points of the dashed lines will appear in the submatrix. In Figure 2.1, the $(0,1)$-level local least squares problem for the center node o_j is

$$\begin{pmatrix}
a & & & & c \\
 & a & & & d \\
 & & a & & b \\
 & & & a & e \\
b & e & c & d & a \\
c & & & & \\
d & c & & & \\
 & d & & & \\
 & b & d & & \\
 & & b & & \\
 & & e & b & \\
 & & & e & \\
e & & & c &
\end{pmatrix}
\begin{pmatrix} x_W \\ x_N \\ x_E \\ x_S \\ x_C \end{pmatrix}
=
\begin{pmatrix}
0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0
\end{pmatrix}, \tag{2.5}$$

which yields

$$(m_{o_j})_i = \begin{cases} x_W & \text{if } i = o_j - 1, \\ x_N & \text{if } i = o_j + 5, \\ x_E & \text{if } i = o_j + 1, \\ x_S & \text{if } i = o_j - 5, \\ x_C & \text{if } i = o_j, \\ 0 & \text{otherwise.} \end{cases}$$

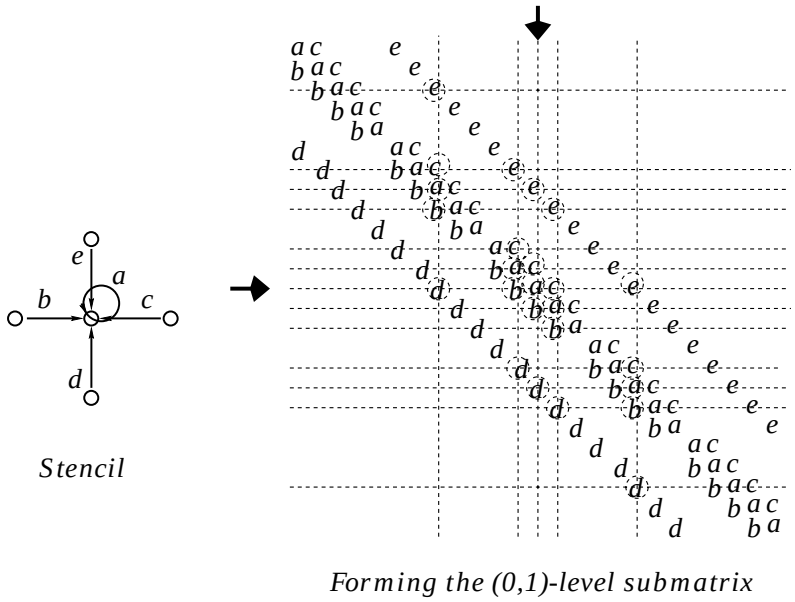


FIG. 2.1. *A stencil and the construction of a submatrix.*

2.4. Simplified SAI smoothers. Computing these local least squares solutions is an easy task and can be implemented in parallel effectively. For any constant coefficient PDE on a regular mesh, we may further simplify the computations as follows. We observe that the local least squares problems, and hence the least square solutions, corresponding to the interior points are identical (cf. (2.5)). However, they may be different near the boundary. We simply use the least squares solutions obtained from the interior points for the solutions at the boundary. Hence, we need to solve only one small least squares problem to obtain the entire approximate inverse. Our tests indicate that this simplified approach does not bring any noticeable penalty in performance for regular PDEs.

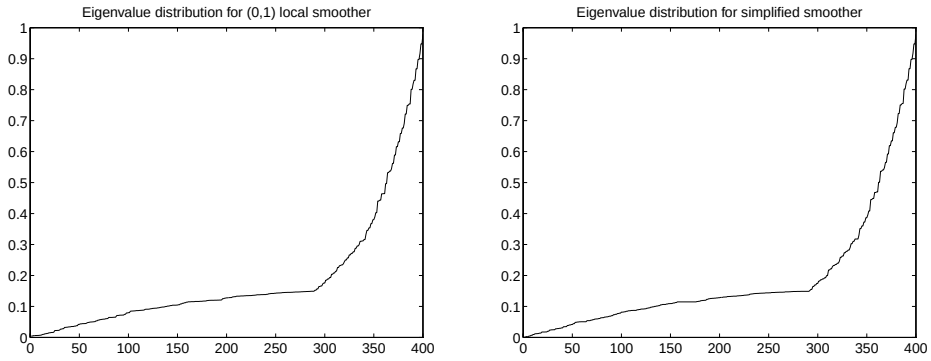
As an illustration, consider solving the following Poisson equation on a 20×20 rectangular grid:

$$\Delta u(x, y) = 1, \qquad (x, y) \in \Omega.$$

In Figure 2.2, we present the plots of the eigenvalues of the iteration matrices associated with the (0,1)-level SAI smoother and its simplified version. The eigenvalues in Figure 2.2(a) were calculated by MATLAB using the *eig* function and those in Figure 2.2(b) by the analytic formula (5.2) derived in section 5. There are no visible differences.

Finally, we would like to mention a number of other techniques which can also be adopted to save the cost in computing the SAI. A follow-up paper will discuss these in lengths. For example,

- the shape of the Green's function for different positions inside the solution region does not vary much for a PDE with constant coefficients; when an unstructured mesh is used, we may compute one approximation of the discrete Green's function accurately and use interpolation to obtain the approximation for all other mesh points;



(a) $(0,1)$-level SAI smoother (b) Simplified smoother

FIG. 2.2. *Comparison of the eigenvalue distributions.*

- for smooth variable coefficient PDEs, the use of one local least squares solution for several of its neighbors is feasible;
- for anisotropic problems, a higher level SAI smoother is required; however, the cost of the local least-squares problems grows rapidly; an a priori drop tolerance technique [37] can significantly reduce the computations.

3. Computation complexity. The use of SAI smoothers consists of two phases: the setup phase for constructing the sparse approximate inverse by solving the n independent least squares problems (2.2) and the application phase, which applies the SAI matrix to a vector. We analyze the complexity of each phase in the following sections.

3.1. Setup phase. The computational cost involved depends on the size of the least squares matrix $A_{k,l}$, which in turn depends on the size of $|\mathcal{L}_k(o_j)|$ and the number of levels k used. In general, it is hard to estimate $|\mathcal{L}_k|$ for a general sparse matrix A. Thus, we restrict our discussion to matrices arising from 5-point stencil discretizations on rectangular grids. First, we have the formula for $|\mathcal{L}_k|$.

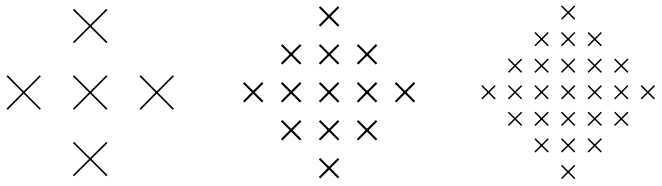
THEOREM 3.1. *For nodes o_j away from the boundary, the number of the first k level nodes is*

$$|\mathcal{L}_k(o_j)| = 2k^2 + 6k + 5.$$

Proof. The formula can be easily derived by examining the first few levels of neighboring nodes as shown in Figure 3.1. First, we observe that there are $2k+3$ rows for the level k pattern. Second, the number of nodes in the central row increases by 2 for each increasing level, and hence is equal to $2k + 3$. The number of nodes in the upper $k + 1$ rows starts from 1 and increases by 2 for each row below, and hence it has a total of $(k + 1)^2$ nodes. Thus, altogether, we have

$$|\mathcal{L}_k(o_j)| = 2(k + 1)^2 + (2k + 3) = 2k^2 + 6k + 5. \qquad \square$$

Although $|\mathcal{L}_k|$ grows quadratically with respect to k, we remark that our numerical experiments in section 6 show that small values of k are sufficient to give rise to an effective SAI smoother. In fact, $k = 1$ is good enough for smooth coefficient PDEs.



(a) level=0 (b) level=1 (c) level=2

FIG. 3.1. *Different level of stencil nodes.*

The following is a well-known result [16] for estimating the complexity of the QR factorization method for solving least squares problems. For simplicity, we assume that $|\mathcal{L}_k(o_j)| \equiv |\mathcal{L}_k|$ is constant for all o_j.

THEOREM 3.2. *The complexity of solving* (2.4) *is*

$$2|\mathcal{L}_l||\mathcal{L}_k|^2 - \frac{2}{3}|\mathcal{L}_k|^3.$$

For general purpose, (0,1)-level SAI smoother is sufficient. In that case, $|\mathcal{L}_0| = 1$ and $|\mathcal{L}_1| = 5$. Thus we need about 10 flops to solve each least squares problem (2.4), and hence the total cost is about 2 Gauss–Seidel iterations.

There are several ways to reduce the cost of the setup phase. For instance, we may employ the simplified version of SAI smoothers described in section 2.4 for constant coefficient PDEs. We need to solve only one least squares problem to obtain the entire approximate inverse, and its cost is clearly negligible.

In many applications where we may have to solve the linear system $Ax^{(i)} = b^{(i)}$ repeatedly with different right-hand sides, the setup cost becomes less significant. Moreover, in a parallel computation environment, the setup phase can be implemented easily in parallel.

3.2. Application phase. Applying an SAI to a vector is simply a matrix-vector multiplication. Thus, its complexity depends on the number of elements per row, which is determined by $|\mathcal{L}_k|$.

THEOREM 3.3. *The complexity of applying the SAI smoother is*

$$n(2k^2 + 6k + 5).$$

Remark. For the recommended (0,1)-level SAI smoother, the complexity is the same as one step of Gauss–Seidel iteration. In a parallel computation environment, the application of the SAI smoother can be done in parallel easily, whereas Gauss–Seidel is a very sequential process.

4. Adaptivity and anisotropic problems. In addition to simple construction and easy parallel implementation, SAI smoothers offer the capability to improve their quality by adjusting the number of levels (i.e., the sparsity of the approximate inverse) used. Relaxation methods, for instance, Gauss–Seidel, do not have such a feature, which means users have to look for other smoothers whenever relaxation smoothers fail to give fast multigrid convergence. For SAI smoothers, we may tune a parameter—the number of levels—to improve the resulting multigrid convergence; in the extreme case when the number of levels equals the matrix size, the SAI is the exact inverse, yielding

one step multigrid convergence. Thus, SAI smoothers are more robust, flexible and can be applied to a larger class of problems.

The potential disadvantage of using higher level SAI smoothers is the higher cost involved. However, like ILU(k) methods, the issue is not just cost per iteration but also whether the resulting multigrid iteration converges. After all, a costly convergent method is still better than inexpensive nonconvergent methods. Anisotropic problems usually have many relatively small entries in both the matrix A and its inverse. We may significantly reduce the cost by discarding these small entries to obtain a sparser approximate inverse. The resulting higher level SAI smoothers can be shown to be inexpensive and yet very effective for anisotropic problems (section 6).

We remark that there are two typical techniques for multigrid to solve anisotropic problems. One technique, similar to ours, is to use a better smoother, for instance, block relaxation methods. Another approach is to use semicoarsening, a special coarsening technique. Both techniques are usually used for regular grids only and are most effective if the direction of isotropy is in either the x or y direction. Special techniques are needed to handle variable anisotropic problems. Our SAI smoothers, on the other hand, are algebraic and do not require any information about the direction of isotropy. Exactly the same procedure can be applied to variable anisotropic problems as well as to irregular grids. We also mention that there exist counterparts of block relaxation smoothers and semicoarsening for irregular grids. They are generally more complicated in that they require heuristic algorithms to determine the lines of different isotropy. Finally, we remark that our SAI smoothers are not meant to be the ultimate method for solving constant coefficient anisotropic problems on regular grids where standard techniques work well. Rather, they are proposed as robust smoothers which are simple to use, applicable to both regular and irregular grids, effective for both simple and tough (e.g., anisotropic) PDE problems, and easily parallelizable.

5. Smoothing factor analysis. We present an analysis of the smoothing factors for constant coefficient PDEs on two-dimensional rectangular grids (equispaced) in this section. For simplicity, we consider only the simplified (0,1)-level SAI smoothers, which have been shown in section 2.4 to be only a small perturbation of the original (0,1)-level SAI smoothers.

The smoothing factor analysis is based on the following theorem.

THEOREM 5.1. *Given two tridiagonal matrices*

$$B = \text{tridiag}(b, a', c)_{p\times p}, \qquad C = \text{tridiag}(d, a'', e)_{q\times q},$$

where b, c, d, e are nonnegative. Then the eigenvalues of the matrix

$$B \otimes I_{q\times q} + I_{p\times p} \otimes C \tag{5.1}$$

are

$$a + 2\sqrt{bc}\cos\theta_k + 2\sqrt{de}\cos\theta_j, \qquad 1 \le k \le p \;\; 1 \le j \le q, \tag{5.2}$$

where $a = a' + a''$, $\theta_k = \frac{k\pi}{p+1}$, $\theta_j = \frac{j\pi}{q+1}$. The corresponding eigenvectors are

$$\left\{ \left(\frac{c}{b}\right)^{\frac{n-s}{2}} \left(\frac{e}{d}\right)^{\frac{n-t}{2}} \sin(s\theta_k)\sin(t\theta_j) \right\}, \qquad 1 \le s \le p,\ 1 \le t \le q.$$

Proof. This result can be shown by using the well-known analysis for the eigenvalues and eigenvectors of a tridiagonal matrix; for example, see [42, 34]. ∎

Unlike Fourier analysis for PDEs with periodic boundary condition, Theorem 5.1 shows that the sine functions $\{\sin\frac{ks\pi}{m+1}\sin\frac{jt\pi}{n+1}\}$ are not the eigenvectors for general constant coefficient PDEs with Dirichlet boundary condition; the eigenvectors may depend on the coefficients a, b, c, d, and e. To give further insight into Theorem 5.1, we continue the analysis further with the assumption that A is symmetric; i.e., $b = c$ and $d = e$.

Denote by M the simplified SAI of the matrix A. Then M can also be written in the form of (5.1).

THEOREM 5.2. *The eigenvalues of the iteration matrix $I - MA$ are*

$$1 - ((x_C + 2x_W\cos\theta_k + 2x_N\cos\theta_j)\ (a + 2b\cos\theta_k + 2d\cos\theta_j))\,,$$

where $1 \le k \le p, \quad 1 \le j \le q$.

Proof. Consider the left-hand side of (2.5). By the symmetry of A, if we switch x_W and x_E, we will get exactly the same coefficient least squares matrix. Thus $x_E = x_W$. Similarly, we also have $x_N = x_S$. Hence M is also symmetric. By Theorem 5.1, both A and M have the same set of eigenvectors and so they can be diagonalized at the same time. The eigenvalues of the iteration matrix are then given by Theorem 5.1. ∎

Remark. This analysis can also be generalized to three-dimensional problems.

5.1. Local mode analysis. We carry out the classical local mode analysis [5] to analyze the smoothing efficiency of the (0,1)-level SAI smoothers for Poisson equations. By a direct computation and Theorem 5.1, we have the following two lemmas.

LEMMA 5.3. *Suppose A is the standard Laplacian operator. Then the least squares solution of* (2.5) *is*

$$x_W = x_N = x_E = x_S = \frac{3}{61}, \qquad x_C = \frac{17}{61}.$$

LEMMA 5.4. *Let M be the $(0,1)$-level sparse approximate inverse of the standard Laplacian operator. Then the eigenvalues of M are*

$$-\frac{17}{61} - \frac{6}{61}\cos\theta_k - \frac{6}{61}\cos\theta_j.$$

Remark. Lemma 5.4 shows that $M < 0$, which implies M is invertible.

As a consequence of Theorem 5.2, we have the following result.

LEMMA 5.5. *The eigenvalues of the iteration matrix of the $(0,1)$-level smoother are*

$$\lambda(\theta_k,\theta_j) = 1 - \left(\frac{17}{61} + \frac{6}{61}\cos\theta_k + \frac{6}{61}\cos\theta_j\right)(4 - 2\cos\theta_k - 2\cos\theta_j). \tag{5.3}$$

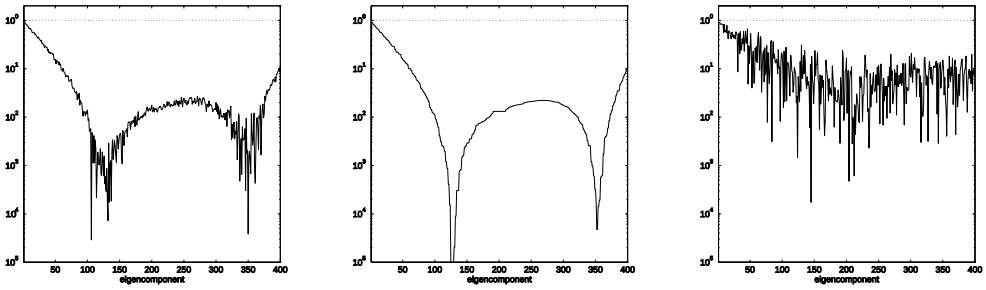
Now, we show the smoothing factor of our SAI smoother. Define the high frequencies:

$$\Theta_H = \left\{(\theta_k,\theta_j) : \frac{p}{2} \le k \le p, \ \ \frac{q}{2} \le j \le q\right\}.$$

We have the following smoothing factor result.

THEOREM 5.6. *Let ρ be the smoothing factor defined as*

$$\rho \equiv \max\{|\lambda(\theta_k,\theta_j)| : (\theta_k,\theta_j) \in \Theta_H\},$$



(a) (0,1)-level smoother (b) Simplified smoother (c) GS smoother

FIG. 5.1. *Errors after two smoothing steps of different smoothers. The dotted line denotes the initial error.*

where $\lambda(\theta_k, \theta_j)$ *is as in Lemma* 5.5. *Then* $\rho \leq \frac{21}{61} \approx \frac{1}{3}$.

Proof. Write $c_k = \cos\theta_k$ and $c_j = \cos\theta_j$. For $(\theta_k, \theta_j) \in \Theta_H$, we have

$$-1 < c_k, c_j < 0. \tag{5.4}$$

Expand the formula in (5.3) and obtain

$$\lambda(\theta_k, \theta_j) = \frac{12}{61}(c_k + c_j)^2 + \frac{10}{61}(c_k + c_j) - \frac{7}{61}.$$

Let $x = c_k + c_j$. By (5.4), $-2 < x < 0$. By a simple calculation, the polynomial $p(x) = \frac{12}{61}x^2 + \frac{10}{61} - \frac{7}{61}$ has a maximum in absolute value at $x = -2$ for $-2 \leq x \leq 0$. Thus

$$\rho \leq \max\{|p(x)| : -2 \leq x \leq 0\} = |p(-2)| = \frac{21}{61}. \qquad \square$$

Thus, the SAI smoothers reduce the high frequencies by a factor of almost 1/3, and hence they are effective smoothers for smooth coefficient PDEs. As a comparison, similar calculations can be carried out for Gauss–Seidel with natural and red-black ordering [22, 40]. The smoothing factors are 1/2 and 1/4, respectively. In section 6, we show that the SAI smoothers are also effective for tough PDEs such as anisotropic problems.

5.2. Spectral analysis. In addition to the above Fourier analysis, we compare the smoothing efficiency of our SAI smoothers and the Gauss–Seidel smoother using a spectral analysis. Figure 5.1 shows the errors after two smoothing steps of different smoothers. The x-axis represents the eigencomponents with respect to the Laplacian matrix, and the y-axis represents their magnitudes. For Laplacian matrices, the small and large eigenvalues correspond to the low and high frequencies, respectively. From the plots, there is no significant difference between the (0,1)-level local smoother and the simplified smoother. Moreover, they both dampen high frequencies more effectively than the Gauss–Seidel smoother.

6. Numerical results. In this section, we demonstrate the effectiveness of the proposed SAI smoothers. Example 1 illustrates that our SAI smoothers perform essentially the same as GS smoothers, if the latter work. It shows that the additional easy

parallel implementation feature of the SAI smoothers do not affect convergence. Besides, for some problems, SAI smoothers may converge while Gauss–Seidel smoothers do not.

In addition to the gain of parallel efficiency, our SAI smoothers have the capability of improving themselves by adjusting a parameter. When relaxation smoothers do not work, we can do nothing to make them work. For SAI smoothers, we may increase the number of levels, which in turn increases the density of the sparse approximate inverse, improving their quality for hard problems. We illustrate this point by solving anisotropic coefficient PDE problems in Example 2. We show that the cost of higher level SAI smoothers is only about twice as expensive per iteration, whereas the reduction in the number of iterations is much more than double. Moreover, the SAI smoothers also work well for unstructured grid problems where the standard multigrid techniques for anisotropic problems on regular grids cannot be applied. Thus our SAI smoothers are more robust.

In all the examples, Dirichlet boundary conditions are used, and the right-hand-side function is $f(x) = 1$. In the multigrid procedure, a V-cycle is used with two presmoothing and two postsmoothing steps. Linear interpolation is used for structured grid problems, and a specialized energy-minimizing interpolation [38, 39] is used for unstructured grid problems. The number of multigrid levels is such that the coarsest grid is 3×3 for structured square grid problems and a total of four levels for unstructured grid problems. The iteration was terminated when the relative residual norm was less than 10^{-8}. We are using zeros as the initial guess in all cases. Actually, we may report better numbers of iterations if a random initial guess is used. The results are summarized in table form, where the number of V-cycles and the average convergence rate of the last 10 iterations are shown.

Example 1. We compare the performance of two Gauss–Seidel smoothers with the (0,1)-level SAI smoother and its simplified version, the simplified (0,1)-level SAI smoother by several smooth and rough coefficient PDEs on a 33×33 square grid. The equation for the variable coefficient problem is

$$((1 + x^2)u_x)_x + u_{yy} + \tan^2 y u_y = -100x^2.$$

The equation for the helical spring problem [15] is

$$u_{xx} + u_{yy} + \frac{3}{5 - y}u_x - 2G\lambda = 0,$$

where G and λ are some constants. The discontinuous coefficient PDE is

$$(a(x, y)u_x)_x + (b(x, y)u_y)_y + u_x + u_y = \sin(\pi xy),$$

where the coefficients $a(x, y)$ and $b(x, y)$ are defined as

$$a(x, y) = b(x, y) = \begin{cases} 10^{-3} & (x, y) \in [0, 0.5] \times [0.5, 1], \\ 10^{3} & (x, y) \in [0.5, 1] \times [0, 0.5], \\ 1 & \text{otherwise.} \end{cases}$$

Table 6.1 shows the convergence results. For smooth coefficient problems, all the methods have similar convergence behavior. Since the (0,1)-level is used, the cost per iteration is also the same for all the methods. Besides, we observe that the simplified (0,1)-level SAI performs essentially the same as its original version, while its setup cost involves only one least squares solve.

TABLE 6.1

Smooth and rough coefficient PDEs on a 33×33 *mesh. GS, Gauss–Seidel with natural ordering; GS(rb), Gauss–Seidel with red-black ordering; SAI,* $(0,1)$*-level SAI smoother; SSAI, simplified* $(0,1)$*-level SAI smoother.* ∞ *indicates divergence.*

Problems	Iteration				Convergence rate			
	GS	GS(rb)	SAI	SSAI	GS	GS(rb)	SAI	SSAI
Poisson	9	7	9	9	0.11	0.07	0.11	0.11
Variable coeff.	13	10	12	17	0.22	0.15	0.19	0.35
Helical spring	12	9	12	12	0.20	0.12	0.19	0.20
Discont. coeff.	∞	∞	22	∞	∞	∞	0.40	∞

For the discontinuous coefficient problem, the two Gauss–Seidel smoothers diverge while our (0,1)-level SAI smoother converges. The failure of the simplified SAI smoother is expected since the Green's function of a single point in the interior may not be a good approximation to those with a region of totally different coefficients.

Example 2. We show how we may improve the quality of SAI by using a higher level SAI when (0,1)-level SAI does not work well. We consider two model anisotropic coefficient PDEs.

Problem 1: The single direction anisotropic problem is

$$\begin{aligned} 100u_{xx} + u_{yy} &= 1, \qquad (x,y) \in \Omega, \\ u|_\Gamma &= 0. \end{aligned}$$

Problem 2: This has a more sophisticated anisotropy structure in both the x and y directions:

$$a(x,y)u_{xx} + b(x,y)u_{yy} = 1,$$

where the coefficients $a(x,y)$ and $b(x,y)$ are defined as

$$a(x,y) = \begin{cases} 100 & (x,y) \in [0,0.5] \times [0,0.5] \text{ or } [0.5,1] \times [0.5,1], \\ 1 & \text{otherwise}, \end{cases}$$

$$b(x,y) = \begin{cases} 100 & (x,y) \in [0,0.5] \times [0.5,1] \text{ or } [0.5,1] \times [0,0.5], \\ 1 & \text{otherwise}. \end{cases}$$

The results are shown in Tables 6.2 and 6.3. The first three rows show the convergence results for Problem 1 on 32×32, 64×64, and 128×128 grids, respectively. Similarly, rows 4 to 6 show the results for Problem 2. The last two rows show the results for Problem 1 on two unstructured grids shown in Figure 6.1.

As is well known, multigrid with Gauss–Seidel smoother is very slow for these kinds of problems. Similar results are also obtained for Gauss–Seidel with red-black ordering, and hence they are omitted. For Problem 1 on regular grids, a standard technique[2] for improving the multigrid convergence is to use line (block) relaxation methods. As indicated in the table, the block Jacobi smoother is quite effective for small grids but eventually slows down for bigger grids. Moreover, the block Jacobi smoother for Problem 1 does not work for Problem 2 since the latter has different isotropies in both the x and y directions. We have to manually change the two line relaxation smoothing in the x direction into one line relaxation in the x direction and

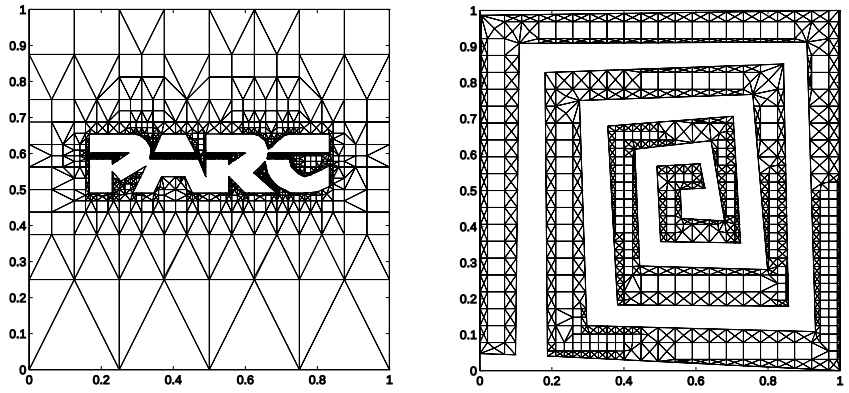
[2]Another technique is to use *semicoarsening*. Since this approach uses a different coarsening rather than an improved smoother, we do not compare this method with our SAI smoother.

TABLE 6.2
Convergence results for anisotropic problems on different grids. GS, Gauss–Seidel with natural ordering; BJ, Block Jacobi. See text for the definitions of SAI$(3,\epsilon)$ *and SAI*$(4,\epsilon)$. $\epsilon = 0.0008$ *for the regular grid problems and* 0.0004 *for the unstructured grid problems.*

Grids	Iteration			
	GS	BJ	SAI(3,ϵ)	SAI(4,ϵ)
Problem 1 (32×32)	> 100	9	25	18
Problem 1 (64×64)	> 100	31	33	24
Problem 1 (128×128)	> 100	> 100	37	27
Problem 2 (32×32)	81	16	15	12
Problem 2 (64×64)	> 100	29	28	22
Problem 2 (128×128)	> 100	67	39	32
Parc	92	–	27	22
Spiral	74	–	18	14

TABLE 6.3
Convergence rates for anisotropic problems on different grids. GS, Gauss–Seidel with natural ordering; BJ, Block Jacobi. See text for the definitions of SAI$(3,\epsilon)$ *and SAI*$(4,\epsilon)$. $\epsilon = 0.0008$ *for the regular grid problems and* 0.0004 *for the unstructured grid problems.*

Grids	Convergence rate			
	GS	BJ	SAI(3,ϵ)	SAI(4,ϵ)
Problem 1 (32×32)	0.89	0.13	0.52	0.40
Problem 1 (64×64)	0.91	0.62	0.61	0.50
Problem 1 (128×128)	0.92	0.88	0.64	0.53
Problem 2 (32×32)	0.82	0.34	0.32	0.23
Problem 2 (64×64)	0.89	0.57	0.55	0.47
Problem 2 (128×128)	0.91	0.66	0.60	0.54
Parc	0.86	–	0.55	0.47
Spiral	0.79	–	0.37	0.27



(a) parc (b) spiral

FIG. 6.1. *The unstructured grids for Example* 2.

one line relaxation in the y direction. The results are still shown under the column BJ. For unstructured grid problems, blocks defined along the direction of the anisotropy are hard, if not impossible, to determine. Thus we do not test the block Jacobi smoother in this case. On the whole, the convergence of the block Jacobi smoother is

TABLE 6.4

Cost of applying different smoothers relative to Gauss–Seidel. Under each smoother, the first column shows the cost per iteration relative to one GS iteration, and the second column shows the total cost = cost per iteration × total number of iterations.

Grids	GS		BJ		SAI(3,ϵ)		SAI(4,ϵ)	
Problem 1 (32×32)	1	> 100	2	18	1.50	38	1.57	28
Problem 1 (64×64)	1	> 100	2	62	1.65	54	1.68	40
Problem 1 (128×128)	1	> 100	2	> 100	1.72	64	1.74	47
Problem 2 (32×32)	1	81	2	32	1.64	25	2.02	24
Problem 2 (64×64)	1	> 100	2	58	1.72	48	1.92	42
Problem 2 (128×128)	1	> 100	2	> 100	1.76	69	1.86	60
Parc	1	92	–	–	1.58	43	1.90	42
Spiral	1	74	–	–	1.19	21	1.30	18

quite reasonable; but it is essentially a regular grid technique.

The previous (0,1)-level SAI smoother is not very effective in this case. We improve the performance by using higher level SAI smoothers (section 2). For higher level SAI, however, the approximate inverse is much denser. We control the amount of fill-in by dropping small elements. SAI(k, ϵ) denotes the $(k, k+1)$-level SAI smoother where the elements in the approximate inverse whose absolute values are below ϵ are dropped. We remark that we may also drop small elements in the matrix A before we compute the approximate inverse. Our experience indicates that the difference between the two dropping strategies is small and hence we consider only the former approach.

Tables 6.2 and 6.3 show that the higher level SAI smoothers perform essentially the same as the block Jacobi smoother for small grids and better for larger grids. More importantly, in contrast to line relaxation smoothers, the construction of the SAI smoothers does not require any information about the lines of different isotropy. Hence the exact same construction procedure can be applied to both Problems 1 and 2. The higher level SAI allows us to capture the anisotropy easily by adjusting only one number, the level of fill-in, with no need to track geometrically the anisotropic directions, which are often hard to determine. Furthermore, it does not matter whether it is a regular grid or irregular grid problem. For the two unstructured grid problems, the SAI smoothers converge much faster than the Gauss–Seidel smoother. We remark that the apparent increase in the number of multigrid iterations with respect to the mesh size is probably due to the fact that the multigrid preconditioner used in each grid size problem is slightly different since the entries dropped in constructing the SAI smoothers are different. Mesh size independent convergence should be observed if no dropping strategy is used.

Table 6.4 shows the costs of different smoothers relative to Gauss–Seidel. Under each smoother, the first column shows the cost per iteration of the smoother relative to one Gauss–Seidel iteration; the second column shows the total cost = cost per iteration × total number of iterations. The cost of each smoother is estimated as follows. The cost of one Gauss–Seidel iteration is estimated as the number of nonzeros in the matrix A. Thus, the relative cost of the SAI smoothers is estimated as the ratio of the number of nonzeros in the sparse approximate inverse to the number of nonzeros in the matrix A. The cost of inverting the diagonal block is about $8n$, assuming each block is a symmetric positive definite tridiagonal matrix [16]. Thus the cost of the block Jacobi smoother is about twice that of Gauss–Seidel.

From the table, it is clear that the higher cost of higher level SAI smoothers re-

sults in a lower total cost due to the much better improvement in the total number of iterations. Hence it is worth the effort to use higher level SAI smoothers for difficult problems such as anisotropic coefficient PDEs.

We note that we may adjust/improve the SAI smoothers locally according to the local behavior of the PDE coefficients. For instance, we may use a higher level SAI approximation on the subdomains with anisotropic PDE coefficients and a simple SAI approximation on the subdomains with isotropic coefficients. This local adaptivity feature is another advantage over methods based on incomplete factorizations. However, for Example 2, the coefficient is anisotropic throughout the entire domain, and higher level SAI approximation is needed everywhere.

Acknowledgments. Both authors would like to thank Tony Chan for his comments and discussions on the development of the sparse approximate inverse smoothers, and Li-Tien Cheng for his careful proofreading. The second author would also like to acknowledge the support and hospitality of the Computer Science Department of the University of Waterloo; this work was initiated during a visit there.

REFERENCES

[1] R. E. Bank and C. C. Douglas, *Sharp estimates for multigrid rates of convergence with general smoothing and acceleration*, SIAM J. Numer. Anal., 22 (1985), pp. 617–633.

[2] M. W. Benson, *Iterative Solution of Large Scale Linear Systems*, M. Sc. Thesis, Lakehead University, Thunder Bay, ON, Canada, 1973.

[3] M. W. Benson and P. O. Frederickson, *Iterative solution of large scale linear systems arising in certain multidimensional approximation problems*, Utilitas Math., 22 (1982), pp. 127–120.

[4] M. Benzi, C. D. Meyer, and M. Tůma, *A sparse approximate inverse preconditioner for the conjugate gradient method*, SIAM J. Sci. Comput., 17 (1996), pp. 1135–1149.

[5] A. Brandt, *Multi-level adaptive solutions to boundary-value problems*, Math. Comp., 31 (1977), pp. 333–390.

[6] R. Bridson and W.-P. Tang, *Ordering, anisotropy, and factored sparse approximate inverses*, SIAM J. Sci. Comput., 21 (1999), pp. 867–882.

[7] R. H. Chan, T. F. Chan, and W. L. Wan, *Multigrid for Differential-Convolution Problems Arising from Image Processing*, CAM Report 97-20, UCLA, Los Angeles, CA, 1997.

[8] T. Chan, W. -P. Tang, and W. L. Wan, *Wavelet sparse approximate inverse preconditioners,* BIT, 37 (1997), pp. 644–650.

[9] E. Chow and Y. Saad, *Approximate inverse techniques for block-partitioned matrices*, SIAM J. Sci. Comput., 18 (1997), pp. 1657–1675.

[10] S. S. Clift and W.-P. Tang, *Weighted graph based ordering techniques for preconditioned conjugate gradient methods*, BIT, 35 (1995), pp. 30–47.

[11] J. Cosgrove, J. Diaz, and A. Griewank, *Approximate inverse preconditionings for sparse linear systems*, Internat. J. Comput. Math., 44 (1992), pp. 91–110.

[12] D. F. D'Azevedo, P. A. Forsyth, and W.-P. Tang, *Towards a cost effective ILU preconditioner with high level fill*, BIT, 32 (1992), pp. 442–463.

[13] G. E. Forsythe and W. R. Wasow, *Finite Difference Methods for Partial Differential Equations,* John Wiley, New York, 1960.

[14] A. George and J. Liu, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1981.

[15] L. Collatz, *The Numerical Treatment of Differential Equations*, 3rd ed., Springer-Verlag, New York, 1966.

[16] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1989.

[17] N. I. M. Gould and J. A. Scott, *On Approximate-Inverse Preconditioners*, Tech. Report RAL-TR-95-026, Computing and Information System Department, Atlas Center, Rutherford Appleton Lab., Oxfordshire UK, June, 1995.

[18] M. Grote and H. Simon, *Parallel preconditioning and approximate inverse on the connection machine*, in Proceedings of the Scalable High Performance Computing Conference (SHPCC), Williamsburg, VA, IEEE Computer Science Press, Los Alamitos, CA, 1992, pp. 76–83.

[19] M. Grote and H. D. Simon, *Parallel preconditioning and approximate inverses on the connection machine*, in Proceedings of the Sixth SIAM Conference on Parallel Processing for Scientific Computing II, R. F. Sincovec, D. E. Keyey, M. R. Leuze, L. R. Petzold, and D. A. Reed, eds., SIAM, Philadelphia, 1993, pp. 519–523.

[20] M. J. Grote and T. Huckle, *Parallel preconditioning with sparse approximate inverses*, SIAM J. Sci. Comput., 18 (1997), pp. 838–853.

[21] I. Gustafsson, *A class of first order factorization methods*, BIT, 18 (1978), pp. 142–156.

[22] W. Hackbusch, *Multi-Grid Methods and Applications*, Springer-Verlag, Berlin, 1985.

[23] P. W. Hemker, *The incomplete LU-decomposition as a relaxation method in multi-grid algorithms*, in Boundary and Interior Layers—Computational and Asymptotic Methods, J. H. Miller, ed., Boole Press, Dublin, 1980, pp. 306–311.

[24] T. Huckle and M. Grote, *A New Approach to Parallel Preconditioning with Sparse Approximate Inverses*, SCCM Report, Stanford University, Stanford, CA, 1994.

[25] T. Huckle, *Sparse approximate inverses and multigrid methods*, in Proceedings of the Sixth SIAM Conference on Applied Linear Algebra, Snowbird, UT, 1997.

[26] A. Jameson, *Solution of the Euler equation for two-dimensional flow by a multigrid method*, Appl. Math. Comput., 13 (1983), pp. 327–355.

[27] I. E. Kaporin, *New convergence results and preconditioning strategies for the conjugate gradient method*, Numer. Linear Algebra Appl., 1 (1994), pp. 179–210.

[28] L. Y. Kolotilina , *Explicit preconditioning of systems of linear algebraic equations with dense matrices*, J. Soviet Math., 43 (1988), pp. 2566–2573.

[29] L. Kolotilina, A. Nikishin, and A. Yeremin, *Factorized sparse approximate inverse (FSAI) preconditionings for solving* 3*D FE systems on massively parallel computers.* II, in Proceedings of the IMACS International Symposium, Iterative Methods in Linear Algebra, Brussels, R. Beauwens and P. Groen, eds., North-Holland, Amsterdam, 1992, pp. 311–312.

[30] L. Yu. Kolotilina and A. Yu. Yeremin, *Factorized sparse approximate inverse preconditionings* I. *Theory*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 45–58.

[31] J. Lifshitz, A. Nikishin, and A. Yeremin, *Sparse approximate inverse (FSAI) preconditionings for solving* 3*D CFD problems on massively parallel computers*, in Proceedings of the IMACS International Symposium, Iterative Methods in Linear Algebra, Brussels, R. Beauwens and P. Groen, eds., North-Holland, Amsterdam, 1992, pp. 83–84.

[32] G. Meurant, *A review on the inverse of symmetric tridiagonal and block tridiagonal matrices*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 707–728.

[33] J. Rice and R. Boisvert, *Solving Elliptic Problems Using ELLPACK*, Springer-Verlag, New York, 1985.

[34] U. Schumann and R. A. Sweet, *Fast Fourier transforms for direct solution of Poisson's equation with staggered boundary conditions*, J. Comput. Phys., 75 (1988), pp. 123–137.

[35] W.-P. Tang, *Schwarz Splitting and Template Operators*, Ph.D. thesis, Stanford University, Stanford, CA, 1987.

[36] W.-P. Tang, *Template Operators and Exponential Decay,* Tech. Report CS-90-46, University of Waterloo, ON, Canada, 1990.

[37] W.-P. Tang, *Toward an effective sparse approximate inverse preconditioner,* SIAM J. Matrix Anal. Appl., 20 (1999), pp. 970–986.

[38] W. L. Wan, *An Energy-Minimizing Interpolation for Multigrid*, CAM Report 97-18, Department of Mathematics, UCLA, Los Angeles, CA, 1997.

[39] W. L. Wan, T. F. Chan, and B. Smith, *An Energy-Minimizing Interpolation for Robust Multigrid*, CAM Report 98-6, UCLA, Los Angeles, CA, 1998.

[40] P. Wesseling, *An Introduction to Multigrid Methods*, John Wiley, New York, 1992.

[41] G. Wittum, *On the robustness of ILU smoothing,* SIAM J. Sci. Comput., 10 (1989), pp. 699–717.

[42] D. Young, *Iterative Solution of Large Linear Systems*, Academic Press, New York, 1971.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 4, pp. 1253–1278
© 2000 Society for Industrial and Applied Mathematics

A MULTILINEAR SINGULAR VALUE DECOMPOSITION*

LIEVEN DE LATHAUWER†, BART DE MOOR†, AND JOOS VANDEWALLE†

Abstract. We discuss a multilinear generalization of the singular value decomposition. There is a strong analogy between several properties of the matrix and the higher-order tensor decomposition; uniqueness, link with the matrix eigenvalue decomposition, first-order perturbation effects, etc., are analyzed. We investigate how tensor symmetries affect the decomposition and propose a multilinear generalization of the symmetric eigenvalue decomposition for pair-wise symmetric tensors.

Key words. multilinear algebra, singular value decomposition, higher-order tensor

AMS subject classifications. 15A18, 15A69

PII. S0895479896305696

1. Introduction. An increasing number of signal processing problems involve the manipulation of quantities of which the elements are addressed by more than two indices. In the literature these higher-order equivalents of vectors (first order) and matrices (second order) are called higher-order tensors, multidimensional matrices, or multiway arrays. For a lot of applications involving higher-order tensors, the existing framework of vector and matrix algebra appears to be insufficient and/or inappropriate. In this paper we present a proper generalization of the singular value decomposition (SVD).

To a large extent, higher-order tensors are currently gaining importance due to the developments in the field of higher-order statistics (HOS): for multivariate stochastic variables the basic HOS (higher-order moments, cumulants, spectra and cepstra) are highly symmetric higher-order tensors, in the same way as, e.g., the covariance of a stochastic vector is a symmetric (Hermitean) matrix. A brief enumeration of some opportunities offered by HOS gives an idea of the promising role of higher-order tensors on the signal processing scene. It is clear that statistical descriptions of random processes are more accurate when, in addition to first- and second-order statistics, they take HOS into account; from a mathematical point of view this is reflected by the fact that moments and cumulants correspond, within multiplication with a fixed scalar, to the subsequent coefficients of a Taylor series expansion of the first, resp., second, characteristic function of the stochastic variable. In statistical nonlinear system theory HOS are even unavoidable (e.g., the autocorrelation of x^2 is a fourth-order moment).

*Received by the editors June 24, 1996; accepted for publication (in revised form) by B. Kagstrom January 4, 1999; published electronically April 18, 2000. This research was partially supported by the Flemish Government (the Concerted Research Action GOA-MIPS, the FWO (Fund for Scientific Research - Flanders) projects G.0292.95, G.0262.97, and G.0240.99, the FWO Research Communities ICCoS (Identification and Control of Complex Systems) and Advanced Numerical Methods for Mathematical Modelling, the IWT Action Programme on Information Technology (ITA/GBO/T23)-IT-Generic Basic Research), the Belgian State, Prime Minister's Office - Federal Office for Scientific, Technical and Cultural Affairs (the Interuniversity Poles of Attraction Programmes IUAP P4-02 and IUAP P4-24) and the European Commission (Human Capital and Mobility Network SIMONET, SC1-CT92-0779). The first author is a post-doctoral Research Assistant with the F.W.O. The second author is a senior Research Associate with the F.W.O. and an Associate Professor at the K.U. Leuven. The third author is a Full Professor at the K.U. Leuven. The scientific responsibility is assumed by the authors.

http://www.siam.org/journals/simax/21-4/30569.html

†K.U. Leuven, E.E. Dept., ESAT–SISTA/COSIC, Kard. Mercierlaan 94, B-3001 Leuven (Heverlee), Belgium (Lieven.DeLathauwer@esat.kuleuven.ac.be, Bart.DeMoor@esat.kuleuven.ac.be, Joos.Vandewalle@esat.kuleuven.ac.be).

Moreover, higher-order cumulants and spectra of a random variable are insensitive to an additive Gaussian perturbation of that variable, which is exploited in higher-order-only techniques, conceptually blind for Gaussian noise. HOS make it possible to solve the source separation (SS) problem by mere exploitation of the statistical independence of the sources, without knowledge of the array manifold; they also contain sufficient information to allow a blind identification of linear filters, without making assumptions on the minimum/nonminimum phase character, etc. (for general aspects of HOS, the interested reader is referred to the tutorial papers [34, 37, 38, 30] and the bibliography [41]; for the use of tensor decompositions as a basic tool in HOS-based signal processing, we refer to [11, 12, 9, 19]).

Higher-order tensors do not merely play an important role in HOS. As a matter of fact they seem to be used in the most various disciplines, like chemometrics, psychometrics, econometrics, image processing, biomedical signal processing, etc. Most often they appear in factor analysis-type problems of multiway datasets [13]. Another, more trivial, use is as a formalism to describe linear vector-matrix, matrix-vector, matrix-matrix, ... transformations, in the same way as matrices describe linear transformations between vector spaces. Interesting also is the fact that higher-order terms in the Taylor series expansion of a multivariate function and higher-order Volterra filter kernels have a tensor form.

On the other hand, one of the most fruitful developments in the world of linear algebra and linear algebra-based signal processing is the concept of the SVD of matrices [21, 35, 44]. In this paper we will discuss a multilinear generalization that has also been investigated in psychometrics as the Tucker model, originally developed to obtain a "method for searching for relations in a body of data," for the case "each person in a group of individuals is measured on each trait in a group of traits by each of a number of methods," or "when individuals are measured by a battery of measures on a number of occasions" [42, 43]. For three-way data, the Tucker model consists of decomposing a real $(I_1 \times I_2 \times I_3)$-tensor $\mathcal{A}$ according to

$$a_{i_1 i_2 i_3} = \sum_{j_1}^{I_1} \sum_{j_2}^{I_2} \sum_{j_3}^{I_3} s_{j_1 j_2 j_3} u^{(1)}_{i_1 j_1} u^{(2)}_{i_2 j_2} u^{(3)}_{i_3 j_3}, \tag{1}$$

in which $u^{(1)}_{i_1 j_1}, u^{(2)}_{i_2 j_2}, u^{(3)}_{i_3 j_3}$ are entries of orthogonal matrices, and $\mathcal{S}$ is a real $(I_1 \times I_2 \times I_3)$-tensor with the property of "all-orthogonality," i.e., $\sum_{i_1 i_2} s_{i_1 i_2 \alpha} s_{i_1 i_2 \beta} = \sum_{i_1 i_3} s_{i_1 \alpha i_3} s_{i_1 \beta i_3} = \sum_{i_2 i_3} s_{\alpha i_2 i_3} s_{\beta i_2 i_3} = 0$ whenever $\alpha \neq \beta$. This decomposition is virtually unknown in the communities of numerical algebra and signal processing; on the other hand, the viewpoint and language in psychometrics is somewhat different from what is common in our field. It is the aim of this paper to derive the tensor decomposition in an SVD terminology, using a notation that is a natural extension of the notation used for matrices. As already mentioned, we will show that the Tucker model is actually a convincing multilinear generalization of the SVD concept itself. From our algebraic point of view, we will ask a number of inevitable questions such as what the geometric link is between the generalized singular vectors/values and the column, row, ... vectors (oriented energy), how the concept of rank lies to the structure of the decomposition, whether the best reduced-rank approximation property carries over, and so on. Our derivations are valid for tensors of arbitrary order and hold for the complex-valued case too. In view of the many striking analogies between the matrix SVD and its multilinear generalization, we use the term *higher-order singular value decomposition* (HOSVD) in this paper; note at this point that the existence of

different multilinear SVD extensions may not be excluded—as a matter of fact, focusing on different properties of the matrix SVD does lead to the definition of different (formally less striking) multilinear generalizations, as we will explain later on.

In our own research the HOSVD has already proved its value. In [14] we showed that the decomposition is fundamentally related to the problem of blind source separation, also known as independent component analysis (ICA). In [18] we used the decomposition to compute an initial value for a tensorial equivalent of the power method, aiming at the computation of the best rank-1 approximation of a given tensor; a high-performant ICA-algorithm was based on this technique. In [19] the HOSVD was used in a dimensionality reduction for higher-order factor analysis-type problems, reducing the computational complexity. The current paper however is the first systematic, elaborated presentation of the HOSVD concept.

The paper is organized as follows. In section 2 we introduce some preliminary material on multilinear algebra, needed for the further developments. The HOSVD model is presented in section 3 and compared to its matrix counterpart. In section 4 we discuss some well-known SVD properties and demonstrate that they have striking higher-order counterparts. In section 5 we investigate how the HOSVD is influenced by symmetry properties of the higher-order tensor; in analogy with the eigenvalue decomposition (EVD) of symmetric (Hermitean) matrices we define a *higher-order eigenvalue decomposition* (HOEVD) for "pair-wise symmetric" higher-order tensors. Section 6 contains a first-order perturbation analysis and section 7 quotes some alternative ways to generalize the SVD.

Before starting with the next section, we would like to add a comment on the notation that is used. To facilitate the distinction between scalars, vectors, matrices, and higher-order tensors, the type of a given quantity will be reflected by its representation: scalars are denoted by lower-case letters (a, b, ...; α, β, ...), vectors are written as capitals (A, B, ...) (italic shaped), matrices correspond to bold-face capitals ($\mathbf{A}$, $\mathbf{B}$, ...), and tensors are written as calligraphic letters ($\mathcal{A}$, $\mathcal{B}$, ...). This notation is consistently used for lower-order parts of a given structure. For example, the entry with row index i and column index j in a matrix $\mathbf{A}$, i.e., $(\mathbf{A})_{ij}$, is symbolized by a_{ij} (also $(A)_i = a_i$ and $(\mathcal{A})_{i_1 i_2 \ldots i_N} = a_{i_1 i_2 \ldots i_N}$); furthermore, the ith column vector of a matrix $\mathbf{A}$ is denoted as A_i, i.e., $\mathbf{A} = (A_1 A_2 \ldots)$. To enhance the overall readability, we have made one exception to this rule: as we frequently use the characters i, j, r, and n in the meaning of indices (counters), I, J, R, and N will be reserved to denote the index upper bounds, unless stated otherwise.

2. Basic definitions.

2.1. Matrix representation of a higher-order tensor. The starting point of our derivation of a multilinear SVD will be to consider an appropriate generalization of the link between the column (row) vectors and the left (right) singular vectors of a matrix. To be able to formalize this idea, we define "matrix unfoldings" of a given tensor, i.e., matrix representations of that tensor in which all the column (row, ...) vectors are stacked one after the other. To avoid confusion, we will stick to one particular ordering of the column (row, ...) vectors; for order three, these unfolding procedures are visualized in Figure 1. Notice that the definitions of the matrix unfoldings involve the tensor dimensions I_1, I_2, I_3 in a cyclic way and that, when dealing with an unfolding of dimensionality $I_c \times I_a I_b$, we formally assume that the index i_a varies more slowly than i_b. In general, we have the following definition.

DEFINITION 1. *Assume an Nth-order tensor* $\mathcal{A} \in \mathbb{C}^{I_1 \times I_2 \times \ldots \times I_N}$. *The matrix unfolding* $\mathbf{A}_{(n)} \in \mathbb{C}^{I_n \times (I_{n+1} I_{n+2} \ldots I_N I_1 I_2 \ldots I_{n-1})}$ *contains the element* $a_{i_1 i_2 \ldots i_N}$ *at the*

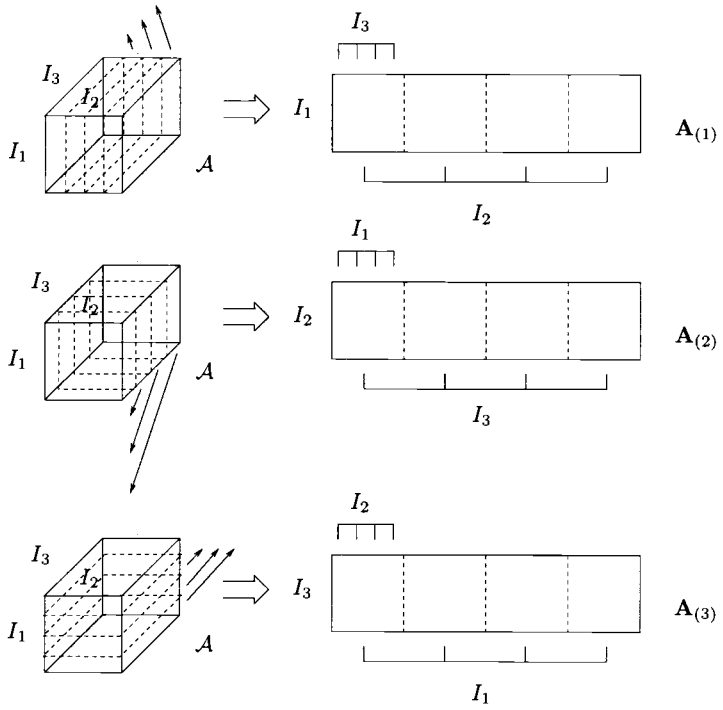


FIG. 1. *Unfolding of the* $(I_1 \times I_2 \times I_3)$*-tensor* $\mathcal{A}$ *to the* $(I_1 \times I_2 I_3)$*-matrix* $\mathbf{A}_{(1)}$*, the* $(I_2 \times I_3 I_1)$*-matrix* $\mathbf{A}_{(2)}$ *and the* $(I_3 \times I_1 I_2)$*-matrix* $\mathbf{A}_{(3)}$ $(I_1 = I_2 = I_3 = 4)$.

position with row number i_n *and column number equal to*

$$(i_{n+1}-1)I_{n+2}I_{n+3}\dots I_N I_1 I_2 \dots I_{n-1} + (i_{n+2}-1)I_{n+3}I_{n+4}\dots I_N I_1 I_2 \dots I_{n-1} + \cdots + (i_N - 1)I_1 I_2 \dots I_{n-1} + (i_1 - 1)I_2 I_3 \dots I_{n-1} + (i_2 - 1)I_3 I_4 \dots I_{n-1} + \cdots + i_{n-1}.$$

Example 1. *Define a tensor* $\mathcal{A} \in \mathbb{R}^{3\times 2\times 3}$ *by* $a_{111} = a_{112} = a_{211} = -a_{212} = 1$, $a_{213} = a_{311} = a_{313} = a_{121} = a_{122} = a_{221} = -a_{222} = 2$, $a_{223} = a_{321} = a_{323} = 4$, $a_{113} = a_{312} = a_{123} = a_{322} = 0$. *The matrix unfolding* $\mathbf{A}_{(1)}$ *is given by*

$$\mathbf{A}_{(1)} = \left(\begin{array}{ccc|ccc} 1 & 1 & 0 & 2 & 2 & 0 \\ 1 & -1 & 2 & 2 & -2 & 4 \\ 2 & 0 & 2 & 4 & 0 & 4 \end{array}\right).$$

2.2. Rank properties of a higher-order tensor. There are major differences between matrices and higher-order tensors when rank properties are concerned. As we will explain in section 3, these differences directly affect the way an SVD generalization could look. As a matter of fact, there is not a unique way to generalize the rank concept.

First, it is easy to generalize the notion of column and row rank. If we refer in general to the column, row, ... vectors of an Nth-order tensor $\mathcal{A} \in \mathbb{C}^{I_1 \times I_2 \times \dots \times I_N}$ as its "n-mode vectors," defined as the I_n-dimensional vectors obtained from $\mathcal{A}$ by varying the index i_n and keeping the other indices fixed, then we have the following definition.

DEFINITION 2. *The n-rank of $\mathcal{A}$, denoted by $R_n = \text{rank}_\text{n}(\mathcal{A})$, is the dimension of the vector space spanned by the n-mode vectors.*

The n-rank of a given tensor can be analyzed by means of matrix techniques.

PROPERTY 1. *The n-mode vectors of $\mathcal{A}$ are the column vectors of the matrix unfolding $\mathbf{A}_{(n)}$ and*

$$\text{rank}_\text{n}(\mathcal{A}) = \text{rank}(\mathbf{A}_{(n)}).$$

A major difference with the matrix case, however, is the fact that the different n-ranks of a higher-order tensor are not necessarily the same, as can easily be verified by checking some examples (see further).

The rank of a higher-order tensor is usually defined in analogy with the fact that a rank-R matrix can be decomposed as a sum of R rank-1 terms [12, 29].

DEFINITION 3. *An Nth-order tensor $\mathcal{A}$ has rank 1 when it equals the outer product of N vectors $U^{(1)}$, $U^{(2)}$, ..., $U^{(N)}$, i.e.,*

$$a_{i_1 i_2 \ldots i_N} = u^{(1)}_{i_1} u^{(2)}_{i_2} \ldots u^{(N)}_{i_N},$$

for all values of the indices.

DEFINITION 4. *The rank of an arbitrary Nth-order tensor $\mathcal{A}$, denoted by $R = \text{rank}(\mathcal{A})$, is the minimal number of rank-1 tensors that yield $\mathcal{A}$ in a linear combination.*

(With respect to the definition of a rank-1 tensor, a remark on the notation has to be made. For matrices, the vector-vector outer product of $U^{(1)}$ and $U^{(2)}$ is denoted as $U^{(1)} \cdot U^{(2)^T}$; to avoid an ad hoc notation based on "generalized transposes" (in which the fact that column vectors are transpose-free would not have an inherent meaning), we will further denote the outer product of $U^{(1)}$, $U^{(2)}$, ..., $U^{(N)}$ by $U^{(1)} \circ U^{(2)} \circ \cdots \circ U^{(N)}$.)

A second difference between matrices and higher-order tensors is the fact that the rank is not necessarily equal to an n-rank, even when all the n-ranks are the same. From the definitions it is clear that always $R_n \leqslant R$.

Example 2. *Consider the $(2 \times 2 \times 2)$-tensor $\mathcal{A}$ defined by*

$$\begin{cases} a_{111} = a_{221} = a_{112} = 1 \\ a_{211} = a_{121} = a_{212} = a_{122} = a_{222} = 0. \end{cases}$$

It follows that $R_1 = R_2 = 2$ but $R_3 = 1$.

Example 3. *Consider the $(2 \times 2 \times 2)$-tensor $\mathcal{A}$ defined by*

$$\begin{cases} a_{211} = a_{121} = a_{112} = 1 \\ a_{111} = a_{222} = a_{122} = a_{212} = a_{221} = 0. \end{cases}$$

The 1-rank, 2-rank, and 3-rank are equal to 2. The rank, however, equals 3, since

$$\mathcal{A} = X_2 \circ Y_1 \circ Z_1 + X_1 \circ Y_2 \circ Z_1 + X_1 \circ Y_1 \circ Z_2,$$

in which

$$X_1 = Y_1 = Z_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad X_2 = Y_2 = Z_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

is a decomposition in a minimal linear combination of rank-1 tensors (a proof is given in [19]*).*

2.3. Scalar product, orthogonality, norm of higher-order tensors. In the HOSVD definition of section 3, the structure constraint of diagonality of the matrix of singular values in the second-order case will be replaced by a number of geometrical conditions. This requires a generalization of the well-known definitions of scalar product, orthogonality, and Frobenius-norm to tensors of arbitrary order. These generalizations can be defined in a straightforward way as follows.

DEFINITION 5. *The scalar product* $\langle \mathcal{A}, \mathcal{B} \rangle$ *of two tensors* $\mathcal{A}, \mathcal{B} \in \mathbb{C}^{I_1 \times I_2 \times \ldots \times I_N}$ *is defined as*

$$\langle \mathcal{A}, \mathcal{B} \rangle \stackrel{\text{def}}{=} \sum_{i_1} \sum_{i_2} \cdots \sum_{i_N} b^*_{i_1 i_2 \ldots i_N} a_{i_1 i_2 \ldots i_N},$$

in which $*$ *denotes the complex conjugation.*

DEFINITION 6. *Arrays of which the scalar product equals* 0 *are orthogonal.*

DEFINITION 7. *The Frobenius-norm of a tensor* $\mathcal{A}$ *is given by*

$$\|\mathcal{A}\| \stackrel{\text{def}}{=} \sqrt{\langle \mathcal{A}, \mathcal{A} \rangle}.$$

2.4. Multiplication of a higher-order tensor by a matrix. Like for matrices, the HOSVD of a higher-order tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ will be defined by looking for orthogonal coordinate transformations of $\mathbb{R}^{I_1}, \mathbb{R}^{I_2}, \ldots, \mathbb{R}^{I_N}$ that induce a particular representation of the higher-order tensor. In this section we establish a notation for the multiplication of a higher-order tensor by a matrix. This will allow us to express the effect of basis transformations.

Let us first have a look at the matrix product $\mathbf{G} = \mathbf{U} \cdot \mathbf{F} \cdot \mathbf{V}^T$, involving matrices $\mathbf{F} \in \mathbb{R}^{I_1 \times I_2}$, $\mathbf{U} \in \mathbb{R}^{J_1 \times I_1}$, $\mathbf{V} \in \mathbb{R}^{J_2 \times I_2}$, and $\mathbf{G} \in \mathbb{R}^{J_1 \times J_2}$. To avoid working with "generalized transposes" in the multilinear case, we observe that the relationship between $\mathbf{U}$ and $\mathbf{F}$ and the relationship between $\mathbf{V}$ (not $\mathbf{V}^T$) and $\mathbf{F}$ are in fact completely similar: in the same way as $\mathbf{U}$ makes linear combinations of the rows of $\mathbf{F}$, $\mathbf{V}$ makes linear combinations of the columns of $\mathbf{F}$; in the same way as the columns of $\mathbf{F}$ are multiplied by $\mathbf{U}$, the rows of $\mathbf{F}$ are multiplied by $\mathbf{V}$; in the same way as the columns of $\mathbf{U}$ are associated with the column space of $\mathbf{G}$, the columns of $\mathbf{V}$ are associated with the row space of $\mathbf{G}$. This typical relationship will be denoted by means of the $\times_n$-symbol: $\mathbf{G} = \mathbf{F} \times_1 \mathbf{U} \times_2 \mathbf{V}$. (For complex matrices the product $\mathbf{U} \cdot \mathbf{F} \cdot \mathbf{V}^H$ is consequently denoted as $\mathbf{F} \times_1 \mathbf{U} \times_2 \mathbf{V}^*$.) In general, we have the following definition.

DEFINITION 8. *The* n*-mode product of a tensor* $\mathcal{A} \in \mathbb{C}^{I_1 \times I_2 \times \ldots \times I_N}$ *by a matrix* $\mathbf{U} \in \mathbb{C}^{J_n \times I_n}$*, denoted by* $\mathcal{A} \times_n \mathbf{U}$*, is an* $(I_1 \times I_2 \times \cdots \times I_{n-1} \times J_n \times I_{n+1} \cdots \times I_N)$*-tensor of which the entries are given by*

$$(\mathcal{A} \times_n \mathbf{U})_{i_1 i_2 \ldots i_{n-1} j_n i_{n+1} \ldots i_N} \stackrel{\text{def}}{=} \sum_{i_n} a_{i_1 i_2 \ldots i_{n-1} i_n i_{n+1} \ldots i_N} u_{j_n i_n}.$$

The n-mode product satisfies the following properties.

PROPERTY 2. *Given the tensor* $\mathcal{A} \in \mathbb{C}^{I_1 \times I_2 \times \ldots \times I_N}$ *and the matrices* $\mathbf{F} \in \mathbb{C}^{J_n \times I_n}$, $\mathbf{G} \in \mathbb{C}^{J_m \times I_m}$ $(n \neq m)$*, one has*

$$(\mathcal{A} \times_n \mathbf{F}) \times_m \mathbf{G} = (\mathcal{A} \times_m \mathbf{G}) \times_n \mathbf{F} = \mathcal{A} \times_n \mathbf{F} \times_m \mathbf{G}.$$

PROPERTY 3. *Given the tensor* $\mathcal{A} \in \mathbb{C}^{I_1 \times I_2 \times \ldots \times I_N}$ *and the matrices* $\mathbf{F} \in \mathbb{C}^{J_n \times I_n}$, $\mathbf{G} \in \mathbb{C}^{K_n \times J_n}$*, one has*

$$(\mathcal{A} \times_n \mathbf{F}) \times_n \mathbf{G} = \mathcal{A} \times_n (\mathbf{G} \cdot \mathbf{F}).$$

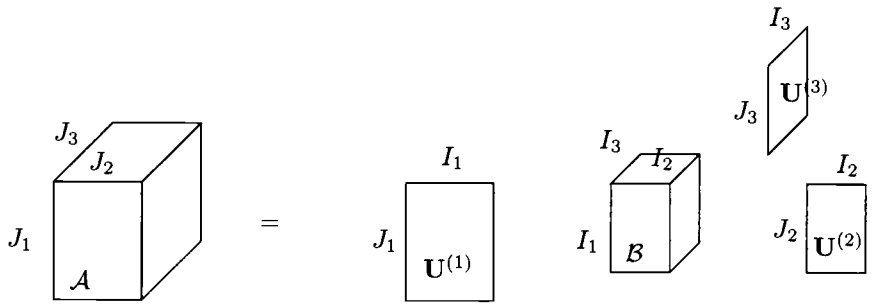


FIG. 2. *Visualization of the multiplication of a third-order tensor* $\mathcal{B} \in \mathbb{C}^{I_1 \times I_2 \times I_3}$ *with matrices* $\mathbf{U}^{(1)} \in \mathbb{C}^{J_1 \times I_1}$, $\mathbf{U}^{(2)} \in \mathbb{C}^{J_2 \times I_2}$, *and* $\mathbf{U}^{(3)} \in \mathbb{C}^{J_3 \times I_3}$.

Figure 2 visualizes the equation $\mathcal{A} = \mathcal{B} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \mathbf{U}^{(3)}$ for third-order tensors $\mathcal{A} \in \mathbb{C}^{J_1 \times J_2 \times J_3}$ and $\mathcal{B} \in \mathbb{C}^{I_1 \times I_2 \times I_3}$. Unlike the conventional way to visualize second-order matrix products, $\mathbf{U}^{(2)}$ has not been transposed, for reasons of symmetry. Multiplication with $\mathbf{U}^{(1)}$ involves linear combinations of the "horizontal matrices" (index i_1 fixed) in $\mathcal{B}$. Stated otherwise, multiplication of $\mathcal{B}$ with $\mathbf{U}^{(1)}$ means that every column of $\mathcal{B}$ (indices i_2 and i_3 fixed) has to be multiplied from the left with $\mathbf{U}^{(1)}$. Similarly, multiplication with $\mathbf{U}^{(2)}$, resp., $\mathbf{U}^{(3)}$, involves linear combinations of matrices, obtained by fixing i_2, resp., i_3. This can be considered as a multiplication, from the left, of the vectors obtained by fixing the indices i_3 and i_1, resp., i_1 and i_2. Visualization schemes like Figure 2 have proven to be very useful to gain insight in tensor techniques.

The n-mode product of a tensor and a matrix is a special case of the inner product in multilinear algebra and tensor analysis [32, 26]. In the literature it often takes the form of an Einstein summation convention. Without going into details, this means that summations are written in full, but that the summation sign is dropped for the index that is repeated. This is of course the most versatile way to write down tensor equations, and in addition, a basis-independent interpretation can be given to Einstein summations. On the other hand, the formalism is only rarely used in signal processing and numerical linear algebra, whereas using the $\times_n$-symbol comes closer to the common way of dealing with matrix equations. It is our experience that the use of the $\times_n$-symbol demonstrates more clearly the analogy between matrix and tensor SVD, and that, more in general, a conceptual insight in tensor decompositions is easier induced by means of the $\times_n$-notation and visualizations like Figure 2 than by the use of element-wise summations.

3. A multilinear SVD. In this section an SVD model is proposed for Nth-order tensors. To facilitate the comparison, we first repeat the matrix decomposition in the same notation, as follows.

THEOREM 1 (matrix SVD). *Every complex* $(I_1 \times I_2)$*-matrix* $\mathbf{F}$ *can be written as the product*

$$\mathbf{F} = \mathbf{U}^{(1)} \cdot \mathbf{S} \cdot \mathbf{V}^{(2)^H} = \mathbf{S} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{V}^{(2)^*} = \mathbf{S} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)}, \tag{2}$$

in which

1. $\mathbf{U}^{(1)} = \left(U_1^{(1)} U_2^{(1)} \ldots U_{I_1}^{(1)}\right)$ *is a unitary* $(I_1 \times I_1)$*-matrix,*

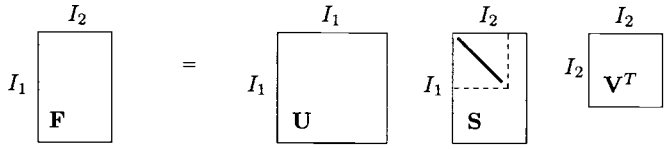


FIG. 3. *Visualization of the matrix SVD.*

2. $\mathbf{U}^{(2)} = \left(U_1^{(2)} U_2^{(2)} \dots U_{I_2}^{(2)}\right) (= \mathbf{V}^{(2)^*})$ *is a unitary* $(I_2 \times I_2)$*-matrix,*

3. $\mathbf{S}$ *is an* $(I_1 \times I_2)$*-matrix with the properties of*

(i) *pseudodiagonality*:

$$\mathbf{S} = \operatorname{diag}(\sigma_1, \sigma_2, \dots, \sigma_{\min(I_1, I_2)}), \tag{3}$$

(ii) *ordering*:

$$\sigma_1 \geqslant \sigma_2 \geqslant \dots \geqslant \sigma_{\min(I_1, I_2)} \geqslant 0. \tag{4}$$

The σ_i *are singular values of* $\mathbf{F}$ *and the vectors* $U_i^{(1)}$ *and* $U_i^{(2)}$ *are, resp., an ith left and an ith right singular vector. The decomposition is visualized in Figure* 3.

Now we state the following theorem.

THEOREM 2 (Nth-order SVD). *Every complex* $(I_1 \times I_2 \times \dots \times I_N)$*-tensor* $\mathcal{A}$ *can be written as the product*

$$\mathcal{A} = \mathcal{S} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \cdots \times_N \mathbf{U}^{(N)}, \tag{5}$$

in which

1. $\mathbf{U}^{(n)} = \left(U_1^{(n)} U_2^{(n)} \dots U_{I_n}^{(n)}\right)$ *is a unitary* $(I_n \times I_n)$*-matrix,*

2. $\mathcal{S}$ *is a complex* $(I_1 \times I_2 \times \dots \times I_N)$*-tensor of which the subtensors* $\mathcal{S}_{i_n=\alpha}$*, obtained by fixing the nth index to* α*, have the properties of*

(i) *all-orthogonality: two subtensors* $\mathcal{S}_{i_n=\alpha}$ *and* $\mathcal{S}_{i_n=\beta}$ *are orthogonal for all possible values of* n*,* α *and* β *subject to* $\alpha \neq \beta$:

$$\langle \mathcal{S}_{i_n=\alpha}, \mathcal{S}_{i_n=\beta} \rangle = 0 \quad \textit{when} \quad \alpha \neq \beta, \tag{6}$$

(ii) *ordering*:

$$\|\mathcal{S}_{i_n=1}\| \geqslant \|\mathcal{S}_{i_n=2}\| \geqslant \dots \geqslant \|\mathcal{S}_{i_n=I_n}\| \geqslant 0 \tag{7}$$

for all possible values of n.

The Frobenius-norms $\|\mathcal{S}_{i_n=i}\|$*, symbolized by* $\sigma_i^{(n)}$*, are* n*-mode singular values of* $\mathcal{A}$ *and the vector* $U_i^{(n)}$ *is an ith* n*-mode singular vector. The decomposition is visualized for third-order tensors in Figure* 4.

Discussion. Applied to a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, Theorem 2 says that it is always possible to find orthogonal transformations of the column, row, and 3-mode space such that $\mathcal{S} = \mathcal{A} \times_1 \mathbf{U}^{(1)^T} \times_2 \mathbf{U}^{(2)^T} \times_3 \mathbf{U}^{(3)^T}$ is all-orthogonal and ordered (the new basis vectors are the columns of $\mathbf{U}^{(1)}$, $\mathbf{U}^{(2)}$, and $\mathbf{U}^{(3)}$). All-orthogonality means that the different "horizontal matrices" of $\mathcal{S}$ (the first index i_1 is kept fixed, while the two other indices, i_2 and i_3, are free) are mutually orthogonal with respect to the scalar product

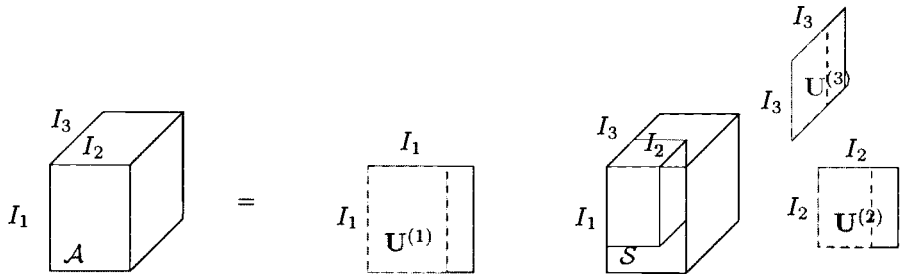


FIG. 4. *Visualization of the HOSVD for a third-order tensor.*

of matrices (i.e., the sum of the products of the corresponding entries vanishes); at the same time, the different "frontal" matrices (i_2 fixed) and the different "vertical" matrices (i_3 fixed) should be mutually orthogonal as well. The ordering constraint imposes that the Frobenius-norm of the horizontal (frontal, resp., vertical) matrices does not increase as the index i_1 (i_2, resp., i_3) is increased. While the orthogonality of $\mathbf{U}^{(1)}$, $\mathbf{U}^{(2)}$, $\mathbf{U}^{(3)}$, and the all-orthogonality of $\mathcal{S}$ are the basic assumptions of the model, the ordering condition should be regarded as a convention, meant to fix a particular ordering of the columns of $\mathbf{U}^{(1)}$, $\mathbf{U}^{(2)}$, and $\mathbf{U}^{(3)}$ (or the horizontal, frontal, and vertical matrices of $\mathcal{S}$, stated otherwise).

Comparison of the matrix and tensor theorem shows a clear analogy between the two cases. First, the left and right singular vectors of a matrix are generalized as the n-mode singular vectors. Next, the role of the singular values is taken over by the Frobenius-norms of the $(N-1)$th-order subtensors of the "core tensor" $\mathcal{S}$; notice at this point that in the matrix case, the singular values also correspond to the Frobenius-norms of the rows and the columns of the "core matrix" $\mathbf{S}$. For Nth-order tensors, N (possibly different) sets of n-mode singular values are defined; in this respect, remember from section 2.2 that an Nth-order tensor can also have N different n-rank values. The essential difference is that $\mathcal{S}$ is in general a full tensor, instead of being pseudodiagonal (this would mean that nonzero elements could only occur when the indices $i_1 = i_2 = \cdots = i_N$). Instead, $\mathcal{S}$ obeys the condition of all-orthogonality; here we notice that in the matrix case $\mathbf{S}$ is all-orthogonal as well: due to the diagonal structure, the scalar product of two different rows or columns also vanishes. We also remark that, by definition, the n-mode singular values are positive and real, like in the matrix case. On the other hand the entries of $\mathcal{S}$ are not necessarily positive in general; they can even be complex, when $\mathcal{A}$ is a complex-valued tensor.

One could wonder whether imposing the condition of pseudodiagonality on the core tensor $\mathcal{S}$ would not be a better way to generalize the SVD of matrices. The answer is negative: in general, it is impossible to reduce higher-order tensors to a pseudodiagonal form by means of orthogonal transformations. This is easily shown by counting degrees of freedom: pseudodiagonality of a core tensor containing I nonzero elements would imply that the decomposition would exhibit not more than $I\,(\sum I_n + 1 - N(I+1)/2)$ degrees of freedom, while the original tensor contains $I_1 I_2 \ldots I_N$ independent entries. Only in the second-order case both quantities are equal for $I = \min(I_1, I_2)$—only in the second-order case, the condition of pseudodiagonality makes sense. However, we will prove that relaxation of the pseudodiagonality condition

to all-orthogonality yields a decomposition that always exists. As a matter of fact, "relaxation" is a too hard term: Property 5 in section 4 will show that the matrix SVD itself could have been defined by requiring the rows/columns of the matrix $\mathbf{S}$ to be mutually orthogonal; apart from some trivial normalization conventions, the resulting decomposition is exactly the same as the conventional one, obtained via the pseudodiagonality condition.

Equivalent representations. A *matrix representation* of the HOSVD can be obtained by unfolding $\mathcal{A}$ and $\mathcal{S}$ in model equation (5):

$$\mathbf{A}_{(n)} = \mathbf{U}^{(n)} \cdot \mathbf{S}_{(n)} \cdot \left(\mathbf{U}^{(n+1)} \otimes \mathbf{U}^{(n+2)} \otimes \cdots \otimes \mathbf{U}^{(N)} \otimes \mathbf{U}^{(1)} \otimes \mathbf{U}^{(2)} \otimes \cdots \otimes \mathbf{U}^{(n-1)}\right)^T, \tag{8}$$

in which $\otimes$ denotes the Kronecker product [2, 40]. (The Kronecker product of two matrices $\mathbf{F} \in \mathbb{C}^{I_1 \times I_2}$ and $\mathbf{G} \in \mathbb{C}^{J_1 \times J_2}$ is defined according to

$$\mathbf{F} \otimes \mathbf{G} \stackrel{\text{def}}{=} (f_{i_1 i_2}\mathbf{G})_{1 \leqslant i_1 \leqslant I_1; 1 \leqslant i_2 \leqslant I_2}.)$$

Notice that the conditions (6) and (7) imply that $\mathbf{S}_{(n)}$ has mutually orthogonal rows, having Frobenius-norms equal to $\sigma_1^{(n)}, \sigma_2^{(n)}, \ldots, \sigma_{I_n}^{(n)}$. Let us define a diagonal matrix $\mathbf{\Sigma}^{(n)} \in \mathbb{R}^{I_n \times I_n}$ and a column-wise orthonormal matrix $\mathbf{V}^{(n)} \in \mathbb{C}^{I_{n+1}I_{n+2}\ldots I_N I_1 I_2 \ldots I_{n-1} \times I_n}$ according to

$$\mathbf{\Sigma}^{(n)} \stackrel{\text{def}}{=} \operatorname{diag}(\sigma_1^{(n)}, \sigma_2^{(n)}, \ldots, \sigma_{I_n}^{(n)}), \tag{9}$$

$$\mathbf{V}^{(n)^H} \stackrel{\text{def}}{=} \tilde{\mathbf{S}}_{(n)} \cdot \left(\mathbf{U}^{(n+1)} \otimes \mathbf{U}^{(n+2)} \otimes \cdots \otimes \mathbf{U}^{(N)} \otimes \mathbf{U}^{(1)} \otimes \mathbf{U}^{(2)} \otimes \cdots \otimes \mathbf{U}^{(n-1)}\right), \tag{10}$$

in which $\tilde{\mathbf{S}}_{(n)}$ is a normalized version of $\mathbf{S}_{(n)}$, with the rows scaled to unit-length

$$\mathbf{S}_{(n)} = \mathbf{\Sigma}^{(n)} \cdot \tilde{\mathbf{S}}_{(n)}. \tag{11}$$

Expressing (8) in terms of $\mathbf{\Sigma}^{(n)}$ and $\mathbf{V}^{(n)}$ shows that, at a matrix level, the HOSVD conditions lead to an SVD of the matrix unfoldings

$$\mathbf{A}_{(n)} = \mathbf{U}^{(n)} \cdot \mathbf{\Sigma}^{(n)} \cdot \mathbf{V}^{(n)^H} \tag{12}$$

($1 \leqslant n \leqslant N$). Below we will show that, on the other hand, the left singular matrices of the different matrix unfoldings of $\mathcal{A}$ correspond to unitary transformations that induce the HOSVD structure. This strong link ensures that the HOSVD inherits all the classical column/row space properties from the matrix SVD (see section 4).

The dyadic decomposition could be generalized by expressing the HOSVD model as an *expansion of mutually orthogonal rank-1 tensors*,

$$\mathcal{A} = \sum_{i_1} \sum_{i_2} \cdots \sum_{i_N} s_{i_1 i_2 \ldots i_N}\, U_{i_1}^{(1)} \circ U_{i_2}^{(2)} \circ \cdots \circ U_{i_N}^{(N)}, \tag{13}$$

in which the coefficients $s_{i_1 i_2 \ldots i_N}$ are the entries of an ordered all-orthogonal tensor $\mathcal{S}$. The orthogonality of the rank-1 terms follows from the orthogonality of the n-mode singular vectors. In connection with the discussion on pseudodiagonality versus all-orthogonality, we remark that the summation generally involves $r_1 r_2 \ldots r_N$ terms (instead of $\min(I_1, I_2, \ldots, I_N)$), in which r_n is the highest index for which $\|\mathcal{S}_{i_n = r_n}\| >$

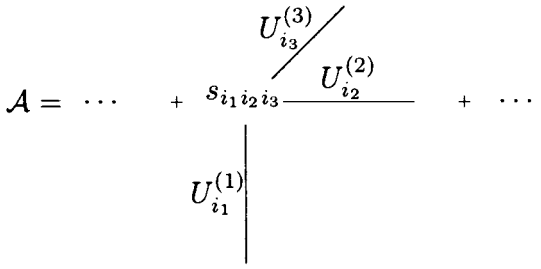


FIG. 5. *Visualization of a triadic decomposition.*

0 in (7). In Figure 5 this decomposition is visualized for third-order tensors. Where the dyadic decomposition expresses a matrix in an essentially unique way as a minimal linear combination of products of column and row vectors that are each mutually orthonormal, the meaning of (13) is less outspoken. For example, the matrix $\mathbf{U}'^{(n)} = \mathbf{U}^{(n)} \cdot \mathbf{Q}$, in which $\mathbf{Q}$ is a unitary matrix, together with the tensor $\mathcal{S}' = \mathcal{S} \times_n \mathbf{Q}^H$ still leads to an expansion in $r_1 r_2 \ldots r_N$ mutually orthogonal rank-1 tensors (however, $\mathcal{S}'$ is not all-orthogonal in general). The unitary matrix $\mathbf{Q}$ could even be chosen in such a way that it induces zero entries in $\mathcal{S}'$, thereby decreasing the number of terms in the rank-1 expansion (e.g., the unitary factor of a QR-decomposition of $\mathbf{S}_{(n)}$ induces $r_n(r_n - 1)/2$ zeros).

Proof. The derivation of Theorem 2 establishes the connection between the HOSVD of a tensor $\mathcal{A}$ and the matrix SVD of its matrix unfoldings. It is given in terms of real-valued tensors; the complex case is completely analogous but more cumbersome from a notational point of view.

Consider two $(I_1 \times I_2 \times \cdots \times I_N)$ tensors $\mathcal{A}$ and $\mathcal{S}$, related by

$$\mathcal{S} = \mathcal{A} \times_1 \mathbf{U}^{(1)^T} \times_2 \mathbf{U}^{(2)^T} \cdots \times_N \mathbf{U}^{(N)^T}, \tag{14}$$

in which $\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \ldots, \mathbf{U}^{(N)}$ are orthogonal matrices. Equation (14) can be expressed in a matrix format as

$$\mathbf{A}_{(n)} = \mathbf{U}^{(n)} \cdot \mathbf{S}_{(n)} \cdot \left(\mathbf{U}^{(n+1)} \otimes \mathbf{U}^{(n+2)} \cdots \mathbf{U}^{(N)} \otimes \mathbf{U}^{(1)} \otimes \mathbf{U}^{(2)} \cdots \mathbf{U}^{(n-1)} \right)^T. \tag{15}$$

Now consider the particular case where $\mathbf{U}^{(n)}$ is obtained from the SVD of $\mathbf{A}_{(n)}$ as

$$\mathbf{A}_{(n)} = \mathbf{U}^{(n)} \cdot \mathbf{\Sigma}^{(n)} \cdot \mathbf{V}^{(n)^T}, \tag{16}$$

in which $\mathbf{V}^{(n)}$ is orthogonal and $\mathbf{\Sigma}^{(n)} = \operatorname{diag}(\sigma_1^{(n)}, \sigma_2^{(n)}, \ldots, \sigma_{I_n}^{(n)})$, where

$$\sigma_1^{(n)} \geqslant \sigma_2^{(n)} \geqslant \cdots \geqslant \sigma_{I_n}^{(n)} \geqslant 0. \tag{17}$$

We call r_n the highest index for which $\sigma_{r_n}^{(n)} > 0$. Taking into account that the Kronecker factor in (15) is orthogonal, comparison of (15) and (16) shows that

$$\mathbf{S}_{(n)} = \mathbf{\Sigma}^{(n)} \cdot \mathbf{V}^{(n)^T} \cdot \left(\mathbf{U}^{(n+1)} \otimes \mathbf{U}^{(n+2)} \cdots \mathbf{U}^{(N)} \otimes \mathbf{U}^{(1)} \otimes \mathbf{U}^{(2)} \cdots \mathbf{U}^{(n-1)} \right). \tag{18}$$

This equation implies, for arbitrary orthogonal matrices $\mathbf{U}^{(1)}$, $\mathbf{U}^{(2)}$,..., $\mathbf{U}^{(n-1)}$, $\mathbf{U}^{(n+1)}, \ldots, \mathbf{U}^{(N)}$, that

$$\langle \mathcal{S}_{i_n=\alpha}, \mathcal{S}_{i_n=\beta} \rangle = 0 \qquad \text{when} \qquad \alpha \neq \beta, \tag{19}$$

and

$$\|\mathcal{S}_{i_n=1}\| = \sigma_1^{(n)} \geqslant \|\mathcal{S}_{i_n=2}\| = \sigma_2^{(n)} \geqslant \cdots \geqslant \|\mathcal{S}_{i_n=I_n}\| = \sigma_{I_n}^{(n)} \geqslant 0, \tag{20}$$

and, if $r_n < I_n$,

$$\|\mathcal{S}_{i_n=r_n+1}\| = \sigma_{r_n+1}^{(n)} = \cdots = \|\mathcal{S}_{i_n=I_n}\| = \sigma_{I_n}^{(n)} = 0. \tag{21}$$

By constructing the matrices $\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \ldots, \mathbf{U}^{(n-1)}, \mathbf{U}^{(n+1)}, \ldots, \mathbf{U}^{(N)}$ in a similar way as $\mathbf{U}^{(n)}$, $\mathcal{S}$ can be made to satisfy all the conditions of the HOSVD theorem. On the other hand, as can be seen from (12), all the matrices $\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \ldots, \mathbf{U}^{(N)}$ and tensors $\mathcal{S}$ satisfying the HOSVD theorem can be found by means of the SVD of $\mathbf{A}_{(1)}$, $\mathbf{A}_{(2)}$, ..., $\mathbf{A}_{(N)}$, where $\mathcal{S}$ follows from (14). □

Computation. Equation (12) and the preceding proof actually indicate how the HOSVD of a given tensor $\mathcal{A}$ can be computed: the n-mode singular matrix $\mathbf{U}^{(n)}$ (and the n-mode singular values) can directly be found as the left singular matrix (and the singular values) of an n-mode matrix unfolding of $\mathcal{A}$ (any matrix of which the columns are given by the n-mode vectors can be resorted to, as the column ordering is of no importance). Hence computing the HOSVD of an Nth-order tensor leads to the computation of N different matrix SVDs of matrices with size $(I_n \times I_1 I_2 \ldots I_{n-1} I_{n+1} \ldots I_N)(1 \leqslant n \leqslant N)$.

Afterwards, the core tensor $\mathcal{S}$ can be computed by bringing the matrices of singular vectors to the left side of (5):

$$\mathcal{S} = \mathcal{A} \times_1 \mathbf{U}^{(1)^H} \times_2 \mathbf{U}^{(2)^H} \cdots \times_N \mathbf{U}^{(N)^H}. \tag{22}$$

This can be computed in a matrix format, e.g.,

$$\mathbf{S}_{(n)} = \mathbf{U}^{(1)^H} \cdot \mathbf{A}_{(n)} \cdot \left(\mathbf{U}^{(n+1)} \otimes \mathbf{U}^{(n+2)} \otimes \cdots \otimes \mathbf{U}^{(N)} \otimes \mathbf{U}^{(1)} \otimes \mathbf{U}^{(2)} \otimes \cdots \otimes \mathbf{U}^{(n-1)} \right). \tag{23}$$

Equations (12) and (23) essentially form a square-root version of the "operational procedures," discussed in [43]. As such they are numerically more reliable, especially for ill-conditioned tensors, i.e., tensors for which $\sigma_1^{(n)} \gg \sigma_{R_n}^{(n)}$, for one or more values of n [22].

Example 4. *Consider the* $(3 \times 3 \times 3)$ *tensor* $\mathcal{A}$ *defined by a matrix unfolding* $\mathbf{A}_{(1)}$, *equal to*

$$\left(\begin{array}{rrr|rrr|rrr} 0.9073 & 0.7158 & -0.3698 & 1.7842 & 1.6970 & 0.0151 & 2.1236 & -0.0740 & 1.4429 \\ 0.8924 & -0.4898 & 2.4288 & 1.7753 & -1.5077 & 4.0337 & -0.6631 & 1.9103 & -1.7495 \\ 2.1488 & 0.3054 & 2.3753 & 4.2495 & 0.3207 & 4.7146 & 1.8260 & 2.1335 & -0.2716 \end{array}\right).$$

The 1*-mode singular vectors are the columns of the left singular matrix of* $\mathbf{A}_{(1)}$*; in the same way,* $\mathbf{U}^{(2)}$ *and* $\mathbf{U}^{(3)}$ *can be obtained:*

$$\mathbf{U}^{(1)} = \left(\begin{array}{rrr} 0.1121 & -0.7739 & -0.6233 \\ 0.5771 & 0.5613 & -0.5932 \\ 0.8090 & -0.2932 & 0.5095 \end{array} \right),$$

$$\mathbf{U}^{(2)} = \begin{pmatrix} 0.4624 & 0.0102 & 0.8866 \\ 0.8866 & -0.0135 & -0.4623 \\ -0.0072 & -0.9999 & 0.0152 \end{pmatrix},$$

$$\mathbf{U}^{(3)} = \begin{pmatrix} 0.6208 & -0.4986 & 0.6050 \\ -0.0575 & -0.7986 & -0.5992 \\ 0.7819 & 0.3371 & -0.5244 \end{pmatrix}.$$

The core tensor of the HOSVD then follows from application of (23); *its unfolding* $\mathbf{S}_{(1)}$ *is equal to*

$$\left(\begin{array}{rrr|rrr|rrr} 8.7088 & 0.0489 & -0.2797 & 0.1066 & 3.2737 & 0.3223 & -0.0033 & -0.1797 & -0.2222 \\ -0.0256 & 3.2546 & -0.2853 & 3.1965 & -0.2130 & 0.7829 & 0.2948 & -0.0378 & -0.3704 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{array}\right).$$

The core tensor is all-orthogonal: the rows of $\mathbf{S}_{(1)}$ *are mutually orthogonal, but so also are the matrices formed by columns* 1/2/3, 4/5/6, *and* 7/8/9, *as well as the three matrices formed by columns* 1/4/7, 2/5/8, *and* 3/6/9. *This boils down to orthogonality of, resp., the "horizontal," "frontal," and "vertical" matrices of* $\mathcal{A}$. *The core tensor is also ordered: its matrices are put in order of decreasing Frobenius-norm. The Frobenius-norms give the n-mode singular values of* $\mathcal{A}$*:*

$$\begin{array}{ll} \textit{mode } 1: & 9.3187, 4.6664, 0, \\ \textit{mode } 2: & 9.3058, 4.6592, 0.5543, \\ \textit{mode } 3: & 9.2822, 4.6250, 1.0310. \end{array}$$

4. Properties. Many properties of the matrix SVD have a very clear higher-order counterpart, because of the strong link between the HOSVD of a higher-order tensor and the SVDs of its matrix unfoldings. In this section, we list the multilinear equivalents of a number of classical matrix SVD properties. The basic link itself is repeated as Property 12. The proofs are outlined at the end of the section.

PROPERTY 4 (uniqueness).

(i) *The n-mode singular values are uniquely defined.*

(ii) *When the n-mode singular values are different, then the n-mode singular vectors are determined up to multiplication with a unit-modulus factor. When* $U_\alpha^{(n)}$ *is multiplied by* $e^{j\theta}$, *then* $\mathcal{S}_{i_n=\alpha}$ *has to be multiplied by the inverse factor* $e^{-j\theta}$.

The n-mode singular vectors corresponding to the same n-mode singular value can be replaced by any unitary linear combination. The corresponding subtensors $\{\mathcal{S}_{i_n=\alpha}\}$ *have to be combined in the inverse way. Formally,* $\mathbf{U}^{(n)}$ *can be replaced by* $\mathbf{U}^{(n)} \cdot \mathbf{Q}$, *in which* $\mathbf{Q}$ *is a block-diagonal matrix, consisting of unitary blocks, where the block-partitioning corresponds to the partitioning of* $\mathbf{U}^{(n)}$ *in sets of n-mode singular vectors with identical n-mode singular value. At the same time* $\mathcal{S}$ *has to be replaced by* $\mathcal{S} \times_n \mathbf{Q}^H$.

For real-valued tensors uniqueness is up to the sign, resp., multiplication with an orthogonal matrix.

The first property implies that the HOSVD shows essentially the same uniqueness properties as the matrix SVD. The only difference consists of the fact that Theorem 2 contains weaker normalization conventions. The equivalent situation for matrices would be to allow that $\mathbf{S}$ consists of diagonal blocks, corresponding to the different singular values, in which each block consists of a unitary matrix, multiplied with the singular value under consideration.

PROPERTY 5 (generalization). *The tensor SVD of a second-order tensor boils down (up to the underdetermination) to its matrix SVD.*

From the discussion in section 3 it is clear that the HOSVD is a formal equivalent of the matrix SVD. Moreover, according to Property 5, it is a true generalization in the sense that, when Theorem 2 is applied to matrices (second-order tensors), it leads to the classical matrix SVD. (Note however that, by convention, the 2-mode singular vectors are *defined* as the complex conjugates of the right matrix singular vectors.) As such, Theorem 2 really establishes a multilinear SVD framework, containing the matrix decomposition in the special case of second-order tensors.

PROPERTY 6 (n-rank). *Let the HOSVD of $\mathcal{A}$ be given as in Theorem 2, and let r_n be equal to the highest index for which $\|\mathcal{S}_{i_n=r_n}\| > 0$ in (7); then one has*

$$R_n = \text{rank}_{\mathtt{n}}(\mathcal{A}) = r_n.$$

The fact that the number of nonvanishing singular values of a given matrix equals its (column/row) rank carries over to the n-mode singular values and the n-rank values of a given tensor (recall from section 2.2 that the n-rank values are not necessarily the same). Like for matrices, this link even holds in a numerical sense, as will be shown by the perturbation analysis in section 6: the number of significant n-mode singular values of a given tensor equals its numerical n-rank. In a matrix context, this property is of major concern for the estimation of "problem dimensionalities," like the estimation of the number of sources in the source separation problem, the estimation of filter lengths in identification, the estimation of the number of harmonics in the harmonic retrieval problem, etc. [45, 46]. Property 6 may play a similar role in multilinear algebra. Let us illustrate this by means of a small example. Consider the most elementary relationship in multivariate statistics, in which an I-dimensional stochastic vector X consists of a linear mixture of $J \leqslant I$ stochastic components. Whereas the number of components is usually estimated as the number of significant eigenvalues of the covariance matrix of X, the number of skew or kurtic components might as well be estimated as the number of significant n-mode singular values of the third-order, resp., fourth-order, cumulant tensor of X [17].

Finally, we remember from section 2.2 that knowledge of the n-rank values of a given tensor does not allow us in general to make precise statements about the rank of that tensor. With this respect, other SVD generalizations might be more interesting (see section 7).

PROPERTY 7 (structure). *Let the HOSVD of $\mathcal{A}$ be given as in Theorem 2; then one has*

(i) *the n-mode vector space* $\mathtt{R}(\mathbf{A}_{(n)}) = \text{span}(U_1^{(n)}, \ldots, U_{R_n}^{(n)})$,

(ii) *the orthogonal complement of* $\mathtt{R}(\mathbf{A}_{(n)})$, *the left n-mode null space* $\mathtt{N}(\mathbf{A}_{(n)}^H) = \text{span}(U_{R_n+1}^{(n)}, \ldots, U_{I_n}^{(n)})$.

In the same way as the left and right singular vectors of a matrix give an orthonormal basis for its column and row space (and their orthogonal complements), the n-mode singular vectors of a given tensor yield an orthonormal basis for its n-mode vector space (and its orthogonal complement). For matrices, the property was the starting-point for the development of subspace algorithms [45, 46, 47]; Property 7 allows for an extension of this methodology in multilinear algebra.

PROPERTY 8 (norm). *Let the HOSVD of $\mathcal{A}$ be given as in Theorem 2; then the following holds.*

$$\begin{aligned}\|\mathcal{A}\|^2 &= \sum_{i=1}^{R_1} \left(\sigma_i^{(1)}\right)^2 = \cdots = \sum_{i=1}^{R_N} \left(\sigma_i^{(N)}\right)^2 \\ &= \|\mathcal{S}\|^2.\end{aligned}$$

In multilinear algebra as well as in matrix algebra, the Frobenius-norm is unitarily invariant. As a consequence, the well-known fact that the squared Frobenius-norm of a matrix equals the sum of its squared singular values can be generalized.

DEFINITION 9 (n-mode oriented energy). *The n-mode oriented energy of an Nth-order tensor $\mathcal{A} \in \mathbb{C}^{I_1 \times I_2 \times \ldots \times I_N}$ in the direction of a unit-norm vector X, denoted by $\texttt{OE}_\texttt{n}(X, \mathcal{A})$, is the oriented energy of the set of n-mode vectors, i.e.,*

$$\texttt{OE}_\texttt{n}(X, \mathcal{A}) \stackrel{\text{def}}{=} \|X^H \mathbf{A}_{(n)}\|^2.$$

The concept of oriented energy of a given matrix and the link with the SVD of that matrix, which form the basis of SVD-based signal separation algorithms, can easily be generalized as well. Whereas Property 8 merely states that the squared Frobenius-norm of a tensor (i.e., the "energy" contained in the tensor) equals the sum of the squared n-mode singular values, the HOSVD actually gives a pretty detailed geometrical picture of the energy distribution over the n-mode vector space. Definition 9 generalizes the definition of oriented energy. We now state the following property.

PROPERTY 9 (oriented energy). *The directions of extremal n-mode oriented energy correspond to the n-mode singular vectors, with extremal energy value equal to the corresponding squared n-mode singular value.*

This means that the n-mode vectors mainly contain contributions in the direction of $U_1^{(n)}$; this particular direction accounts for an amount of $\sigma_1^{(n)^2}$ with respect to the total amount of energy in the tensor. Next, the n-mode oriented energy reaches an extremum in the direction of $U_2^{(n)}$, perpendicular to $U_1^{(n)}$, for a value of $\sigma_2^{(n)^2}$, and so on. Discarding the components of the n-mode vectors in the direction of an n-mode singular vector $U_{i_n}^{(n)}$ (e.g., the one corresponding to the smallest n-mode singular value) to obtain a tensor $\hat{\mathcal{A}}$ introduces an error $\|\mathcal{A} - \hat{\mathcal{A}}\|^2 = \sigma_{i_n}^{(n)^2}$.

PROPERTY 10 (approximation). *Let the HOSVD of $\mathcal{A}$ be given as in Theorem 2 and let the n-mode rank of $\mathcal{A}$ be equal to R_n $(1 \leqslant n \leqslant N)$. Define a tensor $\hat{\mathcal{A}}$ by discarding the smallest n-mode singular values $\sigma_{I'_n+1}^{(n)}, \sigma_{I'_n+2}^{(n)}, \ldots, \sigma_{R_n}^{(n)}$ for given values of I'_n $(1 \leqslant n \leqslant N)$, i.e., set the corresponding parts of $\mathcal{S}$ equal to zero. Then we have*

$$\|\mathcal{A} - \hat{\mathcal{A}}\|^2 \leqslant \sum_{i_1=I'_1+1}^{R_1} \sigma_{i_1}^{(1)^2} + \sum_{i_2=I'_2+1}^{R_2} \sigma_{i_2}^{(2)^2} + \cdots + \sum_{i_N=I'_N+1}^{R_N} \sigma_{i_N}^{(N)^2}. \tag{24}$$

This property is the higher-order equivalent of the link between the SVD of a matrix and its best approximation, in a least-squares sense, by a matrix of lower rank. However, the situation is quite different for tensors. By discarding the smallest n-mode singular values, one obtains a tensor $\hat{\mathcal{A}}$ with a column rank equal to I'_1, row rank equal to I'_2, etc. However, this tensor is in general not the best possible approximation under the given n-mode rank constraints (see, e.g., Example 5). Nevertheless, the ordering assumption (7) implies that the "energy" of $\mathcal{A}$ is mainly concentrated in the part corresponding to low values of $i_1, i_2, \ldots, i_N$. Consequently, if $\sigma_{I'_n}^{(n)} \gg \sigma_{I'_n+1}^{(n)}$ (e.g., I'_n corresponds to the numerical n-rank of $\mathcal{A}$; the smaller n-mode singular values are not significant (see also Property 6)), $\hat{\mathcal{A}}$ is still to be considered as a good approximation of $\mathcal{A}$. The error is bounded as in (24). For procedures to enhance a given approximation, we refer to [28, 17].

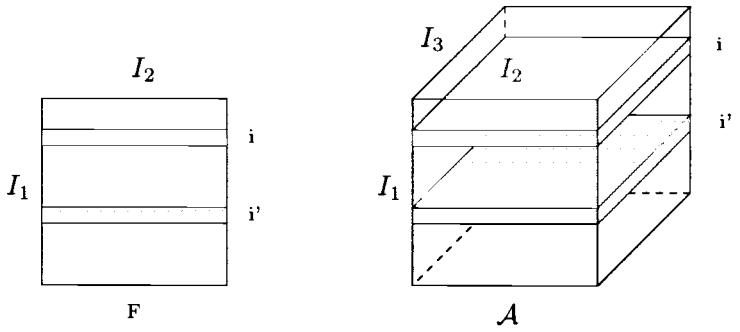


FIG. 6. *Construction of* $\mathbf{H}^{(1)}$ *for* (a) *a matrix and* (b) *a third-order tensor.*

PROPERTY 11 (link between HOSVD and matrix EVD). *Let the HOSVD of* $\mathcal{A}$ *be given as in Theorem* 2. *Define* $\mathbf{H}^{(n)} \stackrel{\text{def}}{=} \mathbf{A}_{(n)} \cdot \mathbf{A}_{(n)}^H$, *i.e.,* $\mathbf{H}^{(n)}$ *contains on position* (i, i') *the scalar product* $\langle \mathcal{A}_{i_n=i}, \mathcal{A}_{i_n=i'} \rangle$. *If the EVD of* $\mathbf{H}^{(n)}$ *is given by*

$$\mathbf{H}^{(n)} = \mathbf{U}^{(n)} \cdot \mathbf{D}^{(n)} \cdot \mathbf{U}^{(n)^H},$$

then $\mathbf{U}^{(n)}$ *contains the n-mode singular vectors of* $\mathcal{A}$. *Moreover the scalar product* $\langle \mathcal{S}_{i_n=i}, \mathcal{S}_{i_n=i'} \rangle$ *is the* (i, i')*th element of* $\mathbf{D}^{(n)}$. *Different subtensors of* $\mathcal{S}$ *are mutually orthogonal.*

In the same way as the SVD of a matrix $\mathbf{F}$ is related to the EVD of the Hermitean (real symmetric) positive (semi)definite matrices $\mathbf{F}\mathbf{F}^H$ and $\mathbf{F}^H\mathbf{F}$, the HOSVD of a higher-order tensor $\mathcal{A}$ is related to the EVD of Hermitean (real symmetric) positive (semi)definite matrices, constructed from $\mathcal{A}$. The construction is clarified in Figure 6: just like the entries of $\mathbf{F}\mathbf{F}^H$ consist of the mutual scalar products of the rows of $\mathbf{F}$, the matrix $\mathbf{H}^{(1)}$ in Property 11 is computed from the scalar products of the "horizontal matrices" of $\mathcal{A}$, in the third-order case. This property can, e.g., be useful for interpretations in a statistical context, where $\mathbf{H}^{(n)}$ corresponds to the sample correlation matrix of the n-mode vectors of $\mathcal{A}$.

PROPERTY 12 (link between HOSVD and matrix SVD). *Let the HOSVD of* $\mathcal{A}$ *be given as in Theorem* 2. *Then*

$$\mathbf{A}_{(n)} = \mathbf{U}^{(n)} \cdot \mathbf{\Sigma}^{(n)} \cdot \mathbf{V}^{(n)^H}$$

is an SVD of $\mathbf{A}_{(n)}$, *where the diagonal matrix* $\mathbf{\Sigma}^{(n)} \in \mathbb{R}^{I_n \times I_n}$ *and the column-wise orthonormal matrix* $\mathbf{V}^{(n)} \in \mathbb{C}^{I_{n+1}I_{n+2}\ldots I_N I_1 I_2 \ldots I_{n-1} \times I_n}$ *are defined according to*

$$\mathbf{\Sigma}^{(n)} \stackrel{\text{def}}{=} \text{diag}(\sigma_1^{(n)}, \sigma_2^{(n)}, \ldots, \sigma_{I_n}^{(n)}),$$

$$\mathbf{V}^{(n)^H} \stackrel{\text{def}}{=} \tilde{\mathbf{S}}_{(n)} \cdot \left(\mathbf{U}^{(n+1)} \otimes \mathbf{U}^{(n+2)} \ldots \mathbf{U}^{(N)} \otimes \mathbf{U}^{(1)} \otimes \mathbf{U}^{(2)} \cdots \otimes \mathbf{U}^{(n-1)} \right)^T,$$

in which $\tilde{\mathbf{S}}_{(n)}$ *is a normalized version of* $\mathbf{S}_{(n)}$, *with the rows scaled to unit-length*

$$\mathbf{S}_{(n)} = \mathbf{\Sigma}^{(n)} \cdot \tilde{\mathbf{S}}_{(n)}.$$

Proof. The proofs are established by a further analysis of the derivation in section 3; the key idea behind this derivation is explained in Property 12.

Property 4 is proved by investigating the uniqueness conditions.

Property 5 follows by applying the same procedure to second-order tensors.

Property 6 follows from the combination of Property 12 and Property 1. In the same way the deduction of Property 7 and Property 9 is trivial.

Property 8: We have $\|\mathcal{A}\|^2 = \|\mathbf{A}_{(n)}\|^2$ which, in turn, equals $\sum_{i=1}^{r_n} \|\mathcal{S}_{i_n=i}\|^2 = \|\mathcal{S}\|^2$.

Property 10: We have

$$\begin{aligned}
\|\mathcal{A} - \hat{\mathcal{A}}\|^2 &= \sum_{i_1=1}^{R_1}\sum_{i_2=1}^{R_2}\cdots\sum_{i_N=1}^{R_N} s_{i_1 i_2 \ldots i_N}^2 - \sum_{i_1=1}^{I'_1}\sum_{i_2=1}^{I'_2}\cdots\sum_{i_N=1}^{I'_N} s_{i_1 i_2 \ldots i_N}^2 \\
&= \sum_{i_1=I'_1+1}^{R_1}\sum_{i_2=I'_2+1}^{R_2}\cdots\sum_{i_N=I'_N+1}^{R_N} s_{i_1 i_2 \ldots i_N}^2 \\
&\leqslant \sum_{i_1=I'_1+1}^{R_1}\sum_{i_2=1}^{R_2}\cdots\sum_{i_N=1}^{R_N} s_{i_1 i_2 \ldots i_N}^2 + \sum_{i_1=1}^{R_1}\sum_{i_2=I'_2+1}^{R_2}\cdots\sum_{i_N=1}^{R_N} s_{i_1 i_2 \ldots i_N}^2 \\
&\quad + \cdots + \sum_{i_1=1}^{R_1}\sum_{i_2=1}^{R_2}\cdots\sum_{i_N=I'_N+1}^{R_N} s_{i_1 i_2 \ldots i_N}^2 \\
&= \sum_{i_1=I'_1+1}^{R_1} {\sigma_{i_1}^{(1)}}^2 + \sum_{i_2=I'_2+1}^{R_2} {\sigma_{i_2}^{(2)}}^2 + \cdots + \sum_{i_N=I'_N+1}^{R_N} {\sigma_{i_N}^{(N)}}^2.
\end{aligned}$$

Property 11 follows from the link between matrix SVD and HOSVD, combined with the relation between matrix SVD and EVD. ∎

Example 5. *Consider the tensor $\mathcal{A}$ and its HOSVD components, which are given in Example* 4.

From the n-mode singular values, we see that $R_2 = R_3 = 3$, while $R_1 = 2$. Clearly, the column space of $\mathcal{A}$ is only two-dimensional. The first two columns of $\mathbf{U}^{(1)}$ form an orthonormal basis for this vector space; the third 1*-mode singular vector is orthogonal to the column space of $\mathcal{A}$.*

The sum of the squared n-mode singular values is equal to 108.6136 *for all three modes;* 108.6136 *is the squared Frobenius-norm of $\mathcal{A}$.*

Discarding $\sigma_3^{(2)}$ and $\sigma_3^{(3)}$, i.e., replacing $\mathcal{S}$ by a tensor $\hat{\mathcal{S}}$, having a matrix unfolding $\hat{\mathbf{S}}_{(1)}$ equal to

$$\left(\begin{array}{rrr|rrr|rrr}
8.7088 & 0.0489 & 0.0000 & 0.1066 & 3.2737 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\
-0.0256 & 3.2546 & 0.0000 & 3.1965 & -0.2130 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\
0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000
\end{array}\right)$$

gives an approximation $\hat{\mathcal{A}}$ for which $\|\mathcal{A} - \hat{\mathcal{A}}\| = 1.0880$. On the other hand, the tensor $\mathcal{A}'$ that best matches $\mathcal{A}$ while having three n-ranks equal to 2 *is defined by the unfolding*

$$\left(\begin{array}{rrr|rrr|rrr}
0.8188 & 0.8886 & -0.0784 & 1.7051 & 1.7320 & -0.0274 & 1.7849 & 0.2672 & 1.7454 \\
1.0134 & -0.8544 & 2.1455 & 1.9333 & -1.5390 & 3.9886 & -0.2877 & 1.5266 & -2.0826 \\
2.1815 & 0.0924 & 2.4019 & 4.3367 & 0.3272 & 4.6102 & 1.8487 & 2.1042 & -0.2894
\end{array}\right).$$

For this tensor, we have that $\|\mathcal{A} - \mathcal{A}'\| = 1.0848$.

5. A multilinear symmetric EVD. Many applications show highly symmetric higher-order tensors. As an example, higher-order moments and cumulants of a real random vector are invariant under arbitrary index permutations. The same holds, e.g., for the symmetric tensorial representation of homogeneous polynomials [12]. For matrices, this symmetry is reflected by the fact that the left and right singular vectors are, up to the sign, identical; this leads to a particular form of the EVD, in which the eigenmatrix is an orthogonal (unitary) matrix. In this section, we will investigate if similar properties hold for higher-order tensors as well.

First, we prove that, if a tensor is invariant to (or mapped onto its complex conjugate by) a permutation of its indices, this symmetry is also reflected by the HOSVD components.

THEOREM 3. *Let the HOSVD of a tensor* $\mathcal{A} \in \mathbb{C}^{I_1 \times I_2 \times \cdots \times I_N}$ *be denoted as in Theorem* 2. *Consider a permutation* $\mathtt{P}$ *of its indices and the decomposition of this permutation in a sequence of permutation cycles* $\mathtt{C}_1$, $\mathtt{C}_2, \ldots$. *We assume, without losing generality (possibly by redefining the ordering of the indices of* $\mathcal{A}$*), that*

$$\begin{aligned}\mathtt{P}(i_1, i_2, \ldots, i_N) &= (\mathtt{C}_1(i_1, i_2, \ldots, i_{n_1}), \mathtt{C}_2(i_{n_1+1}, i_{n_1+2}, \ldots, i_{n_2}), \ldots) \\ &= (i_2, i_3, \ldots, i_{n_1}, i_1; i_{n_1+2}, \ldots, i_{n_2}, i_{n_1+1}; \ldots).\end{aligned}$$

If $a_{i_1 i_2 \ldots i_N} = a_{\mathtt{P}(i_1 i_2 \ldots i_N),}$, *then one has*

$$\mathbf{U}^{(1)} = \mathbf{U}^{(2)} = \cdots = \mathbf{U}^{(n_1)};\ \mathbf{U}^{(n_1+1)} = \mathbf{U}^{(n_1+2)} = \cdots = \mathbf{U}^{(n_2)};\ \ldots.$$

If $a_{i_1 i_2 \ldots i_N} = a^*_{\mathtt{P}(i_1 i_2 \ldots i_N)}$, *then one has*

$$\mathbf{U}^{(1)} = \mathbf{U}^{(2)^*} = \mathbf{U}^{(3)} = \cdots;\ \mathbf{U}^{(n_1+1)} = \mathbf{U}^{(n_1+2)^*} = \mathbf{U}^{(n_1+3)} = \cdots;\ \ldots.$$

For permutation cycles with an odd number of elements, the matrix of singular vectors is real.

The core tensor $\mathcal{S}$ *exhibits the same permutation symmetry as* $\mathcal{A}$.

Proof. Construct the higher-order tensors $\mathcal{A}'$, $\mathcal{A}''$, ... by repeatedly permuting the indices by $\mathtt{P}$:

$$a'_{i_1, i_2, \ldots, i_N} \overset{\text{def}}{=} a_{\mathtt{P}(i_1, i_2, \ldots, i_N)}; \qquad a''_{i_1, i_2, \ldots, i_N} \overset{\text{def}}{=} a_{\mathtt{P}(\mathtt{P}(i_1, i_2, \ldots, i_N))}; \qquad \ldots.$$

For real symmetry, we have that $\mathcal{A} = \mathcal{A}' = \mathcal{A}'' = \ldots$; for Hermitean symmetry, we have that $\mathcal{A} = (\mathcal{A}')^* = \mathcal{A}'' = \ldots$. The same equalities hold for the matrix unfoldings $\mathbf{A}_{(n_1)}$, $\mathbf{A}'_{(n_1)}$, $\mathbf{A}''_{(n_1)}$, ..., as well as for $\mathbf{A}_{(n_2)}$, $\mathbf{A}'_{(n_2)}$, $\mathbf{A}''_{(n_2)}$, ..., etc. Hence the same symmetry is shown by the left singular matrices of these matrix unfoldings, which correspond to $\mathbf{U}^{(1)}$, $\mathbf{U}^{(2)}$, ..., resp., $\mathbf{U}^{(n_1+1)}$, $\mathbf{U}^{(n_1+2)}$, ..., etc.

If $\mathtt{P}$ corresponds to Hermitean symmetry, and, e.g., $\mathtt{C}_1$ is a cycle with an odd number of elements, then it can be shown that $\mathbf{U}^{(1)} = \mathbf{U}^{(1)^*}$ by repeating the construction above n_1 times.

For an analysis of the symmetry of $\mathcal{S}$, we write down an element-wise form of (22) (we take the example of Hermitean symmetry):

$$s_{j_1 j_2 \ldots j_N} = \sum_{i_1 i_2 \ldots i_N} a_{i_1 i_2 \ldots i_N} u^{(n_1)^*}_{i_1 j_1} u^{(n_1)}_{i_2 j_2} u^{(n_1)^*}_{i_3 j_3} \ldots u^{(n_2)^*}_{i_{n_1+1} j_{n_1+1}} u^{(n_2)}_{i_{n_1+2} j_{n_1+2}} u^{(n_2)^*}_{i_{n_1+3} j_{n_1+3}} \ldots. \tag{25}$$

Permutation of the indices by $\mathtt{P}$ yields the following element:

$$s_{\mathtt{P}(j_1 j_2 \ldots j_N)} = \sum_{i_1 i_2 \ldots i_N} a_{i_1 i_2 \ldots i_N} u^{(n_1)^*}_{i_1 j_2} u^{(n_1)}_{i_2 j_3} u^{(n_1)^*}_{i_3 j_4} \ldots u^{(n_2)^*}_{i_{n_1+1} j_{n_1+2}} u^{(n_2)}_{i_{n_1+2} j_{n_1+3}} u^{(n_2)^*}_{i_{n_1+3} j_{n_1+4}} \ldots. \tag{26}$$

Invoking the Hermitean symmetry of $\mathcal{A}$, and comparing (26) to (25), it follows that $s_{\mathtt{P}(j_1 j_2 \ldots j_N)} = s^*_{j_1 j_2 \ldots j_N}$. ∎

A consequence of Theorem 3 is that the i-mode singular vectors, for different i, are fully related under a condition that we define as *pair-wise symmetry*.

DEFINITION 10. *A higher-order tensor $\mathcal{A} \in \mathbb{C}^{I \times I \times \ldots \times I}$ is called pair-wise symmetric when, for every pair of indices (i_{n_1}, i_{n_2}), there exists a permutation $\mathtt{P}_{n_1 n_2}$ such that $i_{n_1} = \mathtt{P}_{n_1 n_2}(i_{n_2})$ and either $a_{\mathtt{P}_{n_1 n_2}(i_1 i_2 \ldots i_N)} = a_{i_1 i_2 \ldots i_N}$ or $a_{\mathtt{P}_{n_1 n_2}(i_1 i_2 \ldots i_N)} = a^*_{i_1 i_2 \ldots i_N}$.*

Pair-wise symmetric higher-order tensors are a tensorial equivalent of symmetric and Hermitean matrices. Theorem 3 and Definition 10 lead in a very natural way to a generalization of the orthogonal (unitary) EVD of symmetric (Hermitean) matrices. An HOEVD for pair-wise symmetric higher-order tensors can be defined as follows.

THEOREM 4 (Nth-order pair-wise symmetric EVD). *Every pair-wise symmetric $(I \times I \times \cdots \times I)$-tensor $\mathcal{A}$ can be written as the product*

$$\mathcal{A} = \mathcal{S} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \cdots \times_N \mathbf{U}^{(N)}, \tag{27}$$

in which the following hold.

1. *$\mathcal{S}$ is an all-orthogonal $(I \times I \times \cdots \times I)$-tensor with the same pair-wise symmetry as $\mathcal{A}$.*

2. *$\mathbf{U}^{(1)} = \mathbf{U} = (U_1 U_2 \ldots U_I)$ is a unitary $(I \times I)$-matrix, equal to $\mathbf{U}^{(n)}$ or $\mathbf{U}^{(n)^*}$ (for all n, $1 \leqslant n \leqslant N$), depending on the symmetry of $\mathcal{A}$ as stated in Theorem 3.*

The Frobenius-norms of the subtensors of $\mathcal{S}$, obtained by fixing one index to i, are the Nth-order eigenvalues and are symbolized by λ_i. The vectors U_i are the Nth-order eigenvectors.

Example 6. *Consider a complex third-order tensor $\mathcal{A}$ that satisfies $a_{\mathtt{P}(i_1 i_2 i_3)} = a_{i_2 i_3 i_1} = a^*_{i_1 i_2 i_3}$. According to Theorem 3, and taking into account that $\mathtt{P}$ is a cycle with three elements, the HOSVD of $\mathcal{A}$ is given by*

$$\mathcal{A} = \mathcal{S} \times_1 \mathbf{U} \times_2 \mathbf{U} \times_3 \mathbf{U},$$

where $\mathbf{U}$ is real and $\mathcal{S}$ satisfies the same symmetries as $\mathcal{A}$. In fact, it is pretty obvious in this example that $\mathbf{U}$ is real: $a_{i_1 i_2 i_3} = a^*_{i_2 i_3 i_1} = a_{i_3 i_1 i_2} = a^*_{i_1 i_2 i_3}$ implies that $\mathcal{A}$ itself, and hence the whole decomposition, is real.

Example 7. *A common way to deal with the complex nature of a random vector X in the definition of its fourth-order cumulant tensor $\mathcal{C}$ is the following element-wise definition:*

$$\begin{aligned} c_{i_1 i_2 i_3 i_4} &\stackrel{\text{def}}{=} \operatorname{cum}(x_{i_1}, x^*_{i_2}, x^*_{i_3}, x_{i_4}) \\ &\stackrel{\text{def}}{=} \mathtt{E}\{\tilde{x}_{i_1} \tilde{x}^*_{i_2} \tilde{x}^*_{i_3} \tilde{x}_{i_4}\} - \mathtt{E}\{\tilde{x}_{i_1} \tilde{x}^*_{i_2}\}\mathtt{E}\{\tilde{x}^*_{i_3} \tilde{x}_{i_4}\} \\ &\quad - \mathtt{E}\{\tilde{x}_{i_1} \tilde{x}^*_{i_3}\}\mathtt{E}\{\tilde{x}^*_{i_2} \tilde{x}_{i_4}\} - \mathtt{E}\{\tilde{x}_{i_1} \tilde{x}_{i_4}\}\mathtt{E}\{\tilde{x}^*_{i_2} \tilde{x}^*_{i_3}\}, \end{aligned} \tag{28}$$

in which $\mathtt{E}$ denotes the expectation operator and $\tilde{x}$ equals $x - \mathtt{E}\{x\}$.

The permutation symmetries of $\mathcal{C}$ can be generated by the permutation pair $\mathtt{P}_1$ and $\mathtt{P}_2$, satisfying

$$c_{\mathtt{P}_1(i_1 i_2 i_3 i_4)} = c_{i_4 i_2 i_3 i_1} = c_{i_1 i_2 i_3 i_4}, \tag{29}$$

$$c_{\mathtt{P}_2(i_1 i_2 i_3 i_4)} = c_{i_2 i_1 i_4 i_3} = c^*_{i_1 i_2 i_3 i_4}. \tag{30}$$

Equation (30) *implies that $\mathbf{U}^{(1)} = \mathbf{U}^{(2)^*}$ and that $\mathbf{U}^{(3)} = \mathbf{U}^{(4)^*}$; on the other hand* (29) *implies that $\mathbf{U}^{(1)} = \mathbf{U}^{(4)}$. Hence the HOSVD of $\mathcal{C}$ reveals the structure of* (28),

$$\mathcal{C} = \mathcal{S} \times_1 \mathbf{U} \times_2 \mathbf{U}^* \times_3 \mathbf{U}^* \times_4 \mathbf{U},$$

with $\mathcal{S}$ satisfying (29) *and* (30) *as well.*

6. First-order perturbation analysis. In this section, the first-order derivatives of the n-mode singular values and vectors with respect to the elements of the original tensor are computed. The expressions can be used to analyze the numerical sensitivity and the statistical performance of HOSVD-based signal processing algorithms (e.g., [14]), especially for high signal-to-noise ratios: if the output of these algorithms can be considered as a function of the n-mode singular values and vectors, then the derivatives of the output could be computed via the chain rule.

An important application area is linked with the discussion on pseudodiagonality versus all-orthogonality (see section 3). In several applications, one has the a priori knowledge that a particular tensor can be made pseudoorthogonal by means of unitary transformations (e.g., cumulants of a stochastic vector with statistically independent components, after unitary transformation, like in a typical source separation context; cumulants of a stochastic vector with statistically independent components are pseudodiagonal). As pseudodiagonality is a special case of all-orthogonality, these unitary transformations coincide with the matrices of n-mode singular vectors. However, due to an imperfect knowledge of the tensor under consideration, caused by measurement errors, etc., the HOSVD of the estimated tensor will generally yield a core tensor that is all-orthogonal yet not exactly pseudodiagonal. To investigate such effects, first-order perturbation analysis results are required.

Due to the relationship between the matrix and tensor SVD, a perturbation analysis of the HOSVD can be based on existing results that have been developed in numerical linear algebra. Let $\mathcal{A}$ be a real or complex higher-order tensor, the elements of which are analytic functions of a real parameter p. Like for matrices, an n-mode singular value and vector are analytic functions of p in the neighborhood of a nominal parameter p_0 where that singular value is isolated and nonzero. In case of a zero n-mode singular value the n-mode singular value function need not be differentiable, let alone analytic, due to the definition of an n-mode singular value as a norm being a nonnegative number: the differentiability of a smooth function that becomes negative in a neighborhood of p_0 may be lost by flipping the negative parts around the zero axis. A similar effect appears when several n-mode singular value functions cross each other, due to the ordering constraint on the HOSVD: the differentiability of the smooth functions that cross each other may be lost by combining the function parts in order of magnitude. However, one can prove that analyticity can be guaranteed by dropping the sign and ordering convention. With this goal we define, in analogy to the "unordered-unsigned singular value decomposition" (USVD) of [20] an *unordered-unsigned higher-order singular value decomposition* (UHOSVD).

THEOREM 5 (UHOSVD). *If the elements of $\mathcal{A} \in \mathbb{C}^{I_1 \times I_2 \times \cdots \times I_N}$ are analytic functions of a real parameter p, then there exist real analytic functions $f_i^{(n)} : \mathbb{R} \to \mathbb{R}$ $(1 \leqslant i \leqslant I_n)$ such that, for all $p \in \mathbb{R}$,*

$$\{\sigma_i^{(n)}(p)\,(1 \leqslant i \leqslant I_n)\} = \{|f_i^{(n)}|\,(1 \leqslant i \leqslant I_n)\}.$$

The functions $f_i^{(n)}$ are called unordered, unsigned n-mode singular value functions.

The analytic continuation of a set of n-mode singular vectors at a nominal parameter value, where they correspond to a multiple n-mode singular value, will be denoted by the term *preferred n-mode singular vectors*. They are given by the following theorem.

THEOREM 6 (preferred n-mode singular vectors). *Let the HOSVD of $\mathcal{A}$ be given as in Theorem* 2*, the ordering of the n-mode singular vectors being of no importance.*

Let $\sigma^{(n)}$ be an n-mode singular value of multiplicity m and let an SVD of $\mathbf{A}_{(n)}$ be given by

$$\mathbf{A}_{(n)} = \mathbf{U}^{(n)} \cdot \mathbf{\Sigma}^{(n)} \cdot \mathbf{V}^{(n)^H} = \left(\mathbf{U}_1^{(n)} \mathbf{U}_2^{(n)}\right) \cdot \begin{pmatrix} \sigma^{(n)} \mathbf{I}_m & 0 \\ 0 & \mathbf{\Sigma}_2 \end{pmatrix} \cdot \begin{pmatrix} \mathbf{V}_1^{(n)^H} \\ \mathbf{V}_2^{(n)^H} \end{pmatrix}, \tag{31}$$

in which $\mathbf{U}^{(n)}$, $\mathbf{\Sigma}^{(n)}$, and $\mathbf{V}^{(n)}$ are defined in accordance to Property 12. *In the partitioning, $\mathbf{U}_1^{(n)}$ and $\mathbf{V}_1^{(n)}$ correspond to the singular value $\sigma^{(n)}$.*

Consider a first-order perturbation of $\mathcal{A}$ as $\mathcal{A} + \epsilon\mathcal{B}$, $\epsilon \in \mathbb{R}$, and define the matrices $\mathbf{B}_{j_1 j_2}$, $j_1, j_2 = 1, 2$, as

$$\mathbf{B}_{j_1 j_2} \stackrel{\text{def}}{=} \mathbf{U}_{j_1}^{(n)^H} \cdot \mathbf{B}_{(n)} \cdot \mathbf{V}_{j_2}^{(n)}.$$

There are two cases.

1. $\sigma^{(n)} \neq 0$: *consider the eigenvalue problem*

$$\frac{\mathbf{B}_{11} + \mathbf{B}_{11}^H}{2} = \mathbf{X} \cdot \mathbf{\Lambda} \cdot \mathbf{X}^H.$$

Then the m preferred n-mode singular vectors are given by the columns of $\mathbf{U}_1^{(n)} \cdot \mathbf{X}$. (*The m preferred right singular vectors in* (31) *are given by the columns of $\mathbf{V}_1^{(n)} \cdot \mathbf{X}$.*)

2. $\sigma^{(n)} = 0$: *consider an SVD of $\mathbf{B}_{11}$:*

$$\mathbf{B}_{11} = \mathbf{X} \cdot \mathbf{\Lambda} \cdot \mathbf{Y}^H.$$

Then the m preferred n-mode singular vectors are given by the columns of $\mathbf{U}_1^{(n)} \cdot \mathbf{X}$. (*The m preferred right singular vectors in* (31) *are given by the columns of $\mathbf{V}_1^{(n)} \cdot \mathbf{Y}$.*)

In analogy with [27, 20], we state now the following perturbation theorem.

THEOREM 7 (first-order perturbation of the HOSVD). *Consider a higher-order tensor $\mathcal{A}(\epsilon)$, the elements of which are analytic functions of a real parameter ϵ. Represent the first-order approximation of the Taylor series expansion of $\mathcal{A}(\epsilon)$ around $\epsilon = 0$ by $\mathcal{A} + \epsilon\,\mathcal{B}$. Let the HOSVD of $\mathcal{A}$ be given as in Theorem* 2. *Let an SVD of $\mathbf{A}_{(n)}$ be given by*

$$\mathbf{A}_{(n)} = \mathbf{U}^{(n)} \cdot \mathbf{\Sigma}^{(n)} \cdot \mathbf{V}^{(n)^H},$$

in which $\mathbf{U}^{(n)}$, $\mathbf{\Sigma}^{(n)}$, and $\mathbf{V}^{(n)}$ are defined in accordance to Property 12. *In case of multiple n-mode singular values, $\mathbf{U}^{(n)}$ and $\mathbf{V}^{(n)}$ contain preferred basis vectors. Define the matrix $\mathbf{B}$ as*

$$\mathbf{B} \stackrel{\text{def}}{=} \mathbf{U}^{(n)^H} \cdot \mathbf{B}_{(n)} \cdot \mathbf{V}^{(n)}.$$

Then one has the following.

1. *The slopes of the n-mode unordered-unsigned singular value functions are given by the diagonal elements of $\mathbf{B}$.*

2. *The first-order approximation of the Taylor series expansion of the n-mode singular vector functions is given by $\mathbf{U}^{(n)} + \epsilon\,\mathbf{U}^{(n)} \cdot \mathbf{\Omega}$, in which $\mathbf{\Omega}$ is a skew-Hermitean matrix defined by*

$$\begin{aligned}
\omega_{ij} &\stackrel{\text{def}}{=} 0 \\
&\quad \textit{if } i = j, \\
\omega_{ij} &\stackrel{\text{def}}{=} (\sigma_i^{(n)} b_{ij}^* + \sigma_j^{(n)} b_{ji})/(\sigma_i^{(n)^2} - \sigma_j^{(n)^2}) \\
&\quad \textit{if } \sigma_i^{(n)} \neq \sigma_j^{(n)}, \\
\omega_{ij} &\stackrel{\text{def}}{=} \mathtt{lim}_{\epsilon \to 0} \left[(\sigma_i^{(n)}(\epsilon) b_{ij}^* + \sigma_j^{(n)}(\epsilon) b_{ji})/(\sigma_i^{(n)^2}(\epsilon) - \sigma_j^{(n)^2}(\epsilon)) \right] \\
&\quad \textit{if } \sigma_i^{(n)}(\epsilon) = \sigma_j^{(n)}(\epsilon) \textit{ for the isolated value } \epsilon = 0, \\
\omega_{ij} &\stackrel{\text{def}}{=} -\omega_{ji} \stackrel{\text{def}}{=} \textit{arbitrary} \\
&\quad \textit{if } \sigma_i^{(n)}(\epsilon) = \sigma_j^{(n)}(\epsilon) \textit{ in a neighborhood of } \epsilon = 0 \ (i \neq j).
\end{aligned}$$

7. Related tensor decompositions. The SVD generalization we present in this paper is so clearly analogous to the SVD of matrices since the i-mode vectors play exactly the same role in the tensor decomposition as column and row vectors do for the matrix SVD. However, it is still possible to focus on other properties of the matrix SVD when looking for an equivalent tensor decomposition. It appears that for higher-order tensors different properties may raise different decompositions. Here we list three alternative approaches. The analogy with the matrix case is less striking than for the HOSVD. Depending on the context, one can choose the appropriate approach to address a certain problem.

7.1. Linear mapping between higher-order spaces. In the same way as there exists an isomorphic link between matrix algebra and the algebra of linear vector mappings, a higher-order tensor can be regarded as a formal representation of a linear mapping between a matrix and a vector space, a matrix and a matrix space, a matrix and a higher-order tensor space, etc. For example, assuming a basis in the space of N_1th-order and N_2th-order tensors, a linear transformation of $\mathcal{B} \in \mathbb{C}^{J_1 \times J_2 \times \cdots \times J_{N_1}}$ to $\mathcal{C} \in \mathbb{C}^{I_1 \times I_2 \times \cdots \times I_{N_2}}$ can be represented by $\mathcal{A} \in \mathbb{C}^{I_1 \times \cdots I_{N_2} \times J_1 \times \cdots \times J_{N_1}}$, according to the element-wise equation

$$c_{i_1 i_2 \ldots i_{N_2}} = \sum_{j_1 j_2 \ldots j_{N_1}} a_{i_1 i_2 \ldots i_{N_2} j_1 j_2 \ldots j_{N_1}} b_{j_1 j_2 \ldots j_{N_1}}$$

for a particular choice of N_1 summation indices.

Obviously, a tensorial SVD equivalent is the SVD of this linear mapping. We notice a link with the HOSVD: $\mathbf{U}^{(n)}$, $\mathbf{\Sigma}^{(n)}$, and $\mathbf{V}^{(n)}$ in (12) are the SVD components of $\mathcal{A}$, interpreted as a linear mapping from $\mathbb{C}^{I_{n+1} \times I_{n+2} \times \cdots \times I_N \times I_1 \times I_2 \times \cdots I_{n-1}}$ to $\mathbb{C}^{I_n}$, defined by summation over the indices $i_{n+1}, i_{n+2}, \ldots, i_N, i_1, i_2, \ldots, i_{n-1}$.

7.2. Optimal tensor diagonalization by unitary transformations. The fact that a generic higher-order tensor cannot be diagonalized by unitary transformations could be remedied by weakening the condition of the diagonality of the core tensor to a condition of "maximal diagonality." Formally, given a higher-order tensor $\mathcal{A} \in \mathbb{C}^{I_1 \times I_2 \times \cdots \times I_N}$, this new problem consists of the determination of unitary matrices $\mathbf{U}^{(1)} \in \mathbb{C}^{I_1 \times I_1}$, $\mathbf{U}^{(2)} \in \mathbb{C}^{I_2 \times I_2}$, $\ldots$, $\mathbf{U}^{(N)} \in \mathbb{C}^{I_N \times I_N}$ such that the least-squares diagonality criterion $\sum_i |s_{ii \ldots i}|^2$ is optimized, where the core tensor $\mathcal{S}$ is still defined by (5).

For problems dealing with tensors that can theoretically be diagonalized (see also section 6), forcing the diagonal structure may be more robust (at the expense of a

higher computational cost) than the computation of the HOSVD, in which the deviation from diagonality is simply considered as a perturbation effect. For algorithms, we refer to [11, 16], exploring Jacobi-type procedures.

7.3. Canonical decomposition. A major difference between matrices and higher-order tensors is that the HOSVD does not provide precise information about rank-related issues. With this respect, the "canonical decomposition" (CANDECOMP), or "parallel factors" model (PARAFAC) may be more informative [10, 24, 1, 31, 39].

DEFINITION 11 (CANDECOMP). *A canonical decomposition or parallel factors decomposition of a tensor* $\mathcal{A} \in \mathbb{C}^{I_1 \times I_2 \times \cdots \times I_N}$ *is a decomposition of* $\mathcal{A}$ *as a linear combination of a minimal number of rank-*1 *terms:*

$$\mathcal{A} = \sum_{r}^{R} \lambda_r \, U_r^{(1)} \circ U_r^{(2)} \circ \cdots \circ U_r^{(N)}. \tag{32}$$

This decomposition is important for applications, as the different rank-1 terms can often be related to different "mechanisms" that have contributed to the higher-order tensor; in addition, sufficiently mild uniqueness conditions enable the actual computation of these components (without imposing orthogonality constraints, as in the matrix case).

However, the practical determination of a tensor rank is a much harder problem than the determination of its n-ranks, since it involves the tensor as a global quantity, rather than as a collection of n-mode vectors. A fortiori, the computation of the CANDECOMP is much harder than the computation of the HOSVD, and many problems remain to be solved. By way of illustration, the determination of the maximal rank value over the set of $(I_1 \times I_2 \times \cdots \times I_N)$-tensors is still an open problem in the literature (it is not bounded by $\min(I_1, I_2, \ldots, I_N)$). We refer to [4] for a tutorial on the current state-of-the-art. In addition, [12] gives an overview of some partial results obtained for symmetric tensors.

8. Conclusion. In this paper the problem of generalizing the SVD of matrices to higher-order tensors is investigated. The point of departure is that the multilinear equivalent should address the so-called n-mode vectors in a similar way as the SVD does for column and row vectors. It is shown that this implies that the role of the matrix of singular values has to be assumed by a tensor with the property of all-orthogonality. The resulting decomposition, which is referred to as the HOSVD, is always possible for real or complex Nth-order tensors, and when applied to matrices, it reduces—up to some trivial indeterminacies—to the matrix SVD. In the psychometric literature, the decomposition is known as the Tucker model.

From the fact that n-mode vectors and column (row) vectors play the same role in HOSVD, resp., SVD, it follows that the matrices of singular vectors can be computed from the sets of n-mode vectors in the same way as in the second-order case. In other words, the HOSVD of an Nth-order tensor boils down to N matrix SVDs. As a consequence, several matrix properties have a very clear higher-order counterpart. Uniqueness, link with EVD, directions and values of extremal oriented energy, etc. are investigated and explicit expressions for a first-order perturbation analysis are presented.

The link between tensor symmetry and HOSVD is also investigated: if a higher-order tensor is transformed into itself or into its complex conjugate by a permutation of the indices, then this symmetry is reflected by the HOSVD components. It turns

out that the most adequate way to analyze this relation is the decomposition of the permutation in a sequence of permutation cycles. When every pair of indices is connected by some cycle, a property which is denoted as pair-wise symmetry, then the different matrices of higher-order singular vectors are either equal or complex conjugated. In analogy to the HOEVD of a real symmetric or Hermitean matrix the HOEVD of a pair-wise symmetric higher-order tensor is defined. Its computation, for a pair-wise symmetric Nth-order $(I \times I \times \cdots \times I)$-tensor, reduces to the SVD of an $(I \times I^{N-1})$-matrix unfolding. It remains an open problem if and how the high symmetry of this matrix can be exploited to speed up the calculation of its SVD.

The link with the SVD of matrices makes the HOSVD the proper tool for the analysis of n-mode vector space properties. However, the fact that the core tensor is in general a full tensor, instead of being pseudodiagonal, results in two main differences with the matrix case. First, the HOSVD of a given tensor allows us to estimate its n-ranks, but these values give only a rough lower-bound on the rank; more in general, the HOSVD does not allow for an interpretation in terms of minimal expansions in rank-1 terms. This shortcoming is inherent in the structure of multilinear algebra; with respect to rank-related issues, the CANDECOMP may be informative. Second, discarding the smallest n-mode singular values does not necessarily lead to the best possible approximation with reduced n-rank values $(1 \leqslant n \leqslant N)$. However, if there is a large gap between the n-mode singular values, then the approximation can still be qualified as accurate; if necessary, it can be used as a good starting value for an additional optimization algorithm. Finally, we remark that in some applications one knows a priori that the core tensor is pseudodiagonal. As pseudodiagonality is a special case of all-orthogonality, analysis procedures may then be based on the HOSVD, up to perturbation effects. If a higher accuracy is mandatory, it might be preferable to determine unitary transformations that explicitly make the higher-order tensor as diagonal as possible, rather than all-orthogonal.

REFERENCES

[1] C.J. Appellof and E.R. Davidson, *Strategies for analyzing data from video fluoromatric monitoring of liquid chromatographic effluents*, Analytical Chemistry, 53 (1981), pp. 2053–2056.

[2] R.E. Bellman, *Matrix Analysis*, McGraw-Hill, NY, 1978.

[3] R.M. Bowen and C.-C. Wang, *Introduction to Vectors and Tensors. Linear and Multilinear Algebra*, Plenum Press, NY, 1980.

[4] R. Bro, *PARAFAC. Tutorial and applications*, Chemom. Intell. Lab. Syst., Special Issue 2nd Internet Conf. in Chemometrics (INCINC'96), 38 (1997), pp. 149–171.

[5] J.-F. Cardoso, *Eigen-structure of the fourth-order cumulant tensor with application to the blind source separation problem*, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Albuquerque, NM, 1990, pp. 2655–2658.

[6] J.-F. Cardoso, *Super-symmetric decomposition of the fourth-order cumulant tensor. Blind identification of more sources than sensors*, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Toronto, Canada, 1991, pp. 3109–3112.

[7] J.-F. Cardoso, *Fourth-order cumulant structure forcing. Application to blind array processing.* in Proceedings of the IEEE Signal Processing Workshop on Statistical Signal and Array Processing, 1992, pp. 136–139.

[8] J.-F. Cardoso and P. Comon, *Tensor-based independent component analysis*, in Signal Processing V: Theories and Applications, Proc. EUSIPCO-90, L. Torres, ed., Elsevier, Amsterdam, 1990, pp. 673–676.

[9] J.-F. Cardoso and P. Comon, *Independent component analysis, a survey of some algebraic methods*, in Proceedings of the IEEE International Symposium on Circuits and Systems, Atlanta, GA, 1996, pp. 93–96.

[10] J.D. Carroll and S. Pruzansky, *The CANDECOMP-CANDELINC family of models and methods for multidimensional data analysis*, in Research Methods for Multimode Data Analysis, H.G. Law, C.W. Snyder, J.A. Hattie, and R.P. McDonald, eds., Praeger, NY, 1984, pp. 372–402.
[11] P. Comon, *Independent component analysis, a new concept?*, Signal Process., Special Issue on Higher Order Statistics, 36 (1994), pp. 287–314.
[12] P. Comon and B. Mourrain, *Decomposition of quantics in sums of powers of linear forms*, Signal Process., Special Issue on Higher Order Statistics, 53 (1996), pp. 93–108.
[13] R. Coppi and S. Bolasco, eds., *Multiway Data Analysis*, Elsevier, Amsterdam, 1989.
[14] L. De Lathauwer, B. De Moor, and J. Vandewalle, *Blind source separation by higher-order singular value decomposition*, in Signal Processing VII: Theories and Applications, Proc. EUSIPCO-94, Edinburgh, UK, 1994, pp. 175–178.
[15] L. De Lathauwer, B. De Moor, and J. Vandewalle, *Independent component analysis based on higher-order statistics only*, in Proceedings of the IEEE Signal Processing Workshop on Statistical Signal and Array Processing, Corfu, Greece, 1996, pp. 356–359.
[16] L. De Lathauwer, B. De Moor, and J. Vandewalle, *Blind source separation by simultaneous third-order tensor diagonalization*, in Signal Processing VIII: Theories and Applications, Proc. EUSIPCO-96, Trieste, Italy, 1996, pp. 2089–2092.
[17] L. De Lathauwer, B. De Moor, and J. Vandewalle, *Dimensionality reduction in higher-order-only ICA*, in Proceedings of the IEEE Signal Processing Workshop on HOS, Banff, Alberta, Canada, 1997, pp. 316–320.
[18] L. De Lathauwer, P. Comon, B. De Moor, and J. Vandewalle, *Higher-order power method—Application in independent component analysis*, in Proceedings of the International Symposium on Nonlinear Theory and Its Applications, Las Vegas, UT, 1995, pp. 91–96.
[19] L. De Lathauwer, *Signal Processing Based on Multilinear Algebra*, Ph.D. thesis, K.U. Leuven, E.E. Dept.-ESAT, Belgium, 1997.
[20] B. De Moor and S. Boyd, *Analytic Properties of Singular Values and Vectors*, Tech. Report 1989-28, SISTA, E.E. Dept.-ESAT, K.U. Leuven, 1989.
[21] E.F. Deprettere, Ed., *SVD and Signal Processing. Algorithms, Applications and Architectures*, North-Holland, Amsterdam, 1988.
[22] G.H. Golub and C.F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[23] W. Greub, *Multilinear Algebra*, Springer-Verlag, NY, 1978.
[24] R.A. Harshman and M.E. Lundy, *The PARAFAC model for three-way factor analysis and multidimensional scaling*, in Research Methods for Multimode Data Analysis, H.G. Law, C.W. Snyder, J.A. Hattie, and R.P. McDonald, eds., Praeger, NY, 1984, pp. 122–215.
[25] J. Hastad, *Tensor rank is NP-complete*, J. Algorithms, 11 (1990), pp. 644–654.
[26] D.C. Kay, *Theory and Problems of Tensor Calculus*, McGraw-Hill, New York, 1988.
[27] T. Kato, *A Short Introduction to Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1982.
[28] P.M. Kroonenberg, *Three-Mode Principal Component Analysis*, DSWO Press, Leiden, The Netherlands, 1983.
[29] J.B. Kruskal, *Rank, decomposition, and uniqueness for* 3*-way and* N*-way arrays*, in Multiway Data Analysis, R. Coppi and S. Bolasco, eds., North-Holland, Amsterdam, 1989, pp. 7–18.
[30] J.L. Lacoume, M.Gaeta, and P.O. Amblard, *From order* 2 *to HOS: New tools and applications*, in Signal Processing VI: Theories and Applications, Proc. EUSIPCO-92, J. Vandewalle and R. Boite, eds., Elsevier, Amsterdam, 1992, pp. 91–98.
[31] S.E. Leurgans, R.T. Ross, and R.B. Abel, *A decomposition for three-way arrays*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1064–1083.
[32] M. Marcus, *Finite Dimensional Multilinear Algebra*, Dekker, New York, 1975.
[33] P. McCullagh, *Tensor Methods in Statistics*, Chapman and Hall, London, UK, 1987.
[34] J.M. Mendel, *Tutorial on higher-order statistics (spectra) in signal processing and system theory: theoretical results and some applications*, Proc. IEEE, 79 (1991), pp. 278–305.
[35] M. Moonen and B. De Moor, eds., *SVD and Signal Processing*, III. *Algorithms, Applications and Architectures*, Elsevier, Amsterdam, 1995.
[36] D.G. Nathcott, *Multilinear Algebra*, Cambridge University Press, Cambridge, UK, 1984.
[37] C.L. Nikias, *Higher order spectra in signal processing*, in Signal Processing V: Theories and Applications, Proc. EUSIPCO-90, L. Torres, ed., Elsevier, Amsterdam, 1990, pp. 35–41.
[38] C.L. Nikias and J.M. Mendel, *Signal processing with higher-order spectra*, IEEE Signal Process. Mag., 10 (1993), pp. 10–37.
[39] P. Paatero, *A weighted non-negative least squares algorithm for three-way "PARAFAC" factor*

analysis, Chemom. Intell. Lab. Syst., Special Issue, 2nd Internet Conf. in Chemometrics (INCINC'96), 38 (1997), pp. 223–242.

[40] P.A. Regalia and S.K. Mitra, *Kronecker products, unitary matrices and signal processing applications*, SIAM Rev., 31 (1989), pp. 586–613.

[41] A. Swami and G. Giannakis, *Bibliography on HOS*, Signal Process., 60 (1997), pp. 65–126.

[42] L.R. Tucker, *The extension of factor analysis to three-dimensional matrices*, in Contributions to Mathematical Psychology, H. Gulliksen and N. Frederiksen, eds., Holt, Rinehart & Winston, New York, 1964, pp. 109–127.

[43] L.R. Tucker, *Some mathematical notes on three-mode factor analysis*, Psychometrika, 31 (1966), pp. 279–311.

[44] R. Vaccaro, ed., *SVD and Signal Processing*, II. *Algorithms*, *Applications and Architectures*, Elsevier, Amsterdam, 1991.

[45] J. Vandewalle and D. Callaerts, *Singular value decomposition: A powerful concept and tool in signal processing*, in Mathematics in Signal Processing II, Clarendon Press, Oxford, 1990, pp. 539–559.

[46] J. Vandewalle and B. De Moor, *On the use of the singular value decomposition in identification and signal processing*, in NATO ASI Series: Numerical Linear Algebra, Digital Signal Processing and Parallel Algorithms, F70 (1991), pp. 321–360.

[47] P. Van Overschee and B. De Moor, *Subspace Identification for Linear Systems. Theory*, *Implementation*, *Applications*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 4, pp. 1279–1299
© 2000 Society for Industrial and Applied Mathematics

AN AUGMENTED CONJUGATE GRADIENT METHOD FOR SOLVING CONSECUTIVE SYMMETRIC POSITIVE DEFINITE LINEAR SYSTEMS*

JOCELYNE ERHEL† AND FRÉDÉRIC GUYOMARC'H†

Abstract. Many scientific applications require one to solve successively linear systems $Ax = b$ with different right-hand sides b and a symmetric positive definite matrix A. The conjugate gradient method applied to the first system generates a Krylov subspace which can be efficiently recycled thanks to orthogonal projections in subsequent systems. A modified conjugate gradient method is then applied with a specific initial guess and initial descent direction and a modified descent direction during the iterations. This paper gives new theoretical results for this method and proposes a new version. Numerical experiments show the efficacy of our method even for quite different right-hand sides.

Key words. conjugate gradient, Krylov subspace, orthogonal projection

AMS subject classifications. 65F10, 65F35

PII. S0895479897330194

1. Introduction.

In this paper we are concerned with the solution of several symmetric linear systems of the form

$$Ax^{(i)} = b^{(i)},$$

where A is a $N \times N$ symmetric positive definite matrix. If A is not too large, a direct method is competitive since it only requires one factorization for several triangular solves. However, if A is large, an iterative method such as the conjugate gradient (CG) or the preconditioned conjugate gradient (PCG) is usually necessary because of memory constraints [4, 25]. The aim of this paper is to speed up the convergence of CG or PCG in order to reduce the CPU cost. The method should keep memory requirements low and vectorize or parallelize easily. Here we assume that the different right-hand sides $b^{(i)}$ are computed sequentially. The main idea is to solve the first system by CG and to recycle in the subsequent systems the Krylov subspace $\mathcal{K}_m(A, s_0)$ generated in the seed system.

1.1. Some examples.

This situation arises, for instance, when a new right-hand side depends upon previous solutions. Many applications lead to such problems.

- The asymptotic development of solutions of thick plates problems requires solving successive symmetric positive definite linear systems. The matrix is the same, but each solution at a given order of development depends upon the previous order [15].
- Nonlinear systems are often solved by an approximate Newton method where the approximate Jacobian matrix is frozen during several iterations. Each iteration then requires solving a linear system with a matrix unchanged and with a varying right-hand side which depends upon the previous iteration. Such nonlinear systems appear, for example, in implicit discretizations of differential equations.

*Received by the editors November 14, 1997; accepted for publication (in revised form) by D. P. O'Leary September 10, 1999; published electronically April 18, 2000.
http://www.siam.org/journals/simax/21-4/33019.html
†INRIA–Rennes, IRISA, Campus Universitaire de Beaulieu, 35042 Rennes Cedex, France (Jocelyne.Erhel@inria.fr, Frederic.Guyomarch@inria.fr, http://www.irisa.fr/aladin/).

- To compute generalized eigenvalues close to a given value σ, one often uses the so-called shift-and-invert method. This implies solving at each iteration of the eigenvalue solver linear systems of the form $(A - \sigma B)x^{(i)} = b^{(i)}$.
- When nonlinear least squares problems are solved by gradient methods, each iteration requires the computation of the gradient or a preconditioned gradient. If the preconditioning is a given and fixed linear operator, then computing the preconditioned gradient will involve solving a new linear system with a fixed matrix (the preconditioning matrix) and a varying right-hand side. This example arises in a wave-scattering problem studied in [3], where the gradient is preconditioned by a Laplacian.

1.2. Previous work. This problem of solving successive linear systems has been studied by several authors.

When the different right-hand sides are known a priori, then block CG methods are efficient [20]. In [6], block CG is combined with a seed projection method which generates a Krylov subspace for the so-called seed system and projects the residuals of other systems onto this subspace. Similar ideas are developed in [13, 21, 28]. This method is also extended to varying matrices [5].

Here we assume that the different right-hand sides $b^{(i)}$ are computed sequentially. A first idea is to use previous systems to derive an initial guess for the current system. In [12], the current right-hand side is A-projected onto the subspace spanned by the l previous approximate solutions, where l is a user-chosen parameter. When A is symmetric and possibly indefinite, a similar idea to the seed projection is introduced in [22] and further analyzed in [24]. This method computes an initial approximation by using a projection onto the Krylov subspace $\mathcal{K}_m(A, s_0)$. Saad [24] shows that this so-called restarted Lanczos–Galerkin method is in some sense equivalent to a block Lanczos method. Van der Vorst [30] gives a more stable formulation to compute the initial guess and extends the method to problems of the form $f(A)x = b$.

A second idea is to use information from previous systems in the iterations. In [22] and [24], the classical Lanczos procedure is modified so that the new Lanczos vector is orthogonal to $\mathcal{K}_m(A, s_0)$. This modified Lanczos method starts with the initial approximation defined previously. It appears sufficient to orthogonalize the current Lanczos vector against the last vector in $\mathcal{K}_m(A, s_0)$. This method is no longer equivalent to a block Lanczos method.

If the matrix is symmetric positive definite, the same approach can be applied to the CG method. It has the usual advantage of CG compared to Lanczos implementation. Indeed, the projected matrix is implicitly factored by recurrences at each iteration rather than explicitly at convergence. This modified CG method is applied in [23, 11, 2] in a domain decomposition framework. The small interface problem is solved by CG and several systems are solved successively because the initial problem is nonlinear. This modified CG method implements the two ideas presented above, since it uses the first Krylov subspace $\mathcal{K}_m(A, s_0)$ by computing an initial approximation and by forcing orthogonality against $\mathcal{K}_m(A, s_0)$ during iterations. But here, the current direction is orthogonalized against all previous directions to avoid loss of orthogonality due to rounding errors. Numerical experiments show a significant improvement over the classical CG method.

This idea is somewhat related to deflation methods. A deflation has been applied to CG in [19] for boundary value problems, and more recently in [26] for general symmetric positive definite systems. Similar ideas have been also applied to nonsymmetric problems using the GMRES algorithm [17, 7, 8, 1, 10, 9, 18, 16].

1.3. Contribution. In this paper, we present two algorithms called InitCG and AugCG. They are both CG forms for use when the symmetric matrix is positive definite and are derived from Lanczos-based algorithms described in [24].

InitCG computes an initial approximate solution as in [22] and [24] but uses the stable formulation proposed in [30] and uses the CG method instead of the Lanczos method. As in [24] for Lanczos, we prove that during the first iterations InitCG is in some sense equivalent to a block CG algorithm. Our numerical experiments show that InitCG is efficient if the second right-hand side is close to the first one.

AugCG includes a Krylov correction at each iteration. It is similar to the modified Lanczos method of [24] but here too uses the CG method. It is also similar to the modified CG method used in [11] for domain decomposition methods but uses a recurrence property rather than a full orthogonalization.

The main contribution of the paper is to prove an error bound giving the asymptotic rate of convergence of our AugCG algorithm. As far as we know, this result is new, although some lines of the proof can be found in [19].

Here we use the framework from [14] to design a balanced projection method defined by a solution space condition and a Petrov–Galerkin condition. Thanks to this formalism, it is straightforward to prove the convergence of the method. Then we prove that this balanced projection method possesses a four-term recurrence, so that the overhead introduced is very small.

Our AugCG method introduces at each step k the subspace $\mathcal{K}_{m,k}(A, s_0, r_0) = \mathcal{K}_m(A, s_0) + Span\{r_0, r_1, \ldots, r_k\}$, where $r_0, r_1, \ldots, r_k$ are the previous residuals. Our objective is to characterize this subspace, which is not a Krylov subspace. To achieve this, we introduce the matrix H of the A-orthogonal projection onto $\mathcal{K}_m(A, s_0)^{\perp_A}$. The main result, which is new, is that $Span\{r_0, r_1, \ldots, r_k\}$ is the Krylov subspace $\mathcal{K}_k(H^*AH, r_0)$. This allows us to derive an error bound involving the condition number of the matrix H^*AH.

To prove this result, we use a polynomial formalism with polynomials in the two variables A and H. We give another proof, which was first outlined in [19] and developed in [26], based on the fact that AugCG is equivalent to the CG method preconditioned by the symmetric positive semidefinite matrix HH^* and with the same initial approximation. Since the preconditioning is here singular, the usual results on PCG do not directly apply but must be proved in this context.

Our numerical experiments illustrate the behavior of AugCG algorithm. We observe that AugCG has a better asymptotic convergence than CG or InitCG. We did not investigate the numerical stability of AugCG. It may happen that full orthogonalization is required, as in domain decomposition methods.

The paper is organized as follows. Section 2 reviews the properties of CG needed in the new algorithms. Section 3 defines the algorithm InitCG and studies its convergence. Section 4 presents the algorithm AugCG and studies its convergence thanks to a polynomial formalism and to a preconditioning formalism. Section 5 deals with some practical considerations, such as complexity, memory requirements, numerical stability, and preconditioning. Finally, section 6 presents our numerical results and section 7 gives concluding remarks.

1.4. Notation. We introduce some notation which will be useful throughout the paper. We denote (x, y) the scalar product between the two vectors x and y of $\mathbb{R}^N$; the transpose of the matrix B is denoted by B^*. $Span(B)$ denotes the subspace spanned by the column vectors of B. $\mathbb{R}[X]$ denotes the set of polynomials in one variable; $\mathbb{R}_j[X]$ denotes the set of polynomials in one variable of degree at most j;

$\mathbb{R}\langle X, Y\rangle$ denotes the set of polynomials in two noncommutative variables.

2. Background. Let A be a symmetric positive definite matrix of $\mathbb{R}^{N,N}$ and let

$$Ay = c, \tag{2.1}$$

$$Ax = b \tag{2.2}$$

be two successive linear systems to solve.

We assume that a classical CG algorithm is used to solve the first system (2.1). We do the analysis for CG, but the same applies for PCG. To set the notation, the algorithm is given below.

ALGORITHM 1. CG1.

* iterative solution of $Ay = c$;
* Initialization;
choose y_0;
$s_0 = c - Ay_0$;
$w_0 = s_0$;
* Iteration;
for $j = 0, 1, \ldots$ **until** convergence **do**
  $\gamma_j = (s_j, s_j)/(w_j, Aw_j)$;
  $y_{j+1} = y_j + \gamma_j w_j$;
  $s_{j+1} = s_j - \gamma_j A w_j$;
  $\delta_{j+1} = (s_{j+1}, s_{j+1})/(s_j, s_j)$;
  $w_{j+1} = s_{j+1} + \delta_{j+1} w_j$;
end do

Let $S_j = (s_0, s_1, \ldots, s_j)$ and $W_j = (w_0, w_1, \ldots, w_j)$ be the set of residuals and conjugate directions generated. Recall that

$$S_j^* S_j = \Delta_j,\ W_j^* A W_j = D_j,\ Span(S_j) = Span(W_j) = \mathcal{K}_j(A, s_0), \tag{2.3}$$

where Δ_j and D_j are diagonal matrices and $\mathcal{K}_j(A, s_0)$ is the Krylov subspace of dimension $j + 1$ generated by the initial residual s_0.

We also recall the following result, which introduces the projections H_j and H_j^*.

DEFINITION 2.1. *Let $H_j = I - W_j D_j^{-1} (AW_j)^*$ be the matrix of the A-orthogonal projection onto $\mathcal{K}_j(A, s_0)^{\perp_A}$; $H_j^* = I - (AW_j) D_j^{-1} W_j^*$ is the A^{-1}-orthogonal projection onto $\mathcal{K}_j(A, s_0)^{\perp}$.*

THEOREM 2.2. *The residual s_j satisfies the relations $s_{j+1} = H_j^* s_0 \quad (j \geq 0)$ and $H_m^* s_j = s_j$ for $m < j$.*

Proof. $s_1 = H_0^* s_0$ since $w_0 = s_0$. We prove the first relation by induction. We have $s_{j+1} = s_j - \gamma_j A w_j$ and $(s_j, s_j) = (w_j, s_j) = (w_j, s_0)$ so that $s_{j+1} = H_{j-1}^* s_0 - Aw_j\ 1/(w_j, Aw_j)\ w_j^* s_0 = H_j^* s_0$.

For $m < j$, $W_m^* s_j = 0$ so that $H_m^* s_j = s_j$. ∎

3. InitCG algorithm. We now want to use this information to speed up the solution of the second system (2.2). We will use m vectors, with m given, and from now on omit the index m in S, W, D, H, H^*.

A first idea, introduced in [22, 24, 30] for the Lanczos method, is to choose an initial guess x_0 such that the initial residual $r_0 = b - Ax_0$ is orthogonal to the Krylov subspace $\mathcal{K}_m(A, s_0)$. We thus want to enforce the orthogonality conditions

$$W^* r_0 = 0. \tag{3.1}$$

PROPOSITION 3.1. *Let x_{-1} be any initial approximate solution and $r_{-1} = b - Ax_{-1}$. To guarantee* (3.1)*, we must choose*

$$x_0 = x_{-1} + WD^{-1}W^*r_{-1}, \quad r_0 = b - Ax_0 = H^*r_{-1}. \tag{3.2}$$

Proof. $W^*r_0 = 0$ is equivalent to $r_0 \in Im(H^*)$; in other words

$$W^*r_0 = 0 \Leftrightarrow \exists u, \quad r_0 = H^*u.$$

Thus (3.1) implies that $r_0 = b - Ax_0 = u - AWD^{-1}W^*u$. Multiplying by A^{-1}, we get $x_0 = A^{-1}(b-u) + WD^{-1}W^*u$. Let $x_{-1} = A^{-1}(b-u)$, then $u = b - Ax_{-1}$, so that

$$x_0 = x_{-1} + WD^{-1}W^*r_{-1}, \text{ with } r_{-1} = b - Ax_{-1}.$$

Conversely, if x_0 satisfies (3.2), then $r_0 = H^*r_{-1}$. ∎

As far as x_{-1} is concerned, we can choose, for example, $x_{-1} = y_0$. Then $r_{-1} = b - Ax_{-1} = b - c + s_0$ and $r_0 = H^*r_{-1} = H^*(b-c) + s_m$.

We can also choose $x_{-1} = y_j$, where $j > m$ is the number of iterations in the first system. Then $r_{-1} = b - Ax_{-1} = b - c + s_j$ and $r_0 = H^*r_{-1} = H^*(b-c) + s_j$. This second solution is interesting because $\|s_j\|_{A^{-1}} \leq \|s_m\|_{A^{-1}}$. Anyway, the quality of the initial solution will depend upon $\|H^*(b-c)\|_{A^{-1}}$.

We then use the CG algorithm as usual, starting with $p_0 = r_0$. We get the following algorithm, which we call InitCG.

ALGORITHM 2. INITCG.

* iterative solution of $Ax = b$;
* Initialization;
choose x_{-1};
$r_{-1} = b - Ax_{-1}$;
$x_0 = x_{-1} + WD^{-1}W^*r_{-1}$;
$r_0 = b - Ax_0$;
$p_0 = r_0$;
* Iteration;
for $k = 0, 1, \ldots$ **until** convergence **do**
  $\alpha_k = (r_k, r_k)/(p_k, Ap_k)$;
  $x_{k+1} = x_k + \alpha_k p_k$;
  $r_{k+1} = r_k - \alpha_k Ap_k$;
  $\beta_{k+1} = (r_{k+1}, r_{k+1})/(r_k, r_k)$;
  $p_{k+1} = r_{k+1} + \beta_{k+1}p_k$;
end do

InitCG is still a CG method. However, during the first $m/2$ iterations, this algorithm is equivalent to a block CG method.

THEOREM 3.2. *Let n be the number of iterations in the second resolution. As long as $n \leq m/2$, the CG algorithm applied to the linear system $Ax = b$, started with the initial guess x_0 and initial residual r_0 given by* (3.2)*, where x_{-1} is any vector, is mathematically equivalent to the block CG algorithm with block size of* 2 *started with the block $[s_0, r_0]$.*

Of course, this result has no practical interest since r_0 is not known before W is built, so that the block CG should start after the first system has been solved. Anyway, it has a theoretical interest. For the first $m/2$ iterations, we get an accelerated scheme compared to a classical CG method. To prove this theorem, we follow and detail the proof given in [24] for the Lanczos method.

Proof. Let $S_j = (s_0, s_1, \ldots, s_j)$ and $R_k = (r_0, r_1, \ldots, r_k)$. We first show that $Span(S_j)$ is orthogonal to $Span(R_k)$ for all (j,k) such that $j + k \leq m$.

Let s and r be two vectors of $Span(S_j)$ and $Span(R_k)$, respectively. Thanks to (2.3), there exist a polynomial $Q \in \mathbb{R}_j[X]$ and a polynomial $P \in \mathbb{R}_k[X]$ such that $s = Q(A)s_0$ and $r = P(A)r_0$. Therefore,

$$s^* r = (Q(A)s_0)^*(P(A)r_0)$$

and we only have to prove the orthogonality for monomials such as $A^i s_0$ and $A^l r_0$ with $i \leq j$ and $l \leq k$. This property comes from the choice of the initial residual r_0; indeed, since A is symmetric,

$$(A^i s_0)^*(A^l r_0) = (A^{i+l} s_0)^* r_0.$$

Now, since $i + l \leq j + k \leq m$, $A^{i+l} s_0 \in \mathcal{K}_m(A, s_0) = Span(W)$, and by construction $W^* r_0 = 0$, so we conclude that $(A^i s_0)^*(A^i r_0) = 0$ and consequently $s^* r = 0$.

Now, we consider a block CG algorithm started with the block $[s_0, r_0]$, with r_0 defined by (3.2).

We denote by $[s_k, r_k]$ the block of size 2 of residuals and by $[w_k, p_k]$ the block of size 2 of descent directions. The block CG algorithm, with no preconditioning, gives

$$[s_{k+1}, r_{k+1}] = [s_k, r_k] - A[w_k, p_k]\eta_k,$$

where

$$\eta_k = ([w_k, p_k]^* A[w_k, p_k])^{-1}[s_k, r_k]^*[s_k, r_k].$$

As long as $k + 1 \leq m/2$, so that $2k + 1 \leq m$, we have $w_k^* A p_k = p_k^* A w_k = 0$ since $Aw_k \in Span(S_{k+1})$ and $p_k \in Span(R_k)$. Hence

$$[w_k, p_k]^* A[w_k, p_k] = \begin{pmatrix} w_k^* A w_k & 0 \\ 0 & p_k^* A p_k \end{pmatrix}.$$

Also $s_k^* r_k = 0$ since $s_k \in Span(S_k)$ and $r_k \in Span(R_k)$, so

$$[s_k, r_k]^*[s_k, r_k] = \begin{pmatrix} s_k^* s_k & 0 \\ 0 & r_k^* r_k \end{pmatrix}.$$

Hence we get

$$\eta_k = \begin{pmatrix} s_k^* s_k / w_k^* A w_k & 0 \\ 0 & r_k^* r_k / p_k^* A p_k \end{pmatrix} = \begin{pmatrix} \gamma_k & 0 \\ 0 & \alpha_k \end{pmatrix}$$

and

$$\begin{cases} s_{k+1} = s_k - \gamma_k A w_k, \\ r_{k+1} = r_k - \alpha_k A p_k, \end{cases}$$

which is exactly what is given by Algorithms 1 and 2. The proof is similar for the computation of $[w_k, p_k]$. ∎

4. AugCG algorithm.

4.1. Construction of AugCG. Another way to improve the convergence is to keep the orthogonality condition throughout the iterations. This method was developed for the Lanczos algorithm in [22, 24]. It was also defined for the CG algorithm in [19, 23, 11].

We still denote by x_k the current approximate solution, by r_k the current residual $r_k = b - Ax_k$, by $e_k = x - x_k$ the current error where x is the exact solution, and by p_k the current descent direction.

The idea here is to build the method with two subspaces: the Krylov subspace $\mathcal{K}_m(A, s_0)$ already computed and the subspace $Span(r_0, r_1, \ldots, r_k)$. A condition to satisfy is that the residual r_{k+1} must be orthogonal to both subspaces and that the descent direction p_{k+1} must be A-conjugate to both subspaces. These orthogonality conditions will appear to be satisfied by a recurrence relation.

We thus start with x_0 and r_0 as defined by (3.2). We must start with a descent direction A-orthogonal to $Span(W)$, so that we must choose

$$p_0 = (I - WD^{-1}(AW)^*)r_0 = Hr_0. \tag{4.1}$$

Our method is a projection method onto a subspace which is not a Krylov subspace. This subspace is defined by

$$\mathcal{K}_{m,k}(A, s_0, r_0) = \mathcal{K}_m(A, s_0) + Span(r_0, r_1, \ldots, r_k) = Span(W) + Span(R_k). \tag{4.2}$$

Our projection method is then defined by the solution space condition

$$x_{k+1} - x_k \in \mathcal{K}_{m,k}(A, s_0, r_0) \tag{4.3}$$

and by the Petrov–Galerkin condition

$$(r_{k+1}, z) = 0 \quad \text{for all } z \in \mathcal{K}_{m,k}(A, s_0, r_0). \tag{4.4}$$

The algorithm, which we call augmented conjugate gradient (AugCG), defined by (3.2), (4.1), (4.2), (4.3), and (4.4), is a balanced projection method (BPM) based on the matrix $B = A$ as defined in [14]. Since A is assumed to be HPD, we therefore have immediately the following result.

THEOREM 4.1. *Let A be a symmetric positive definite matrix. The algorithm AugCG applied to the linear system $Ax = b$ will not break down at any step. The approximate solution x_k is the unique minimizer of the error $\|e_k\|_A$ over the solution space $\mathcal{K}_{m,k}(A, s_0, r_0)$ and there exists $\epsilon > 0$ independent of e_0 such that for all k*

$$\|e_k\|_A \leq (1 - \epsilon)\ \|e_{k-1}\|_A.$$

Proof. See Theorems 2.4, 2.6, and 2.7 in [14]. ∎

We will now prove that AugCG can be implemented with a four-term recurrence. The solution space condition (4.3) will be satisfied as usual in the CG. The Petrov–Galerkin condition (4.4) will be satisfied simply thanks to three orthogonality conditions: the current residual r_{k+1} is orthogonal to r_k, and the current descent direction p_{k+1} is A-conjugate to p_k and to the last vector w_m. This recurrence means that the complexity of the inner loop is similar to the classical CG. Our new method adds merely one dot-product and one vector update at each iteration. We will prove that the following algorithm implements the projection method AugCG.

ALGORITHM 3. AUGCG.
* iterative solution of $Ax = b$;
* Initialization;
choose x_{-1};
$r_{-1} = b - Ax_{-1}$;
$x_0 = x_{-1} + WD^{-1}W^* r_{-1}$;
$r_0 = b - Ax_0$;
$p_0 = (I - WD^{-1}(AW)^*)r_0$;
* Iteration;
for $k = 0, 1, \ldots$ **until** convergence **do**
$\alpha_k = (r_k, r_k)/(p_k, Ap_k)$;
$x_{k+1} = x_k + \alpha_k p_k$;
$r_{k+1} = r_k - \alpha_k Ap_k$;
$\beta_{k+1} = (r_{k+1}, r_{k+1})/(r_k, r_k)$;
$\mu_{k+1} = (r_{k+1}, Aw_m)/(w_m, Aw_m)$;
$p_{k+1} = r_{k+1} + \beta_{k+1} p_k - \mu_{k+1} w_m$;
end do

As shown in Algorithm 3, we introduce a new scalar μ_{k+1} and a new update for the descent direction p_{k+1}. The initialization step guarantees the relations (3.2) and (4.1).

THEOREM 4.2. *Algorithm* 3 *implements the BPM AugCG defined by* (3.2), (4.1), (4.2), (4.3), *and* (4.4). *More precisely,*

$$z^* r_{k+1} = 0 \quad \text{and} \quad z^* Ap_{k+1} = 0 \quad \textit{for all } z \in \mathcal{K}_{m,k}(A, s_0, r_0).$$

Proof. $p_0 \in Span(W) + Span(R_0)$ and by induction $p_k \in Span(W) + Span(R_k)$ so that Algorithm 3 satisfies the solution space condition (4.3). Indeed,

$$\mathcal{K}_{m,k}(A, s_0, r_0) = Span(W) + Span(R_k) = Span(W) + Span(P_k).$$

As for the orthogonality conditions, we first prove by induction that $W^* r_k = 0$ and $W^* Ap_k = 0$. This is true for $k = 0$. Now we assume it is true at step k. So

$$W^* r_{k+1} = W^* r_k - \alpha_k W^* Ap_k = 0$$

and

$$W^* Ap_{k+1} = W^* A(r_{k+1} - \mu_{k+1} w_m) + \beta_{k+1} W^* Ap_k$$

so that we have to prove that $W^* A(r_{k+1} - \mu_{k+1} w_m) = 0$. To get this, we show that

$$Hr_{k+1} = r_{k+1} - \mu_{k+1} w_m.$$

Indeed,

$$Hr_{k+1} = r_{k+1} - \mu_{k+1} w_m - \sum_{j=1}^{j=m-1} \frac{(r_{k+1}, Aw_j)}{(w_j, Aw_j)} w_j.$$

For $j \leq m-1$, $Aw_j \in Span(w_0, \ldots, w_{j+1}) \subset Span(W)$ and $W^* r_{k+1} = 0$ so that $(r_{k+1}, Aw_j) = 0$. Consequently, $Hr_{k+1} = r_{k+1} - \mu_{k+1} w_m$ and $W^* Ap_{k+1} = 0$.

Now we have to prove by induction that $r_j^* r_{k+1} = 0$ and $p_j^* Ap_{k+1} = 0$ for all $j \leq k$.

This is true for $k = 0$; assuming it is true for k, then

$$r_j^* r_{k+1} = r_j^* r_k - \alpha_k r_j^* A p_k.$$

We first assume that $j \leq k-1$. Then $r_j^* r_k = 0$ by induction. Also $r_j \in Span(W) + Span(P_j)$ so that $r_j^* A p_k = 0$ by induction and by the property $W^* A p_k = 0$ already proved, so that $r_j^* r_{k+1} = 0$. For $j = k$, we have

$$r_k^* r_{k+1} = r_k^* r_k - \alpha_k r_k^* A p_k,$$

but $r_k^* A p_k = p_k^* A p_k - \beta_k p_{k-1}^* A p_k + \mu_k w_m^* A p_k = p_k^* A p_k$ so that by definition of α_k we get $r_k^* r_{k+1} = 0$.

So it remains to prove $p_j^* A p_{k+1} = 0$. We first assume that $j \leq k-1$. Then

$$p_j^* A p_{k+1} = p_j^* A r_{k+1} + \beta_{k+1} p_j^* A p_k - \mu_{k+1} p_j^* A w_m,$$

but $p_j^* A r_{k+1} = r_{k+1}^* A p_j = r_{k+1}^* (r_j - r_{j+1})/\alpha_j = 0$ by the result proved just above, $p_j^* A p_k = 0$ by induction, and $p_j^* A w_m = 0$ by the previous result, so that $p_j^* A p_{k+1} = 0$. For $j = k$ we have

$$p_k^* A p_{k+1} = r_{k+1}^* A p_k + \beta_{k+1} p_k^* A p_k - \mu_{k+1} p_k^* A w_m,$$

but $r_{k+1}^* A p_k = r_{k+1}^* (r_k - r_{k+1})/\alpha_{k+1} = -r_{k+1}^* r_{k+1}/\alpha_{k+1}$ and $p_k^* A w_m = 0$, so that by definition of α_{k+1} and β_{k+1} we get $p_k^* A p_{k+1} = 0$.

This completes the proof. ∎

4.2. Polynomial formalism. A polynomial version of the AugCG algorithm will enable us to characterize the subspace $\mathcal{K}_{m,k}$. We first express r_k and p_k with polynomials in two variables, A and H. More precisely, we have the following result.

THEOREM 4.3. *For $k \geq 0$, there exist two polynomials in two variables $\Phi_k(X,Y)$ and $\Psi_k(X,Y)$ in $\mathbb{R}\langle X,Y\rangle$ such that $r_k = \Phi_k(A,H)r_0$ and $p_k = \Psi_k(A,H)r_0$.*

Furthermore $\Phi_0(X,Y) = 1$, $\Psi_0(X,Y) = Y$, and we have the following recurrence relations in the polynomial space $\mathbb{R}\langle X,Y\rangle$:

$$\begin{aligned} \Phi_{k+1}(X,Y) &= \Phi_k(X,Y) - \alpha_k X \Psi_k(X,Y), \\ \Psi_{k+1}(X,Y) &= Y\Phi_{k+1}(X,Y) + \beta_{k+1}\Psi_k(X,Y). \end{aligned}$$

Proof. Since $p_0 = Hr_0$, we define immediately $\Phi_0(X,Y) = 1$ and $\Psi_0(X,Y) = Y$. The recurrences are deduced from Algorithm 3: it is easy to rewrite the definition of r_{k+1} as $\Phi_{k+1}(A,H)r_0 = \Phi_k(A,H)r_0 - \alpha_k A\Psi_k(A,H)r_0$ which gives the recurrence relation for Φ_k. Recalling that $Hr_{k+1} = r_{k+1} - \mu_{k+1}w_m$ we rewrite the definition of p_{k+1} as

$$\Psi_{k+1}(A,H)r_0 = H\Phi_{k+1}(A,H)r_0 + \beta_{k+1}\Psi_k(A,H)r_0$$

which gives the recurrence relation for Ψ_k. ∎

We can simplify the polynomial expression by introducing polynomials in only one variable, XY. Indeed, we have the following result.

THEOREM 4.4. *For $k \geq 0$, there exist polynomials of degree k in one variable $\Xi_k(Z)$ and $\Omega_k(Z)$ in $\mathbb{R}_k[Z]$ such that*

$$\Phi_k(X,Y) = \Xi_k(XY) \quad \textit{and} \quad \Psi_k(X,Y) = Y\Omega_k(XY). \tag{4.5}$$

Proof. We will prove the theorem by induction: It is true for $k = 0$, since $\Phi_0(X,Y) = 1$ and $\Psi_0(X,Y) = Y$. Assuming it is true at step k, then

$$\begin{aligned}
\Phi_{k+1}(X,Y) &= \Phi_k(X,Y) - \alpha_k X\Psi_k(X,Y) \\
&= \Xi_k(XY) - \alpha_k XY\Omega_k(XY) \\
&= \Xi_{k+1}(XY), \\
\Psi_{k+1}(X,Y) &= Y\Phi_{k+1}(X,Y) + \beta_{k+1}\Psi_k(X,Y) \\
&= Y\Xi_{k+1}(XY) + \beta_{k+1}Y\Omega_k(XY) \\
&= Y(\Xi_{k+1}(XY) + \beta_{k+1}\Omega_k(XY)) \\
&= Y\Omega_{k+1}(XY).
\end{aligned}$$

Clearly, by induction, $\Xi_{k+1}(Z)$ and $\Omega_{k+1}(Z)$ are of degree $k+1$. ∎

This simple polynomial formulation leads to a characterization of the subspace $Span(R_k)$ as a Krylov subspace. This result, which is new as far as we know, is important since it allows us to derive an asymptotic error bound.

THEOREM 4.5. *For $k \geq 0$, $Span(r_0, \ldots, r_k) = \mathcal{K}_k(H^*AH, r_0)$ and*

$$\mathcal{K}_{m,k}(A, s_0, r_0) = \mathcal{K}_m(A, s_0) \overset{\perp_A}{\oplus} \mathcal{K}_k(H^*AH, r_0).$$

Proof. Since the polynomial $\Xi_k(Z)$ is of degree k, polynomials $(\Xi_i(Z))_{i=0\ldots k}$ are independent and build a basis of $\mathbb{R}_k[Z]$. Therefore, since $r_k = \Xi_k(AH)r_0$,

$$Span(r_0, \ldots, r_k) = \mathcal{K}_k(AH, r_0).$$

It is easy to show that $H^*AH = AH$ so that by induction $(H^*AH)^i r_0 = (AH)^i r_0$ and

$$Span(r_0, \ldots, r_k) = \mathcal{K}_k(H^*AH, r_0).$$

The characterization of $\mathcal{K}_{m,k}$ follows immediately. ∎

We can use classical results of conjugate gradient theory: algorithm AugCG computes the minimum over $\mathcal{K}_{m,k}$ which is no larger than the minimum over $\mathcal{K}_k(H^*AH, r_0)$.

COROLLARY 4.6. *Let κ_1 be the condition number of H^*AH; the error at iteration k in algorithm AugCG satisfies*

$$\|e_k\|_A \leq 2\|e_0\|_A \left(\frac{\sqrt{\kappa_1} - 1}{\sqrt{\kappa_1} + 1}\right)^k.$$

This result means that the asymptotic rate of convergence of AugCG is the same as or better than that for the classical CG because the condition number κ_1 of H^*AH is no larger than the condition number κ_0 of A.

4.3. Preconditioned formalism. We briefly present another way to get the same result, which is developed in [19] for the orthodir version of CG and in [26] for any W. We still consider the projections H and H^*. We will show that algorithm AugCG is equivalent to CG preconditioned by the symmetric positive semidefinite matrix HH^* and started by (3.2).

Indeed, AugCG can be rewritten in the following form.

ALGORITHM 4. AUGCG—PRECONDITIONED FORM.

* iterative solution of $Ax = b$;
* Initialization;
choose x_{-1};
$r_{-1} = b - Ax_{-1}$;
$x_0 = x_{-1} + WD^{-1}W^*r_{-1}$;
$r_0 = b - Ax_0 = H^*r_{-1}$;
$z_0 = Hr_0 = HH^*r_0$;
$p_0 = z_0$;
* Iteration;
for $k = 0, 1, \ldots$ **until** convergence **do**
  $\alpha_k = (r_k, z_k)/(p_k, Ap_k)$;
  $x_{k+1} = x_k + \alpha_k p_k$;
  $r_{k+1} = r_k - \alpha_k Ap_k$;
  $z_{k+1} = Hr_{k+1} = HH^*r_{k+1}$;
  $\beta_{k+1} = (r_{k+1}, z_{k+1})/(r_k, z_k)$;
  $p_{k+1} = z_{k+1} + \beta_{k+1}p_k$;
end do

Compared to Algorithm 3, we introduced the matrix H using the fact that $H^*r_k = r_k$, $p_0 = Hr_0$, and $r_{k+1} - \mu_{k+1}w_m = Hr_{k+1}$; we also used the equality $(r_k, z_k) = (r_k, r_k)$, which comes from $(r_k, z_k) = (r_k, Hr_k) = (H^*r_k, r_k) = (r_k, r_k)$.

Therefore, algorithm AugCG is nothing else than PCG with the symmetric positive semidefinite preconditioning matrix HH^* and with the initial guess x_0. This is not classical since usually the preconditioning matrix is nonsingular.

As for any PCG, even with a singular preconditioner, we can prove by recurrence that $r_k \in \mathcal{K}_k(AH, r_0)$ and $p_k \in H\mathcal{K}_k(AH, r_0)$. These properties are stated in Theorem 4.4 in a polynomial form.

Though the preconditioner is semidefinite, the algorithm is well defined, converges and does not break down, because r_0 is chosen orthogonal to W. These properties are stated in theorem 3.2. Therefore, the theory on PCG applies, the algorithm can be rewritten as CG applied to $H^*AHy = H^*b$, and the asymptotic convergence rate is governed by the condition number of H^*AH. This result is stated in Corollary 4.6.

5. Practical considerations.

5.1. Complexity and memory requirements of InitCG and AugCG. We first consider operations count. The initial guess computation in InitCG and AugCG is the most time-consuming overhead. The computation of x_0 in InitCG and AugCG is a BLAS2 operation (projection onto a subspace of size m). The computation of p_0 in AugCG is also a BLAS2 operation.

On the other hand, iterations in InitCG have the same cost as iterations of CG and each iteration of AugCG adds merely one dot-product and one vector update (BLAS1 operations), both of which are very cheap. We emphasize that neither InitCG nor AugCG induces new matrix-vector products. Hence the global overhead introduced should be easily balanced by the reduction in the number of iterations. Moreover, all operations are of type BLAS2 or BLAS1 so they easily vectorize or parallelize.

Let us examine now the memory requirements. Here too the initial guess computation in InitCG and AugCG is the most expensive. It requires storing the m vectors W. In contrast, the iterations of AugCG need no more than two extra vectors (w_m

and Aw_m). Hence it seems reasonable to use secondary storage for W if the main memory is not large enough.

If we do not rely on secondary storage, the question of the choice of m arises. This will be analyzed in numerical experiments.

5.2. Numerical stability of InitCG and AugCG. Now, we must also analyze the effect of rounding errors. Quite often in Lanczos-type methods, rounding errors lead to a loss of orthogonality, which implies poor convergence or divergence. This situation is analyzed for modified Lanczos method in [22] and [30]. In InitCG and AugCG, we enforce the initial orthogonality conditions (3.1). In AugCG, we also enforce two orthogonality conditions during the iterations.

As far as (3.1) is concerned, we follow the scheme proposed in [30] for the initial residual r_0 which merely implements a modified Gram–Schmidt process. We extend this scheme to the initial descent direction p_0. This ensures numerical stability for both r_0 and p_0.

During iterations of AugCG, the new descent direction p_{k+1} is enforced to be A-orthogonal to the last vector in the previous system w_m and to the previous descent direction p_k. Here too we apply a modified Gram–Schmidt process by first A-orthogonalizing against w_m then A-orthogonalizing against p_k. Anyway, loss of orthogonality can appear during the iterations of AugCG, as observed in [26]. A full orthogonalization, as in [26] and [2], can remedy this phenomenon in most situations. We do not investigate full orthogonalization further here.

5.3. Preconditioned versions InitPCG and AugPCG. We also analyze the application of AugCG to PCG, which we call AugPCG, since in practice we will use a preconditioned version of the algorithm. We assume that PCG has been applied in the first system with the symmetric positive definite preconditioning matrix M^{-1}. So now

$$Span(W) = M^{-1}Span(S) = M^{-1}\mathcal{K}_m(AM^{-1}, s_0).$$

The method AugPCG is defined by the subspace

$$\mathcal{K}_{m,k}(A, s_0, r_0) = M^{-1}(Span(S) + Span(R_k)) = Span(W) + Span(P_k)$$

with the subspace condition (4.3) and the Petrov–Galerkin condition (4.4). The results proved above still hold, with the following characterization of $Span(R_k)$:

$$Span(R_k) = \mathcal{K}_k(AHM^{-1}, r_0).$$

Therefore, AugPCG can be viewed as CG preconditioned by both HH^* and M^{-1}, and we have the following implementation.

ALGORITHM 5. AUGPCG.

* iterative solution of $Ax = b$;
* save vectors w_j and Aw_j in secondary storage $j = 1, \ldots, m-1$;
* save vectors w_m and Aw_m in main memory;
* save quantities (w_j, Aw_j) in a vector of length m;
* Initialization;
choose x_{-1};
$x_0 = x_{-1}$;
$r_0 = b - Ax_0$;
for $j = 1$ **to** m **do**

$\sigma_j = (r_0, w_j)/(w_j, Aw_j)$;
$x_0 = x_0 + \sigma_j w_j$;
$r_0 = r_0 - \sigma_j A w_j$;
end do;
$z_0 = M^{-1} r_0$;
for $j = 1$ **to** m **do**
$z_0 = z_0 - (z_0, Aw_j)/(w_j, Aw_j) w_j$;
end do;
$p_0 = z_0$;
* Iteration;
for $k = 0, 1, \ldots$ **until** convergence **do**
$\alpha_k = (r_k, z_k)/(p_k, Ap_k)$;
$x_{k+1} = x_k + \alpha_k p_k$;
$r_{k+1} = r_k - \alpha_k A p_k$;
$z_{k+1} = M^{-1} r_{k+1}$;
$\mu_{k+1} = (z_{k+1}, Aw_m)/(w_m, Aw_m)$;
$z_{k+1} = z_{k+1} - \mu_{k+1} w_m$;
$\beta_{k+1} = (r_{k+1}, z_{k+1})/(r_k, z_k)$;
$p_{k+1} = z_{k+1} + \beta_{k+1} p_k$;
end do

InitPCG starts with the same x_0, r_0 but with $p_0 = M^{-1} r_0$. During iterations of InitPCG, $z_{k+1} = M^{-1} r_{k+1}$, the same formula as in PCG.

6. Numerical experiments. We present some numerical experiments which study the efficiency of our InitCG and AugCG methods. We report here results on some test problems; other examples can be found in [26].

We implemented the method in Matlab, so we give only convergence results, since CPU times would not give realistic comparisons. We solve a first system $Ay = c$, then a second system $Ax = b$. We give numerical experiments for different matrices A.

In all our examples, we choose

$$c = Ae, \text{ where } e = (1, 1, \ldots, 1),$$
$$y_0 = 0, \text{ hence } s_0 = c,$$
$$x_{-1} = y_j, \text{ where } j \text{ is the total number of iterations done in the first system,}$$
$$x_0 = x_{-1} + WD^{-1}W^* r_{-1}.$$

We then study two cases for the right-hand side b. In the first case, $b^{(1)}$ is "close" to c:

$$b^{(1)} = \nu c + \epsilon e + AWe,$$

so that $H^* b^{(1)} = \nu H^* c + \epsilon H^* e = \nu s_m + \epsilon H^* e$ and

$$r_0^{(1)} = (\nu - 1) s_m + s_j + \epsilon H^* e.$$

In the second case, $b^{(2)}$ is "far" from c:

$$b^{(2)} = e,$$

so that

$$r_0^{(2)} = H^*(e - Ae) + s_j.$$

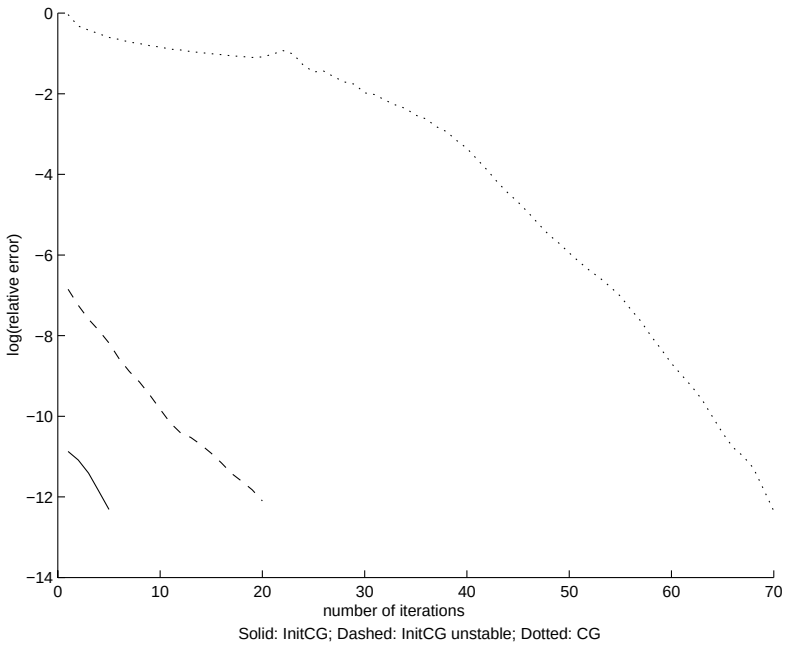


FIG. 6.1. *Results for the Laplacian matrix of order* 900 *with right-hand side* $b^{(1)}$ *(Example* 1*).*

The stopping criterion for the second system is

$$\|r_k\| \leq tol \quad \|b\|,$$

where *tol* is specified for each example.

All figures plot the relative residuals $\|r_k\|/\|b\|$ versus the number of iterations.

Figure 6.1 compares CG, InitCG-unstable, and InitCG. Figures 6.2–6.6 compare CG, InitCG, and AugCG or PCG, InitPCG, and AugPCG. In Figures 6.2–6.6,

- the dotted line plots CG or PCG started with $x_{-1}, r_{-1} = b - Ax_{-1}$,
- the dashed line plots InitCG or InitPCG started with $x_0, r_0 = b - Ax_0 = H^* r_{-1}$,
- and the solid line plots AugCG or AugPCG started with x_0.

6.1. Example 1: Efficiency and numerical stability of InitCG. We choose first the matrix of the Laplacian with Dirichlet boundary conditions discretized by a centered 5-point finite difference scheme on a square grid of size 32×32 so that the matrix is of size 900.

We take here $b = b^{(1)}$, with $\nu = 10$ and $\epsilon = 10^{-2}$. We choose $tol = 10^{-12}$. Here $m = 65$ and $j = 68$, which corresponds to $\|s_j\| \leq 10^{-12}\|c\|$.

We compare CG and InitCG, using either classical Gram–Schmidt (CGS) or modified Gram–Schmidt (MGS) to implement the projection H^*. Figure 6.1 shows the convergence curves to solve $Ax = b$ for CG started with x_{-1} (dotted); for InitCG with CGS, called unstable (dashed); and for InitCG with MGS (solid).

Because m is rather large, the initial error is very small in InitCG. Although CG starts with $x_{-1} = y_j$, InitCG converges much faster than CG thanks to the projection H^*, with $r_0 = H^* r_{-1}$.

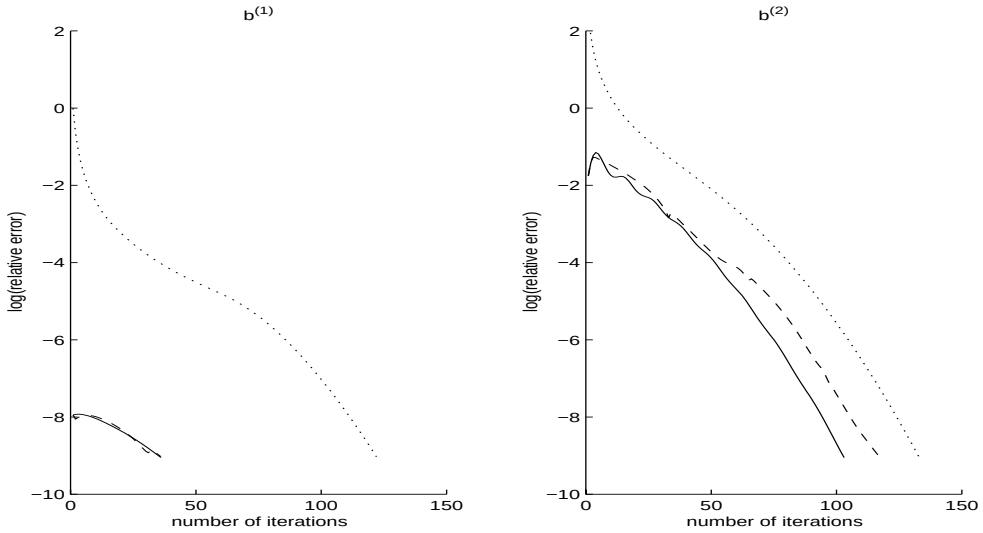


FIG. 6.2. *Results for the matrix* diag([1 : 500]) *(Example 2). Dotted line: CG; dashed line: InitCG; solid line: AugCG.*

If m is small, then both formulations CGS and MGS yield the same result. But here, with $m = 65$, CGS is unstable, as expected, and we will use MGS in what follows, in both InitCG and AugCG.

6.2. Example 2: Impact of right-hand side. Since we study only the convergence behavior, we choose here a diagonal matrix for which we can easily choose the eigenvalues. This is quite usual, as noted in [30, 29]. We choose $A = \text{diag}([1 : 500])$.

Figure 6.2 shows the results for $b^{(1)}$ and $b^{(2)}$. For both right-hand sides, $m = 30$ and $tol = 10^{-9}$. For $b = b^{(1)}$, $\nu = 1$ and $\epsilon = 10^{-2}$.

InitCG (dashed line) and AugCG (solid line) are similar for the "close" right-hand side $b^{(1)}$ and much faster than CG (dotted line); they save about 90 iterations thanks to $\|r_0\|/\|b\|$, which is much smaller than $\|r_{-1}\|/\|b\|$. In some cases, in particular when tol is not too small, AugCG can converge in 1 or 2 iterations [2].

InitCG is not as efficient as AugCG for the "far" right-hand side $b^{(2)}$ because $\|H^*(b-c)\|$ is too large. InitCG and AugCG start with a better approximation than CG, because $b^{(2)}$ is probably related somehow to c. InitCG and AugCG are similar during the first iterations, as expected, then AugCG has a better asymptotic rate than InitCG. Also AugCG converges faster than CG, as expected. In this example, AugCG saves about m iterations.

6.3. Example 3: Impact of size m. We study here the impact of the size m on the efficiency of AugPCG. We use the matrix S1RMQ4M1, from the Cylshell group of independent sets and generators.[1] The matrix is of size $N = 5489$, with 133950 nonzero entries and a condition number equal to 1.81×10^6. The matrix is preconditioned by incomplete Cholesky IC(1).

We choose $b = b^{(2)}$ and $tol = 10^{-9}$.

Figure 6.3 plots the convergence curve of PCG (dotted line) and AugPCG (solid lines) for m varying from 10 to 50.

[1] http://math.nist.gov/MatrixMarket/data

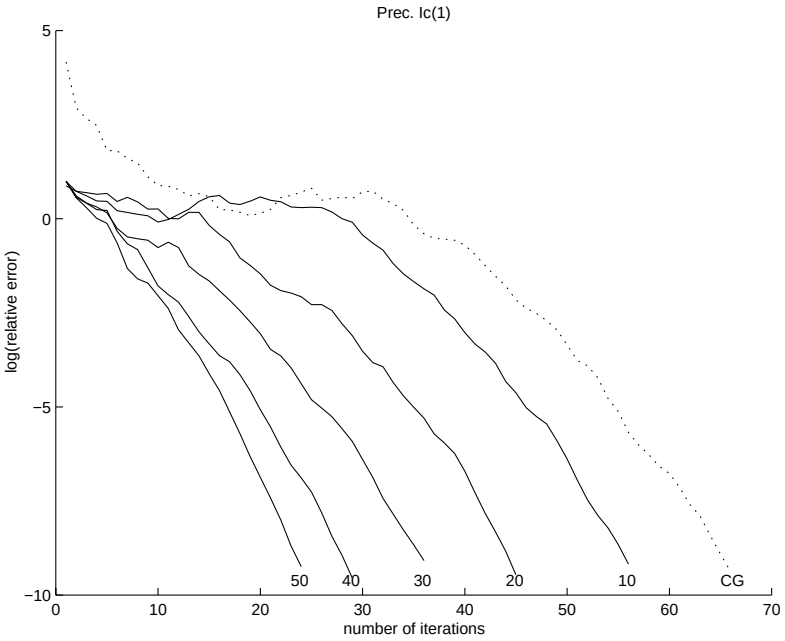


FIG. 6.3. *Results for the matrix S1RMQ4M1 with the right-hand side $b^{(2)}$ (Example 3). Dotted line: PCG; solid line: AugPCG with varying m.*

Up to $m = 40$, AugPCG saves about m iterations, then it saves slightly less than m iterations.

6.4. Example 4: Impact of preconditioning with the close right-hand side $b^{(1)}$. We take in this example the matrix S2RMQ4M1 from the same collection. The matrix is of size $N = 5,489$ with $134,420$ nonzero entries and with a condition number equal to 1.77×10^8.

We choose $b = b^{(1)}$ with $\nu = 1$ and $\epsilon = 10^{-3}$. We choose $m = 20$ and $tol = 10^{-6}$. Figure 6.4 shows the results with four incomplete Cholesky preconditioners with a varying fill-in: IC(1), IC(2), IC(3), and IC(4).

Here InitPCG is quite efficient because both right-hand sides are "close." The initial relative residual $\|r_0\|/\|b\|$ in InitPCG and AugPCG is much smaller than the relative residual $\|r_{-1}\|/\|b\|$ in PCG. It is almost equal to $tol = 10^{-6}$.

AugPCG saves about 36 iterations with IC(1) and about 20 iterations for the other three preconditioners. For IC(1), the gain in the initial residual is relatively more important because PCG converges slowly.

With all four preconditioners, the initial large gain is partly lost afterward because the residual increases. It seems that the convergence curves of AugPCG and PCG have similar patterns, with a shift of about m iterations.

6.5. Example 5: Impact of preconditioning with the far right-hand side $b^{(2)}$. We choose the same matrix S2RMQ4M1 and the parameters $b = b^{(2)}$, $m = 20$, and $tol = 10^{-6}$. Figure 6.5 shows the results for the preconditioners IC(1), IC(2), IC(3), and IC(4).

InitPCG is not efficient because the right-hand side $b^{(2)}$ is too "far" from c.

With all four preconditioners, AugPCG saves about 15 iterations, slightly less

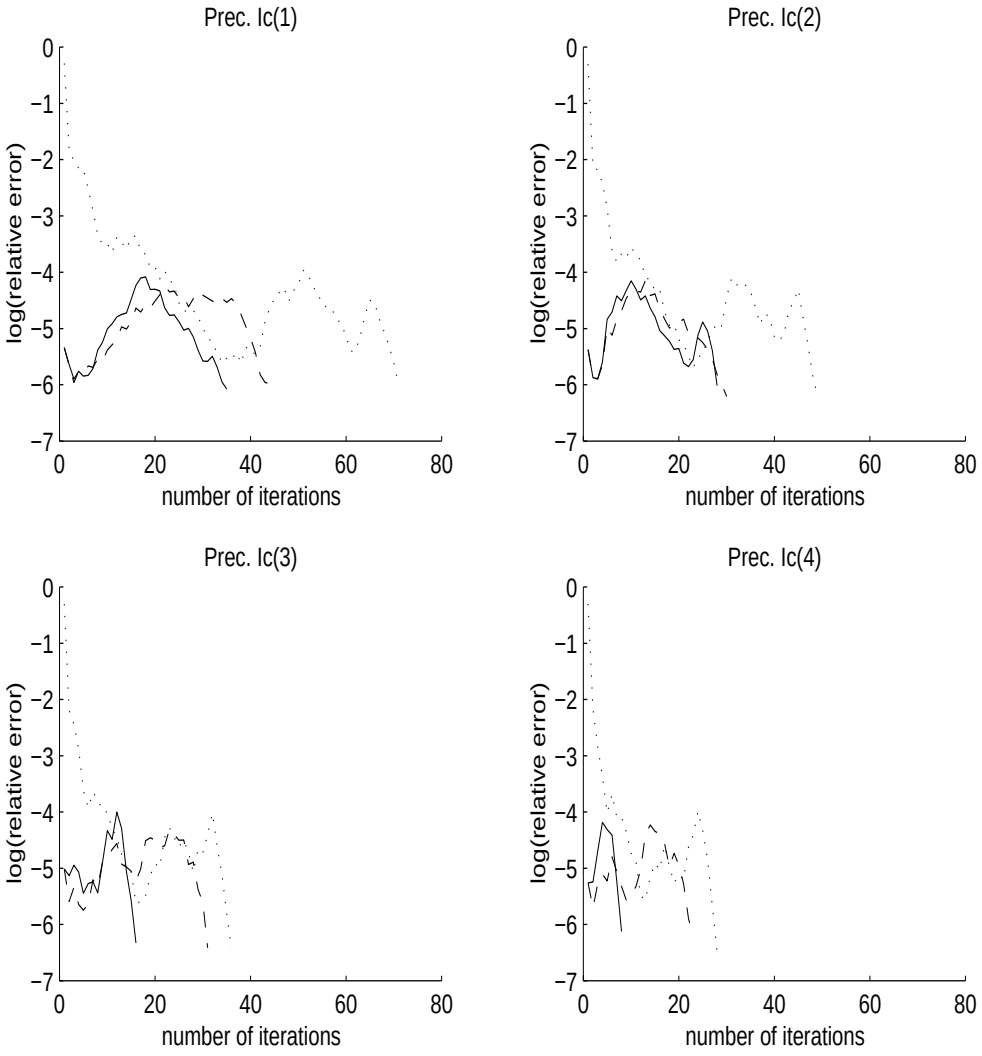


FIG. 6.4. *Results for the matrix S2RMQ4M1 with the right-hand side* $b^{(1)}$ *(Example* 4*). Dotted line: PCG; dashed line: InitPCG; solid line: AugPCG.*

than m equal to 20. So, as in Example 4, AugPCG remains efficient if the preconditioner improves. As in Example 4, AugPCG and PCG have similar patterns with a shift of almost m iterations.

6.6. Example 6: Relative impact of preconditioning and size m. We study here an example where CG converges slowly. We show the impact of preconditioning and of m. We choose the matrix S3RMT3M1 from the same collection. This matrix is of size $N = 5,489$ with $111,579$ nonzero entries. The condition number is 2.48×10^{10}.

We choose $b = b^{(2)}$ and $tol = 10^{-6}$. Figure 6.6 shows the results for $m = 20$ with the preconditioner IC(1), $m = 20$ with the preconditioner IC(2), $m = 50$ and IC(1), $m = 100$ and IC(1).

In all cases, InitPCG is not efficient at all because the right-hand side is too far.

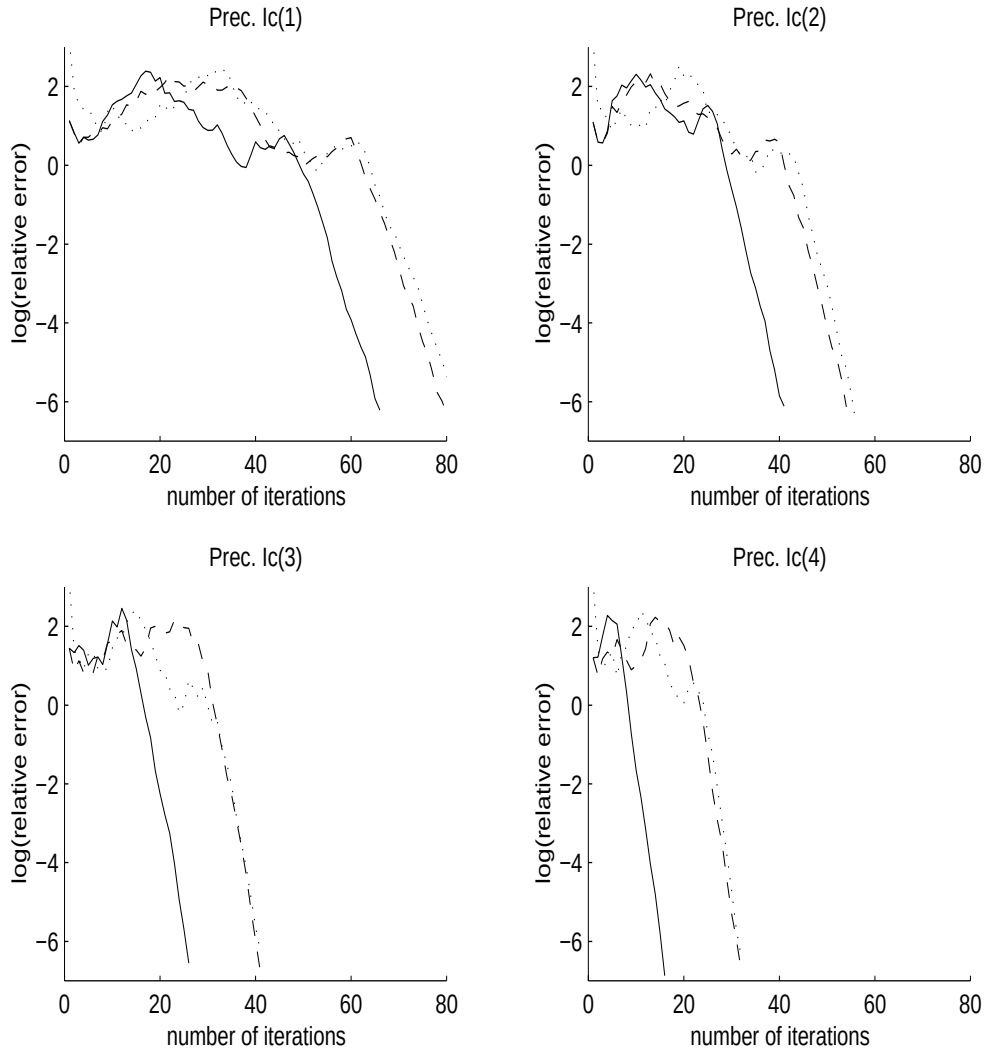


FIG. 6.5. *Results for the matrix S2RMQ4M1 with the right-hand side* $b^{(2)}$ *(Example* 5*). Dotted line: PCG; dashed line: InitPCG; solid line: AugPCG.*

Here, the initial error is of the same order as the error in CG.

With IC(1) and $m = 20$, AugPCG is not very efficient because PCG did not converge well in the seed system.

There are two ways to get a better efficiency. We first improve the preconditioner. With IC(2) and $m = 20$, AugPCG saves 15 iterations compared to PCG.

We then increase the number of vectors retained, while keeping the poor preconditioning IC(1). Whereas $m = 20$ does not improve convergence, $m = 50$ and $m = 100$ are quite efficient.

If the seed system converges fast, then AugPCG is efficient and m can be small. If the seed system converges slowly, then m must be large for an efficient acceleration, but m is restricted by memory constraints. Indeed, the initial guess computation requires the storage of m full vectors of length N. If secondary storage is used, the

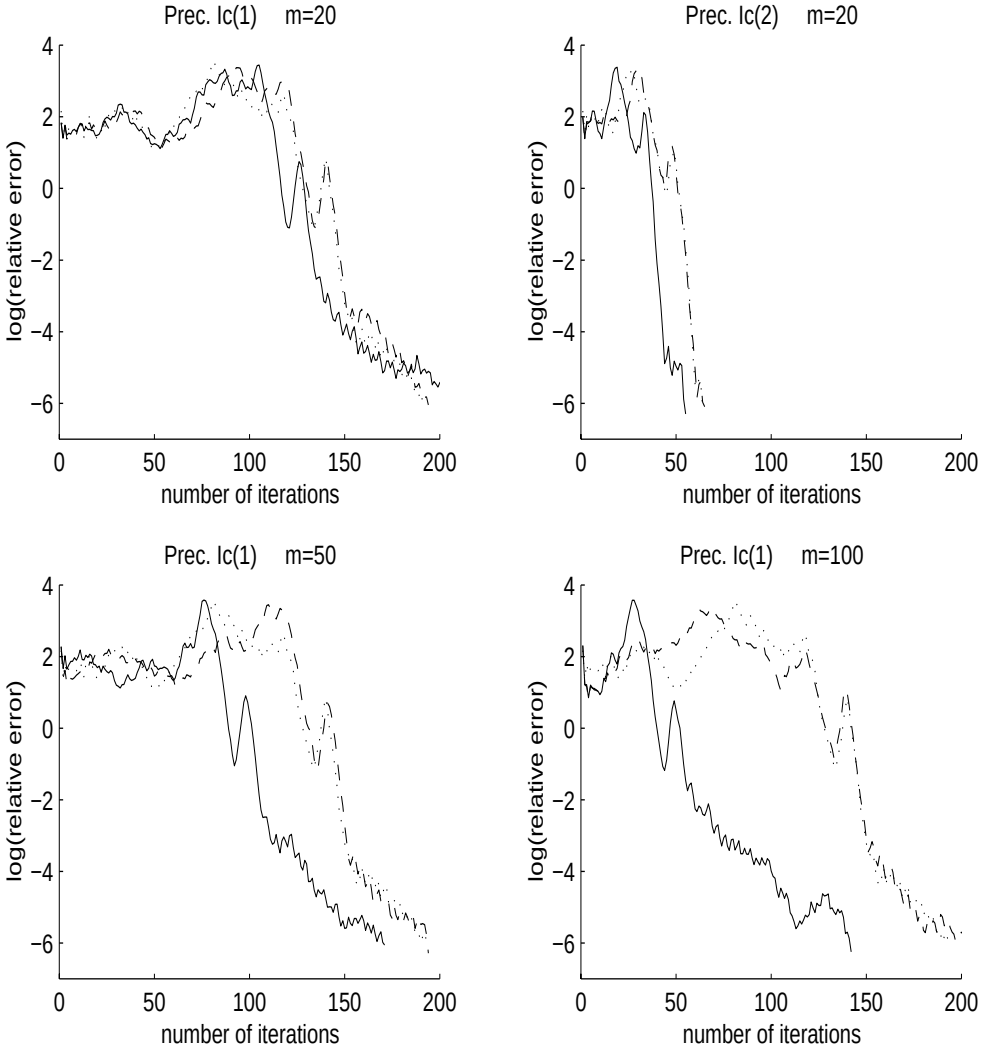


FIG. 6.6. *Results for the matrix S3RMT3M1 with the right-hand side* $b^{(2)}$ *(Example* 6*). Dotted line: CG; dashed line: InitCG; solid line: AugCG.*

gain in iterations must balance the overhead due to I/O [3]. As we mentioned in section 5, the initial guess computation involves BLAS2 operations of order m, so that the CPU cost increases with m. This overhead depends upon the relative cost of each iteration.

The optimal choice of m yielding a minimal CPU time is still an open question. It is probably better to use a good preconditioner with a moderate m. Also it seems that m must be large enough to capture the smallest eigenvalues of A, which mostly affect the convergence of CG [29, 27].

7. Conclusion. This papers considers the problem of solving $Ax = b$ once another system $Ay = c$ has already been solved, where A is symmetric positive definite. The CG applied to $Ay = c$ generates a Krylov subspace $K_m(A, s_0)$, which is used in the second system. The method InitCG computes an initial guess x_0 and residual r_0

using a stable formulation of the A^{-1}-orthogonal projection H^* onto $\mathcal{K}_m(A, s_0)^{\perp}$. The method AugCG not only starts with x_0, but also modifies the descent direction p_k for $k \geq 0$ using a cheap formulation of the A-orthogonal projection H onto $\mathcal{K}_m(A, s_0)^{\perp_A}$.

This paper shows that during the first $m/2$ iterations, InitCG is equivalent to a block CG method started with the 2-block $[s_0, r_0]$. It also shows that convergence of AugCG is governed by the condition number of H^*AH. Numerical experiments, using CG or PCG, demonstrate the efficiency of the method.

Though this method is described here for only two systems, it can be easily extended to more by involving all previous Krylov subspaces in the current system. However, the main drawback of this approach is the rapidly increasing memory requirement. But secondary storage can be used for large systems because it is used only in the initialization part of AugCG. We implemented AugCG with secondary storage in a wave-scattering problem [3]. Also, memory overhead is not crucial when AugCG is applied to the interface problem in a domain decomposition framework [2]. A generalization of AugCG based on a deflation approach was investigated in [26] as well as in [19].

Acknowledgment. The authors wish to thank the referees for their helpful comments and suggestions.

REFERENCES

[1] J. Baglama, D. Calvetti, G. H. Golub, and L. Reichel, *Adaptively preconditioned GMRES algorithms*, SIAM J. Sci. Comput., 20 (1998), pp. 243–269.

[2] M. Brieu, *Homogénéisation et endommagement de composites élastomère par techniques de calcul parallèle*, Ph.D. thesis, Ecole Normale Supérieure de Cachan, France, 1999.

[3] M.-O. Bristeau and J. Erhel, *Augmented conjugate gradient. Application in an iterative process for the solution of scattering problems*, Numer. Algorithms, 18 (1998), pp. 71–90.

[4] A. Bruaset, *A Survey of Preconditioned Iterative Methods*, Pitman Res. Notes Math. Ser. 328, Longman Scientific and Technical, Harlow, UK, 1995.

[5] T. Chan and M. Ng, *Galerkin Projection Methods for Solving Multiple Linear Systems*, Tech. Report 96-31, UCLA, Los Angeles, CA, 1996.

[6] T. F. Chan and W. L. Wan, *Analysis of projection methods for solving linear systems with multiple right-hand sides*, SIAM J. Sci. Comput., 18 (1997), pp. 1698–1721.

[7] A. Chapman and Y. Saad, *Deflated and augmented Krylov subspace techniques*, Numer. Linear Algebra Appl., 4 (1997), pp. 43–66.

[8] R. Choquet, *Etude de la methode de Newton-GMRES—Application aux equations de Navier-Stokes compressibles*, Ph.D. thesis, Université de Rennes 1, France, 1995.

[9] J. Erhel and K. Burrage, *On the performance of various adaptive preconditioned GMRES strategies*, Numer. Linear Algebra Appl., 5 (1998), pp. 101–121.

[10] J. Erhel, K. Burrage, and B. Pohl, *Restarted GMRES preconditioned by deflation*, J. Comput. Appl. Math., 69 (1996), pp. 303–318.

[11] C. Farhat, L. Crivelli, and F.-X. Roux, *Extending substructure based iterative solvers to multiple load and repeated analyses*, Comput. Methods Appl. Mech. Engrg., 117 (1994), pp. 195–209.

[12] P. F. Fisher, *Projection Techniques for Iterative Solution of $Ax = b$ with Successive Right-Hand Sides*, ICASE Tech. Report 93-90, NASA Langley Research Center, Hampton, VA, 1993.

[13] P. Joly, *Résolution de systèmes linéaires avec plusieurs seconds membres par la méthode du gradient conjugué*, Tech. Report R-91012, Université Pierre et Marie Curie, Paris, France, 1991.

[14] W. D. Joubert and T. A. Manteuffel, *Iterative Methods for Nonsymmetric Linear Systems*, Academic Press, New York, 1990, pp. 149–171.

[15] R. Kail, F. Léné, and Y.-H. D. Roeck, *An asymptotic model of laminated plates*, Rev. Européenne Élém. Finis, 6 (1997), pp. 23–42.

[16] C. Le Calvez, *Accélération de méthodes de Krylov pour la résolution de systèmes linéaires creux sur machines parallèles*, Ph.D. thesis, Université de Lille, France, 1998.

[17] R. B. Morgan, *A restarted GMRES method augmented with eigenvectors*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1154–1171.
[18] R. B. Morgan, *Implicitly restarted GMRES and Arnoldi methods for nonsymmetric systems of equations*, SIAM J. Matrix Anal. Appl., to appear.
[19] R. A. Nicolaides, *Deflation of conjugate gradients with applications to boundary value problems*, SIAM J. Numer. Anal., 24 (1987), pp. 355–365.
[20] D. P. O'Leary, *The block conjugate gradient algorithm and related methods*, Linear Algebra Appl., 29 (1980), pp. 293–322.
[21] M. Papadrakakis and S. Smerou, *A new implementation of the Lanczos method in linear problems*, Internat. J. Numer. Methods in Engrg., 29 (1990), pp. 141–159.
[22] B. Parlett, *A new look at the Lanczos algorithm for solving symmetric systems of linear equations*, Linear Algebra Appl., 29 (1980), pp. 323–346.
[23] C. Rey, *Développement d'algorithmes parallèles de résolution en calcul non-linéaire de structures hétérogènes: Application au cas de la butée acier-élastomère.*, Ph.D. thesis, Ecole Normale Supérieure de Cachan, France, 1994.
[24] Y. Saad, *On the Lanczos method for solving symmetric linear systems with several right-hand sides*, Math. Comp., 178 (1987), pp. 651–662.
[25] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing Company, Boston, 1996.
[26] Y. Saad, M. Yeung, J. Erhel, and F. Guyomarc'h, *A deflated version of the conjugate gradient algorithm*, SIAM J. Sci. Comput., to appear.
[27] G. Sleijpen and A. van der Sluis, *Further results on the convergence behavior of conjugate-gradients and Ritz values*, Linear Algebra Appl., 246 (1996), pp. 233–278.
[28] C. Smith, A. Peterson, and R. Mittra, *A conjugate gradient algorithm for the treatment of multiple incident electromagnetic fields*, IEEE Trans. Antennas and Propagation, 37 (1989), pp. 1490–1493.
[29] A. van der Sluis and H. van der Vorst, *The rate of convergence of conjugate gradients*, Numer. Math., 48 (1986), pp. 543–560.
[30] H. van der Vorst, *An iterative method for solving $f(A)x = b$ using Krylov subspace information obtained for the symmetric positive definite matrix A*, J. Comput. Appl. Math., 18 (1987), pp. 249–263.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 4, pp. 1300–1317
© 2000 Society for Industrial and Applied Mathematics

CONSTRAINT PRECONDITIONING FOR INDEFINITE LINEAR SYSTEMS*

CARSTEN KELLER†, NICHOLAS I. M. GOULD‡, AND ANDREW J. WATHEN†

Abstract. The problem of finding good preconditioners for the numerical solution of indefinite linear systems is considered. Special emphasis is put on preconditioners that have a 2×2 block structure and that incorporate the $(1,2)$ and $(2,1)$ blocks of the original matrix. Results concerning the spectrum and form of the eigenvectors of the preconditioned matrix and its minimum polynomial are given. The consequences of these results are considered for a variety of Krylov subspace methods. Numerical experiments validate these conclusions.

Key words. preconditioning, indefinite matrices, Krylov subspace methods

AMS subject classifications. 65F10, 65F15, 65F50

PII. S0895479899351805

1. Introduction.

In this paper, we are concerned with investigating a new class of preconditioners for indefinite systems of linear equations of a sort which arise in constrained optimization as well as in least-squares, saddle-point, and Stokes problems. We attempt to solve the indefinite linear system

$$\underbrace{\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}}_{\mathcal{A}} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \underbrace{\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}}_{b}, \tag{1.1}$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric and $B \in \mathbb{R}^{m\times n}$. Throughout the paper we shall assume that $m \le n$ and that $\mathcal{A}$ is nonsingular, in which case B must be of full rank.

Example 1.1 (quadratic programming problem). Consider the problem of minimizing a function of n variables subject to m linear equality constraints on the variables, i.e.,

$$\begin{aligned} &\underset{x}{\text{minimize}}\ f(x) = \tfrac{1}{2}x^T Ax - c^T x \\ &\text{subject to } Bx = d. \end{aligned} \tag{1.2}$$

Any finite solution to (1.2) is a stationary point of the Lagrangian function

$$L(x,\lambda) = \tfrac{1}{2}x^T Ax - c^T x + \lambda^T(Bx - d),$$

where the λ_i are referred to as *Lagrangian multipliers*. On differentiating L with respect to x and λ the solution of (1.2) is readily seen to satisfy $n + m$ linear equations of the form (1.1), with $x_1 = x$, $x_2 = \lambda$, $b_1 = c$, and $b_2 = d$. For this application these are known as the Karush–Kuhn–Tucker (KKT) conditions.

Example 1.2 (saddle-point problems). Mixed finite element approximations of variational problems are expressible in the following common form.

*Received by the editors February 1, 1999; accepted for publication (in revised form) by M. L. Overton August 31, 1999; published electronically April 18, 2000.
http://www.siam.org/journals/simax/21-4/35180.html
†Oxford University Computing Laboratory, Parks Road, Oxford OX1 3QD, UK (ck@comlab.ox.ac.uk, wathen@comlab.ox.ac.uk).
‡Rutherford Appleton Laboratory, Chilton, Didcot, OX11 0QX, UK (N.I.M.Gould@rl.ac.uk).

Find u and p in certain spaces such that

$$a(u,v) + b(v,p) = \langle f, v\rangle, \qquad b(u,q) = \langle g, q\rangle. \tag{1.3}$$

For appropriate v and q (1.3) leads to systems of the form (1.1), where x_1 are the coefficients of the approximation of u and x_2 are the coefficients of the approximation of p with respect to chosen bases. See, for example, Quarteroni and Valli [23, Chapters 7, 9].

Krylov subspace methods, which are among the most important iterative methods currently available, apply techniques that involve orthogonal projections onto subspaces of the form

$$\mathcal{K}(\mathcal{A}, b) \equiv \text{span}\{b, \mathcal{A}b, \mathcal{A}^2 b, \dots, \mathcal{A}^{n-1}b, \dots\}.$$

The most common schemes that use this idea are the method of conjugate gradients (CG) for symmetric positive definite matrices, the method of minimum residuals (MINRES) for symmetric and possibly indefinite matrices, and the generalized minimum residual method (GMRES) for unsymmetric matrices, although many other methods are available; see, for example, Greenbaum [13].

One common feature of the above-mentioned methods is that the solution of the linear system (1.1) is found within $n+m$ iterations in exact arithmetic; see Joubert and Manteuffel [15, p. 152]. For very large (and possibly sparse) linear systems this upper limit on the number of iterations is often not practical. The idea of preconditioning attempts to improve on the spectral properties, i.e., the clustering of the eigenvalues, such that the total number of iterations required to solve the system to within some tolerance is decreased substantially.

In this paper, we are specifically concerned with nonsingular preconditioners of the form

$$\mathcal{G} = \begin{bmatrix} G & B^T \\ B & 0 \end{bmatrix}, \tag{1.4}$$

where $G \in \mathbb{R}^{n\times n}$ approximates, but is not the same as, A. The inclusion of the exact representation of the $(1,2)$ and $(2,1)$ matrix blocks in the preconditioner, which are often associated with constraints (see Example 1.1), leads one to hope for a more favorable distribution of the eigenvalues of the (left-)preconditioned linear system

$$\mathcal{G}^{-1}\mathcal{A}x = \mathcal{G}^{-1}b. \tag{1.5}$$

Since these blocks are unchanged from the original system, we shall call $\mathcal{G}$ a *constraint preconditioner*. A preconditioner of the form $\mathcal{G}$ has recently been used by Lukšan and Vlček [16] in the context of constrained nonlinear programming problems; see also Coleman [7], Polyak [21], and Gould, Hribar, and Nocedal [12]. Here, we derive arguments that confirm and extend some of the results in [16] and highlight the favorable features of a preconditioner of the form $\mathcal{G}$. Note that Golub and Wathen [11] recently considered a symmetric preconditioner of the form (1.4) for problems of the form (1.1), where A is nonsymmetric.

We comment that for certain partial differential equation problems, which give rise to linear systems of the form (1.1), extremely effective positive definite preconditioners have been proposed. These make use of particular structures in the underlying problems but they are not necessarily of the form (1.4). For example, for Stokes problems

describing slow viscous flow of an incompressible fluid see Silvester and Wathen [26]. The preconditioning discussed in this paper does not assume any underlying structure except for the algebraic block structure indicated in (1.1).

In section 2, we determine the eigensolution distribution of the preconditioned system and give lower and upper bounds for the eigenvalues of $\mathcal{G}^{-1}\mathcal{A}$ in the case when the submatrix G is positive definite. Section 3 describes the convergence behavior of a Krylov subspace method such as GMRES, section 4 investigates possible implementation strategies, while in section 5 we give numerical results to support the theory developed in this paper.

2. Preconditioning $\mathcal{A}$. For symmetric (and in general normal) matrix systems, the convergence of an applicable iterative method is determined by the distribution of the eigenvalues of the coefficient matrix. In particular, it is desirable that the number of distinct eigenvalues, or at least the number of clusters, be small, because in this case convergence will be rapid. To be more precise, if there are only a few distinct eigenvalues, then optimal methods like CG, MINRES, or GMRES will terminate (in exact arithmetic) after a small and precisely defined number of steps. We prove a result of this type below. For nonnormal systems, convergence, as opposed to termination, is not so readily described; see Greenbaum [13, p. 5].

2.1. Eigenvalue distribution. The eigenvalues of the preconditioned coefficient matrix $\mathcal{G}^{-1}\mathcal{A}$ may be derived by considering the generalized eigenvalue problem

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \lambda \begin{bmatrix} G & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \tag{2.1}$$

Let $QR = \begin{bmatrix} Y & Z \end{bmatrix} \begin{bmatrix} R^T & 0^T \end{bmatrix}^T$ be an orthogonal factorization of B^T, where $R \in \mathbb{R}^{m\times m}$ is upper triangular, $Y \in \mathbb{R}^{n\times m}$, and $Z \in \mathbb{R}^{n\times(n-m)}$ is a basis for the nullspace of B. Premultiplying (2.1) by the nonsingular and square matrix

$$\begin{bmatrix} Z^T & 0 \\ Y^T & 0 \\ 0 & I \end{bmatrix}$$

and postmultiplying by its transpose gives

$$\begin{bmatrix} Z^TAZ & Z^TAY & 0 \\ Y^TAZ & Y^TAY & R \\ 0 & R^T & 0 \end{bmatrix} \begin{bmatrix} x_z \\ x_y \\ y \end{bmatrix} = \lambda \begin{bmatrix} Z^TGZ & Z^TGY & 0 \\ Y^TGZ & Y^TGY & R \\ 0 & R^T & 0 \end{bmatrix} \begin{bmatrix} x_z \\ x_y \\ y \end{bmatrix}, \tag{2.2}$$

with $x = Zx_z + Yx_y$ and where we made use of the equalities $BZ = 0$ and $R = (BY)^T$. Performing a simultaneous sequence of row and column interchanges on both matrices in (2.2) reveals two lower block-triangular matrices

$$\tilde{\mathcal{A}} = \begin{bmatrix} R^T & 0 & 0 \\ Z^TAY & Z^TAZ & 0 \\ Y^TAY & Y^TAZ & R \end{bmatrix}, \qquad \tilde{\mathcal{G}} = \begin{bmatrix} R^T & 0 & 0 \\ Z^TGY & Z^TGZ & 0 \\ Y^TGY & Y^TGZ & R \end{bmatrix},$$

and thus the preconditioned coefficient matrix $\mathcal{G}^{-1}\mathcal{A}$ is similar to

$$\mathcal{P} = \tilde{\mathcal{G}}^{-1}\tilde{\mathcal{A}} = \begin{array}{c} \begin{matrix} m & (n-m) & m \end{matrix} \\ \begin{bmatrix} I & 0 & 0 \\ \Theta & (Z^TGZ)^{-1}(Z^TAZ) & 0 \\ \Upsilon & \Gamma & I \end{bmatrix} \end{array} \begin{array}{c} m \\ (n-m) \\ m \end{array}. \tag{2.3}$$

Here, the precise forms of Θ, Υ, and Γ are irrelevant for the argument that follows; they are in general nonzero. We just proved the following theorem.

THEOREM 2.1. *Let $\mathcal{A} \in \mathbb{R}^{(n+m)\times(n+m)}$ be a symmetric and indefinite matrix of the form*

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix},$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric and $B \in \mathbb{R}^{m\times n}$ is of full rank. Assume Z is an $n \times (n-m)$ basis for the nullspace of B. Preconditioning $\mathcal{A}$ by a matrix of the form

$$\mathcal{G} = \begin{bmatrix} G & B^T \\ B & 0 \end{bmatrix},$$

where $G \in \mathbb{R}^{n\times n}$ is symmetric, $G \neq A$, and $B \in \mathbb{R}^{m\times n}$ is as above, implies that the matrix $\mathcal{G}^{-1}\mathcal{A}$ has

1. *an eigenvalue at 1 with multiplicity $2m$, and*
2. *$n-m$ eigenvalues which are defined by the generalized eigenvalue problem $Z^TAZx_z = \lambda Z^TGZx_z$.*

Note that, if either Z^TAZ is positive definite or we choose G so that Z^TGZ is positive definite, then the indefinite constrained preconditioner applied to the indefinite linear system (1.1) yields the preconditioned matrix $\mathcal{P}$ which has real eigenvalues.

Remark 2.2. In the above argument, we assumed that B has full row rank and consequently applied an orthogonal factorization of B^T, which resulted in an upper triangular matrix $R \in \mathbb{R}^{m\times m}$. If B does not have full row rank, i.e., $\text{rank}(B) = m-k$ for some integer $k \leq m$, then k zero rows and columns can be deleted from both matrices in (2.2), thus giving a reduced system of dimension $(n+m-k)\times(n+m-k)$. This removal of the redundant information does not impose any restriction on the proposed preconditioner, since all mathematical arguments apply equivalently to the reduced system of equations.

2.2. Eigenvector distribution. We mentioned above that the termination of a Krylov subspace method is related to the location of the eigenvalues and the number of corresponding linearly independent eigenvectors. In order to establish the association between eigenvectors and eigenvalues, we expand the general eigenvalue problem (2.2), yielding

$$Z^TAZx_z + Z^TAYx_y = \lambda\left[Z^TGZx_z + Z^TGYx_y\right], \tag{2.4}$$

$$Y^TAZx_z + Y^TAYx_y + Ry = \lambda\left[Y^TGZx_z + Y^TGYx_y + Ry\right], \tag{2.5}$$

$$R^Tx_y = \lambda R^Tx_y. \tag{2.6}$$

From (2.6), it may be deduced that either $\lambda = 1$ or $x_y = 0$. In the former case, (2.4) and (2.5) simplify to

$$Z^TAZx_z + Z^TAYx_y = Z^TGZx_z + Z^TGYx_y,$$
$$Y^TAZx_z + Y^TAYx_y = Y^TGZx_z + Y^TGYx_y,$$

which can consequently be written as

$$Q^TAQw = Q^TGQw, \tag{2.7}$$

where $Q = \begin{bmatrix} Y & Z \end{bmatrix}$ and $w = \begin{bmatrix} x_y^T & x_z^T \end{bmatrix}^T$. Since Q is orthogonal, the general eigenvalue problem (2.7) is equivalent to considering

$$Aw = \sigma Gw, \tag{2.8}$$

where $w \neq 0$ if and only if $\sigma = 1$. There are m linearly independent eigenvectors $\begin{bmatrix} 0^T & 0^T & y^T \end{bmatrix}^T$ corresponding to $w = 0$, and a further i $(1 \leq i \leq n)$ linearly independent eigenvectors (corresponding to eigenvalues $\sigma = 1$ of (2.8)).

Now, suppose $\lambda \neq 1$, in which case $x_y = 0$. Equations (2.4) and (2.5) yield

$$Z^T AZ x_z = \lambda Z^T GZ x_z, \tag{2.9}$$

$$Y^T AZ x_z + Ry = \lambda \left[Y^T GZ x_z + Ry \right]. \tag{2.10}$$

The generalized eigenvalue problem (2.9) defines $n - m$ eigenvalues, where j $(1 \leq j \leq n - m)$ of these are not equal to 1 and for which two cases have to be distinguished. If $x_z \neq 0$, y must satisfy

$$\left[Y^T AZ - \lambda Y^T GZ \right] x_z = (\lambda - 1) Ry,$$

from which follows that the corresponding eigenvectors are defined by $\begin{bmatrix} x_z^T & 0^T & y^T \end{bmatrix}^T$. If $x_z = 0$, we deduce from (2.10) that

$$Ry = \lambda Ry,$$

and hence that $y = 0$ since $\lambda \neq 1$. As $\begin{bmatrix} x_z^T & x_y^T & y^T \end{bmatrix}^T = 0$ in this case, no extra eigenvectors arise.

Summarizing the above, it is evident that $\mathcal{P}$ has $m + i + j$ eigenvectors. We now show that, under realistic assumptions, these eigenvectors are in fact linearly independent.

THEOREM 2.3. *Let $\mathcal{A} \in \mathbb{R}^{(n+m)\times(n+m)}$ be a symmetric and indefinite matrix of the form*

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix},$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric and $B \in \mathbb{R}^{m\times n}$ is of full rank. Assume the preconditioner $\mathcal{G}$ is defined by a matrix of the form

$$\mathcal{G} = \begin{bmatrix} G & B^T \\ B & 0 \end{bmatrix},$$

where $G \in \mathbb{R}^{n\times n}$ is symmetric, $G \neq A$, and $B \in \mathbb{R}^{m\times n}$ is as above. Let Z denote an $n \times (n - m)$ basis for the nullspace of B and suppose that $Z^T GZ$ is positive definite. The preconditioned matrix $\mathcal{G}^{-1}\mathcal{A}$ has $n + m$ eigenvalues as defined by Theorem 2.1 and $m + i + j$ linearly independent eigenvectors. There are

1. *m eigenvectors of the form $\begin{bmatrix} 0^T & 0^T & y^T \end{bmatrix}^T$ that correspond to the case $\lambda = 1$;*
2. *i $(0 \leq i \leq n)$ eigenvectors of the form $\begin{bmatrix} x_z^T & x_y^T & y^T \end{bmatrix}^T$ arising from $Aw = \sigma Gw$ with $w = \begin{bmatrix} x_y^T & x_z^T \end{bmatrix}^T$ linearly independent, $\sigma = 1$, and $\lambda = 1$; and*

(3) j $(0 \le j \le n-m)$ *eigenvectors of the form* $\begin{bmatrix} x_z^T & 0^T & y^T \end{bmatrix}^T$ *that correspond to the case* $\lambda \neq 1$.

Proof. To prove that the $m+i+j$ eigenvectors of $\mathcal{P}$ are linearly independent, we need to show that

$$\begin{bmatrix} 0 & \cdots & 0 \\ 0 & \cdots & 0 \\ y_1^{(1)} & \cdots & y_m^{(1)} \end{bmatrix} \begin{bmatrix} a_1^{(1)} \\ \vdots \\ a_m^{(1)} \end{bmatrix} + \begin{bmatrix} x_{z_1}^{(2)} & \cdots & x_{z_i}^{(2)} \\ x_{y_1}^{(2)} & \cdots & x_{y_i}^{(2)} \\ y_1^{(2)} & \cdots & y_i^{(2)} \end{bmatrix} \begin{bmatrix} a_1^{(2)} \\ \vdots \\ a_i^{(2)} \end{bmatrix} + \begin{bmatrix} x_{z_1}^{(3)} & \cdots & x_{z_j}^{(3)} \\ 0 & \cdots & 0 \\ y_1^{(3)} & \cdots & y_j^{(3)} \end{bmatrix} \begin{bmatrix} a_1^{(3)} \\ \vdots \\ a_j^{(3)} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \tag{2.11}$$

implies that the vectors $a^{(k)}$ $(k = 1, \ldots, 3)$ are zero vectors. Multiplying (2.11) by $\mathcal{A}$ and $\mathcal{G}^{-1}$, and recalling that in the previous equation the first matrix arises from the case $\lambda_k = 1$ $(k = 1, \ldots, m)$, the second matrix from the case $\lambda_k = 1$ and $\sigma_k = 1$ $(k = 1, \ldots, i)$, and the last matrix from $\lambda_k \neq 1$ $(k = 1, \ldots, j)$, gives

$$\begin{bmatrix} 0 & \cdots & 0 \\ 0 & \cdots & 0 \\ y_1^{(1)} & \cdots & y_m^{(1)} \end{bmatrix} \begin{bmatrix} a_1^{(1)} \\ \vdots \\ a_m^{(1)} \end{bmatrix} + \begin{bmatrix} x_{z_1}^{(2)} & \cdots & x_{z_i}^{(2)} \\ x_{y_1}^{(2)} & \cdots & x_{y_i}^{(2)} \\ y_1^{(2)} & \cdots & y_i^{(2)} \end{bmatrix} \begin{bmatrix} a_1^{(2)} \\ \vdots \\ a_i^{(2)} \end{bmatrix} + \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_j \end{bmatrix} \begin{bmatrix} x_{z_1}^{(3)} & \cdots & x_{z_j}^{(3)} \\ 0 & \cdots & 0 \\ y_1^{(3)} & \cdots & y_j^{(3)} \end{bmatrix} \begin{bmatrix} a_1^{(3)} \\ \vdots \\ a_j^{(3)} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{2.12}$$

Subtracting (2.11) from (2.12) we obtain

$$\begin{bmatrix} \lambda_1 - 1 & & \\ & \ddots & \\ & & \lambda_j - 1 \end{bmatrix} \begin{bmatrix} x_{z_1}^{(3)} & \cdots & x_{z_j}^{(3)} \\ 0 & \cdots & 0 \\ y_1^{(3)} & \cdots & y_j^{(3)} \end{bmatrix} \begin{bmatrix} a_1^{(3)} \\ \vdots \\ a_j^{(3)} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix},$$

which simplifies to

$$\begin{bmatrix} x_{z_1}^{(3)} & \cdots & x_{z_j}^{(3)} \\ 0 & \cdots & 0 \\ y_1^{(3)} & \cdots & y_j^{(3)} \end{bmatrix} \begin{bmatrix} a_1^{(3)} \\ \vdots \\ a_j^{(3)} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{2.13}$$

since $\lambda_k \neq 1$ $(k = 1, \ldots, j)$.

The assumption that $Z^T G Z$ is positive definite implies that $x_{z_k}^{(3)}$ $(k = 1, \ldots, j)$ in (2.13) are linearly independent and thus that $a_k^{(3)} = 0$ $(k = 1, \ldots, j)$. Similarly, $a_k^{(2)} = 0$ $(k = 1, \ldots, i)$ follows from the linear independence of $\begin{bmatrix} x_{z_k}^{(2)^T} & x_{y_k}^{(2)^T} \end{bmatrix}^T$ $(k = 1, \ldots, i)$, and thus (2.11) simplifies to

$$\begin{bmatrix} 0 & \cdots & 0 \\ 0 & \cdots & 0 \\ y_1^{(1)} & \cdots & y_m^{(1)} \end{bmatrix} \begin{bmatrix} a_1^{(1)} \\ \vdots \\ a_m^{(1)} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}.$$

However, $y_k^{(1)}$ $(k = 1, \ldots, m)$ are linearly independent and thus $a_k^{(1)} = 0$ $(k = 1, \ldots, m)$. ∎

Remark 2.4. Note that the result of Theorem 2.3 remains true if $Z^T(\gamma A + \sigma G)Z$ is positive definite for some scalars γ and σ; see Parlett [20, p. 343] for details.

To show that the eigenvector bounds of Theorem 2.3 can in fact be attained, consider the following two examples.

Example 2.5 (minimum bound). Consider the matrices

$$\mathcal{A} = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 2 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad \mathcal{G} = \begin{bmatrix} 1 & 3 & 0 \\ 3 & 4 & 1 \\ 0 & 1 & 0 \end{bmatrix},$$

so that $m = 1$ and $n = 2$. The preconditioned matrix $\mathcal{P}$ has an eigenvalue at 1 with multiplicity 3, but only one eigenvector arising from case (1) in Theorem 2.3. This eigenvector may be taken to be $\begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T$.

Example 2.6 (maximum bound). Let $\mathcal{A} \in \mathbb{R}^{3\times 3}$ be defined as in Example 2.5, but assume $\mathcal{G} = \mathcal{A}$. The preconditioned matrix $\mathcal{P}$ has an eigenvalue at 1 with multiplicity 3 and clearly a complete set of eigenvectors. These may be taken to be the columns of the identity matrix.

2.3. Eigenvalue bounds. It is apparent from the calculations in the previous section that the eigenvalue at 1 with multiplicity $2m$ is independent of the choice of G in the preconditioner. On the contrary, the $n - m$ eigenvalues that are defined by (2.9) are highly sensitive to the choice of G. If G is a close approximation of A, we can expect a more favorable distribution of eigenvalues and consequently may expect faster convergence of an appropriate iterative method. In order to determine a good factorization of A, it will be helpful to find intervals in which the $n - m$ eigenvalues are located. If G is a positive definite matrix, one possible approach is provided by Cauchy's interlace theorem.

THEOREM 2.7 (Cauchy's interlace theorem). *Suppose $\mathcal{T} \in \mathbb{R}^{n\times n}$ is symmetric and*

$$\mathcal{T} = \begin{bmatrix} H & \star \\ \star & \star \end{bmatrix},$$

where $H \in \mathbb{R}^{m\times m}$ with $m < n$. Label the eigenpairs of $\mathcal{T}$ and H as

$$\begin{aligned} \mathcal{T}z_i &= \alpha_i z_i, \quad i = 1, \ldots, n, \quad \alpha_1 \le \alpha_2 \le \cdots \le \alpha_n, \\ Hy_i &= \lambda_i y_i, \quad i = 1, \ldots, m, \quad \lambda_1 \le \lambda_2 \le \cdots \le \lambda_m. \end{aligned}$$

Then

$$\alpha_k \le \lambda_k \le \alpha_{k+(n-m)}, \qquad k = 1, \ldots, m.$$

Proof. See Parlett [20, p. 203]. ∎

The applicability of Theorem 2.7 is verified by recalling the definitions of Q and Z given in the previous section and by considering the generalized eigenvalue problems

$$Q^T AQv = \alpha Q^T GQv \tag{2.14}$$

and

$$Z^T AZw = \lambda Z^T GZw. \tag{2.15}$$

Since G is positive definite, so is Q^TGQ, and we may therefore write

$$Q^TGQ = \begin{bmatrix} Z^TGZ & Z^TGY \\ Y^TGZ & Y^TGY \end{bmatrix} = \underbrace{\begin{bmatrix} L & 0 \\ R & S \end{bmatrix}}_{M} \underbrace{\begin{bmatrix} L^T & R^T \\ 0 & S^T \end{bmatrix}}_{M^T},$$

where $LL^T = Z^TGZ$, $R = Y^TGZL^{-T}$, and $SS^T = Y^TGY - RR^T$. Rewriting (2.14) and (2.15) gives

$$M^{-1}Q^TAQM^{-T}u = \alpha u \tag{2.16}$$

and

$$L^{-1}Z^TAZL^{-T}z = \lambda z, \tag{2.17}$$

where $u = M^Tv$ and $z = L^Tw$.

Now, since the matrix $M^{-1}Q^TAQM^{-T}$ is similar to $G^{-1}A$, (2.16) defines the same eigenvalues α_i $(i = 1, \ldots, n)$ as $G^{-1}A$. We may therefore apply Theorem 2.7 directly. The result is that the $n-m$ eigenvalues λ_i of (2.9) satisfy $\alpha_k \leq \lambda_k \leq \alpha_{k+m}$ $(k = 1, \ldots, n-m)$. In particular, the λ_i are bounded by the extreme eigenvalues of $G^{-1}A$ so that the λ_i will necessarily be clustered if G is a good approximation of A. Furthermore, a good preconditioner G for A implies that Z^TGZ is at least as good a preconditioner for Z^TAZ. To show that the preconditioner Z^TGZ can in fact be much better, consider the following example, taken from the CUTE collection [5].

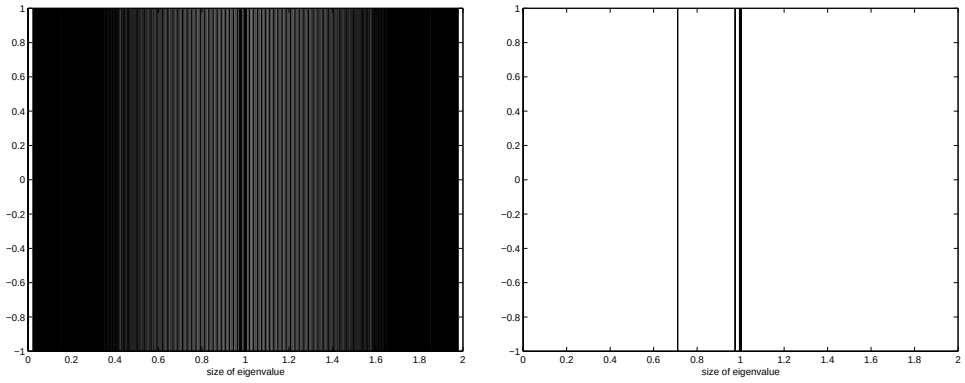
Example 2.8. Consider the convex quadratic programming problem *BLOWEYC*, which may be formulated as

$$\text{minimize } u(s)^TAu(s) + u(s)^Tw(s) - v(s)^TAu(s) - 2.0v(s)^Tw(s) - u(s)^Tv(s)$$
$$\text{subject to } Aw(s) = u(s), \quad u(s) \in [-1, 1], \quad \text{and} \quad \int_0^1 u(s)\, ds = 0.4.$$

Selecting a size parameter of 500 discretization intervals defines a set of linear equations of the form (1.1), where $n = 1002$ and $m = 502$. Letting G be the diagonal of A, we may deduce from the above theory that the extreme eigenvalues of $G^{-1}A$ give lower and upper bounds for the $n-m$ eigenvalues defined by the general eigenvalue problem (2.9). In Figure 2.1(a), the 1002 eigenvalues of $G^{-1}A$ are drawn as vertical lines, whereas Figure 2.1(b) displays the 500 eigenvalues of $(Z^TGZ)^{-1}Z^TAZ$.

The spectrum of Figure 2.1(a) is equivalent to a graph of the entire spectrum of $\mathcal{P}$, but with an eigenvalue at 1 of multiplicity 502 removed. Rounded to two decimal places, the numerical values of the two extreme eigenvalues of $G^{-1}A$ are 0.02 and 1.98, whereas the extreme eigenvalues of $(Z^TGZ)^{-1}Z^TAZ$ are given by 0.71 and 1. Note that for this example a large number of eigenvalues of $G^{-1}A$ are clustered in the approximate intervals $[0.02, 0.38]$ and $[1.65, 1.97]$. The eigenvalue distribution in Figure 2.1(b) reveals that there is one eigenvalue near 0.71 and a group of eigenvalues near 1. It follows that any appropriate iterative method which solves (1.5) can be expected to converge in a very small number of steps; this is verified by the numerical results presented in section 5.

It is readily seen from Example 2.8 that in this case the bounds provided by Theorem 2.7 are not descriptive in that there is significantly more clustering of the eigenvalues than implied by the theorem.



(a) Eigenvalues of $G^{-1}A$ (b) Eigenvalues of $(Z^TGZ)^{-1}Z^TAZ$

FIG. 2.1. *Continuous vertical lines represent the eigenvalues of* (a) $G^{-1}A$ *and* (b) $(Z^TGZ)^{-1}Z^TAZ$.

3. Convergence. In the context of this paper, the convergence of an iterative method under preconditioning is influenced not only by the spectral properties of the coefficient matrix, but also by the relationship between the dimensions n and m. In particular, it follows from Theorem 2.1 that in the special case when $n = m$ the preconditioned linear system (1.5) has only one eigenvalue at 1 with multiplicity $2n$. For $m < n$, matrix (2.3) gives an eigenvalue at 1 with multiplicity $2m$ and $n - m$ (generally distinct) eigenvalues whose value may or may not be equal to 1. Before we examine how these results determine upper bounds on the number of iterations of an appropriate Krylov subspace method, we recall the definition of the *minimum polynomial* of a matrix.

DEFINITION 3.1. *Let $\mathcal{A} \in \mathbb{R}^{(n+m)\times(n+m)}$. The monic polynomial f of minimum degree such that $f(\mathcal{A}) = 0$ is called the minimum polynomial of $\mathcal{A}$.*

The importance of this definition becomes apparent when considering subsequent results and by recalling that similar matrices have the same minimum polynomial.

Krylov subspace theory states that iteration with any method with an optimality property such as GMRES will terminate when the degree of the minimum polynomial is attained; see Saad and Schultz [25, Proposition 2] and also Campbell et al. [6] in the case of GMRES. (To be precise, the number may be smaller in special cases where b is a combination of a few eigenvectors that affect the "grade" of $\mathcal{A}$ with respect to b.) In particular, the degree of the minimum polynomial is equal to the dimension of the corresponding Krylov subspace (for general b) (see Saad [24, Proposition 6.1]), and so the following theorems are relevant.

THEOREM 3.2. *Let $\mathcal{A} \in \mathbb{R}^{(n+m)\times(n+m)}$ be a symmetric and indefinite matrix of the form*

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix},$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric and $B \in \mathbb{R}^{m\times n}$ is of full rank. Let $m = n$. If $\mathcal{A}$ is preconditioned by a matrix of the form

$$\mathcal{G} = \begin{bmatrix} G & B^T \\ B & 0 \end{bmatrix},$$

where $G \in \mathbb{R}^{n\times n}$, $G \neq A$, and $B \in \mathbb{R}^{m\times n}$ is as above, then the Krylov subspace $\mathcal{K}(\mathcal{P}, b)$ is of dimension at most 2 for any b.

Proof. Writing the preconditioned system (2.3) in its explicit form, we observe that $\mathcal{P}$ is in fact given by

$$\begin{bmatrix} I & 0 \\ \Upsilon & I \end{bmatrix}, \tag{3.1}$$

where Υ is nonzero if and only if $A \neq G$. To show that the dimension of the corresponding Krylov subspace is at most 2, we need to determine the minimum polynomial of the system. It is evident from (3.1) that the eigenvalues of $\mathcal{P}$ are all 1 and $\mathcal{P} - I \neq 0$. However, $(\mathcal{P} - I)^2 = 0$ and so the minimum polynomial is of order 2. ∎

Remark 3.3. It is of course possible in the case $n = m$ to solve the (square) constrained equation $Bx_1 = b_2$ and then to obtain $x_2 = B^{-T}(b_1 - Ax_1)$. This gives motivation for why the result of Theorem 3.2 is independent of G.

Remark 3.4. The important consequence of Theorem 3.2 is that termination of an iteration method such as GMRES will occur in at most two steps for any choice of b, even though the preconditioned matrix is not diagonalizable (unless $A = G$).

THEOREM 3.5. *Let $\mathcal{A} \in \mathbb{R}^{(n+m)\times(n+m)}$ be a symmetric and indefinite matrix of the form*

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix},$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric and $B \in \mathbb{R}^{m\times n}$ is of full rank. Assume $m < n$ and $\mathcal{A}$ is nonsingular. Furthermore, assume $\mathcal{A}$ is preconditioned by a matrix of the form

$$\mathcal{G} = \begin{bmatrix} G & B^T \\ B & 0 \end{bmatrix},$$

where $G \in \mathbb{R}^{n\times n}$ is symmetric, $G \neq A$, and $B \in \mathbb{R}^{m\times n}$ is as above. If Z^TGZ is positive definite, where Z is an $n \times (n - m)$ basis for the nullspace of B, then the dimension of the Krylov subspace $\mathcal{K}(\mathcal{P}, b)$ is at most $n - m + 2$.

Proof. From the eigenvalue derivation in section 2.1, it is evident that the characteristic polynomial of the preconditioned linear system (1.5) is

$$(\mathcal{P} - I)^{2m} \prod_{i=1}^{n-m} (\mathcal{P} - \lambda_i I).$$

To prove the upper bound on the dimension of the Krylov subspace, we need to show that the order of the minimum polynomial is less than or equal to $n - m + 2$.

Expanding the polynomial $(\mathcal{P} - I)\prod_{i=1}^{n-m}(\mathcal{P} - \lambda_i I)$ of degree $n-m+1$, we obtain a matrix of the form

$$\begin{bmatrix} 0 & 0 & 0 \\ \left[\prod_{i=1}^{n-m}(S - \lambda_i I)\right]\Theta & (S - I)\prod_{i=1}^{n-m}(S - \lambda_i I) & 0 \\ \Phi_{n-m} & \Gamma\prod_{i=1}^{n-m}(S - \lambda_i I) & 0 \end{bmatrix}, \tag{3.2}$$

where $S = (Z^TGZ)^{-1}Z^TAZ$ and

$$\Phi_{n-m} = \Upsilon \prod_{i=1}^{n-m} (I - \lambda_i I) + \Gamma \left[\prod_{i=1}^{n-m-1} (S - \lambda_i I)\right]\Theta + \Gamma\Psi_{n-m}.$$

Here, Ψ_{n-m} is defined by the recursive formula

$$\Psi_{n-m} = \left[\Psi_{n-m-1} + \left[\prod_{i=1}^{n-m-2} (S - \lambda_i I) \right] \Theta \right] (I - \lambda_{n-m} I) \qquad (n - m > 2),$$

with base cases $\Psi_1 = 0$ and $\Psi_2 = \Theta(I - \lambda_2 I)$.

Note that the $(2,1)$, $(2,2)$, and $(3,2)$ entries of matrix (3.2) are in fact zero, since the λ_i $(i = 1, \ldots, n-m)$ are the eigenvalues of S, which is similar to a symmetric matrix and is thus diagonalizable. Thus, (3.2) may be written as

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \Phi_{n-m} & 0 & 0 \end{bmatrix}, \tag{3.3}$$

and what remains is to distinguish two different cases for the value of Φ_{n-m}, that is, $\Phi_{n-m} = 0$ and $\Phi_{n-m} \neq 0$. In the former case, the order of the minimum polynomial of $\mathcal{P}$ is less than or equal to $n-m+1$, and thus the dimension of the Krylov subspace $\mathcal{K}(\mathcal{P}, b)$ is of the same order. In the latter case, the dimension of $\mathcal{K}(\mathcal{P}, b)$ is less than or equal to $n-m+2$ since multiplication of (3.3) by another factor $(\mathcal{P} - I)$ gives the zero matrix. ∎

The upper bound on the dimension of the Krylov subspace, as stated in Theorem 3.5, can be reduced in the special case when $(Z^T G Z)^{-1}(Z^T A Z)$ has repeated eigenvalues. This result is stated in Theorem 3.7. The following (randomly generated) example shows that the bound in Theorem 3.5 is attainable.

Example 3.6. Let $A \in \mathbb{R}^{6\times 6}$ and $B^T \in \mathbb{R}^{6\times 2}$ be given by

$$A = \begin{bmatrix} 2.69 & 1.62 & 1.16 & 1.60 & 0.81 & -1.97 \\ 1.62 & 6.23 & -1.90 & 1.89 & 0.90 & 0.05 \\ 1.16 & -1.90 & 4.01 & -0.16 & -0.16 & -1.60 \\ 1.60 & 1.89 & -0.16 & 1.45 & 0.01 & -0.89 \\ 0.81 & 0.90 & -0.16 & 0.01 & 1.94 & 0.38 \\ -1.97 & 0.05 & -1.60 & -0.89 & 0.38 & 5.38 \end{bmatrix}, \quad B^T = \begin{bmatrix} 0 & -0.59 \\ -0.59 & 0 \\ 0 & 2.00 \\ 0 & 0 \\ -0.02 & 0 \\ 0.33 & 0.17 \end{bmatrix},$$

and assume that $G = \operatorname{diag}(A)$. For the above matrices the $(3,1)$ entry of (3.3) is

$$\Phi_4 = \begin{bmatrix} 0 & -0.07 \\ -0.22 & -0.02 \end{bmatrix}.$$

It follows that the minimum polynomial is of order 6, and thus the bound given in Theorem 3.5 is sharp.

THEOREM 3.7. *Let $\mathcal{A} \in \mathbb{R}^{(n+m)\times(n+m)}$ be a symmetric and indefinite matrix of the form*

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix},$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric and $B \in \mathbb{R}^{m\times n}$ is of full rank. Assume $m < n$, $\mathcal{A}$ is nonsingular, and $\mathcal{A}$ is preconditioned by a matrix of the form

$$\mathcal{G} = \begin{bmatrix} G & B^T \\ B & 0 \end{bmatrix},$$

where $G \in \mathbb{R}^{n\times n}$ is symmetric, $G \neq A$, and $B \in \mathbb{R}^{m\times n}$ is as above. Furthermore, let Z be an $n \times (n-m)$ basis for the nullspace of B and assume $(Z^TGZ)^{-1}(Z^TAZ)$ has k $(1 \leq k \leq n-m)$ distinct eigenvalues λ_i $(1 \leq i \leq k)$ of respective multiplicity μ_i, where $\sum_{i=1}^{k} \mu_i = n-m$. Then the dimension of the Krylov subspace $\mathcal{K}(\mathcal{P}, b)$ is at most $k+2$.

Proof. The proof is similar to the one for Theorem 3.5. In the case when $(Z^TGZ)^{-1}(Z^TAZ)$ has k distinct eigenvalues of multiplicity μ_i, we may, without loss of generality, write the characteristic polynomial of $\mathcal{P}$ as

$$(\mathcal{P}-I)^{2m-2}\left[\prod_{i=1}^{k}(\mathcal{P}-\lambda_i I)^{\mu_i-1}\right](\mathcal{P}-I)\underbrace{(\mathcal{P}-I)\left[\prod_{i=1}^{k}(\mathcal{P}-\lambda_i I)\right]}_{(\dagger)}.$$

Expanding (†), we obtain the matrix

$$\begin{bmatrix} 0 & 0 & 0 \\ \left[\prod_{i=1}^{k}(S-\lambda_i I)\right]\Theta & (S-I)\prod_{i=1}^{k}(S-\lambda_i I) & 0 \\ \Phi_k & \Gamma\prod_{i=1}^{k}(S-\lambda_i I) & 0 \end{bmatrix}, \tag{3.4}$$

where $S = (Z^TGZ)^{-1}Z^TAZ$ and

$$\Phi_k = \Upsilon\prod_{i=1}^{k}(I-\lambda_i I) + \Gamma\left[\prod_{i=1}^{k-1}(S-\lambda_i I)\right]\Theta + \Gamma\Psi_k.$$

Here, Ψ_k is given by the recursive formula

$$\Psi_k = \left[\Psi_{k-1} + \left[\prod_{i=1}^{k-2}(S-\lambda_i I)\right]\Theta\right](I-\lambda_k I) \qquad (k>2),$$

with base cases $\Psi_1 = 0$ and $\Psi_2 = \Theta(I-\lambda_2 I)$.

Note that the $(2,1)$, $(2,2)$, and $(3,2)$ blocks of matrix (3.4) are in fact zero. It follows that, for $\Phi_k \neq 0$, a further multiplication of (3.4) by $(\mathcal{P}-I)$ gives the zero matrix and thus the dimension of Krylov subspace $\mathcal{K}(\mathcal{P}, b)$ is less then or equal to $k+2$. ∎

To verify that the bound in Theorem 3.7 is attainable, consider the following example.

Example 3.8. Let $A \in \mathbb{R}^{4\times4}$, $G \in \mathbb{R}^{4\times4}$, and $B^T \in \mathbb{R}^{4\times1}$ be given by

$$A = \begin{bmatrix} 6 & 0 & 0 & 0 \\ 0 & 6 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}, \quad G = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{2} \end{bmatrix},$$

and $B = \begin{bmatrix} 0 & 0 & 10^{-3} & 10^{-3} \end{bmatrix}$. Then two of the $n-m$ eigenvalues that are defined by the generalized eigenvalue problem (2.9) are distinct and given by $[2, 4]$. It follows that the $(3,1)$ entry of (3.4) is nonzero with

$$\Phi_3 = \begin{bmatrix} -1071.41 \end{bmatrix},$$

and so the minimum polynomial is of order 4.

4. Implementation. There are various strategies that can be used to implement the proposed preconditioner, two of which are used in the numerical results in section 5. The first strategy applies the standard (preconditioned) GMRES algorithm [25], where the preconditioner step is implemented by means of a symmetric indefinite factorization of (1.4). Such a factorization of the preconditioner may be much less demanding than the factorization of the initial coefficient matrix if G is a considerably simpler matrix than A. The second approach, discussed in the next section, is based on an algorithm that solves a reduced linear system [12].

4.1. CG on a reduced linear system. In [12], Gould, Hribar, and Nocedal propose a CG-like algorithm to solve equality constrained quadratic programming problems such as the one described in Example 1.1. The algorithm is based on the idea of computing an implicit basis Z which spans the nullspace of B. The nullspace basis is then used to remove the constraints from the system of equations, thus allowing the application of the CG method to the (positive definite) reduced system.

Assume that $W_{zz} = Z^T G Z$ is a symmetric and positive definite preconditioner matrix of dimension $(n-m)\times(n-m)$ and Z is an $n\times(n-m)$ matrix. The algorithm can then be stated as follows.

ALGORITHM 4.1. PRECONDITIONED CG FOR A REDUCED SYSTEM.

(1) Choose an initial point x_1 satisfying $Bx_1 = b_2$.

(2) Compute

$$\begin{aligned} r &= Ax_1 - b_1, \\ r &\leftarrow r - B^T y, \qquad (4.1)\\ g &= Z W_{zz}^{-1} Z^T r, \qquad (4.2)\\ p &= -g. \end{aligned}$$

(3) Repeat the following steps until $|\,(r^+)^T g^+\,| \to 0$ is satisfied:

$$\begin{aligned} \alpha &= r^T g / p^T A p, \\ x_1 &\leftarrow x_1 + \alpha p, \\ r^+ &= r + \alpha A p, \\ r^+ &\leftarrow r^+ - B^T y, \qquad (4.3)\\ g^+ &= Z W_{zz}^{-1} Z^T r^+, \qquad (4.4)\\ \beta &= (r^+)^T g^+ / r^T g, \\ p &\leftarrow -g^+ + \beta p, \\ g &\leftarrow g^+, \\ r &\leftarrow r^+. \end{aligned}$$

The vectors y in (4.1) and (4.3) are arbitrary, but in practice they are chosen to try to make r and r^+ close to $B^T y$. The computation of the preconditioned residual in (4.4) is often the most expensive computational factor in the algorithm. Gould, Hribar, and Nocedal [12] suggest avoiding the explicit use of the nullspace Z, but instead computing g^+ by applying a symmetric indefinite factorization of

$$\begin{bmatrix} G & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} g^+ \\ v^+ \end{bmatrix} = \begin{bmatrix} r^+ \\ 0 \end{bmatrix}. \qquad (4.5)$$

This factorization also determines the vector y in (4.3), where $y = v^+$. The vector y in (4.1) and the preconditioned residual (4.2) are determined in a similar manner.

In practice, (4.5) can often be factored efficiently by using the MA27 package of the Harwell Subroutine Library when G is a simple matrix block, whereas the direct application of MA27 to the original system (1.1) is limited by space requirements as well as time for large enough systems [8]. In this context, the factorization consists of three separate routines, the first two of which analyze and factorize the matrix in (4.5). They need to be executed only once in step (1) of Algorithm 4.1. Repeated calls to the third routine within MA27 apply forward- and backward-substitutions to find the initial point x_1 in step (1), solve for y in (4.1) and g in (4.2), and also find y in (4.3) and g^+ in (4.4).

Remark 4.2. The computation of the projected residual g^+ is often accompanied by significant roundoff errors if this vector is much smaller than the residual r^+. *Fixed precision iterative refinement* is used in (4.3) to redefine r^+ so that its norm is closer to that of g^+. In the same manner, (4.1) is used to refine the initial residual r. The result is a dramatic reduction of the roundoff errors in the projection operation; see Gould, Hribar, and Nocedal [12].

Remark 4.3. The current implementations, which apply the proposed preconditioner, require an explicit factorization of the indefinite matrix $\mathcal{G}$. For larger systems, the complexity of the two methods introduced in this section can grow substantially for increasing problem dimensions and may not be feasible in comparison to the complexity of methods such as (block-) diagonal and triangular preconditioners. The authors are currently investigating a new implementation approach which is mainly based on an adaptation of the direct solver MA27. Preliminary numerical results show the beneficial features of the proposed approach.

5. Numerical results. We now present the results of numerical experiments that reinforce the analysis given in previous sections. The test problems employed are partly matrices that arise in linear and nonlinear optimization (see, for instance, Example 1.1) and partly randomized sparse matrices which arise in areas such as those described by Example 1.2. In particular, the problems in Table 5.1 are of the form (1.3), with A representing a *stiffness matrix*, whereas the examples in Table 5.2 are of the same form, but with A being a *mass matrix* or a *discrete Laplacian*. The four test examples of Table 5.3 are taken from the *constrained and unconstrained testing environment (CUTE)* [5], where the individual problems have the following characteristics.

BLOWEYC is the convex quadratic programming problem

$$\text{minimize } u(s)^T Au(s) + u(s)^T w(s) - v(s)^T Au(s) - 2.0v(s)^T w(s) - u(s)^T v(s)$$
$$\text{subject to } \; Aw(s) = u(s), \;\; u(s) \in [-1, 1], \;\; \text{and } \int_0^1 u(s)\, ds = 0.4,$$

which is a simplification of a problem given in [4]. It is part of a modeling exercise, and the actual value of the solution is not used in a genuine practical application.

*CVXQP*1 is a convex quadratic program whose constraints are linear and which has continuous first and second order derivatives. It is a purely academic problem and has been constructed specifically for test purposes.

*MOSARQP*2 is also a convex quadratic problem which was first proposed by Morales-Pérez and Sargent [17]. Once again, it is a purely academic problem which has been constructed specifically for test purposes.

*FIT*2*P* is a dual version of a problem that originates from a model for fitting linear inequalities to data by minimization of a sum of piecewise linear penalties.

TABLE 5.1
Random test problems with A being a stiffness matrix ($G = \text{diag}(A)$).

		Stiff I	Stiff II	Stiff III	Stiff IV	Stiff V	Stiff VI
	n	578	2178	8450	8450	33282	33282
	m	32	128	236	46	128	512
	Nonzero entries in A	2316	9740	39948	39948	161804	161804
	Nonzero entries in B	427	1871	3600	685	7726	15830
MINRES	# of iterations	174	387	639	515	1579	1571
	Time in seconds	0.4	3.0	18.1	12.7	170.4	175.5
PGMRES	# of iterations	46	87	228	242	313	319
	Time in seconds	0.2	3.5	92.4	104.2	695.7	730.6
RCG	# of iterations	35	72	197	214	294	295
	Time in seconds	0.1	1.0	7.7	7.6	44.8	48.7
MA27	Time in seconds	0.1	0.3	1.7	2.1	7.2	11.8

As indicated throughout, all matrices are of the form

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}, \tag{5.1}$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric, $B \in \mathbb{R}^{m\times n}$ has full rank, and $m \le n$. Four different approaches to finding solutions to (1.1) are compared: three iterative algorithms based on Krylov subspaces, and the direct solver MA27, which applies a sparse variant of Gaussian elimination; see Duff and Reid [8]. To investigate possible favorable aspects of preconditioning, it makes sense to compare unpreconditioned with preconditioned solution strategies. The indefinite nature of matrix (5.1) suggests the use of MINRES in the unpreconditioned case. Of course, positive definite preconditioning could be employed with MINRES; see, for example, Murphy, Golub, and Wathen [18], or Silvester and Wathen [26], and Wathen and Silvester [27] in case of the Stokes problem. However, this is not done here. We refer to Battermann and Heinkenschloss [3] for some numerical results of this type.

As outlined in section 4, we employ two slightly different strategies in order to implement the preconditioner $\mathcal{G}$. The first method applies a standard (full) GMRES($\mathcal{A}$) code (PGMRES in Tables 5.1, 5.2, and 5.3), which is mathematically equivalent to MINRES($\mathcal{A}$) for symmetric matrices $\mathcal{A}$, whereas the second approach implements Algorithm 4.1 (RCG in Tables 5.1, 5.2 and 5.3). The choice $G = \text{diag}(A)$ in the preconditioner is made for both PGMRES and RCG.

All tests were performed on a SUN Ultra SPARCII-300MHz (ULTRA-30) workstation with 245 MB physical RAM and running SunOS release 5.5.1. Programs were written in standard Fortran 77 using the SUN WorkShop *f77* compiler (version 4.2) with the *-0* optimization flag set. In order to deal with large sparse matrices, we implemented an index storage format that only stores nonzero matrix elements; see Press, Teukolsky, and Vetterling [22]. The termination criterion for all iterative methods was taken to be a residual vector of order less than 10^{-6} in the 2-norm.

As part of its analysis procedure, MA27 accepts the pattern of some coefficient matrix and chooses pivots for the factorization and solution phases of subsequent routines. The amount of pivoting is controlled by the special threshold parameter u ($-1/2 \le u \le 1/2$). Modifying u within its positive range influences the accuracy of the resulting solution, whereas a negative value prevents any pivoting; see Duff and Reid [8]. In this context, the early construction of some of the test examples with the default value $u = 0.1$ was accompanied by difficulties in the form of memory limitations. We met the trade-off between less use of memory and solutions of high

TABLE 5.2
Random test problems with A being a mass matrix ($G = \mathrm{diag}(A)$).

		Mass I	Mass II	Mass III	Mass IV	Mass V	Mass VI
	n	578	2178	8450	8450	33282	33282
	m	32	128	236	46	128	512
	Nonzero entries in A	3128	13400	55448	55448	225560	225560
	Nonzero entries in B	427	1871	3600	686	7726	15830
MINRES	# of iterations	135	312	162	67	473	845
	Time in seconds	0.3	2.9	6.8	3.5	65.1	117.9
PGMRES	# of iterations	11	11	11	11	12	12
	Time in seconds	0.1	0.7	2.6	2.4	21.0	31.3
RCG	# of iterations	9	10	11	11	12	12
	Time in seconds	0.1	0.7	2.6	2.3	20.3	26.6
MA27	Time in seconds	0.1	0.4	6.5	4.5	–	24.1

TABLE 5.3
CUTE test problems ($G = \mathrm{diag}(A)$).

		BLOWEYC	CVXQP1	MOSARQP2	FIT2P
	n	1002	100	930	13525
	m	502	100	30	3000
	Nonzero entries in A	3004	672	1020	13525
	Nonzero entries in B	2503	295	148	50284
MINRES	# of iterations	363	> 200	51	180
	Time in seconds	1.8	–	0.1	14.6
PGMRES	# of iterations	2	2	6	1
	Time in seconds	0.4	0.1	0.2	13.2
RCG	# of iterations	2	2	4	1
	Time in seconds	0.3	0.1	0.1	12.1
MA27	Time in seconds	0.2	0.1	0.1	11.5

enough accuracy by choosing the parameter value $u = 0.01$ for the results in Tables 5.1, 5.2, and 5.3. However, this resulted in a floating point exception when using MA27 with one of the larger examples in Table 5.2, which is indicated by a dash.

The numerical results suggest that the inclusion of the $(1, 2)$ and $(2, 1)$ block of $\mathcal{A}$ into the preconditioner, together with $G = \mathrm{diag}(A)$, results in a considerable reduction of iterations, where the appropriate bounds of Theorems 3.2, 3.5, and 3.7 are attained in all cases. Specifically, Theorem 3.2 applies in the context of problem CVXQP1.

Test problems Stiff III to Stiff VI in Table 5.1 emphasize the storage problems that are associated with the use of long recurrences in the PGMRES algorithm. The time required to find solutions to Stiff III to Stiff VI via the PGMRES algorithm is not comparable to any of the other methods, which is due to the increased storage requirements and the data trafficking involved. A solution to the memory problems is to restart PGMRES after a prescribed number of iterations, but the iteration counts for such restarts would not be comparable with those of full PGMRES.

The examples in Table 5.2 indicate the beneficial strategy of preconditioning the mass matrix A with its diagonal ($G = \mathrm{diag}(A)$); see Wathen [28]. As expected from the theory described in [28], PGMRES, as well as RCG, converge in virtually all cases in only a few number of steps and also outperform the solutions found by the direct solver MA27 as far as the time measurements are concerned.

These measurements for MA27 suggest that the preconditioned CG algorithm, discussed in section 4.1, is a suitable alternative to the direct solver. Whereas both

MINRES and especially PGMRES are in most cases considerably slower than MA27, the timings for RCG are in virtually all cases comparable. For problems of large enough dimension or bandwidth, the resources required by MA27 must become prohibitive, in which case RCG becomes even more competitive.

6. Conclusion. In this paper, we investigated a new class of preconditioner for indefinite linear systems that incorporate the $(1,2)$ and $(2,1)$ blocks of the original matrix. These blocks are often associated with constraints. In our numerical results, we used a simple diagonal matrix G to approximate the $(1,1)$ block of $\mathcal{A}$, even though other approximations, such as an incomplete factorization of A, are possible. We first showed that the inclusion of the constraints into the preconditioner clusters at least $2m$ eigenvalues at 1, regardless of the structure of G. However, unless G represents A exactly, $\mathcal{P}$ does not have a complete set of linearly independent eigenvectors and thus the standard convergence theory for Krylov subspace methods is not readily applicable.

To find an upper bound on the number of iterations required to solve linear systems of the form (1.1) by means of appropriate subspace methods, we used a minimum polynomial argument. Theorem 3.2 considers the special condition $n = m$, in which case termination is guaranteed in two iterations. For $m < n$, Theorem 3.5 gives a general (sharp) upper bound on the dimension of the Krylov subspace, whereas Theorem 3.7 defines a considerably stronger result if some of the $n - m$ eigenvalues, defined by $(Z^T GZ)^{-1}(Z^T AZ)$, are repeated.

In the special case when G is a positive definite matrix block, we were able to apply Cauchy's interlacing theorem in order to give upper and lower bounds for the $n - m$ eigenvalues that are defined by the $(2,2)$ block of matrix (2.3).

To confirm the analytical results in this paper, we used three different subspace methods, MINRES of Paige and Saunders [19] for the unpreconditioned matrix system and RCG of Gould, Hribar, and Nocedal [12] and PGMRES of Saad and Schultz [25] for the preconditioned case. Overall, the results show that the number of iterations is decreased substantially if preconditioning is applied. The Krylov subspaces, which are built during the execution of the two preconditioned implementations, are in theory of equal dimension for any one of the 16 test examples, and thus PGMRES and RCG can be expected to terminate in the same number of steps. However, convergence to any prescribed tolerance may occur for a different number of steps since PGMRES and RCG minimize different quantities. This can be seen in some of the examples. Nevertheless, we note that convergence for both methods is attained much earlier than suggested by the bounds in Theorems 3.2, 3.5, and 3.7.

We have not compared our results with positive definite preconditioning in connection with MINRES—this is known to be a very competitive approach in various applications.

Acknowledgments. The authors would like to thank Gene H. Golub for his insightful comments during the process of the work and an anonymous referee for helpful suggestions.

REFERENCES

[1] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, Cambridge, UK, 1996.
[2] R. Barrett, M. W. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. van der Vorst, *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, SIAM, Philadelphia, 1993.

[3] A. Battermann and M. Heinkenschloss, *Preconditioners for Karush-Kuhn-Tucker matrices arising in the optimal control of distributed systems*, in Optimal Control of Partial Differential Equations, Internat. Ser. Numer. Math. 126, W. Desch, F. Kappel, and K. Kunisch, eds., Birkhäuser, Berlin, 1998, pp. 15–32.

[4] J. F. Blowey and C. M. Elliot, *The Cahn-Hilliard gradient theory for phase separation with non-smooth free energy. Part* II: *Numerical analysis*, European J. Appl. Math., 3 (1992), pp. 147–179.

[5] I. Bongartz, A. Conn, N. Gould, and Ph. Toint, *CUTE: Constrained and unconstrained testing environment*, ACM Trans. Math. Software, 21 (1995), pp. 123–160.

[6] S. L. Campbell, I. C. F. Ipsen, C. T. Kelley, and C. D. Meyer, *GMRES and the minimum polynomial*, BIT, 36 (1996), pp. 664–675.

[7] T. F. Coleman, *Linearly constrained optimization and projected preconditioned conjugate gradients*, in Proceedings of the Fifth SIAM Conference on Applied Linear Algebra, Proc. Appl. Math. 72, J. Lewis, ed., SIAM, Philadelphia, 1994, pp. 188–122.

[8] I.S. Duff and J. K. Reid, *The multifrontal solution of indefinite sparse symmetric linear equations*, ACM Trans. Math. Software, 9 (1983), pp. 302–325.

[9] H. C. Elman, *Perturbation of eigenvalues of preconditioned Navier-Stokes operators*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 733–751.

[10] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1996.

[11] G. H. Golub and A. J. Wathen, *An iteration for indefinite systems and its application to the Navier–Stokes equations*, SIAM J. Sci. Comput., 19 (1998), pp. 530–539.

[12] N. I. M. Gould, M. E. Hribar, and J. Nocedal, *On the Solution of Equality Constrained Quadratic Programming Problems Arising in Optimization*, Technical Report RAL-TR-1998-069, Rutherford Appleton Laboratory, Oxfordshire, UK, 1998.

[13] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, Frontiers in Appl. Math. 17, SIAM, Philadelphia, 1997.

[14] A. Greenbaum, and L. Gurvits, *Max-min properties of matrix factor norms*, SIAM J. Sci. Comput., 15 (1994), pp. 348–358.

[15] W. D. Joubert and T. A. Manteuffel, *Iterative methods for non-symmetric linear systems*, in Iterative Methods for Large Linear Systems, D. R. Kincaid and L. J. Hayes, eds., Academic Press, New York, 1990, pp. 149–171.

[16] L. Lukšan and J. Vlček, *Indefinitely preconditioned inexact Newton method for large sparse equality constrained non-linear programming problems*, Numer. Linear Algebra Appl., 5 (1998), pp. 219–247.

[17] J. L. Morales-Pérez and R. W. H. Sargent, *On the implementation and performance of an interior point method for large sparse convex quadratic programming*, Optim. Methods Softw., 1 (1992), pp. 153–168.

[18] M. F. Murphy, G. H. Golub, and A. J. Wathen, *A note on preconditioning for indefinite linear systems*, Technical Report NA99107, Numerical Analysis Group, University of Oxford, UK, 1999.

[19] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.

[20] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Classics Appl. Math. 20, SIAM, Philadelphia, 1998.

[21] B. T. Polyak, *The conjugate gradient method in extremal problems*, USSR Comput. Math. Math. Phys., 9 (1969), pp. 94–112.

[22] W. H. Press, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in Fortran: The Art of Scientific Computing*, Cambridge University Press, Cambridge, UK, 1992.

[23] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer-Verlag, Berlin, 1997.

[24] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing Company, Boston, 1996.

[25] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[26] D. J. Silvester and A. J. Wathen, *Fast iterative solution of stabilised Stokes systems. Part* II: *Using general block preconditioners*, SIAM J. Numer. Anal., 31 (1994), pp. 1352–1367.

[27] A. J. Wathen and D. J. Silvester, *Fast iterative solution of stabilised Stokes systems. Part* I: *Using simple diagonal preconditioners*, SIAM J. Numer. Anal., 30 (1993), pp. 630–649.

[28] A. J. Wathen, *Realistic eigenvalue bounds for the Galerkin mass matrix*, IMA J. Numer. Anal., 7 (1987), pp. 449–457.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 4, pp. 1318–1323
© 2000 Society for Industrial and Applied Mathematics

SOME PROPERTIES OF SYMMETRIC QUASI-DEFINITE MATRICES*

ALAN GEORGE†, KH. IKRAMOV‡, AND A. B. KUCHEROV‡

Abstract. Symmetric quasi-definite matrices arise in numerous applications, notably in interior point methods in mathematical programming. Several authors have derived various properties of these matrices. This article provides a list of some previously known properties and adds a number of others that are believed to be new.

Key words. eigenvalues, eigenvectors, principal angles between subspaces

AMS subject classification. 65F10

PII. S0895479897329400

1. Introduction.

Following [1] and [8], we say that a symmetric matrix K with real entries is *symmetric quasi-definite* (SQD) if there exists a permutation matrix Π such that $\Pi K \Pi^T$ has the form

$$\begin{pmatrix} H & A^T \\ A & -G \end{pmatrix}, \tag{1.1}$$

where $H \in \mathbb{R}^{n\times n}$ and $G \in \mathbb{R}^{m\times m}$ are symmetric positive definite matrices with $m, n > 0$. For the purposes of this paper, it will be enough to consider the case when K itself (rather than some permutation of it) has the form (1.1).

A number of important properties of SQD matrices have been established in [1] and [8], as follows.

PROPERTY 1. *An SQD matrix K is nonsingular and its inverse is also quasi-definite:*

$$K^{-1} = \begin{pmatrix} E & B^T \\ B & -F \end{pmatrix}, \qquad E \succ 0,\ F \succ 0. \tag{1.2}$$

PROPERTY 2. *An SQD matrix K is strongly factorizable. That is, for every permutation P there exist a diagonal matrix D and a unit lower-triangular matrix L such that*

$$PKP^T = LDL^T. \tag{1.3}$$

Here D in general will have both positive and negative diagonal elements.

PROPERTY 3. *Let*

$$J = I_n \oplus (-I_m). \tag{1.4}$$

*Received by the editors October 30, 1997; accepted for publication (in revised form) by P. E. Gill on October 12, 1999; published electronically April 18, 2000. This work was supported in part by Natural Sciences and Engineering Research Council of Canada grant OGP 000811.
http://www.siam.org/journals/simax/21-4/32940.html
†Department of Computer Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada (jageorge@sparse1.uwaterloo.ca).
‡Moscow State University, Faculty of Computational Mathematics and Cybernetics, Moscow 119899, Russia (ikramov@cmc.msk.su, kucherov@cmc.msk.su).

Then, for any SQD matrix K, the matrix

$$\tilde{K} = KJ = \begin{pmatrix} H & -A^T \\ A & G \end{pmatrix} \tag{1.5}$$

is unsymmetric positive definite; i.e., its symmetric part

$$\frac{1}{2}(\tilde{K} + \tilde{K}^T) = H \oplus G \tag{1.6}$$

is positive definite.

Note that an unsymmetric matrix $\tilde{K}$ with a positive definite symmetric part has its field of values contained in the right half-plane; in particular, all eigenvalues of $\tilde{K}$ have positive real parts.

This paper supplements the list above with additional properties of SQD matrices. Some of them are scattered in the literature and not well known, while others are believed to be new.

Note that the properties above as well as the properties in the main body of the paper are valid, with the evident alterations, for complex Hermitian quasi-definite matrices, which have the form

$$\begin{pmatrix} H & A^* \\ A & -G \end{pmatrix}, \qquad H = H^* \succ 0,\ G = G^* \succ 0.$$

2. Properties of SQD matrices. Scaling A in (1.1) by a parameter ρ yields the family of SQD matrices

$$L(\rho) = \begin{pmatrix} H & \rho A^T \\ \rho A & -G \end{pmatrix}, \tag{2.1}$$

and the eigenvalues of $L(\rho)$ are (naturally) functions of ρ:

$$\lambda_i = \lambda_i(\rho), \qquad i = 1, \ldots, N, \tag{2.2}$$

where

$$N = n + m. \tag{2.3}$$

Since the diagonal block H in (2.1) is positive definite, it follows from the Courant-Fisher theorem [4, p. 179] that $L(\rho)$ has at least n positive eigenvalues. Using the negative definite submatrix $-G$ in a similar way implies that at least m eigenvalues of $L(\rho)$ are negative. These observations together with (2.3) yield the following property.

PROPERTY 4. *For any value of ρ, n eigenvalues $\lambda_i(\rho)$ of (2.1) are positive and m eigenvalues are negative.*

From this point on it will be assumed that the functions

$$\lambda_i(\rho), \qquad i = 1, \ldots, n, \tag{2.4}$$

are positive and the functions

$$\lambda_i(\rho), \qquad i = n+1, \ldots, N, \tag{2.5}$$

are negative. Note that

$$L(-\rho) = J^{-1}L(\rho)J = JL(\rho)J.$$

Thus, $\lambda_i(\rho)$ are even functions of ρ and one can restrict ρ to the interval $0 \le \rho < \infty$.

The following assertion, which was proved in [7], is stated as follows.

PROPERTY 5. *For $\rho \geq 0$, eigenvalues* (2.4) *and* (2.5) *are nondecreasing and nonincreasing functions of ρ, respectively.*

In [5], Thompson gave another proof of Property 5 based on using the classical theorems by Cauchy, Weyl, and Ostrowski. The authors in [5] and [7] obtained their results before Vanderbei coined the term "quasi-definite matrix" and refer to these matrices as "partitioned Hermitian matrices involving a parameter."

Note that inserting a parameter into the diagonal blocks of an SQD matrix

$$\begin{pmatrix} \rho H & A^T \\ A & -\rho G \end{pmatrix}$$

does not lead to such a clear-cut result as Property 5. However, it is possible to obtain some results about the behavior of the eigenvalues of another parameterized family of SQD matrices, namely,

$$K(\rho) = \begin{pmatrix} H + \rho I & A^T \\ A & -G - \rho I \end{pmatrix}. \tag{2.6}$$

To this end, some properties of the eigenvectors of an SQD matrix

$$K = \begin{pmatrix} H & A^T \\ A & -G \end{pmatrix} \tag{2.7}$$

are required. These results are of interest in their own right and appear to be new.

Let λ and μ be a positive and a negative eigenvalue, respectively, of (2.7), with x and y being the corresponding eigenvectors. Partition x and y as

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \qquad y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \qquad x_1, y_1 \in \mathbb{R}^n. \tag{2.8}$$

PROPERTY 6. *For vectors x, y in* (2.8)*, the following inequalities hold:*

$$\|x_1\|_2 > \|x_2\|_2, \tag{2.9}$$

$$\|y_2\|_2 > \|y_1\|_2. \tag{2.10}$$

Proof. Consider (2.9) and let

$$z = Jx = \begin{pmatrix} x_1 \\ -x_2 \end{pmatrix}.$$

Multiplying the relation $Kx = \lambda x$ on the left by z^T then yields

$$z^T K x = z^T \tilde{K} z = \lambda z^T x = \lambda(\|x_1\|_2^2 - \|x_2\|_2^2).$$

Since $z^T \tilde{K} z > 0$ (from Property 3) and $\lambda > 0$, the inequality (2.9) is immediate. The proof of (2.10) is similar and is omitted. □

Relation (2.9) assumes an interesting form for $n = 1$; letting $x^T = (\xi_1, \xi_2, \ldots, \xi_N)$, (2.9) implies

$$|\xi_1|^2 > |\xi_2|^2 + \cdots + |\xi_N|^2.$$

Similarly, for $m = 1$ and $y^T = (\eta_1, \ldots, \eta_{N-1}, \eta_N)$, relation (2.10) gives

$$|\eta_1|^2 + |\eta_2|^2 + \cdots + |\eta_{N-1}|^2 < |\eta_N|^2.$$

Property 6 can be extended from a single eigenvector to an *orthonormal* set $u_1, \ldots, u_k$ of eigenvectors of K corresponding to eigenvalues $\lambda_1, \ldots, \lambda_k$ having the same sign. For definiteness, assume that $\lambda_1, \ldots, \lambda_k$ are positive. Let

$$U = (u_1|\cdots|u_k) = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}, \qquad U_1 \in \mathbb{R}^{k\times k}, \tag{2.11}$$

and

$$\Lambda = \text{diag } (\lambda_1, \ldots, \lambda_k). \tag{2.12}$$

In a manner similar to the case $k = 1$, introduce the matrix

$$V = JU = \begin{pmatrix} U_1 \\ -U_2 \end{pmatrix}.$$

Multiplying the relation $KU = U\Lambda$ on the left by V^T gives

$$V^T K U = V^T \tilde{K} V = (V^T U)\Lambda = (U_1^T U_1 - U_2^T U_2)\Lambda. \tag{2.13}$$

The product of the positive definite matrix Λ and the real symmetric matrix

$$W = U_1^T U_1 - U_2^T U_2$$

is a diagonalizable matrix, all of whose eigenvalues are real [4, Theorem 7.6.3]. According to (2.13), $W\Lambda$ is congruent to the positive definite matrix $\tilde{K}$; hence, the eigenvalues of $W\Lambda$ must belong to the half-plane $\pi_+ = \{z \mid \text{Re } z > 0\}$ and, therefore, are positive. But then W must be positive definite (see again Theorem 7.6.3 in [4]). This yields the following generalization of Property 6.

PROPERTY 7. *Let U, U_1, and U_2 be as in* (2.11). *Then*

$$U_1^T U_1 \succ U_2^T U_2. \tag{2.14}$$

Inequality (2.14) has a number of implications. The most obvious are

$$\|U_1\|_F^2 = \text{tr } (U_1^T U_1) > \text{tr } (U_2^T U_2) = \|U_2\|_F^2, \tag{2.15}$$

and

$$\|U_1\|_2 > \|U_2\|_2. \tag{2.16}$$

Of course these relations are simply particular manifestations of the fact that inequality (2.14) implies the corresponding inequalities between the singular values $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_k$ and $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_k$ of the matrices U_1 and U_2 in (2.14) (see [4, Corollary 7.7.4]):

$$\alpha_i \geq \beta_i, \qquad i = 1, \ldots, k. \tag{2.17}$$

The orthonormality of U implies

$$U_1^T U_1 + U_2^T U_2 = I_k,$$

which gives (see (2.14))

$$U_1^T U_1 \succ \frac{1}{2} I_k.$$

Hence,

$$\alpha_i > \frac{1}{\sqrt{2}}, \qquad i = 1, \ldots, k, \tag{2.18}$$

$$\beta_i < \frac{1}{\sqrt{2}}, \qquad i = 1, \ldots, k.$$

These provide the following result.

PROPERTY 8. *Let U, U_1, and U_2 be as in* (2.11). *Then*

(a) *U_1 is of full column rank and $\|U_1^+\|_2 < \sqrt{2}$;*

(b) $\|U_2\|_2 < \frac{1}{\sqrt{2}}$;

(c) $\|U_2 U_1^+\|_2 < 1$.

One can obtain yet another formulation of the inequalities (2.18). Let $\mathcal{L}_k$ and $\mathcal{M}_k$ be the spans of the vectors $u_1, \ldots, u_k$ in (2.11) and the coordinate vectors $e_1, \ldots, e_k$, respectively.

PROPERTY 9. *All principal angles between the subspaces $\mathcal{L}_k$ and $\mathcal{M}_k$ are less than $\pi/4$.*

See Golub and Van Loan [2, p. 584] for the definition of principal angles between subspaces.

PROPERTY 10. *The positive eigenvalues* (2.4) *of $K(\rho)$ of* (2.6) *are increasing functions of ρ. Its negative eigenvalues* (2.5) *are decreasing functions of ρ.*

Proof. For definiteness, consider the positive eigenvalues (2.4). If, for a given ρ_0, all the eigenvalues $\lambda_i(\rho_0), 1 \le i \le n$, are simple, then their derivatives along the set (2.6) exist and are equal to

$$\lambda_i'(\rho_0) = \frac{(Ju_i, u_i)}{(u_i, u_i)}, \tag{2.19}$$

where u_i is the eigenvector of $K(\rho_0)$ corresponding to $\lambda_i(\rho_0)$ [4, Theorem 6.3.12]. Letting

$$u_i = \begin{pmatrix} u_1^{(i)} \\ u_2^{(i)} \end{pmatrix},$$

(2.9) yields

$$(Ju_i, u_i) = \|u_1^{(i)}\|_2^2 - \|u_2^{(i)}\|_2^2 > 0.$$

Hence, all derivatives in (2.19) are positive, which proves Property 10 for the case being considered.

Assume now that an eigenvalue $\lambda(\rho_0)$ is multiple and $u_1, \ldots, u_k$ is a full orthonormal set of eigenvectors of $K(\rho_0)$ corresponding to $\lambda(\rho_0)$. Introducing the matrix U as in (2.11), consider the product

$$U^T J U = U_1^T U_1 - U_2^T U_2 = W \tag{2.20}$$

instead of the numbers (2.19) (for the explanation, see [6]). Property 7 implies that W is positive definite. This ensures that $\lambda(\rho_0)$ gives rise to (at most) k distinct eigenvalue curves, each increasing at ρ_0. This completes the proof of Property 10. ∎

A similar approach (using condition numbers for simple eigenvalues and their block counterparts (2.20) for multiple ones) was employed in [3] in connection with yet another proof of the van Kempen–Thompson theorem on the eigenvalues of (2.1).

The final observation provides a slight generalization of Property 3. Consider an SQD matrix K and a block diagonal matrix

$$D = R \oplus (-S), \qquad R \in \mathbb{R}^{n\times n}, \qquad S \in \mathbb{R}^{m\times m},$$

which is also SQD; i.e., $R \succ 0, S \succ 0$.

PROPERTY 11. *The matrix*

$$\hat{K} = KD$$

is positive stable. That is, its spectrum is contained in the half-plane

$$\pi_+ = \{z \mid \text{Re } z > 0\}.$$

Proof. Indeed,

$$\hat{K} = (KJ)(JD) = \tilde{K}\tilde{D}$$

and the matrix

$$\tilde{D} = JD = R \oplus S$$

is positive definite. Now $\hat{K}$ is similar to the unsymmetric positive definite matrix

$$\tilde{D}^{\frac{1}{2}}\tilde{K}\tilde{D}^{\frac{1}{2}}.$$

Hence, $\hat{K}$ is positive stable. ∎

REFERENCES

[1] P. E. GILL, M. A. SAUNDERS, AND J. R. SHINNERL, *On the stability of Cholesky factorization for symmetric quasidefinite systems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 35–46.

[2] G. GOLUB AND C. V. LOAN, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.

[3] KH. D. IKRAMOV, *The perturbation theory of eigenvalues as a way of proving matrix theorems*, in Numerical Methods of Analysis, N. S. Bakhvalov and V. A. Morozov, eds., Moscow University Press, Moscow, 1997, pp. 68–78.

[4] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[5] R. C. THOMPSON, *The eigenvalues of a partitioned Hermitian matrix involving a parameter*, Linear Algebra Appl., 9 (1974), pp. 243–260.

[6] G. W. STEWART, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764.

[7] H. VAN KEMPEN, *Variation of the eigenvalues of a special class of Hermitian matrices upon variation of some of its elements*, Linear Algebra Appl., 3 (1970), pp. 263–273.

[8] R. J. VANDERBEI, *Symmetric quasidefinite matrices*, SIAM J. Optim., 5 (1995), pp. 100–113.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 4, pp. 1324–1342
© 2000 Society for Industrial and Applied Mathematics

ON THE BEST RANK-1 AND RANK-$(R_1, R_2, \ldots, R_N)$ APPROXIMATION OF HIGHER-ORDER TENSORS*

LIEVEN DE LATHAUWER†, BART DE MOOR†, AND JOOS VANDEWALLE†

Abstract. In this paper we discuss a multilinear generalization of the best rank-R approximation problem for matrices, namely, the approximation of a given higher-order tensor, in an optimal least-squares sense, by a tensor that has prespecified column rank value, row rank value, etc. For matrices, the solution is conceptually obtained by truncation of the singular value decomposition (SVD); however, this approach does not have a straightforward multilinear counterpart. We discuss higher-order generalizations of the power method and the orthogonal iteration method.

Key words. multilinear algebra, singular value decomposition, higher-order tensor, rank reduction

AMS subject classifications. 15A18, 15A69

PII. S0895479898346995

1. Introduction.

Multilinear algebra is the algebra of higher-order tensors, which are the higher-order equivalents of vectors (first order) and matrices (second order), i.e., quantities of which the elements are addressed by more than two indices. Multilinear algebra is gaining more and more interest, largely due to its applications in the context of higher-order statistics (HOS) [17, 18, 2, 4, 5, 7].

Rank-related issues in multilinear algebra are thoroughly different from their matrix counterparts. Let us first introduce some definitions. A rank-1 tensor is a tensor that consists of the outer product of a number of vectors. For an Nth-order tensor $\mathcal{A}$ and N vectors $U^{(1)}$, $U^{(2)}$, $\ldots$, $U^{(N)}$, this means that $a_{i_1 i_2 \cdots i_N} = u^{(1)}_{i_1} u^{(2)}_{i_2} \cdots u^{(N)}_{i_N}$ for all values of the indices, which will be concisely written as $\mathcal{A} = U^{(1)} \circ U^{(2)} \circ \cdots \circ U^{(N)}$. The n-rank of a higher-order tensor is the obvious generalization of the column (row) rank of matrices: given an $(I_1 \times I_2 \times \cdots \times I_N)$-tensor $\mathcal{A}$, it equals the dimension of its n-mode vector space, i.e., the vector space spanned by the I_n-dimensional vectors obtained from $\mathcal{A}$ by varying the index i_n and keeping the other indices fixed. An important difference with the rank of matrices is that the different n-ranks of a higher-order tensor are not necessarily the same. The n-rank will be denoted as $\text{rank}_n(\mathcal{A})$. An Nth-order tensor of which $\text{rank}_1(\mathcal{A}) = R_1$, $\text{rank}_2(\mathcal{A}) = R_2$, etc., will briefly be called a rank-$(R_1, R_2, \ldots, R_N)$ tensor. This is not to be confused with a "rank-R tensor", by which one generally means a tensor that can be decomposed in a sum of R, but not less than R, rank-1 terms; see, e.g., [16].

This paper is a follow-up of [9], in which we discussed a multilinear generalization of the singular value decomposition (SVD). For convenience, we will refer to this decomposition as the higher-order SVD (HOSVD). The starting point of our dis-

*Received by the editors November 6, 1998; accepted for publication (in revised form) by A. Edelman April 4, 1999; published electronically May 4, 2000. This research was partially supported by the Flemish government through the Concerted Research Actions GOA-MIPS and GOA-MEFISTO-666, the Fund for Scientific Research–Flanders (FWO) projects G.0240.99, and the Belgian State, Prime Minister's Office–Federal Office for Scientific, Technical, and Cultural Affairs, through the Interuniversity Poles of Attraction Programmes IUAP P4-02 and IUAP P4-24. The scientific responsibility is assumed by the authors.

http://www.siam.org/journals/simax/21-4/34699.html

†ESAT—SISTA/COSIC, Katholieke Universiteit Leuven, Kardinaal Mercierlaan 94, B-3001 Leuven (Heverlee), Belgium (Lieven.DeLathauwer@esat.kuleuven.ac.be, Bart.DeMoor@esat.kuleuven.ac.be, Joos.Vandewalle@esat.kuleuven.ac.be, http://www.esat.kuleuven.ac.be/sista).

cussion is that, despite the many analogies between the SVD and the HOSVD, the tensor decomposition does not reflect a simple higher-order equivalent of the classical link between the best rank-R approximation of a given matrix and its truncated SVD. Although truncation of the HOSVD of a given tensor may lead to a good rank-$(R_1, R_2, \ldots, R_N)$ approximation ([9] contains an error bound), it turns out that this tensor is in general not the best possible (least-squares) approximation under the given n-mode rank constraints. This paper reports some research results on the estimation of best rank-$(R_1, R_2, \ldots, R_N)$ approximations.

Important research on this topic has already been carried out by Kroonenberg [13, 14], Kroonenberg and de Leeuw [15], and ten Berghe, de Leeuw, and Kroonenberg [19]. They devised an alternating least-squares (ALS) method to improve the fit, which is known as "three-mode factor analysis" in psychometrics. (The generalization to orders higher than three is briefly indicated in [13].) The basic idea is to optimize, mode per mode and in an iterative way, the components of a factorization of the given tensor; each optimization step essentially involves a best reduced-rank approximation of a positive (semi)definite symmetric matrix. We will discuss this work and present the following refinements and complementary results:

(1) In the fundamental best rank-1 approximation problem, an approximation of a given higher-order tensor can gradually be enhanced by means of a relatively simple higher-order generalization of the power method [12] (see sections 3.1 and 3.2).

(2) The efficiency of the higher-order power algorithm can further be increased by updating two modes at the same time (see section 3.3).

(3) Kroonenberg, de Leeuw, and ten Berghe initialized their algorithm with a truncated HOSVD model, but it was indicated that only local optimization was guaranteed. In section 3.4 we will explicitly show that initializing optimization routines with a truncated HOSVD indeed does not always lead to the global optimum; on the other hand, it is our experience that defective cases are rarely met.

(4) For the elementary case of supersymmetric $(2 \times 2 \times \cdots \times 2)$-tensors, the symmetric stationary points of the higher-order power algorithm can be fully characterized (see section 3.5). This case is important, e.g., with respect to applications in blind source separation [11].

(5) With respect to arbitrary values of $R_1, R_2, \ldots, R_N$, we will present a square-root version of the original algorithm, aiming at a higher accuracy, and interpret it as a multilinear generalization of the technique of orthogonal iterations [12] (see section 4).

For clarity, the concepts of this paper are formulated in terms of real-valued tensors. They can be generalized for complex tensors.

Before starting with the next section, we add a comment on the notation that is used. To facilitate the distinction between scalars, vectors, matrices, and higher-order tensors, the type of a given quantity will be reflected by its representation: scalars are denoted by lower-case letters (a, b, ...; α, β, ...), vectors are written as italic capitals (A, B, ...), matrices correspond to boldface capitals ($\mathbf{A}$, $\mathbf{B}$, ...), and tensors are written as calligraphic letters ($\mathcal{A}$, $\mathcal{B}$, ...). This notation is consistently used for lower-order parts of a given structure. For example, the entry with row index i and column index j in a matrix $\mathbf{A}$, i.e., $(\mathbf{A})_{ij}$, is symbolized by a_{ij} (also $(A)_i = a_i$ and $(\mathcal{A})_{i_1 i_2 \cdots i_N} = a_{i_1 i_2 \cdots i_N}$); furthermore, the ith column vector of a matrix $\mathbf{A}$ is denoted as A_i, i.e., $\mathbf{A} = [A_1 A_2 \cdots]$. To enhance the overall readability, we have made one exception to this rule: as we frequently use the characters i, j, r, and n in the meaning of indices (counters), I, J, R, and N will be reserved to denote (unless stated

otherwise) the index upper bounds.

2. Basic definitions. In this section, we introduce some elementary notations and definitions needed in subsequent developments.

2.1. Multiplication of a higher-order tensor by a matrix. Higher-order power and orthogonal iterations involve a multilinear equivalent of matrix-vector and matrix-matrix multiplications.

Let us first have a look at the matrix product $\mathbf{G} = \mathbf{U} \cdot \mathbf{F} \cdot \mathbf{V}^T$, involving matrices $\mathbf{F} \in \mathbb{R}^{I_1 \times I_2}$, $\mathbf{U} \in \mathbb{R}^{J_1 \times I_1}$, $\mathbf{V} \in \mathbb{R}^{J_2 \times I_2}$, and $\mathbf{G} \in \mathbb{R}^{J_1 \times J_2}$. To avoid working with "generalized transposes" in the multilinear case (in which the fact that 1-mode vectors are transpose-free would not have an inherent meaning), we observe that the relationship between $\mathbf{U}$ and $\mathbf{F}$ and the relationship between $\mathbf{V}$ (not $\mathbf{V}^T$) and $\mathbf{F}$ are in fact completely similar: in the same way that $\mathbf{U}$ makes linear combinations of the rows of $\mathbf{F}$, $\mathbf{V}$ makes linear combinations of the columns of $\mathbf{F}$; in the same way that the columns of $\mathbf{F}$ are multiplied by $\mathbf{U}$, the rows of $\mathbf{F}$ are multiplied by $\mathbf{V}$; in the same way that the columns of $\mathbf{U}$ are associated with the column space of $\mathbf{G}$, the columns of $\mathbf{V}$ are associated with the row space of $\mathbf{G}$. This typical relationship will be denoted by means of the $\times_n$ symbol: $\mathbf{G} = \mathbf{F} \times_1 \mathbf{U} \times_2 \mathbf{V}$.

In general, we make the following definition.

DEFINITION 2.1. *The n-mode product of a tensor $\mathcal{A} \in \mathbb{C}^{I_1 \times I_2 \times \cdots \times I_N}$ by a matrix $\mathbf{U} \in \mathbb{C}^{J_n \times I_n}$, denoted by $\mathcal{A} \times_n \mathbf{U}$, is an $(I_1 \times I_2 \times \cdots \times I_{n-1} \times J_n \times I_{n+1} \cdots \times I_N)$-tensor of which the entries are given by*

$$(\mathcal{A} \times_n \mathbf{U})_{i_1 i_2 \cdots i_{n-1} j_n i_{n+1} \cdots i_N} \stackrel{\text{def}}{=} \sum_{i_n} a_{i_1 i_2 \cdots i_{n-1} i_n i_{n+1} \cdots i_N} u_{j_n i_n}.$$

The n-mode product satisfies the following properties.

PROPERTY 1. *Given the tensor $\mathcal{A} \in \mathbb{C}^{I_1 \times I_2 \times \cdots \times I_N}$ and the matrices $\mathbf{F} \in \mathbb{C}^{J_n \times I_n}$, $\mathbf{G} \in \mathbb{C}^{J_m \times I_m}$, one has*

$$(\mathcal{A} \times_n \mathbf{F}) \times_m \mathbf{G} = (\mathcal{A} \times_m \mathbf{G}) \times_n \mathbf{F} = \mathcal{A} \times_n \mathbf{F} \times_m \mathbf{G}.$$

PROPERTY 2. *Given the tensor $\mathcal{A} \in \mathbb{C}^{I_1 \times I_2 \times \cdots \times I_N}$ and the matrices $\mathbf{F} \in \mathbb{C}^{J_n \times I_n}$, $\mathbf{G} \in \mathbb{C}^{K_n \times J_n}$, one has*

$$(\mathcal{A} \times_n \mathbf{F}) \times_n \mathbf{G} = \mathcal{A} \times_n (\mathbf{G} \cdot \mathbf{F}).$$

2.2. Matrix representation of a higher-order tensor. To be able to express our results in a more common matrix language, we define "matrix unfoldings" of a given tensor, i.e., matrix representations of that tensor in which all the column vectors (row vectors, etc.) are ordered sequentially. To avoid confusion, we will retain one particular ordering of the column (row, etc.) vectors; for order three, these unfolding procedures are represented in Figure 2.1. Notice that the definitions of the matrix unfoldings involve the tensor dimensions I_1, I_2, I_3 in a cyclic way and that, when dealing with an unfolding of dimensionality $I_c \times I_a I_b$, we formally assume that the index i_a varies more slowly than i_b. In general, we make the following definition.

DEFINITION 2.2. *Assume an Nth-order tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$. The matrix unfolding $\mathbf{A}_{(n)} \in \mathbb{R}^{I_n \times (I_{n+1} I_{n+2} \cdots I_N I_1 I_2 \cdots I_{n-1})}$ contains the element $a_{i_1 i_2 \cdots i_N}$ at the position with row number i_n and column number equal to $(i_{n+1} - 1) I_{n+2} I_{n+3} \cdots I_N I_1 I_2 \cdots I_{n-1} + (i_{n+2} - 1) I_{n+3} I_{n+4} \cdots I_N I_1 I_2 \cdots I_{n-1} + \cdots + (i_N - 1) I_1 I_2 \cdots I_{n-1} + (i_1 - 1) I_2 I_3 \cdots I_{n-1} + (i_2 - 1) I_3 I_4 \cdots I_{n-1} + \cdots + i_{n-1}$.*

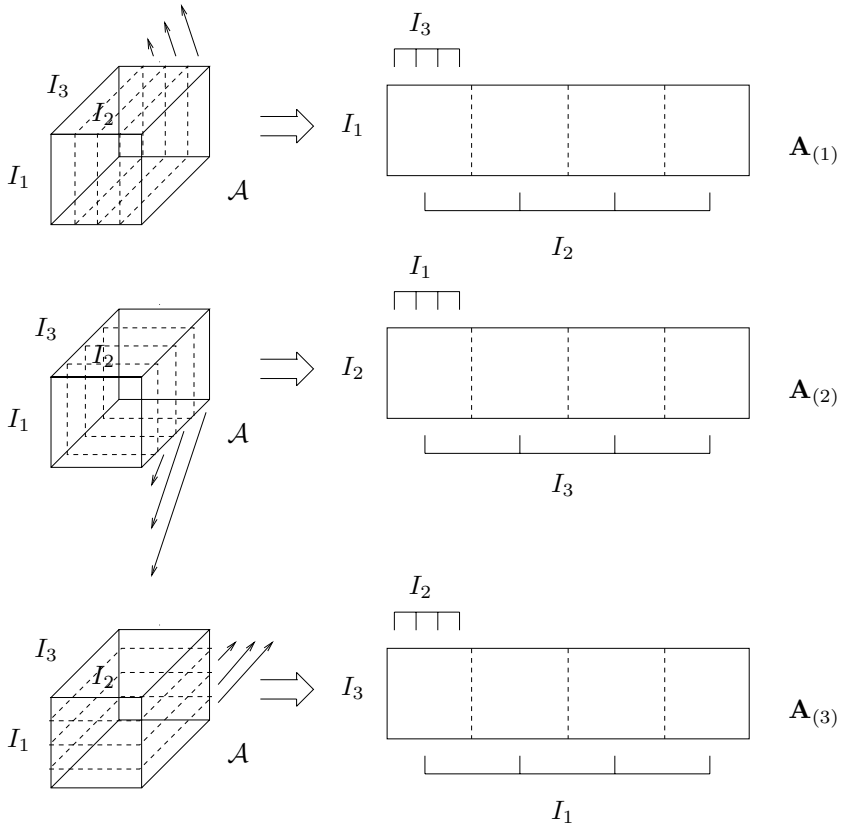


FIG. 2.1. *Unfolding of the* $(I_1 \times I_2 \times I_3)$*-tensor* $\mathcal{A}$ *to the* $(I_1 \times I_2 I_3)$*-matrix* $\mathbf{A}_{(1)}$*, the* $(I_2 \times I_3 I_1)$*-matrix* $\mathbf{A}_{(2)}$*, and the* $(I_3 \times I_1 I_2)$*-matrix* $\mathbf{A}_{(3)}$ $(I_1 = I_2 = I_3 = 4)$.

Example 1. Define a tensor $\mathcal{A} \in \mathbb{R}^{3\times 2\times 3}$ by $a_{111} = a_{112} = a_{211} = -a_{212} = 1$, $a_{213} = a_{311} = a_{313} = a_{121} = a_{122} = a_{221} = -a_{222} = 2$, $a_{223} = a_{321} = a_{323} = 4$, $a_{113} = a_{312} = a_{123} = a_{322} = 0$. The matrix unfolding $\mathbf{A}_{(1)}$ is given by

$$\mathbf{A}_{(1)} = \left(\begin{array}{ccc|ccc} 1 & 1 & 0 & 2 & 2 & 0 \\ 1 & -1 & 2 & 2 & -2 & 4 \\ 2 & 0 & 2 & 4 & 0 & 4 \end{array}\right).$$

2.3. Scalar product and Frobenius norm. The scalar product $\langle \mathcal{A}, \mathcal{B} \rangle$ of two tensors $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is defined in a straightforward way as $\langle \mathcal{A}, \mathcal{B} \rangle \stackrel{\text{def}}{=} \sum_{i_1} \sum_{i_2} \cdots \sum_{i_N} a_{i_1 i_2 \cdots i_N} b_{i_1 i_2 \cdots i_N}$. The Frobenius norm of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is then defined as $\|\mathcal{A}\| \stackrel{\text{def}}{=} \sqrt{\langle \mathcal{A}, \mathcal{A} \rangle}$.

3. Higher-order power iteration. In this section, we investigate how a given tensor can be approximated, in an optimal least-squares sense, by a tensor of rank 1. In section 3.1, the problem is formalized in two different ways and analyzed with the technique of Lagrange multipliers. In section 3.2, we show that an approximation can be improved by means of a higher-order generalization of the power method. A more efficient variant is presented in section 3.3. Section 3.4 deals with the choice of a good

initial value. For supersymmetric $(2\times 2\times\cdots\times 2)$-tensors, the determination of the best rank-1 approximation and the other symmetric stationary points of the higher-order power algorithm is reformulated as a polynomial rooting problem in section 3.5.

Section 3.2 can be considered as a special but important case of the ALS technique for the computation of the decomposition of a given tensor in a minimal sum of rank-1 terms [1, 3]. The idea is to update the components of the approximation mode per mode, which leads to a sequence of linear least-squares problems. However, in the rank-1 case these least-squares problems have a trivial solution, which allows for a faster approach, as explained in section 3.3. More background material on the decomposition of a tensor in a sum of rank-1 terms can be found in [5].

3.1. Best rank-1 approximation. The problem we want to solve can be described mathematically as follows.

Given a real Nth-order tensor $\mathcal{A}\in\mathbb{R}^{I_1\times I_2\times\cdots I_N}$, find a scalar λ and unit-norm vectors $U^{(1)}$, $U^{(2)}$, ..., $U^{(N)}$ such that the rank-1 tensor $\hat{\mathcal{A}}\stackrel{\text{def}}{=}\lambda U^{(1)}\circ U^{(2)}\circ\cdots\circ U^{(N)}$ minimizes the least-squares cost function

$$f(\hat{\mathcal{A}})=\|\mathcal{A}-\hat{\mathcal{A}}\|^2 \tag{3.1}$$

over the manifold of rank-1 tensors.

This constrained optimization problem can be analyzed using the technique of Lagrange multipliers. Therefore, we consider the following combination of f with the constraint terms:

$$\tilde{f}\stackrel{\text{def}}{=}\sum_{i_1 i_2\cdots i_N}(a_{i_1 i_2\cdots i_N}-\lambda u^{(1)}_{i_1}u^{(2)}_{i_2}\cdots u^{(N)}_{i_N})^2+\sum_n\lambda^{(n)}\left(\sum_{i_n}(u^{(n)}_{i_n})^2-1\right), \tag{3.2}$$

in which $\lambda^{(n)}$ $(1\leqslant n\leqslant N)$ are the Lagrange multipliers. Setting the derivative with respect to $u^{(n)}_{i_n}$ equal to zero yields

$$\lambda\sum_{i_1\cdots i_{n-1}i_{n+1}\cdots i_N}a_{i_1 i_2\cdots i_N}u^{(1)}_{i_1}\cdots u^{(n-1)}_{i_{n_1}}u^{(n+1)}_{i_{n+1}}\cdots u^{(N)}_{i_N}$$

$$=\lambda^{(n)}u^{(n)}_{i_n}+\lambda^2 u^{(n)}_{i_n}\sum_{i_1\cdots i_{n-1}i_{n+1}\cdots i_N}(u^{(1)}_{i_1})^2\cdots(u^{(n-1)}_{i_{n_1}})^2(u^{(n+1)}_{i_{n+1}})^2\cdots(u^{(N)}_{i_N})^2. \tag{3.3}$$

Derivation with respect to $\lambda^{(n)}$ and λ, respectively, yields

$$\sum_{i_n}(u^{(n)}_{i_n})^2=1, \tag{3.4}$$

$$\sum_{i_1 i_2\cdots i_N}a_{i_1 i_2\cdots i_N}u^{(1)}_{i_1}u^{(2)}_{i_2}\cdots u^{(N)}_{i_N}=\lambda\sum_{i_1 i_2\cdots i_N}(u^{(1)}_{i_1})^2(u^{(2)}_{i_2})^2\cdots(u^{(N)}_{i_N})^2. \tag{3.5}$$

Combining (3.4) with (3.5) and with the right-hand side of (3.3) yields

$$\sum_{i_1 i_2\cdots i_N}a_{i_1 i_2\cdots i_N}u^{(1)}_{i_1}u^{(2)}_{i_2}\cdots u^{(N)}_{i_N}=\lambda, \tag{3.6}$$

$$\lambda\sum_{i_1\cdots i_{n-1}i_{n+1}\cdots i_N}a_{i_1 i_2\cdots i_N}u^{(1)}_{i_1}\cdots u^{(n-1)}_{i_{n_1}}u^{(n+1)}_{i_{n+1}}\cdots u^{(N)}_{i_N}=(\lambda^2+\lambda^{(n)})u^{(n)}_{i_n}. \tag{3.7}$$

Combining (3.4) with the right-hand side of (3.7), and comparing this to (3.6), yields

$$\sum_{i_1 \cdots i_{n-1} i_{n+1} \cdots i_N} a_{i_1 i_2 \cdots i_N} u^{(1)}_{i_1} \cdots u^{(n-1)}_{i_{n_1}} u^{(n+1)}_{i_{n+1}} \cdots u^{(N)}_{i_N} = \lambda\, u^{(N)}_{i_N}. \tag{3.8}$$

Summarizing, the Lagrange equations correspond to $(1 \leqslant n \leqslant N)$:

$$\mathcal{A} \times_1 U^{(1)^T} \cdots \times_{n-1} U^{(n-1)^T} \times_{n+1} U^{(n+1)^T} \cdots \times_N U^{(N)^T} = \lambda\, U^{(n)}, \tag{3.9}$$

$$\mathcal{A} \times_1 U^{(1)^T} \times_2 U^{(2)^T} \cdots \times_N U^{(N)^T} = \lambda, \tag{3.10}$$

$$\|U^{(n)}\| = 1. \tag{3.11}$$

The best rank-1 approximation problem can also be formulated as follows.

THEOREM 3.1. *Assume a real Nth-order tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots I_N}$, then the minimization of the cost function of* (3.1) *is equivalent to the maximization, over the unit-norm vectors $U^{(1)}, U^{(2)}, \ldots, U^{(N)}$, of the function*

$$g(U^{(1)}, U^{(2)}, \ldots, U^{(N)}) = \left| \mathcal{A} \times_1 U^{(1)^T} \times_2 U^{(2)^T} \cdots \times_N U^{(N)^T} \right|^2. \tag{3.12}$$

If the scalar λ, corresponding to the Frobenius norm of $\hat{\mathcal{A}}$ in (3.1)*, is chosen in accordance with* (3.10)*, then the functions of* (3.1) *and* (3.12) *are related by*

$$f = \|\mathcal{A}\|^2 - g. \tag{3.13}$$

Proof. We have the following:

$$f(\hat{\mathcal{A}}) = \|\mathcal{A} - \hat{\mathcal{A}}\|^2 = \|\mathcal{A}\|^2 - 2\langle \mathcal{A}, \hat{\mathcal{A}} \rangle + \|\hat{\mathcal{A}}\|^2.$$

According to the definition of λ, the value taken by $\langle \mathcal{A}, \hat{\mathcal{A}} \rangle$ equals λ^2. Since $U^{(1)}$, $U^{(2)}$, $\ldots$, $U^{(N)}$ have unit-norm, $\|\hat{\mathcal{A}}\|^2 = \lambda^2$ as well. Combination with the definition of g proves the theorem. ∎

Remark 1. We defined the best rank-1 approximation problem as the minimization of the distance between a given tensor and its approximation on the rank-1 manifold. Theorem 3.1 shows that this is equivalent to the maximization of the norm of the projection of the original tensor onto the rank-1 manifold.

Remark 2. Theorem 3.1 is the higher-order generalization of the fact that the computation of the best rank-1 approximation of a matrix $\mathbf{A}$ is equivalent to the determination of unit-vectors U and V such that $|U^T \mathbf{A} V|^2$ is maximal.

3.2. A power algorithm. For the actual computation of the best rank-1 approximation of $\mathcal{A}$, the Lagrange equations and their derivation can be interpreted in a constructive way. Imagine that the vectors $U^{(1)}, \ldots, U^{(n-1)}, U^{(n+1)}, \ldots, U^{(N)}$ are fixed and that f in (3.1) is merely a quadratic function in the unknown unconstrained vector $\lambda\, U^{(n)}$. Equation (3.9) then simply shows how the optimal $\lambda\, U$ can be computed for fixed $U^{(1)}, \ldots, U^{(n-1)}, U^{(n+1)}, \ldots, U^{(N)}$.

The full set of Lagrange equations $(1 \leqslant n \leqslant N)$ can be linked to an ALS algorithm for the (local) minimization of $f(\hat{\mathcal{A}})$: in each step the estimate of the scalar λ and the estimate of one of the vectors $U^{(1)}, U^{(2)}, \ldots, U^{(N)}$ are optimized, while the other vector estimates are kept constant. The result is shown in step 2 of Algorithm 3.2. We remark that the expression

$$\tilde{U}^{(n)}_{k+1} = \mathcal{A} \times_1 U^{(1)^T}_{k+1} \cdots \times_{n-1} U^{(n-1)^T}_{k+1} \times_{n+1} U^{(n+1)^T}_{k} \cdots \times_N U^{(N)^T}_{k}$$

ALGORITHM 3.2.
HIGHER-ORDER POWER METHOD.
In: $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$.
Out: $\hat{\mathcal{A}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$: estimator of best rank-1 approximation of $\mathcal{A}$.

1. Initial values: $U_0^{(n)}$ is the dominant left singular vector of $\mathbf{A}_{(n)}$ $(2 \leqslant n \leqslant N)$ and/or repeat the algorithm for several initial values.
2. Iterate until convergence:
   - $\tilde{U}_{k+1}^{(1)} = \mathcal{A} \times_2 {U_k^{(2)}}^T \times_3 {U_k^{(3)}}^T \cdots \times_N {U_k^{(N)}}^T$;
     $\lambda_{k+1}^{(1)} = \|\tilde{U}_{k+1}^{(1)}\|$;
     $U_{k+1}^{(1)} = \tilde{U}_{k+1}^{(1)} / \lambda_{k+1}^{(1)}$;
   - $\tilde{U}_{k+1}^{(2)} = \mathcal{A} \times_1 {U_{k+1}^{(1)}}^T \times_3 {U_k^{(3)}}^T \cdots \times_N {U_k^{(N)}}^T$;
     $\lambda_{k+1}^{(2)} = \|\tilde{U}_{k+1}^{(2)}\|$;
     $U_{k+1}^{(2)} = \tilde{U}_{k+1}^{(2)} / \lambda_{k+1}^{(2)}$;
   - $\ldots$
   - $\tilde{U}_{k+1}^{(N)} = \mathcal{A} \times_1 {U_{k+1}^{(1)}}^T \times_2 {U_{k+1}^{(2)}}^T \cdots \times_{N-1} {U_{k+1}^{(N-1)}}^T$;
     $\lambda_{k+1}^{(N)} = \|\tilde{U}_{k+1}^{(N)}\|$;
     $U_{k+1}^{(N)} = \tilde{U}_{k+1}^{(N)} / \lambda_{k+1}^{(N)}$.

   Converged values: $U^{(1)}, U^{(2)}, \ldots, U^{(N)}, \lambda$.
3. $\hat{\mathcal{A}} = \lambda\, U^{(1)} \circ U^{(2)} \circ \cdots \circ U^{(N)}$.

can be written in matrix format as follows:

$$\tilde{U}_{k+1}^{(n)} = \mathbf{A}_{(n)} \cdot (U_{k+1}^{(1)} \otimes \cdots \otimes U_{k+1}^{(n-1)} \otimes U_k^{(n+1)} \otimes \cdots \otimes U_k^{(N)}),$$

in which $\otimes$ represents the Kronecker product.

Clearly this technique is a higher-order extension of the power method for matrices [12]. The termination criterion could be formulated in terms of the accuracy of components (e.g., $|{U_{k+1}^{(n)}}^T U_k^{(n)}| < \epsilon$ for $1 \leqslant n \leqslant N$) or in terms of the quality of the fit (e.g., $\lambda_{k+1}^{(N)} - \lambda_k^{(N)} < \epsilon$). With respect to the initialization, we refer to section 3.4.

Remark 3. For the best approximation of a supersymmetric tensor $\mathcal{A}$ by a supersymmetric rank-1 tensor $\hat{\mathcal{A}} = \lambda\, U \circ U \circ \cdots \circ U$, the derivation is analogous. (Supersymmetric higher-order tensors are tensors that are invariant under arbitrary permutations of their indices; examples are the basic quantities of HOS.) The equivalent of (3.9) is

$$\mathcal{A} \times_1 U^T \cdots \times_{n-1} U^T \times_{n+1} U^T \cdots \times_N U^T = \lambda U. \tag{3.14}$$

This equation can as well be interpreted as an update of the unknown n-mode vector $\lambda U^{(n)}$, while the other vectors $U^{(1)}, \ldots, U^{(n-1)}, U^{(n+1)}, \ldots, U^{(N)}$ are fixed to U. However, such an update breaks the symmetry of the estimate unless λ and U already corresponded to a supersymmetric solution of the Lagrange equation. A symmetric version of the algorithm, based on the iteration

$$\tilde{U}_{k+1} = \mathcal{A} \times_1 U_k^T \times_2 U_k^T \cdots \times_{N-1} U_k^T, \tag{3.15}$$

for which $U_k^{(1)} = U_k^{(2)} = \cdots = U_k^{(N)} = U_k$, is unreliable since it does not necessarily decrease the cost function $f(\hat{\mathcal{A}})$ in a monotonous way, as can easily be verified from

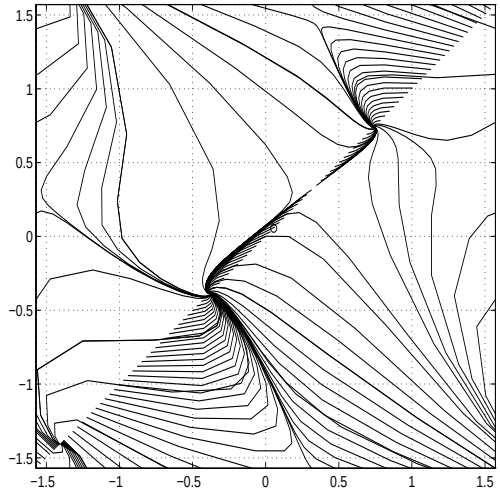


FIG. 3.1. *Depiction of Algorithm* 1 *for a typical example of a supersymmetric tensor in* $\mathbb{R}^{2\times 2\times 2}$. *Abscis*: *the angle* $\theta_k^{(2)}$ *in* $U_k^{(2)} = (\cos\theta_k^{(2)} \ \sin\theta_k^{(2)})^T$ (*in radians*). *Ordinate*: *the angle* $\theta_k^{(3)}$ *in* $U_k^{(3)} = (\cos\theta_k^{(3)} \ \sin\theta_k^{(3)})^T$ (*in radians*). *Both angles are normalized to the interval* $(-\pi/2, +\pi/2]$. *The small circle shows the initial guess obtained by HOSVD. The global optimum is* $(-0.3860, -0.3860)$.

examples. The fact that a power iteration produces unsymmetric intermediate results is not observed for symmetric matrices: at the second-order level, each iteration step involves only the knowledge of one intermediate vector, such that it is simply impossible to compare the different n-mode vector estimates at a given iteration step.

Example 2. Figure 3.1 depicts step 2 of Algorithm 3.2 for a $(2 \times 2 \times 2)$ example. For different choices of θ_0, determining initial vectors $U_0^{(2)} = U_0^{(3)} = (\cos\theta_0 \ \sin\theta_0)^T$, we plotted after each iteration the angle $\theta_k^{(3)}$ in $U_k^{(3)} = (\cos\theta_k^{(3)} \ \sin\theta_k^{(3)})^T$ versus the angle $\theta_k^{(2)}$ in $U_k^{(2)} = (\cos\theta_k^{(2)} \ \sin\theta_k^{(2)})^T$. The angles were normalized to the interval $(-\pi/2, +\pi/2]$. (The sign of the vectors does not matter, as can be seen from the definition of the function g; as far as f is concerned, the sign can be incorporated in the scalar λ.) The tensor $\mathcal{A}$ that we consider is supersymmetric and defined by

$$\begin{cases} a_{111} = 1.5578, & a_{222} = 1.1226, \\ a_{112} = -2.4443, & a_{221} = -1.0982. \end{cases}$$

We observe that Algorithm 3.2 leads to unsymmetric intermediate results not located on the main diagonal of the figure.

We also remark that there are two stable $((-0.3860, -0.3860), (0.7413, 0.7413))$ and two unstable $((-1.4052, -1.4052), (0.3347, 0.3347))$ symmetric stationary points of the higher-order power algorithm. The first three correspond to the three solutions of the Lagrange equation

$$\mathcal{A} \times_2 U^T \times_3 U^T = \lambda U \tag{3.16}$$

on the unit circle $\|U\| = 1$. The solution (0.3347, 0.3347) is special: even after convergence the three substeps of each iteration produce unsymmetric results, but these effects compensate such that the overall power iteration is symmetric. The global optimum is $(-0.3860, -0.3860)$.

3.3. Joint updates in the higher-order power algorithm. In the iteration scheme developed in section 3.2, the estimates of the vectors $U^{(1)}, U^{(2)}, \ldots, U^{(N)}$ are updated one at a time. However, it is possible to update the estimates of two of the vectors in a single iteration step.

Let us assume, for example, that the values of $U^{(1)}, \ldots, U^{(n-1)}, U_k^{(n+2)}, \ldots, U^{(N)}$ are fixed and that $U^{(n)}$ and $U^{(n+1)}$ have to be updated. Theorem 3.1 then implies that the optimal $U^{(n)}$ and $U^{(n+1)}$ can be found as the dominant left and right singular vectors of the matrix $\mathcal{A} \times_1 {U^{(1)}}^T \cdots \times_{n-1} {U^{(n-1)}}^T \times_{n+2} {U^{(n+2)}}^T \cdots \times_N {U^{(N)}}^T$, since $g = |{U^{(n)}}^T \cdot (\mathcal{A} \times_1 {U^{(1)}}^T \cdots \times_{n-1} {U^{(n-1)}}^T \times_{n+2} {U^{(n+2)}}^T \cdots \times_N {U^{(N)}}^T) \cdot U^{(n+1)}|^2$; the corresponding value of g is the squared dominant singular value.

The efficiency of Algorithm 3.2 can be increased by resorting to modern techniques for the estimation of the dominant singular triplet [12]. (Computation of the dominant singular triplet by a matrix power algorithm would correspond to an iteration between substeps n and $n+1$ of step 2 in Algorithm 3.2.) Note that, in the terminology of Algorithm 3.2, $U_k^{(n)}$ and $U_k^{(n+1)}$ can be used as first approximations of $U_{k+1}^{(n)}$ and $U_{k+1}^{(n+1)}$.

3.4. Starting value. Example 2 illustrates a major difference between the second-order and the higher-order problem of best rank-1 approximation: for tensors, the Lagrange equations can have several stable solutions. A descent algorithm, like the higher-order power method, will reveal only the global optimum if the starting point is in the attraction region of this global optimum.

Considering the fact that the best rank-1 approximation of a matrix is obtained by truncating its SVD after the first singular value, it is clear that estimation of $U^{(n)}$ as the dominant left singular vector of the matrix unfolding $\mathbf{A}_{(n)}$ $(1 \leqslant n \leqslant N)$ may yield a good rank-1 approximation. As a matter of fact, this corresponds to truncation of the HOSVD after the first term [9]. Another important difference between matrices and higher-order tensors is that this approximation is not necessarily the globally optimal one. (For example, in Figure 3.1 the estimate obtained by truncation of the HOSVD is indicated by means of a small circle.)

However, in our simulations we have observed that the HOSVD estimate usually belongs to the attraction region of the best rank-1 approximation. Hence we propose to use the HOSVD estimate as the starting value for the power iteration derived above; we assume that this iteration will not cause "jumps" to other attraction regions. However, there is no absolute guarantee: we have been able to generate special cases in which monotonous descent techniques eventually lead to a local optimum with a close-to-optimal fit, but nevertheless different from the global optimum. In this respect, repeating the optimization procedure for several initial values could be envisaged.

Example 3. Consider a tensor $\mathcal{B} \in \mathbb{R}^{2\times2\times2\times2}$, of which the nonzero entries are given by

$$\begin{cases} b_{1111} = 25.1, & b_{1212} = 25.6, \\ b_{2121} = 24.8, & b_{2222} = 23. \end{cases}$$

Due to the symmetries $b_{ijkl} = b_{kjil}$ and $b_{ijkl} = b_{ilkj}$, the best rank-1 approximation of $\mathcal{B}$ has the form $\lambda U^{(1)} \circ U^{(2)} \circ U^{(1)} \circ U^{(2)}$; we write $U^{(1)} = (\cos\alpha \ \ \sin\alpha)^T$ and $U^{(2)} = (\cos\beta \ \ \sin\beta)^T$. The function $\tilde{g}_1 \stackrel{\text{def}}{=} \mathcal{B} \times_1 {U^{(1)}}^T \times_2 {U^{(2)}}^T \times_3 {U^{(1)}}^T \times_4 {U^{(2)}}^T$ corresponds to a weighted arithmetic mean of the nonzero entries of $\mathcal{B}$:

$$\tilde{g}_1(\alpha, \beta) = 25.1 \cos^2 \alpha \cos^2 \beta + 25.6 \cos^2 \alpha \sin^2 \beta + 24.8 \sin^2 \alpha \cos^2 \beta + 23 \sin^2 \alpha \sin^2 \beta.$$

The global maximum of this function is 25.6, reached for $\alpha = 0$ and $\beta = \pi/2$; there are no essentially different local maxima. According to Theorem 3.1, setting $\lambda = 25.6$, $U^{(1)} = (1\ 0)^T$, and $U^{(2)} = (0\ 1)^T$ gives the best rank-1 approximation of $\mathcal{B}$. On the other hand, $\mathcal{B}$ is such that $\mathbf{B}_{(1)}$ and $\mathbf{B}_{(2)}$ each consist of mutually orthogonal rows, arranged in order of decreasing Frobenius norm; hence truncation of the HOSVD yields the vectors $U^{(1)} = (1\ 0)^T$ and $U^{(2)} = (1\ 0)^T$, which belong to the correct attraction region in a trivial way.

Now consider a tensor $\mathcal{A} \in \mathbb{R}^{2\times 2\times 2\times 2}$, equal to $\mathcal{B}$, except for the entries

$$a_{1121} = 0.3, \quad a_{2111} = 0.3.$$

$\mathcal{A}$ has similar symmetries as $\mathcal{B}$. The function $\tilde{g}_2 \stackrel{\text{def}}{=} \mathcal{A} \times_1 {U^{(1)}}^T \times_2 {U^{(2)}}^T \times_3 {U^{(1)}}^T \times_4 {U^{(2)}}^T$, given by

$$\tilde{g}_2(\alpha, \beta) = \tilde{g}_1(\alpha, \beta) + 0.6 \cos^2 \beta \cos \alpha \sin \alpha, \tag{3.17}$$

is shown in Figure 3.2 by means of a mesh and a contour plot. The extra term in (3.17) is small enough such that $\tilde{g}_1$ and $\tilde{g}_2$ are similar—in particular, the global maximum is the same—but it induces a local maximum (namely, $(0.5536, 0)$) on the axis $\beta = 0$. On the contour plot, the zero-gradient points of $\tilde{g}_2$ are indicated by means of "+" marks; the global maximum is indicated by means of a small circle. $\mathbf{A}_{(1)}$ and $\mathbf{A}_{(2)}$ still have nearly orthogonal rows, arranged in order of decreasing Frobenius norm. Truncation of the HOSVD leads to vectors that are still close to $(1\ 0)^T$; this corresponds to the "×" mark on the contour plot. Because of the symmetry of $\tilde{g}_2$ around the axis $\beta = 0$, a gradient ascent starting from the truncated HOSVD will converge not to the global maximum but to the local optimum $(0.5536, 0)$. The same holds for the higher-order power algorithm: after each iteration the intermediate result was plotted as a dot in the contour plot of Figure 3.2. Altering the ordering of the substeps in Algorithm 3.2 yields comparable results. (Starting from the HOSVD guess, the global maximum could be found, however, by alternating, as indicated in section 3.3, between simultaneous updates of the mode-1 and mode-3 vector on one hand, and on the other hand the mode-2 and mode-4 vector, beginning with the former.) It is clear that this example is not an isolated case: the conditions that ensure that the HOSVD truncate is in the wrong attraction region are still satisfied for sufficiently small perturbations of $\mathcal{A}$.

3.5. Best rank-1 approximation of supersymmetric binary tensors. The best rank-1 approximation of supersymmetric $(2 \times 2 \times \cdots \times 2)$-tensors deserves special attention, as it appears that the determination of the symmetric stationary points of the higher-order power algorithm can be reformulated as a polynomial rooting problem in this case. Despite its elementary character, the problem is very relevant: e.g., in [11] we proved that it is intimately connected with the blind separation of a linear mixture of two independent sources; separation of mixtures of a larger size can be achieved by a Jacobi iteration over such elementary cases [4].

We first prove that there can be at most three distinct symmetric solutions to the Lagrange equations for a supersymmetric $(2 \times 2 \times 2)$-tensor; the generalization to arbitrary tensor orders is straightforward. For convenience we use the notation

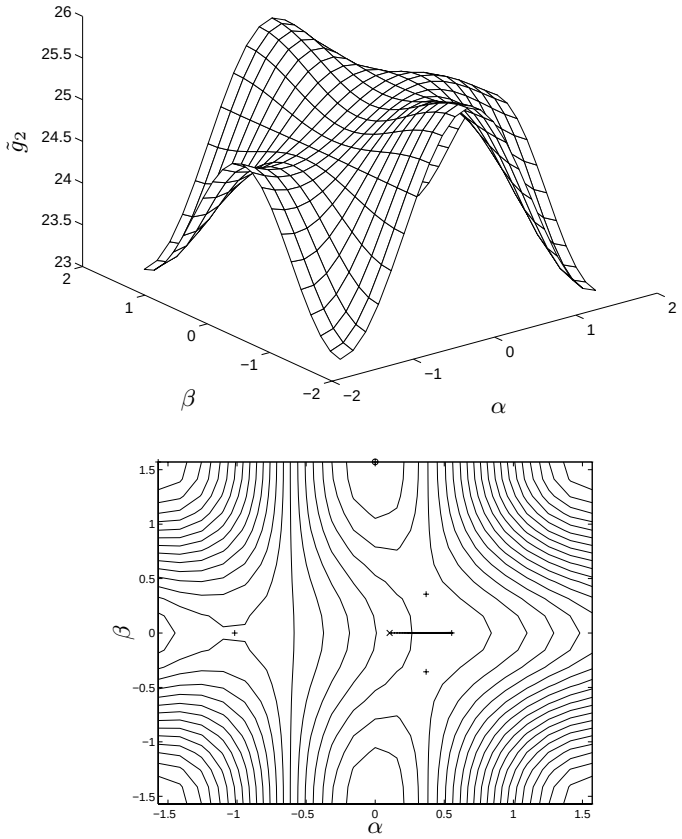


FIG. 3.2. *The quality of the best rank-*1 *approximation of the partially symmetric tensor* $\mathcal{A}$ *in Example* 3*, for different values of* α *and* β*. The "*$\times$*" mark shows the initial guess obtained by HOSVD. The global optimum is indicated by means of a small circle. The "+" marks correspond to the zero-gradient points of the function* $\tilde{g}_2$*, defined in the example. The intermediate results after each iteration cycle of Algorithm* 3.2*, initialized with the first HOSVD components, are plotted as a sequence of dots.*

$U = (c\ s)^T$. Equation (3.16) corresponds to the following set of equations:

$$a_{111}c^2 + 2a_{112}cs + a_{122}s^2 = \lambda c,$$
$$a_{211}c^2 + 2a_{212}cs + a_{222}s^2 = \lambda s.$$

Eliminating λ and claiming that $\|U\|^2 = 1$ yields

$$a_{211}c^3 + (2a_{212} - a_{111})c^2 s + (a_{222} - 2a_{112})cs^2 - a_{122}s^3 = 0,$$
$$c^2 + s^2 = 1.$$

On the unit circle a solution is entirely defined by $t = s/c$, each time corresponding to the vectors $(c\ s)^T$ and $(-c\ -s)^T$, which are not essentially different. The solutions can be found by rooting the polynomial

$$a_{211} + (2a_{212} - a_{111})t + (a_{222} - 2a_{112})t^2 - a_{122}t^3 = 0,$$

which has at most three distinct real roots. The same expression shows where the derivative of $\tilde{g} = \mathcal{A} \times_1 U^T \times_2 U^T \times_3 U^T$ is equal to zero. Similarly, for an Nth-order supersymmetric $(2 \times 2 \times \cdots \times 2)$-tensor, there can be at most N distinct symmetric solutions to the Lagrange equations; these solutions can be found as the roots of an Nth-order polynomial.

On the other hand, it is also possible that a higher-order power iteration yields a stationary solution that is characterized by a symmetric state $U^{(2)} = U^{(3)} = \cdots = U$, but nevertheless produces an intermediate vector $U^{(1)}$ that is different from U; e.g., see the solution (0.3347, 0.3347) in Example 2. Therefore, let us consider solutions to the Lagrange equations of the type

$$\mathcal{A} \times_2 U^T \times_3 U^T = \lambda U^{(1)}, \tag{3.18}$$

$$\mathcal{A} \times_2 {U^{(1)}}^T \times_3 U^T = \lambda U \tag{3.19}$$

for a supersymmetric $(2 \times 2 \times 2)$-tensor $\mathcal{A}$, where $U^{(1)}$ is not necessarily equal to U. Equations (3.18) and (3.19) can be combined as

$$\mathcal{B} \times_2 U^T \times_3 U^T \times_4 U^T = \lambda U, \tag{3.20}$$

in which the tensor $\mathcal{B}$ is defined by $b_{ijkl} = \sum_p a_{ijp} a_{klp}$ for all entries. Although $\mathcal{B}$ itself is not supersymmetric, the solutions to (3.20) can be determined as above; at most four distinct real solutions are possible. Similarly, for an Nth order supersymmetric $(2 \times 2 \times \cdots \times 2)$-tensor, the solutions to the Lagrange equations

$$\mathcal{A} \times_2 U^T \times_3 U^T \cdots \times_N U^T = \lambda U^{(1)}, \tag{3.21}$$

$$\mathcal{A} \times_2 {U^{(1)}}^T \times_3 U^T \cdots \times_N U^T = \lambda U \tag{3.22}$$

can be found by rooting a polynomial of degree $2N - 2$.

The determination of arbitrary unsymmetric stationary points of a higher-order power iteration is much harder than the analysis of the two specific cases above; it generally leads to sets of polynomial equations and will therefore not be considered.

4. Higher-order orthogonal iteration. In this section we generalize the best rank-1 approximation problem of the previous section in the sense that the approximation should now have prespecified mode-1 rank, mode-2 rank, etc. The derivation of the computational procedure follows a similar scheme. In section 4.1, we give two related formal definitions of the approximation problem. Section 4.2 is devoted to the actual computation of the solution.

The derivation follows roughly the same lines as in the work by Kroonenberg [13, 14], Kroonenberg and de Leeuw [15], and ten Berghe, de Leeuw, and Kroonenberg [19], but with some modifications in the practical implementation. On the other hand, the overall presentation is intended to clarify the general linear/multilinear framework underlying the best rank-R/rank-$(R_1, R_2, \ldots, R_N)$ approximation problem. The algorithm that we present is a square-root version of the Kroonenberg algorithm, which increases the accuracy, as is well known [12]. Next, the computation scheme can be based on the calculation of dominant subspaces, rather than individual singular vectors, resulting in a higher efficiency. The algorithm will be interpreted as a multilinear generalization of the technique of orthogonal iterations [12]. An important remark concerns the initialization of the algorithm. Kroonenberg, de Leeuw, and ten Berghe initialized their algorithm with a truncated HOSVD, but it was indicated that only

local optimization was guaranteed. In section 3.4 we have demonstrated that initializing optimization routines with a truncated HOSVD does not always lead to the global optimum, though it is our experience that defective cases are rarely met.

4.1. Best rank-$(R_1, R_2, \ldots, R_N)$ approximation. In this section we generalize the best rank-1 approximation problem of the previous section to the best approximation by a tensor with prespecified mode-1 rank, mode-2 rank, etc. Formally, the problem we want to solve can be formulated as follows.

Given a real Nth-order tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots I_N}$, find a tensor $\hat{\mathcal{A}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots I_N}$, having $\text{rank}_1(\hat{\mathcal{A}}) = R_1$, $\text{rank}_2(\hat{\mathcal{A}}) = R_2, \ldots, \text{rank}_N(\hat{\mathcal{A}}) = R_N$, that minimizes the least-squares cost function

$$f(\hat{\mathcal{A}}) = \|\mathcal{A} - \hat{\mathcal{A}}\|^2. \tag{4.1}$$

The n-rank conditions imply that $\hat{\mathcal{A}}$ can be decomposed as

$$\hat{\mathcal{A}} = \mathcal{B} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \cdots \times_N \mathbf{U}^{(N)}, \tag{4.2}$$

in which $\mathbf{U}^{(1)} \in \mathbb{R}^{I_1 \times R_1}$, $\mathbf{U}^{(2)} \in \mathbb{R}^{I_2 \times R_2}$, $\ldots$, $\mathbf{U}^{(N)} \in \mathbb{R}^{I_N \times R_N}$ each have orthonormal columns and $\mathcal{B} \in \mathbb{R}^{R_1 \times R_2 \times \cdots \times R_N}$.

Actually it is sufficient to determine the matrices $\mathbf{U}^{(1)}$, $\mathbf{U}^{(2)}, \ldots, \mathbf{U}^{(N)}$ for the optimization of f. For any estimate of these matrices, the optimal tensor $\mathcal{B}$ is given by the following theorem.

THEOREM 4.1. *Assume a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots I_N}$ and consider a tensor $\hat{\mathcal{A}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots I_N}$ as in* (4.2)*. For given matrices $\mathbf{U}^{(1)}$, $\mathbf{U}^{(2)}, \ldots, \mathbf{U}^{(N)}$, the tensor $\mathcal{B}$ that optimizes $f(\hat{\mathcal{A}}) = \|\mathcal{A} - \hat{\mathcal{A}}\|^2$ is given by*

$$\mathcal{B} = \mathcal{A} \times_1 \mathbf{U}^{(1)^T} \times_2 \mathbf{U}^{(2)^T} \ldots \times_N \mathbf{U}^{(N)^T}. \tag{4.3}$$

Proof. The optimization of

$$f = \|\mathcal{A} - \mathcal{B} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \cdots \times_N \mathbf{U}^{(N)}\|^2$$

for the unknown elements of $\mathcal{B}$ is merely a classical linear least-squares problem. It corresponds to the least-squares estimation of the solution of the following set of linear equations, which is possibly overdetermined:

$$\mathcal{B} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \cdots \times_N \mathbf{U}^{(N)} = \mathcal{A}.$$

Taking into account that $\mathbf{U}^{(1)}$, $\mathbf{U}^{(2)}, \ldots, \mathbf{U}^{(N)}$ have orthonormal columns, these matrices can be brought to the other side of the equation by (n-mode) multiplication with the transposed matrices. ∎

Remark 4. Theorem 4.1 is the multidimensional equivalent of the optimal choice of λ in (3.10).

Analogous with Theorem 3.1 we state that the best rank-$(R_1, R_2, \ldots, R_N)$ approximation problem can also be formulated as follows.

THEOREM 4.2. *Assume a real Nth-order tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots I_N}$; then the minimization of the cost function of* (4.1) *is equivalent to the maximization, over the matrices $\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \ldots, \mathbf{U}^{(N)}$ having orthonormal columns, of the function*

$$g(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \ldots, \mathbf{U}^{(N)}) = \left\| \mathcal{A} \times_1 \mathbf{U}^{(1)^T} \times_2 \mathbf{U}^{(2)^T} \cdots \times_N \mathbf{U}^{(N)^T} \right\|^2. \tag{4.4}$$

If the tensor $\mathcal{B}$, the Frobenius norm of which equals the Frobenius norm of $\hat{\mathcal{A}}$ in (4.2)*, is chosen in accordance with* (4.3)*, then the functions of* (4.1) *and* (4.4) *are related by*

$$f = \|\mathcal{A}\|^2 - g. \tag{4.5}$$

Proof. We look for the relations between the definition of g and the terms in the following expression for f:

$$f(\hat{\mathcal{A}}) = \|\mathcal{A} - \hat{\mathcal{A}}\|^2 = \|\mathcal{A}\|^2 - 2\langle \mathcal{A}, \hat{\mathcal{A}}\rangle + \|\hat{\mathcal{A}}\|^2.$$

First, the definition of the inner product allows us to write

$$\begin{aligned}\langle \mathcal{A}, \hat{\mathcal{A}}\rangle &= \langle \mathcal{A}, \mathcal{B} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \cdots \times_N \mathbf{U}^{(N)}\rangle \\ &= \langle \mathcal{A} \times_1 \mathbf{U}^{(1)^T} \times_2 \mathbf{U}^{(2)^T} \cdots \times_N \mathbf{U}^{(N)^T}, \mathcal{B}\rangle \\ &= \|\mathcal{B}\|^2.\end{aligned}$$

Next, since $\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \ldots, \mathbf{U}^{(N)}$ have orthonormal columns, they do not affect the Frobenius norm

$$\|\hat{\mathcal{A}}\|^2 = \|\mathcal{B}\|^2.$$

Substitution of the preceding expressions in the equation for f yields

$$f(\hat{\mathcal{A}}) = \|\mathcal{A}\|^2 - \|\mathcal{B}\|^2.$$

Combination with the definition of g proves the theorem. ∎

Remark 5. Basically, the best rank-$(R_1, R_2, \ldots, R_N)$ approximation problem consists of the determination of a reduced n-rank tensor $\hat{\mathcal{A}}$ that explains as much of the "energy" (sum of the squared entries) of a given tensor $\mathcal{A}$ as possible under the given n-rank constraints. Theorem 4.2 implies that this problem is equivalent to the explicit maximization of the energy of the approximation.

In other words, we are looking for a basis of rank-1 tensors $\{U_{r_1}^{(1)} \circ U_{r_2}^{(2)} \circ \cdots \circ U_{r_N}^{(N)}\}$, in which $U_1^{(n)}, U_2^{(n)}, \ldots, U_{R_n}^{(n)}$ are mutually orthonormal $(1 \leqslant n \leqslant N)$, such that the norm of the projection of $\mathcal{A}$ onto this basis is maximal.

4.2. An orthogonal iteration. For the actual computation of the best rank-$(R_1, R_2, \ldots, R_N)$ approximation of $\mathcal{A}$, let us interpret the function g in the following way. Imagine that the matrices $\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(n-1)}, \mathbf{U}^{(n+1)}, \ldots, \mathbf{U}^{(N)}$ are fixed and that g in (4.4) is merely a quadratic expression in the components of the unknown matrix $\mathbf{U}^{(n)}$, consisting of orthonormal columns. We have

$$g = \|\tilde{\mathcal{U}}^{(n)} \times_n \mathbf{U}^{(n)^T}\|^2, \tag{4.6}$$

in which

$$\tilde{\mathcal{U}}^{(n)} \stackrel{\text{def}}{=} \mathcal{A} \times_1 \mathbf{U}^{(1)^T} \cdots \times_{n-1} \mathbf{U}^{(n-1)^T} \times_{n+1} \mathbf{U}^{(n+1)^T} \cdots \times_N \mathbf{U}^{(N)^T}. \tag{4.7}$$

Hence the columns of $\mathbf{U}^{(n)}$ can be found as an orthonormal basis for the dominant subspace of the n-mode space of $\tilde{\mathcal{U}}^{(n)}$.

Repeating this procedure for different mode numbers leads to an ALS algorithm for the (local) minimization of $f(\hat{\mathcal{A}})$: in each step the estimate of one of the matrices $\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \ldots, \mathbf{U}^{(N)}$ is optimized, while the other matrix estimates are kept constant.

ALGORITHM 4.2.
HIGHER-ORDER ORTHOGONAL ITERATION.
In: $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$.
Out: $\hat{\mathcal{A}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$: estimator of best rank-$(R_1, R_2, \ldots, R_N)$ approximation of $\mathcal{A}$.

1. Initial values: $\mathbf{U}_0^{(n)} \in \mathbb{R}^{I_n \times R_n}$, of which the columns form an orthonormal basis for the dominant R_n-dimensional left singular subspace of $\mathbf{A}_{(n)}$ $(2 \leqslant n \leqslant N)$ and/or repeat the algorithm for several initial values.
2. Iterate until convergence:
   - $\tilde{\mathcal{U}}_{k+1}^{(1)} = \mathcal{A} \times_2 \mathbf{U}_k^{(2)^T} \times_3 \mathbf{U}_k^{(3)^T} \cdots \times_N \mathbf{U}_k^{(N)^T}$;
     Maximize over $\mathbf{U}^{(1)} \in \mathbb{R}^{I_1 \times R_1}$ with $\mathbf{U}^{(1)^T}\mathbf{U}^{(1)} = \mathbf{I}$:
     $h(\mathbf{U}^{(1)}) = \|\tilde{\mathcal{U}}_{k+1}^{(1)} \times_1 \mathbf{U}^{(1)^T}\|$;
     $\max(h(\mathbf{U}^{(1)})) = h(\mathbf{U}_{\max}^{(1)}) = \lambda_{k+1}^{(1)}$;
     $\mathbf{U}_{k+1}^{(1)} = \mathbf{U}_{\max}^{(1)}$;
   - $\tilde{\mathcal{U}}_{k+1}^{(2)} = \mathcal{A} \times_1 \mathbf{U}_{k+1}^{(1)^T} \times_3 \mathbf{U}_k^{(3)^T} \cdots \times_N \mathbf{U}_k^{(N)^T}$;
     Maximize over $\mathbf{U}^{(2)} \in \mathbb{R}^{I_2 \times R_2}$ with $\mathbf{U}^{(2)^T}\mathbf{U}^{(2)} = \mathbf{I}$:
     $h(\mathbf{U}^{(2)}) = \|\tilde{\mathcal{U}}_{k+1}^{(2)} \times_2 \mathbf{U}^{(2)^T}\|$;
     $\max(h(\mathbf{U}^{(2)})) = h(\mathbf{U}_{\max}^{(2)}) = \lambda_{k+1}^{(2)}$;
     $\mathbf{U}_{k+1}^{(2)} = \mathbf{U}_{\max}^{(2)}$;
   - $\ldots$
   - $\tilde{\mathcal{U}}_{k+1}^{(N)} = \mathcal{A} \times_1 \mathbf{U}_{k+1}^{(1)^T} \times_2 \mathbf{U}_{k+1}^{(2)^T} \cdots \times_{N-1} \mathbf{U}_{k+1}^{(N-1)^T}$;
     Maximize over $\mathbf{U}^{(N)} \in \mathbb{R}^{I_N \times R_N}$ with $\mathbf{U}^{(N)^T}\mathbf{U}^{(N)} = \mathbf{I}$:
     $h(\mathbf{U}^{(N)}) = \|\tilde{\mathcal{U}}_{k+1}^{(N)} \times_N \mathbf{U}^{(N)^T}\|$;
     $\max(h(\mathbf{U}^{(N)})) = h(\mathbf{U}_{\max}^{(N)}) = \lambda_{k+1}^{(N)}$;
     $\mathbf{U}_{k+1}^{(N)} = \mathbf{U}_{\max}^{(N)}$.

   Converged values: $\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \ldots, \mathbf{U}^{(N)}, \mathcal{B} = \tilde{\mathcal{U}}^{(N)} \times_N \mathbf{U}^{(N)^T}$.
3. $\hat{\mathcal{A}} = \mathcal{B} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \cdots \times_N \mathbf{U}^{(N)}$.

The result is shown in step 2 of Algorithm 4.2. Note that (4.7) can be written in a matrix format as follows:

$$\tilde{\mathbf{U}}_{(n)}^{(n)} = \mathbf{A}_{(n)} \cdot (\mathbf{U}^{(n+1)} \otimes \cdots \otimes \mathbf{U}^{(N)} \otimes \mathbf{U}^{(1)} \otimes \cdots \otimes \mathbf{U}^{(n-1)}).$$

Clearly this technique is a higher-order extension of the orthogonal iteration for matrices [12]. A major difference, in some sense also with the higher-order power method, is that each iteration step involves not only the computation of a multilinear transformation but also the estimation of a dominant subspace.

As far as supersymmetric higher-order tensors are concerned, higher-order orthogonal iterations—like higher-order power iterations—yield unsymmetric intermediate results.

As with higher-order power iterations, it makes sense to initialize the higher-order orthogonal iteration with column-wise orthogonal matrices of which the columns span the space of the dominant left singular vectors of the matrix unfoldings $\mathbf{A}_{(n)}$ $(1 \leqslant n \leqslant N)$; this corresponds to truncation of the HOSVD [9]. We refer to section 3.4.

The complete higher-order orthogonal iteration algorithm, for a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, is presented in Algorithm 4.2.

With respect to Algorithm 4.2, we remark as follows:

(1) It is not required that the columns of $\mathbf{U}_k^{(n)}$ ($k \geqslant 0$ and $1 \leqslant n \leqslant N$) coincide with the left singular vectors of the matrix from which they are derived. It is sufficient to compute an arbitrary orthonormal basis for the R_n-dimensional dominant subspace.

(2) The projection of $\mathbf{U}_k^{(n)}$ on the n-mode space of $\tilde{\mathcal{U}}_{k+1}^{(n)}$ gives a first estimate of the dominant R_n-dimensional subspace in the calculation of $\mathbf{U}_{k+1}^{(n)}$. For a discussion of fast subspace computation methods, we refer to [12, 6].

(3) The stop criterion used in the algorithm for the computation of a dominant subspace, can be based on the value of $h(\mathbf{U}^{(n)})$, which corresponds to the square root of the function g that has to be optimized (see (4.6)). In terms of the accuracy of the components, the termination criterion can take the form of, e.g., $\|{\mathbf{U}_{k+1}^{(n)}}^T \cdot \mathbf{U}_k^{(n)}\|^2 > (1-\epsilon)R_n$ $(1 \leqslant n \leqslant N)$.

(4) In comparison with the situation in section 3.3, jointly updating the estimates of two of the matrices $\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \ldots, \mathbf{U}^{(N)}$ can be quite involved. Let us assume, for example, that the values of $\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(n-1)}, \mathbf{U}_k^{(n+2)}, \ldots, \mathbf{U}^{(N)}$ are fixed and that $\mathbf{U}^{(n)}$ and $\mathbf{U}^{(n+1)}$ have to be updated. Theorem 4.2 then implies that the optimal matrices $\mathbf{U}^{(n)}$ and $\mathbf{U}^{(n+1)}$ can be found as the column-wise orthonormal maximizers of $\|(\mathcal{A} \times_1 {\mathbf{U}^{(1)}}^T \cdots \times_{n-1} {\mathbf{U}^{(n-1)}}^T \times_{n+2} {\mathbf{U}^{(n+2)}}^T \cdots \times_N {\mathbf{U}^{(N)}}^T) \times_n {\mathbf{U}^{(n)}}^T \times_{n+1} {\mathbf{U}^{(n+1)}}^T\|^2$. This involves the estimation of some kind of "simultaneous" dominant subspaces of $R_1 R_2 \cdots R_{n-1} R_{n+2} \cdots R_N$ matrix slices of $\mathcal{A} \times_1 {\mathbf{U}^{(1)}}^T \cdots \times_{n-1} {\mathbf{U}^{(n-1)}}^T \times_{n+2} {\mathbf{U}^{(n+2)}}^T \cdots \times_N {\mathbf{U}^{(N)}}^T$, which is a nontrivial problem if $R_1 R_2 \cdots R_{n-1} R_{n+2} \cdots R_N > 1$.

Example 4. Consider a tensor $\mathcal{A} \in \mathbb{R}^{3 \times 2 \times 2}$, defined by the following matrix unfolding:

$$\mathbf{A}_{(1)} = \left(\begin{array}{cc|cc} 0 & -1 & 1 & 4 \\ 2 & -2 & 3 & -5 \\ 4 & 3 & 5 & -6 \end{array}\right).$$

The left singular matrices of $\mathbf{A}_{(1)}$, $\mathbf{A}_{(2)}$, and $\mathbf{A}_{(3)}$ are given by

$$\mathbf{U}^{(1)} = \begin{pmatrix} -0.2465 & -0.4993 & 0.8306 \\ 0.5217 & 0.6539 & 0.5479 \\ 0.8167 & -0.5684 & -0.0994 \end{pmatrix}, \quad \mathbf{U}^{(2)} = \begin{pmatrix} 0.1715 & 0.9852 \\ 0.9852 & -0.1715 \end{pmatrix},$$

$$\mathbf{U}^{(3)} = \begin{pmatrix} 0.5105 & 0.8599 \\ -0.8599 & 0.5105 \end{pmatrix}.$$

The singular values equal (a) 11.0753, 3.5498, and 3.2768, (b) 10.6975 and 5.6181, and (c) 10.5162, 5.9506, and 0, respectively.

Truncation of the HOSVD after the first term in each mode gives a rank-1 approximation of $\mathcal{A}$ of which the Frobenius norm equals 10.0470. This value is increased to $\lambda = 10.1693$ by a higher-order power iteration (Algorithm 3.2). The evolution of the norm during the iteration is given by the solid line in Figure 4.1. The best rank-1 approximation $\lambda V^{(1)} \circ V^{(2)} \circ V^{(3)}$ is given by

$$V^{(1)} = \begin{pmatrix} -0.2515 \\ 0.6035 \\ 0.7567 \end{pmatrix}, \qquad V^{(2)} = \begin{pmatrix} 0.1344 \\ 0.9909 \end{pmatrix}, \qquad V^{(3)} = \begin{pmatrix} 0.5765 \\ -0.8171 \end{pmatrix}.$$

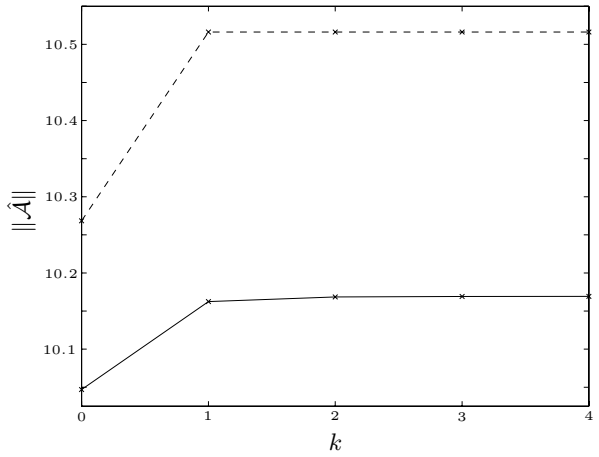


FIG. 4.1. *Evolution of the norm of the approximation $\hat{\mathcal{A}}$ of the tensor $\mathcal{A}$ in Example* 4*, during the higher-order power iteration or higher-order orthogonal iteration, as a function of the iteration step k. Solid line*: *computation of the best rank-*1 *approximation. Dashed line*: *computation of the best rank-*$(2, 2, 1)$ *approximation.*

The best rank-$(2, 2, 1)$ approximation of $\mathcal{A}$, decomposed as $\mathcal{B} \times_1 \mathbf{W}^{(1)} \times_2 \mathbf{W}^{(2)} \times_3 W^{(3)}$ as in (4.2), is given by

$$\mathbf{W}^{(1)} = \begin{pmatrix} -0.2789 & -0.4141 \\ 0.5984 & -0.7806 \\ 0.7511 & 0.4681 \end{pmatrix}, \qquad \mathbf{B}_{(1)} = \left(\begin{array}{cc|c} 10.1473 & 0.0000 \\ 0.0000 & 2.7607 \end{array}\right),$$

$$\mathbf{W}^{(2)} = \begin{pmatrix} 0.0982 & -0.9952 \\ 0.9952 & 0.0982 \end{pmatrix}, \qquad W^{(3)} = \begin{pmatrix} 0.5105 \\ -0.8599 \end{pmatrix}.$$

The Frobenius norm of this approximation equals 10.5162. The evolution of the norm during the higher-order orthogonal iteration (Algorithm 4.2) is given by the dashed line in Figure 4.1.

Example 5. As an alternative for section 4.2, one could think of a generalized deflation approach for the computation of the best rank-$(R_1, R_2, \ldots, R_N)$ approximation, i.e., one could wonder whether the result could not easily be obtained by means of successive rank-1 approximations. In this example we will show that such a procedure is not straightforward.

Consider the supersymmetric $(2 \times 2 \times 2)$-tensor $\mathcal{A}$ of which all the entries are equal to 1 except for $a_{111} = 2$. Obviously, the best rank-$(2, 2, 2)$ approximation is $\mathcal{A}$ itself. The best rank-1 approximation $\lambda_1 U_1 \circ U_1 \circ U_1$ can be computed as explained in section 3.5. Next, we consider the best rank-1 approximation $\lambda_2 U_2 \circ U_2 \circ U_2$ of the residue, and so on. We obtain

$$\begin{array}{lllll} \lambda_1 = 3.2560, & \lambda_2 = 0.5235, & \lambda_3 = -0.3213, & \lambda_4 = -0.1287, & \lambda_5 = 0.0597, \\ \lambda_6 = 0.0264, & \lambda_7 = -0.0118, & \lambda_8 = 0.0053, & \lambda_9 = 0.0024, & \end{array}$$

and so on, and

$$U_1 = \begin{pmatrix} 0.7981 \\ 0.6025 \end{pmatrix}, \quad U_2 = \begin{pmatrix} 0.9186 \\ -0.3952 \end{pmatrix}, \quad U_3 = \begin{pmatrix} 0.0392 \\ -0.9992 \end{pmatrix},$$
$$U_4 = \begin{pmatrix} 0.8733 \\ 0.4872 \end{pmatrix}, \quad U_5 = \begin{pmatrix} 0.8252 \\ -0.5649 \end{pmatrix}, \quad U_6 = \begin{pmatrix} 0.1355 \\ 0.9908 \end{pmatrix},$$
$$U_7 = \begin{pmatrix} 0.9469 \\ 0.3217 \end{pmatrix}, \quad U_8 = \begin{pmatrix} 0.7112 \\ -0.7030 \end{pmatrix}, \quad U_9 = \begin{pmatrix} 0.3107 \\ 0.9505 \end{pmatrix}, \quad \text{etc.}$$

We observe that, in contrast to the matrix case, $\mathcal{A}$ cannot immediately be derived from these rank-1 approximations. Neither is there a straightforward link between the series of rank-1 approximations and the decomposition of $\mathcal{A}$ in a minimal number of rank-1 terms; the latter decomposition is given by $\mathcal{A} = X_1 \circ X_1 \circ X_1 + X_2 \circ X_2 \circ X_2$, in which

$$X_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad X_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

5. Conclusion. There are some remarkable relations, but there are also some striking differences, between the best rank-R approximation of a matrix and the best rank-$(R_1, R_2, \ldots, R_N)$ approximation of an Nth-order tensor:

(1) Unlike the matrix case, the least-squares cost function can show several local optima for higher-order tensors.

(2) For matrices, the best rank-R approximation is obtained by setting the smallest singular values equal to zero while keeping the R largest ones. Truncation of the HOSVD may yield a good tensor approximation, but this approximation is generically suboptimal. However, our simulations suggest that there is still a weak link: it is observed that for well-conditioned problems the HOSVD estimate usually belongs to the valley of the least-squares cost function, corresponding to the global optimum.

(3) One way to obtain the dominant R-dimensional subspaces of the row and column space of a matrix is the power iteration ($R = 1$) or orthogonal iteration ($R > 1$). Higher-order generalizations of these techniques take the form of ALS algorithms, which can be used to enhance the fit between a tensor and a rank-1 or rank-$(R_1, R_2, \ldots, R_N)$ approximation of it.

(4) A higher-order power iteration step basically consists of a multilinear mapping instead of merely a linear transformation. In addition, higher-order orthogonal iterations can involve the estimation of a dominant subspace.

In the literature of psychometrics, the least-squares approximation of a multiway dataset by a dataset with reduced n-mode rank values is known as multimode factor analysis. In this paper, we have further contributed to a linear/multilinear algebraic framework for the best rank-R/rank-$(R_1, R_2, \ldots, R_N)$ approximation problem. We have presented a square-root algorithm, in which an iteration step is based on a multilinear mapping, followed by the estimation of a dominant subspace; this technique was interpreted as a higher-order generalization of the method of orthogonal iterations. We have demonstrated that starting the iteration from a truncated HOSVD does not lead to the global optimum in all possible cases. We have paid special attention to the case of supersymmetric tensors, in view of applications in, for example, HOS. We have extensively discussed the fundamental best rank-1 approximation problem. For supersymmetric $(2 \times 2 \times \cdots \times 2)$-tensors, the determination of the best rank-1

approximation and the other supersymmetric stationary points of the higher-order power algorithm has been reformulated as polynomial rooting problem.

Finally we remark that, as with matrices, the efficiency of higher-order power and orthogonal iterations could further be improved by means of a preprocessing consisting of a finite number of steps in which some tensor entries are set equal to zero. To this end, a higher-order equivalent of the Hessenberg decomposition is proposed in [10]. However this technique is less interesting for tensors than it is for matrices: the relative speed-up becomes smaller for larger tensor order N, as the matrices in which zeros are obtained (e.g., the matrices for which at least one index is equal to one, in the case of square tensors) form a relatively decreasing part of the tensor as N is increased.

REFERENCES

[1] R. Bro, *PARAFAC. Tutorial and applications*, Chemom. Intell. Lab. Systems, 38 (1997), pp. 149–171.

[2] J.-F. Cardoso and P. Comon, *Independent component analysis, a survey of some algebraic methods*, in Proceedings of ISCAS-96, Atlanta, GA, 1996, pp. 93–96.

[3] J. D. Carroll and S. Pruzansky, *The CANDECOMP-CANDELINC family of models and methods for multidimensional data analysis*, in Research Methods for Multimode Data Analysis, H. G. Law, C. W. Snyder, J. A. Hattie, and R. P. McDonald, eds., Praeger, New York, 1984, pp. 372–402.

[4] P. Comon, *Independent component analysis, a new concept?*, Signal Process., 36 (1994), pp. 287–314.

[5] P. Comon and B. Mourrain, *Decomposition of quantics in sums of powers of linear forms*, Signal Process., 53 (1996), pp. 93–108.

[6] P. Comon and G. Golub, *Tracking a few extreme singular values and vectors in signal processing*, Proc. IEEE, 78 (1990), pp. 1327–1343.

[7] L. De Lathauwer, *Signal Processing Based on Multilinear Algebra*, Ph.D. thesis, Katholieke Universiteit Leuven, Leuven, Belgium, 1997.

[8] L. De Lathauwer, B. De Moor, and J. Vandewalle, *Dimensionality reduction in higher-order-only ICA*, in Proceedings of IEEE Signal Processing Workshop on HOS, Banff, Alberta, Canada, IEEE Computer Society, Los Alamitos, CA, 1997, pp. 316–320.

[9] L. De Lathauwer, B. De Moor, and J. Vandewalle, *A multilinear singular value decomposition*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1253–1278.

[10] L. De Lathauwer, B. De Moor, and J. Vandewalle, *A Multilinear Hessenberg Decomposition*, Tech. Report 99-24, SISTA, ESAT, Katholieke Universiteit Leuven, 1999.

[11] L. De Lathauwer, P. Comon, B. De Moor, and J. Vandewalle, *Higher-order power method—Application in independent component analysis*, in Proceedings of NOLTA'95, Las Vegas, NV, 1995, pp. 91–96.

[12] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.

[13] P. M. Kroonenberg, *Three-Mode Principal Component Analysis*, DSWO Press, Leiden, The Netherlands, 1983.

[14] P. M. Kroonenberg, *Three-mode principal component analysis*: *Illustrated with an example from attachment theory*, in Research Methods for Multimode Data Analysis, H. G. Law, C. W. Snyder, J. A. Hattie, and R. P. McDonald, eds., Praeger, New York, 1984, pp. 64–103.

[15] P. M. Kroonenberg and J. de Leeuw, *Principal component analysis of three-mode data by means of alternating least squares algorithms*, Psychometrika, 45 (1980), pp. 69–97.

[16] J. B. Kruskal, *Three-way arrays*: *Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics*, Linear Algebra Appl., 18 (1977), pp. 95–138.

[17] C. L. Nikias and J. M. Mendel, *Signal processing with higher-order spectra*, IEEE Signal Process. Mag., 10 (1993), pp. 10–37.

[18] A. Swami and G. Giannakis, *Bibliography on HOS*, Signal Process., 60 (1997), pp. 65–126.

[19] J. ten Berghe, J. de Leeuw, and P. M. Kroonenberg, *Some additional results on principal components analysis of three-mode data by means of alternating least squares algorithms*, Psychometrika, 52 (1987), pp. 183–191.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 4, pp. 1343–1353
© 2000 Society for Industrial and Applied Mathematics

NONLINEAR MATRIX ITERATIVE PROCESSES AND GENERALIZED COEFFICIENTS OF ERGODICITY*

MARC ARTZROUNI† AND OLIVIER GAVART†

Abstract. A nonlinear matrix iterative process is a linear dynamical system for which there is a nonlinear feedback of the current vector on the entries of the matrix. Stability conditions for an asymptotically exponential solution are studied for such a process in the positive quadrant of $\mathbb{R}^n$. The results hinge on a generalization of the coefficient of ergodicity of a nonnegative matrix.

Key words. nonnegative matrix, coefficient of ergodicity, iterative process, stability

AMS subject classifications. 15A48, 39A10, 39A11

PII. S0895479898348969

1. Introduction.

We let $\mathbb{R}^n_+$ be the positive quadrant of $\mathbb{R}^n$ and $X = (x_i), Y = (y_i)$ be two vectors in $\mathbb{R}^n_+$. The projective distance between X and Y is

$$d(X,Y) = \max_{i,j} \ln \left(\frac{x_i/y_i}{x_j/y_j} \right). \tag{1.1}$$

In the theory of nonnegative matrices the coefficient of ergodicity $\tau(T)$ of a column-allowable $n \times n$ matrix $T = (t_{ij})$ (i.e., a matrix having no zero column) is then defined as

$$\tau(T) = \sup_{X,Y>0; X \neq \lambda Y} \frac{d(X'T, Y'T)}{d(X,Y)}. \tag{1.2}$$

This coefficient (which is between 0 and 1) is also known as *Birkhoff's contraction coefficient* [9] since we have for any X, Y in $\mathbb{R}^n_+$

$$d(X'T, Y'T) \leq \tau(T) d(X,Y). \tag{1.3}$$

The coefficient $\tau(T)$ is useful when studying the linear iterative process $Y(k)' = Y(k-1)'T$ in $\mathbb{R}^n_+$. Indeed, if T is positive (with Perron root the spectral radius $\rho(T)$ and left Perron vector V^*), then T has a coefficient of ergodicity $\tau(T)$ that is less than 1. In this case

$$d(Y(k), V) = d(Y(0)'T^k, V^{*\prime}T^k) \leq \tau(T)^k d(Y(0), V^*) \to 0 \tag{1.4}$$

when $k \to \infty$. Therefore, the vectors $Y(k)$ approach V^* in direction. Asymptotically each component of $Y(k)$ grows as λ^k. We now consider the following definition.

DEFINITION 1.1. *A sequence of positive vectors $Z(k)$ $(k = 0, 1, \ldots)$ is said to be asymptotically exponential in the direction of $Y^* > 0$ with growth rate λ (i.e., $Z(k) \sim \lambda^k Y^*$) if there exists $\alpha > 0$ and a sequence of real vectors $W(k) \to 0$ such that*

$$Z(k) = \alpha \lambda^k (Y^* + W(k)), \quad k = 0, 1, \ldots. \tag{1.5}$$

*Received by the editors November 30, 1998; accepted for publication (in revised form) by U. Helmke August 25, 1999; published electronically May 4, 2000.
http://www.siam.org/journals/simax/21-4/34896.html
†Department of Applied Mathematics, University of Pau, 64000 Pau, France (marc.artzrouni@univ-pau.fr).

A nonautonomous extension of the model $Y(k)' = Y(k-1)'T$ has $Y(k)' = Y(k-1)'T(k)$, where the matrices $T(k)$ now change with the index k (which is typically time) (see [2]). In biological or economics applications, however, there are often feedback mechanisms by which the matrices $T(k)$ are functions not of k but rather of the vector $Y(k-1)$ (see [1], [8]). In this case we have

$$Y(k)' = Y(k-1)'T(Y(k-1)), \qquad k = 1, 2, \ldots. \tag{1.6}$$

The dynamic behavior of such a nonlinear matrix iterative process is of interest, particularly when asymptotically $Y(k) \sim \lambda^k Y^*$ for some λ, Y^*, as in Definition 1.1. The case $\lambda = 1$ has been noted and corresponds to the well-known situation of a converging iterative process (see [4, p. 237]). It is, however, desirable to assess the circumstances under which the exponential behavior $Y(k) \sim Y^*\lambda^k$ occurs with $\lambda \neq 1$. There are two issues.

First, there is the question of the existence of such an equilibrium vector Y^*(which is defined up to a multiplicative constant). If such a vector exists, then for $Y(0) = Y^*$ we must have $Y(k) = \lambda^k Y^*$ for all k. Then

$$Y(k)' = \lambda^k Y^{*\prime} = \lambda^{k-1} Y^{*\prime} T(\lambda^{k-1} Y^*), \qquad k = 1, 2, \ldots, \tag{1.7}$$

and therefore

$$\lambda Y^{*\prime} = Y^{*\prime} T(\lambda^{k-1} Y^*). \tag{1.8}$$

The extension to λ other than 1 implies that the right-hand side of (1.8) must be independent of k. A simple way of achieving this is to assume that $T(Y)$ is a homogeneous matrix function of degree 0, i.e., that $T(cY) = T(Y)$ for any scalar $c > 0$ and vector Y (specifically, each entry $T(Y)_{ij}$ of $T(Y)$ satisfies $T(cY)_{ij} = T(Y)_{ij}$). This homogeneity assumption will be made throughout and is often a natural and realistic assumption in density-dependent biological or demographic models. With this assumption, λ and Y^* must satisfy $\lambda Y^{*\prime} = Y^{*\prime} T(Y^*)$ in order for (1.8) to hold for every k. This means that λ is the spectral radius $\rho(T(Y^*))$ of $T(Y^*)$, with corresponding left Perron vector Y^* [7, p. 493]. When there exist λ and Y^* satisfying $\lambda Y^{*\prime} = Y^{*\prime} T(Y^*)$, we will call Y^* an *equilibrium vector.*

With $X' = (x_1 \quad x_2)$, a typical example of such a homogeneous matrix function is the two-dimensional Leslie matrix

$$T(X) = \begin{pmatrix} \dfrac{a_1 x_1 + b_1 x_2}{c_1 x_1 + d_1 x_2} & s_1 \\ \dfrac{a_2 x_1 + b_2 x_2}{c_2 x_1 + d_2 x_2} & 0 \end{pmatrix} \tag{1.9}$$

with positive parameters a_i, b_i, c_i, d_i. In population dynamics the column of homogeneous functions indicates a feedback of the population vector X on the fertility rates. (Examples of such models can be found in [1] and [8] and more theoretical developments can be found in [5], [6]).

A second question is how to define the exponential behavior $Y(k) \sim \lambda^k Y^*$ in terms of a converging iterative process (as in the case $\lambda = 1$). We will show that $Y(k) \sim \lambda^k Y^*$ holds for large k if $d(Y(k), Y^*)$ approaches 0 exponentially fast. This fast convergence will hinge on a contraction property of $T(X)$ that requires the following definitions.

DEFINITION 1.2. *A projective-distance neighborhood of $Y > 0$ is the set S of all positive vectors X that satisfy $d(X, Y) \leq \varepsilon$ for some $\varepsilon > 0$. This set will be denoted $N(Y, \epsilon)$.*

DEFINITION 1.3. *Given an equilibrium vector Y^*, we say that $T(X)$ is locally contracting with respect to Y^* if there is a projective-distance neighborhood S of Y^* and a positive number $C < 1$ (the coefficient of contraction) such that for any $X \in S$ we have*

$$d(X'T(X), Y^{*\prime}T(Y^*)) = d(X'T(X), Y^{*\prime}) \leq Cd(X, Y^*). \tag{1.10}$$

We will say that $T(X)$ is globally contracting with respect to Y^* if (1.10) *holds for any $X > 0$ (i.e., $S = \mathbb{R}^n_+$).*

The remainder of the paper will be organized as follows. In section 2 we show that this local contraction property implies the exponentially fast convergence of $d(Y(k), Y^*)$ to 0 (as in (1.4)), which in turn ensures that $Y(k) \sim \lambda^k Y^*$ for large k. In section 3 we introduce the generalized coefficient of ergodicity of a matrix T, which is then used to establish a stability condition under which an equilibrium point Y^* is locally contracting. Three numerical examples are given in section 4, and section 5 is devoted to a brief conclusion.

2. Exponential growth and local contractions. We consider the following assumption concerning $T(X)$.

Assumption (A). $T(X)$ is a nonnegative matrix function that is C^1, homogeneous of degree 0 in $\mathbb{R}^n_+$, and column allowable (i.e., for any $X \in \mathbb{R}^n_+$ every column of $T(X)$ has at least one positive entry).

The column-allowability guarantees that for any $Y(0) > 0$ all subsequent iterates $Y(k)' = Y(k-1)'T(Y(k-1))$ will be positive.

In what follows we will need the following technical lemma proved in [2].

LEMMA 2.1. *For any $\varepsilon > 0$ and any two vectors $X = (x_i) > 0$ and $Y = (y_i) > 0$ in $\mathbb{R}^n_+$, then $d(X, Y) \leq \varepsilon$ if and only if there exists a diagonal matrix $M \geq 0$ (with $\operatorname{diag}(M) = \{m(i)\}$) such that $m(i) \leq e^\varepsilon - 1$ $(i = 1, 2, \ldots, n)$ and $X = c(Y + MY)$.*

We can now show that the contraction property defined above implies the asymptotic exponential behavior of $Y(k)$.

THEOREM 2.2. *Assume $T(X)$ satisfies Assumption* (A)*, has an equilibrium vector $Y^* > 0$, and is locally contracting with respect to Y^* in some neighborhood S of Y^*. Then there is a $C < 1$ such that the nonlinear matrix iterative process $Y(k)' = Y(k-1)'T(Y(k-1))$ satisfies, for any $Y(0) \in S$,*

$$d(Y(k), Y^*) \leq C^k d(Y(0), Y^*). \tag{2.1}$$

Furthermore $Y(k) \sim \lambda^k Y^*$ for large k.

Proof. If we let $a \overset{def}{=} d(Y(0)', Y^{*\prime})$, then the local contraction property (1.10) yields $d(Y(k)', Y^{*\prime}) \leq C^k a$, which is (2.1). By Lemma 2.1 each $Y(k)$ can then be written as

$$Y(k) = c(k)(Y^* + M(k)Y^*), \tag{2.2}$$

where $c(k) > 0$ and each diagonal entry of $M(k)$ is bounded by $[\exp(C^k a) - 1]$. Given that $Y(k)' = Y(k-1)'T(Y(k-1))$, we have

$$c(k)(Y^* + M(k)Y^*)' = c(k-1)(Y^* + M(k-1)Y^*)'T[c(k-1)(Y^* + M(k-1)Y^*)]$$

$$= c(k-1)(Y^* + M(k-1)Y^*)'T[(Y^* + M(k-1)Y^*)] \tag{2.3}$$

and therefore

$$\frac{c(k)(Y^* + M(k)Y^*)'}{c(k-1)} = (Y^* + M(k-1)Y^*)'T[(Y^* + M(k-1)Y^*)]. \tag{2.4}$$

A Taylor expansion of $T(X)$ at Y^* yields the existence of a positive number A_1 and of a sequence of real matrices $\Psi(k) = (\psi(k)_{ij})$ (with norm $|||\Psi(k)||| = \max_{i,j} |\psi(k)_{ij}|$) such that for all k

$$T[Y^* + M(k-1)Y^*] = T(Y^*) + \Psi(k-1), \quad |||\Psi(k)||| \le A_1[\exp(C^k a) - 1]. \tag{2.5}$$

From (2.4)–(2.5) it can be seen that there exist $A_2 > 0$ and real vectors $\Phi(k) = (\phi(k)_i)$ (with norm $||\Phi(k)|| = \max_i |\phi(k)_i|$) such that for all k

$$\frac{c(k)}{c(k-1)}Y^* = Y^{*\prime}T(Y^*) + \Phi(k-1) = \lambda Y^* + \Phi(k-1), \quad ||\Phi(k)|| \le A_2[\exp(C^k a) - 1]. \tag{2.6}$$

Hence $c(k)/c(k-1) - \lambda = \eta(k)$ with $\sum_{k=1}^{\infty} |\eta(k)| < \infty$. Therefore the $c(k)$'s grow asymptotically as λ^k in the sense that there exist $\alpha > 0$ and real numbers $\varepsilon(k) \to 0$ such that $c(k) = \alpha\lambda^k(1 + \varepsilon(k))$ for all k. Therefore $Y(k) = \alpha\lambda^k(1 + \varepsilon(k))(Y^* + M(k)Y^*)$, and if we set

$$W(k) \overset{def}{=} [1 + \varepsilon(k)](Y^* + M(k)Y^*) - Y^*, \quad k = 1, 2, \ldots, \tag{2.7}$$

then $Y(k) = \alpha\lambda^k[Y^* + W(k)]$ with $W(k) \to 0$, which shows that $Y(k) \sim \lambda^k Y^*$ for large k.

In the next section we introduce the generalized coefficient of ergodicity of a matrix T, which is then used to establish a stability result.

3. Stability result.

3.1. Generalized coefficient of ergodicity. We begin with some notations. Let $P = (p_{ij})$ be a row-stochastic matrix and define $\Delta(i, j) = \{k : p_{ik} - p_{jk} \ge 0\}$. Then

$$\tau_1(P) \overset{def}{=} \frac{1}{2}\max_{i,j}\sum_{k=1}^{n} |p_{ik} - p_{jk}| = \max_{i,j}\sum_{k\in\Delta(i,j)} (p_{ik} - p_{jk}) \tag{3.1}$$

is a proper coefficient of ergodicity on the set of row-stochastic matrices ($0 \le \tau_1(P) \le 1$; see [9, p. 137].

Now for any nonnegative column-allowable matrix $\Psi = (\psi_{ij})$ and any positive vector $V = (v_i) > 0$, one can construct the row-stochastic matrix function $P(\Psi, V)$ defined as

$$P(\Psi, V)_{ij} \overset{def}{=} \frac{v_j\psi_{ji}}{\sum_{s=1}^{n} v_s\psi_{si}}, \quad i, j = 1, 2, \ldots, n. \tag{3.2}$$

We then call $\tau_1[P(\Psi, V)]$ the *coefficient of ergodicity of the matrix* Ψ *at* V.

We will need to generalize this definition to a *real* matrix Ψ with V still positive. If no entry of $V'\Psi$ is zero, then the denominators appearing in (3.2) are nonzero and the matrix $P(\Psi, V)$ is a *real row-stochastic matrix* (i.e., a real matrix with row sums unity). We now consider the following definition.

DEFINITION 3.1. *Let Ψ be a real matrix and $V > 0$ such that $V'\Psi$ has no zero entry. We let $\Delta(i,j) \stackrel{def}{=} \{k : P(\Psi,V)_{ik} - P(\Psi,V)_{jk} \geq 0\}$, which is nonempty because $P(\Psi,V)$ is a real row-stochastic matrix. We then define*

$$\tau_1[P(\Psi,V)] \stackrel{def}{=} \max_{i,j} \sum_{k\in\Delta(i,j)} (P(\Psi,V)_{ik} - P(\Psi,V)_{jk}) \tag{3.3}$$

as the generalized coefficient of ergodicity of the matrix Ψ at V ($\tau_1[P(\Psi,V)]$ is nonnegative but can be > 1).

We now consider the case of a nonnegative matrix function $T(X)$ ($X \in \mathbb{R}^n_+$) that satisfies Assumption (A). We will need the derived real matrix function $T^*(X)$ defined for $X > 0$ as

$$T^*(X)_{ij} \stackrel{def}{=} T(X)_{ij} + \sum_{k=1}^{n} x_k \frac{\partial T(X)_{kj}}{\partial x_i}. \tag{3.4}$$

Euler's theorem for homogeneous functions implies that for any $k, j \in (1, 2, \ldots, n)$ we have

$$\sum_{i=1}^{n} x_i \frac{\partial T(X)_{kj}}{\partial x_i} = 0. \tag{3.5}$$

The rth component of $X'T^*(X)$ is then $\sum_{i=1}^n x_i T(X)_{ir}$, which is positive because $X > 0$ and $T(X)$ is column allowable. It is then possible to calculate the generalized coefficient of ergodicity of the matrix $T^*(X)$ at X, which leads to the following definition.

DEFINITION 3.2. *Assume $T(X)$ satisfies Assumption (A). We define the generalized coefficient of ergodicity $\tau_1^*(T,X)$ of $T(X)$ at $X = (x_i) > 0$ as the generalized coefficient of ergodicity of the matrix $T^*(X)$ at X, i.e.,*

$$\begin{aligned} \tau_1^*(T,X) &\stackrel{def}{=} \tau_1[P(T^*(X),X)] \\ &= \max_{i,j} \sum_{k\in\Delta(i,j)} (P(T^*(X),X)_{ik} - P(T^*(X),X)_{jk}), \end{aligned} \tag{3.6}$$

where

$$P(T^*(X),X)_{rk} = \frac{x_k \left(T(X)_{kr} + \sum_{i=1}^n x_i \dfrac{\partial T(X)_{ir}}{\partial x_k} \right)}{\sum_{i=1}^n x_i T(X)_{ir}}, \qquad r,k = 1,2,\ldots,n, \tag{3.7}$$

and

$$\Delta(i,j) = \{k : P(T^*(X),X)_{ik} - P(T^*(X),X)_{jk} \geq 0\}. \tag{3.8}$$

In the next subsection we will use these notions to formulate the main stability result of this paper.

3.2. Local stability theorem. For a fixed positive vector $Y = (y_i) > 0$ our ultimate goal is to explore conditions on $T(X)$ under which the function

$$R(X,Y) \stackrel{def}{=} \frac{d(X'T(X), Y'T(Y))}{d(X,Y)} \tag{3.9}$$

remains bounded below 1 for $X = (x_i) > 0$ close to Y for the projective distance but not colinear with Y. When Y is an equilibrium vector Y^*, Theorem 2.2 can then be used to obtain a local stability result.

In Theorem 3.3 we show that $R(X, Y)$ is close to $\tau_1^*[T, Y]$ for X in a sufficiently small projective-distance neighborhood $N(Y, \delta)$ of Y.

THEOREM 3.3. *Assume $T(X)$ satisfies* (A). *Let Y be an arbitrary positive vector. Then for any $\epsilon > 0$ there is $\delta > 0$ such that*

$$\sup_{X \in N(Y,\delta),\ X \neq \lambda Y} R(X, Y) \leq \tau_1^*[T, Y] + \epsilon. \tag{3.10}$$

Proof. In order to study $R(X, Y)$ for fixed Y, we will write the components x_i of X as $x_i = y_i(r + se_i) = ry_i(1 + se_i/r)$, where $r = \min_i x_i/y_i$, $s = \max_i x_i/y_i - \min_i x_i/y_i$ and $E = (e_i)$ is a nonnegative vector with components bounded by 1 with at least one component equal to 0 and one equal to 1. We call Ω the set of such nonnegative vectors:

$$\Omega = \{E = (e_i) : 0 \leq e_i \leq 1, i = 1, 2, \ldots, n; \exists e_r = 0, \exists\ e_s = 1\}. \tag{3.11}$$

We note that $d(X, Y) = \ln(1+s/r)$. Because in (3.9) we are dealing with projective distances (and $T(X)$ is homogeneous of degree 0) we may assume with no loss of generality that $r = 1$. Also, $s > 0$ because X must not be colinear with Y (in order for $R(X, Y)$ to be defined). We can then write $X = Y + sE \cdot Y$ (where $E \cdot Y$ is the componentwise (or Hadamard) product of E by Y). The vector X is thus, without loss of generality, strictly larger than Y componentwise, with its direction beyond Y defined by E and its length by $s > 0$. The advantage of this expression for X is that the denominator $d(X, Y)$ in (3.9) is simply $\ln(1 + s)$.

In order to express the numerator in (3.9) we now define, for all i, k, r,

$$\Theta_i = \sum_{k=1}^{n} y_k T(Y)_{ki}, \quad \Delta(E, s)_{kr} = T(Y + sE \cdot Y)_{kr} - T(Y)_{kr}, \tag{3.12}$$

$$\Gamma_{1,i}(E, s) = \frac{\sum_{k=1}^{n} y_k e_k \Delta(E, s)_{ki}}{\Theta_i}, \quad \Gamma_{2,i}(E, s) = \frac{\sum_{k=1}^{n} y_k \Delta(E, s)_{ki}}{\Theta_i}. \tag{3.13}$$

If we recall the definition $P(T(Y), Y)_{ik} = y_k T(Y)_{ki}/\Theta_i$ of $P(T(Y), Y)$ and use the fact that $X = Y + sE \cdot Y$ and that $T(X)_{ki} = \Delta(E, s)_{ki} + T(Y)_{ki}$, the numerator in (3.9) is

$$\max_{i,j} \ln \left[\frac{\sum_{k=1}^{n} x_k T(X)_{ki} / \sum_{k=1}^{n} y_k T(Y)_{ki}}{\sum_{k=1}^{n} x_k T(X)_{kj} / \sum_{k=1}^{n} y_k T(Y)_{kj}} \right]$$

$$= \max_{i,j} \ln \left[\frac{1 + s \sum_{k=1}^{n} e_k P(T(Y), Y)_{ik} + s\Gamma_{1,i}(E, s) + \Gamma_{2,i}(E, s)}{1 + s \sum_{k=1}^{n} e_k P(T(Y), Y)_{jk} + s\Gamma_{1,j}(E, s) + \Gamma_{2,j}(E, s)} \right]. \tag{3.14}$$

The ratio $R(X, Y)$ is therefore

$$R(Y+sE\cdot Y,Y)=\left[\frac{\max_{i,j}\ln\left[\dfrac{1+s\sum_k e_kP(T(Y),Y)_{ik}+s\Gamma_{1,i}(E,s)+\Gamma_{2,i}(E,s)}{1+s\sum_k e_kP(T(Y),Y)_{jk}+s\Gamma_{1,j}(E,s)+\Gamma_{2,j}(E,s)}\right]}{\ln(1+s)}\right]$$

$$= \frac{\ln\left(\dfrac{1+s\times\max\limits_i a_i(E,s)}{1+s\times\min\limits_j a_j(E,s)}\right)}{\ln(1+s)}, \tag{3.15}$$

where

$$a_r(E,s)=\sum_{q=1}^{n} e_qP(T(Y),Y)_{rq}+\Gamma_{1,r}(E,s)+\Gamma_{2,r}(E,s)/s,\quad r=1,2,\dots,n. \tag{3.16}$$

We will now show that for $s\to 0$ each $a_r(E,s)$ has a limit $a_r^*(E)$. We begin with a Taylor expansion of $\Delta(E,s)_{kr}$ at $s=0$. ($\frac{\partial T(U,V)_{kr}}{\partial s}$ will denote the directional derivative of $T(X)_{kr}$ at U in the direction of V.) Then, for any k,r,

$$\text{for all } s>0,\exists 0\le\theta<1,\quad \text{s.t. } \Delta(E,s)_{kr}=T(Y+sE\cdot Y)_{kr}-T(Y)_{kr}$$

$$= s\frac{\partial T(Y+s\theta E\cdot Y,E\cdot Y)_{kr}}{\partial s}=s\sum_{q=1}^{n} y_qe_q\frac{\partial T(Y+s\theta E\cdot Y)_{kr}}{\partial x_q}, \tag{3.17}$$

where we have used the fact that the directional derivative is the scalar product of the direction vector $E\cdot Y=(y_ie_i)$ by the gradient of the function.

We now recall that $E\in\Omega$, where Ω is the set of nonnegative vectors with components between 0 and 1, with at least one of these components equal to 0 and one equal to 1. Equation (3.17) shows that $\Delta(E,s)_{kr}$ and $\Delta(E,s)_{kr}/s$ are *equicontinuous at* 0, i.e.,

$$\lim_{s\to 0}\Delta(E,s)_{kr}=0,\qquad \lim_{s\to 0}\Delta(E,s)_{kr}/s=\sum_{q=1}^{n} y_qe_q\frac{\partial T(Y)_{kr}}{\partial x_q}, \tag{3.18}$$

where these limits are uniform with respect to k,r, θ, and $E\in\Omega$.

The functions $a_r(E,s)$ in (3.16) are also equicontinuous (uniformly with respect to E and r) at $s=0$ and

$$a_r^*(E)\overset{def}{=}\lim_{s\to 0}a_r(E,s)=\sum_{q=1}^{n} e_qP(T(Y),Y)_{rq}+\frac{\sum_{k=1}^n y_k\sum_{q=1}^n y_qe_q\dfrac{\partial T(Y)_{kr}}{\partial x_q}}{\Theta_r}$$

$$=\frac{\sum_{q=1}^n e_qy_q\left(T(Y)_{qr}+\sum_{k=1}^n y_k\dfrac{\partial T_{kr}(Y)}{\partial x_q}\right)}{\Theta_r}$$

$$=\sum_q e_qP(T^*(Y),Y)_{rq},\quad r=1,2,\dots,n. \tag{3.19}$$

These equicontinuities and the fact that the ratio of the two logarithms in (3.15) approach $\max_{i,j}[a_i^*(E)-a_j^*(E)]$ for $s \to 0$ imply that for any $\epsilon > 0$ there exists $\delta_o > 0$ such that for any $E \in \Omega$ we have

$$s \le \delta_o \Rightarrow \max_{i,j}[a_i^*(E) - a_j^*(E)] - \epsilon \le R(Y + sE \cdot Y, Y) = \frac{\ln\left(\dfrac{1 + s \times \max\limits_i a_i(E,s)}{1 + s \times \min\limits_j a_j(E,s)}\right)}{\ln(1+s)}$$

$$(3.20) \le \max_{i,j}[a_i^*(E) - a_j^*(E)] + \epsilon = \max_{i,j} \sum_{q=1}^{n} e_q[P(T^*(Y),Y)_{iq} - P(T^*(Y),Y)_{jq}] + \epsilon.$$

Therefore

$$\sup_{E\in\Omega, s\le\delta_o} R(Y + sE \cdot Y, Y) \le \sup_{E\in\Omega} \max_{i,j} \sum_{q=1}^{n} e_q[P(T^*(Y),Y)_{iq} - P(T^*(Y),Y)_{jq}] + \epsilon$$

$$= \max_{i,j} \sup_{E\in\Omega} \sum_{q=1}^{n} e_q[P(T^*(Y),Y)_{iq} - P(T^*(Y),Y)_{jq}] + \epsilon. \tag{3.21}$$

Because $P(T^*(Y),Y)$ is row-stochastic, there is necessarily at least one q such that $[P(T^*(Y),Y)_{iq} - P(T^*(Y),Y)_{jq}] \ge 0$ and one q such that $[P(T^*(Y),Y)_{iq} - P(T^*(Y),Y)_{jq}] < 0$ (unless the differences are zero for all q, in which case the sum in (3.21) is 0 regardless of E). The supremum over E is then attained by choosing $e_q = 1$ when $[P(T^*(Y),Y)_{iq} - P(T^*(Y),Y)_{jq}] \ge 0$ and $e_q = 0$ when $[P(T^*(Y),Y)_{iq} - P(T^*(Y),Y)_{jq}] < 0$:

$$\sup_{E\in\Omega, s\le\delta_o} R(Y + sE \cdot Y, Y) \le \max_{i,j} \sum_{q\in\Delta(i,j)} [P(T^*(Y),Y)_{iq} - P(T^*(Y),Y)_{jq}] + \epsilon$$

$$= \tau_1^*[T,Y] + \varepsilon, \tag{3.22}$$

where $\Delta(i,j) = \{k : P(T^*(Y^*),Y^*)_{ik} - P(T^*(Y^*),Y^*)_{jk} \ge 0\}$. For δ in the theorem's statement we set $\delta \overset{def}{=} \ln(1+\delta_o)$. To prove (3.10) we just need to show that any vector $X \in N(Y,\delta)$ is colinear with a vector of the form $Y + sE \cdot Y$ for some $E \in \Omega$ and $s \le \delta_o$. This is the case because it was noted earlier that we could write $x_i = r y_i(1 + s e_i/r)$ with $d(X,Y) = \ln(1 + s/r)$. Therefore if $X \in N(Y,\delta)$, then $\ln(1 + s/r) \le \delta$ and X is colinear with the vector $Y = (y_i + s y_i e_i/r)$ and s/r is less than $e^\delta - 1 = \delta_o$, which completes the proof.

When combined, Theorems 2.2 and 3.3 yield the main result, which follows.

THEOREM 3.4. *Suppose $T(X)$ satisfies Assumption* (A). *Assume there exists an equilibrium vector $Y^* = (y_i^*) > 0$ ($Y^{*\prime} T(Y^*) = \lambda Y^{*\prime}$) such that the generalized coefficient of ergodicity of $T(X)$ at Y^* is less than* 1, *i.e.,*

$$\tau_1^*[T,Y^*] = \max_{i,j} \sum_{k\in\Delta(i,j)} (P(T^*(Y^*),Y^*)_{ik} - P(T^*(Y^*),Y^*)_{jk}) < 1, \tag{3.23}$$

where

$$P(T^*(Y^*),Y^*)_{rk} = \frac{y_k^* \left(T(Y^*)_{kr} + \sum_{i=1}^n y_i^* \dfrac{\partial T(Y^*)_{ir}}{\partial x_k}\right)}{\sum_{i=1}^n y_i^* T(Y^*)_{ir}}, \quad r,k = 1,2,\ldots,n, \tag{3.24}$$

$$\Delta(i,j) = \{k \,:\, P(T^*(Y^*), Y^*)_{ik} - P(T^*(Y^*), Y^*)_{jk} \geq 0\}. \tag{3.25}$$

Then $T(X)$ is locally contracting with respect to Y^* and for any $Y(0)$ in a sufficiently small projective-distance neighborhood of Y^* and the vectors $Y(k)' = Y(k-1)'T(Y(k-1))$ are asymptotically exponential: $Y(k) \sim Y^*\lambda^k$ for large k.

Proof. Theorem 3.3 applied at Y^* with $\epsilon = 0.5(1 - \tau_1^*[T, Y^*])$ shows that $T(X)$ is locally contracting with respect to Y^* with a coefficient of contraction C equal to $0.5(1 + \tau_1^*[T, Y^*]) < 1$. The result follows directly by applying Theorem 2.2.

Remarks.

• At least when $\lambda = 1$ an analogy can be drawn between the condition $\tau_1^*[T, Y^*] < 1$ given here and the classical "spectral radius less than 1" condition for the Jacobian matrix in order for an equilibrium vector to be stable. The quantity $\tau_1^*[T, Y^*]$ is easy to calculate, unlike the spectral radius of a matrix which is computationally of a higher order of complexity. Here the spectral radius criterion is not an option anyway, because the process approaches a whole axis and not one particular vector. This can also be seen by considering in the present context the Jacobian matrix J at an equilibrium vector Y^*. Indeed, it is easy to see that the homogeneity assumption on $T(X)$ implies that the spectral radius of J is in fact larger than or equal to 1.

• In general $\tau_1^*[T, Y^*]$ can take on any positive value. When $T(X)$ is a constant T, however, then $T^*(Y^*) = T$ and $\tau_1^*[T, Y^*] = \tau_1[P(T), Y^*)]$ is ≤ 1.

• Given that $Y^{*\prime}T(Y^*) = \lambda Y^{*\prime}$, the denominator appearing in (3.24) can be replaced with λy_r.

• The ϵ is critical in that $R(Y + sE \cdot Y, Y)$ can very well be an increasing function of s for s increasing beyond 0; i.e., for s arbitrarily small, $R(Y + sE \cdot Y, Y)$ can be larger than $\tau_1[P(T^*(Y^*), Y^*)]$ (even for T constant).

We now briefly give some numerical examples to illustrate this result.

4. Examples.

4.1. The case of constant T. When T is constant then $\tau_1[P(T, Y^*)] \leq 1$ with strict inequality if and only if T' is *scrambling*, i.e., if no two rows of T' are orthogonal. An example is provided by the reducible matrix

$$T = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 2 \end{pmatrix}, \tag{4.1}$$

which has a dominant eigenvalue $\lambda = 2$ with corresponding left eigenvector $Y^{*\prime} = \begin{pmatrix} 2 & 1 & 1 \end{pmatrix}$; T' is scrambling and $\tau_1[P(T, Y^*)] = 0.75$. Therefore $Y(k)' = Y(0)'T^k \sim 2^k Y^{*\prime}$ for large k and any $Y(0) > 0$ close to Y^* (in fact for any $Y(0) > 0$). The strict positivity of $Y(0)$ is the crucial factor here. (Indeed if $Y(0)' = \begin{pmatrix} 1 & 1 & 0 \end{pmatrix}$, then $Y(k)' = \begin{pmatrix} k+1 & 1 & 0 \end{pmatrix}$ for every k.)

This example shows that the iterates $Y(k)' = Y(0)'T^k$ can be asymptotically exponential under conditions much weaker than the positivity (or primitivity) of T. In these latter cases all entries of T^k grow asymptotically like λ^k, and we thus have the stronger result that $Y(k)' = Y(0)'T^k \sim \lambda^k Y^{*\prime}$ for any *nonnegative* $Y(0) \neq 0$ (λ and Y^* are the Perron root and a left Perron vector of T).

4.2. Finitely many equilibria. The matrix function

$$T(X) = \begin{pmatrix} (45x_1 + x_2)/(8x_1 + 0.6x_2) & (0.1x_1 + 2x_2)/(0.5x_1 + 0.7x_2) \\ (55x_1 + 12x_2)/(100x_1 + x_2) & (2x_1 + 9x_2)/(2x_1 + 0.5x_2) \end{pmatrix} \tag{4.2}$$

TABLE 4.1

$Y^{*\prime}$	λ	$\tau_1^*[T, Y^*]$
$\begin{pmatrix} 0.166 & 1 \end{pmatrix}$	11.654	0.447
$\begin{pmatrix} 4.811 & 1 \end{pmatrix}$	5.683	0.846
$\begin{pmatrix} 1.146 & 1 \end{pmatrix}$	5.948	1.156

has three equilibria with values of $(Y^{*\prime}, \lambda, \tau_1^*[T, Y^*])$ given in Table 4.1.

The first two equilibria are stable (since $\tau_1^* < 1$), and the third is unstable.

4.3. Zero or infinitely many equilibria. Consider for $a, b > 0$ and any real α the matrix function

$$T(X) = \begin{pmatrix} a & \dfrac{2 + \cos(x_1/x_2)}{x_1/x_2} \\ \dfrac{2 + \cos(\alpha x_1/x_2)}{x_2/x_1} & b \end{pmatrix}. \tag{4.3}$$

An equilibrium vector $Y^{*\prime} = \begin{pmatrix} y_1^* & y_2^* \end{pmatrix}$ with eigenvalue λ must satisfy $Y^{*\prime}T(Y^*) = \lambda Y^{*\prime}$, i.e.,

$$a + (2 + \cos(\alpha y_1^*/y_2^*)) = \lambda = b + (2 + \cos(y_1^*/y_2^*)). \tag{4.4}$$

For $a = 4, b = 1$, the two equations in (4.4) cannot be satisfied simultaneously and there is no equilibrium.

In what follows we assume $a = 1 = b$. (We note that if $\alpha = 1$, then any $Y^* > 0$ is an equilibrium vector.) When $\alpha \neq 1, -1, 0$ the two infinite sequences (for $p = 1, 2, \ldots$)

$$Y^*(\alpha, p)' = \begin{pmatrix} y_1^*(\alpha, p)_1 & y^*(\alpha, p)_2 \end{pmatrix} \stackrel{def}{=} \begin{pmatrix} 1 & |1 + \alpha|/2p\pi \end{pmatrix}, \tag{4.5}$$

$$Y^*(-\alpha, p)' = \begin{pmatrix} y_1^*(-\alpha, p)_1 & y^*(-\alpha, p)_2 \end{pmatrix} \stackrel{def}{=} \begin{pmatrix} 1 & |1 - \alpha|/2p\pi \end{pmatrix} \tag{4.6}$$

are positive equilibrium vectors that converge to $Y^{**} = \begin{pmatrix} 1 & 0 \end{pmatrix}$(which is not itself an equilibrium vector since it is not positive). The eigenvalues are

$$\lambda(\alpha, p) = 3 + \cos(2p\pi/|1 + \alpha|), \quad \lambda(-\alpha, p) = 3 + \cos(2p\pi/|1 - \alpha|), \tag{4.7}$$

and the corresponding generalized coefficients of ergodicity are

$$\tau_1^*[T, Y^*(\pm\alpha, p)] = \left| 1 + \frac{2p\pi}{|1 \mp \alpha|} \left(\frac{\sin\left(\dfrac{2p\pi}{|1 \mp \alpha|}\right) - \alpha \times \sin\left(\dfrac{2p\pi\alpha}{|1 \mp \alpha|}\right)}{3 + \cos\left(\dfrac{2p\pi\alpha}{|1 \mp \alpha|}\right)} \right) \right|. \tag{4.8}$$

The local stability of each equilibrium vector then depends on whether $\tau_1^*[T, Y^*(\pm\alpha, p)]$ is smaller than 1 or larger than 1. With $\alpha = 0.123$, only 3 among the first 200 equilibrium vectors ($p = 1, 2, \ldots, 100$) have coefficients smaller than 1: $\tau_1^*[T, Y^*(-0.123, 1)] = 0.058; \tau_1^*[T, Y^*(0.123, 7)] = 0.260; \tau_1^*[T, Y^*(0.123, 82)] = 0.077$. For $p \to \infty$, the function $\tau_1^*[T, Y^*(\pm\alpha, p)]$ generally increases with large sine waves of increasing amplitudes. For $p \geq 100$ the pattern for values $\tau_1^*[T, Y^*(\pm\alpha, p)]$ less than 1 is thus complex, but such values seem to be few, with large gaps in between.

In short, we have an accumulation point $Y^{**} = \begin{pmatrix} 1 & 0 \end{pmatrix}$ of the infinite set of equilibrium vectors, of which very few are stable. Given this accumulation, the neighborhoods of the few equilibrium vectors within which the process is stable can be quite small, with many nonstable equilibrium vectors in close proximity. This congestion, combined with observed numerical instabilities in the process $Y(k)' = Y(k-1)'T(Y(k-1))$, make the analytical results presented here all the more useful.

5. Conclusions. Global stability results will need to be investigated and should hinge on the ratio $R(X,Y)$ in (3.9). It is clear that if for an equilibrium vector Y^* there exists $C < 1$ such that

$$\sup_{X>0} R(X, Y^*) < C, \tag{5.1}$$

then Y^* will be a unique, globally stable equilibrium vector. Reasonably simple conditions on $T(X)$ for (5.1) to hold have not been found, however. Another even more general question is whether $\sup_{X,Y>0} R(X,Y)$ is defined, thus extending the definition of the coefficient of ergodicity given in (1.2). If then $\sup_{X,Y>0} R(X,Y) < K < 1$, it will be natural to ask whether contraction-type arguments used in proper metric spaces could be employed here to infer the existence of a single globally stable equilibrium vector.

Acknowledgments. The authors wish to thank two anonymous referees whose comments substantially improved the paper.

REFERENCES

[1] M. Artzrouni, *Les processus itératifs en dynamique des populations et la théorie d'Easterlin.* Math. Sci. Hum., 76 (1981), pp. 33–46.
[2] M. Artzrouni, *On the dynamics of the linear process $Y(k) = A(k)Y(k-1)$ with irreducible matrices $A(k)$*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 822–833.
[3] M. Artzrouni and X. Li, *A note on the coefficient of ergodicity of a column-allowable matrix*, Linear Algebra Appl., 214 (1995), pp. 93–101.
[4] H. Caswell, *Matrix Population Models*, Sinauer, Sunderland, MA, 1989.
[5] T. Fujimoto and U. Krause, *Stable inhomogeneous iterations of nonlinear positive operators on Banach spaces*, SIAM J. Math. Anal., 25 (1994), pp. 1195–1202.
[6] T. Fujimoto and U. Krause, *Asymptotic properties for inhomogeneous iterations of nonlinear operators*, SIAM J. Math. Anal., 19 (1988), pp. 841–853.
[7] R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.
[8] P. Samuelson, *An economists's non-linear model of self-generated fertility waves*, Population Studies, 30 (1976), pp. 243–247.
[9] E. Seneta, *Non-negative Matrices and Markov Chains,* 2nd ed., Springer-Verlag, Berlin, 1981.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 4, pp. 1354–1388
© 2000 Society for Industrial and Applied Mathematics

LOCALLY X MATRICES, SPECTRAL DISTRIBUTIONS, PRECONDITIONING, AND APPLICATIONS*

STEFANO SERRA CAPIZZANO†

Abstract. Starting from a set $X = \{X_n\}_n$ of Hermitian positive definite $n \times n$ matrices, we constructively define a class $\mathcal{LL}(X)$ of "locally X" matrices which can be viewed as the range of a special sequence of linear normally positive operators. Regarding the spectra of these matrices and of the related preconditioned matrices, we prove some Szegő-style ergodic formulas. These results allow one to define a very general procedure for devising optimal and superlinear preconditioners. As special cases, we can deal with matrices coming from the discretization of elliptic and semi-elliptic differential equations defined on multidimensional domains as well as matrices coming from optimization problems connected with graph theory.

Key words. linear positive operators, finite difference and finite element methods, Toeplitz and locally Toeplitz structures, graph, $(0, 1, -1)$ incidence matrix

AMS subject classifications. 15A12, 65F10, 47B36, 28D

PII. S0895479898338366

1. Introduction and premises.

We consider a set $X = \{X_n\}_n$ of Hermitian positive definite $n \times n$ matrices, a set $Z = \{Z_n\}_n$ of unitary $n \times n$ matrices, a bounded domain (nonempty open connected set) $\Omega \subset \mathbf{R}^d$, and a sequence $\mathcal{G} = \{\mathcal{G}_n\}_n$ of grids over $K = \bar{\Omega}$. We suppose that $\mathcal{G}$ is quasi-uniform with respect to K (Definition 2.1) and that X is "local" with respect to $(Z, \mathcal{G})$ in the sense that $Z_n^H X_n Z_n$ is a sparse matrix whose nonzero entries are related to the geometrical distribution of $\mathcal{G}$ over K.

We then define a sequence $\mathcal{LL}(X) = \{\mathcal{LL}_n(X_n)\}_n$ of linear spaces $\mathcal{LL}_n(X_n)$ of Hermitian "locally X" matrices. Indeed, given the space $C(K)$ of the continuous real-valued functions over $K = \bar{\Omega}$ we constructively define special sequences of matrices $\{X_n(a)\}_n$, where $X_n(\cdot)$ is a linear Hermitian matrix-valued operator, a belongs to $C(K)$, and $X_n(1) = X_n$: the union over $a \in C(K)$ of the matrices $X_n(a)$ yields the space $\mathcal{LL}_n(X_n)$. In this way a sequence $Y = \{Y_n\}_n \in \mathcal{LL}(X)$ if there exists a sequence $\{a_n\}_n$ of continuous functions over K such that $Y_n = X_n(a_n)$. Here we are interested in sequences $\{Y_n = X_n(a_n)\}_n$, where $a_n \equiv a$ is fixed with respect to n.

For these matrices we prove the following crucial properties:

- The map $X_n(\cdot) : C(K) \to \mathcal{LL}_n(X_n)$ is a linear (normally) positive operator (NLPO) in the sense stated in Proposition 2.1.
- Given $a \in C(K)$ for which $\{x \in K : a(x) = 0\}$ satisfies some assumptions, the matrices $\{X_n(a)\}_n$ are such that

$$\lim_{n\to\infty} \frac{N^s(X_n(a))}{n} = \frac{\alpha_s(a)}{\text{meas}(K)},$$

where $N^s(X_n(a))$ is the number of the positive, negative, or zero eigenvalues according to $s =$ "+" or $s =$ "−" or $s =$ "0." Moreover $\alpha_s(a)$ denotes the Lebesgue measure of the subset of K where a is positive, negative, or zero

*Received by the editors May 6, 1998; accepted for publication (in revised form) by L. Reichel April 25, 1999; published electronically May 11, 2000.
http://www.siam.org/journals/simax/21-4/33836.html
†Dipartimento di Energetica, Via Lombroso 6/17, 50100 Firenze and Universita di Pisa, Dipartimento di Informatica, Corso Italia 40, 56100 Pisa, Italy (serra@mail.dm.unipi.it).

according to $s =$ "+" or $s =$ "−" or $s =$ "0" and meas$(\cdot)$ is the Lebesgue measure on $\mathbf{R}^d$ (see Theorems 3.4 and 3.5).

Now, fix a and b in the space $C(K)$ so that b is positive and let us introduce new sequences of matrices $\{\mathcal{P}_n(a;b)\}_n$ and $\{\mathcal{P}_n^*(a)\}_n$ with

$$\mathcal{P}_n(a;b) = X_n^{-1}(b)X_n(a) \quad \text{and} \quad \mathcal{P}_n^*(a) = [\hat{D}_n(a)X_n(1)\hat{D}_n(a)]^{-1}X_n(a),$$

where $\hat{D}_n(a) = Z_n D_n^{1/2}(a) Z_n^H$, $D_n(a)$ is the diagonal matrix obtained extracting the main diagonal of $Y_n(a) = Z_n^H X_n(a) Z_n$, and Z_n is the nth unitary matrix of Z.

For these new sequences of matrices we prove localization results and some ergodic Szegő-style relations concerning the asymptotic distribution of the eigenvalues. More precisely, the following facts are proved.

- If a and b are nonnegative and equivalent, i.e., $r \cdot a \le b \le R \cdot a$ with $0 < r \le R < \infty$, then the (nonzero) eigenvalues of $\mathcal{P}_n(a;b)$ belong to $[r, R]$ (see Theorems 3.1 and 3.2).
- If a and b are positive, then we prove that $\{\mathcal{P}_n(a;b)\}_n$ is distributed as a/b; i.e., for any real continuous function F with bounded support we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} F(\lambda_i^{(n)}) = \frac{1}{\text{meas}(K)} \int_K F(a(x)/b(x))dx, \tag{1}$$

  where the set $\{\lambda_i^{(n)}\}_i$ is the complete set of the eigenvalues of $\mathcal{P}_n(a;b)$ (see Theorem 3.7 and section 3.3).
- If a is positive we prove that $\{c^* \cdot \mathcal{P}_n^*(a)\}_n$ is distributed as the constant 1; i.e., for any real continuous function F with bounded support we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} F(\lambda_i^{(n)}) = \frac{1}{\text{meas}(K)} \int_K F(1)dx, \tag{2}$$

  where the set $\{\lambda_i^{(n)}\}_i$ is the complete set of the eigenvalues of $c^* \cdot \mathcal{P}_n^*(a)$ (see Theorem 4.1) where c^* is a suitable scaling number.

Notice that (2) is equivalent to writing that the eigenvalues of $\{c^* \cdot \mathcal{P}_n^*(a)\}_n$ are clustered at 1. (For the notion of *general* and *proper clusters* recalled in Definition 3.2 refer to [44].) Therefore, when the cluster is proper, the matrices $\{\hat{D}_n(a)X_n(1)\hat{D}_n(a)\}_n$ are *superlinear* preconditioners [2, 31] for $\{X_n(a)\}_n \in \mathcal{LL}(X)$ in the sense of Remark 3.2. Moreover the first item (Theorems 3.1 and 3.2) tells us that the preconditioning sequence $\{X_n(b)\}_n$ is optimal in the Axelsson sense [3] (see Definition 3.1), and consequently the number of preconditioned conjugate gradient (PCG) iterations in order to reach the solution within a preassigned accuracy is bounded by an absolute constant independent of the size n (refer to Remark 3.1).

From an abstract point of view it should be noted that the localization results are direct consequences of the concepts of linearity and positivity of the involved matrix operators (compare the results concerning Toeplitz matrices [11, 28, 30], finite difference matrices [31, 36], or abstract classes of matrices [37, 33, 14]).

On the other hand, the "locality" is a sufficient ingredient for the distributional results. It is interesting to compare the present constructive approach with the more abstract formulation considered in [37]: actually, in that paper, the distributional results for the preconditioned matrix sequences are obtained by using the "crucial" assumption that the *nonpreconditioned* ones admit a distribution in the Weyl–Tyrtyshnikov sense [44]. Here we show that the "locality" is enough to infer a similar distributional

behavior on the preconditioned matrices. In addition, what the concept of "locality" reveals is that it is not *necessary* that the matrices $\{X_n(a)\}_n$ (the nonpreconditioned matrices) have an associated distribution. This further nice but unnecessary property can hold if the basic sequence $\{X_n(1) = X_n\}_n$ has an arbitrary distribution. Finally we want to stress that this more specific situation occurs in the case of the locally Toeplitz matrix sequences introduced and studied by Tilli [43]. This fact is not surprising because the Toeplitz structures, which generate in some sense [43] the locally Toeplitz ones, have a further underlying distribution, which is the classical one referred to and analyzed by Szegő [15], Parter [25], Avram [1], Tyrtyshnikov [44, 45], and Tilli [42] over the last 50 years.

The paper is organized as follows. Section 2 is devoted to the main notations and to the general results concerning linear and positive operators in the spirit of [33, 14]. Section 3 is divided into two main parts. In the first we briefly resume localization results for the preconditioned structures and asymptotic results on the extremal behavior and on the conditioning of such classes of matrices. In the second, we derive the ergodic and the density results by following the same approach used in [36] for the case of sequences of finite difference (FD) matrices. Section 4 is devoted to the clustering analysis of sequences $\{\mathcal{P}_n^*(a)\}_n$. To conclude, in section 5 we give some applications involving matrices coming from (semi)elliptic (nonlinear) boundary value problems (BVPs) discretized by FD and finite element (FE) methods and matrices arising from optimization problems and graph theory [9]. In section 6 we make a comparison with other iterative solvers used for BVPs, and finally, in section 7 we propose a further extension of the notion of "locality" that allows one to treat in a unified way "sample" matrices (as FD and graph matrices) and "local integral" matrices (as FE matrices).

1.1. Premise. In order to motivate the construction of locally X matrices, we describe an example that can be regarded as a "model" for our definitions.

Let us consider the following second order BVP:

$$\begin{cases} -\left(a(x)u_x\right)_x = f\left(x\right) & \text{on } \Omega = (0,1), \\ \text{Dirichlet boundary condition on } \partial\,\Omega, \end{cases} \tag{3}$$

that discretized by using centered finite differences of precision order two and stepsize $h = (n+1)^{-1}$ leads to an $n \times n$ linear system $X_n(a)\mathbf{u} = \mathbf{f}$, where

$$X_n(a) = \begin{bmatrix} a_{\frac{1}{2}} + a_{\frac{3}{2}} & -a_{\frac{3}{2}} & & \\ -a_{\frac{3}{2}} & a_{\frac{3}{2}} + a_{\frac{5}{2}} & -a_{\frac{5}{2}} & \\ & \ddots & \ddots & \ddots \\ & & -a_{n-\frac{1}{2}} & a_{n-\frac{1}{2}} + a_{n+\frac{1}{2}} \end{bmatrix}, \quad a_t = a\left(\frac{t}{n+1}\right). \tag{4}$$

Let $X_n = X_n(1)$ be the Toeplitz matrix discretizing problem (3) with $a = 1$. The matrices $X_n(a)$ can be expressed as

$$\sum_{i=1}^{n+1} a_{i-1/2} Q_n(i), \tag{5}$$

where the matrices $Q_n(i)$ are symmetric nonnegative definite dyads given by

$$Q_n(i) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ & -1 & 1 & \\ 0 & 0 & 0 \end{bmatrix}, \quad i = 2, \ldots, n,$$

and $Q_n(1) = \mathbf{e}_1\mathbf{e}_1^T$, $Q_n(n+1) = \mathbf{e}_n\mathbf{e}_n^T$.

Therefore the matrix

$$X_n(1) = \sum_{i=1}^{n+1} Q_n(i) \tag{6}$$

is the sum of all the dyads $Q_n(i)$ and $X_n(a)$ is a weighted sum of the same dyads according to the weights $a_{i-1/2}$. Moreover, each dyad has a "local structure" with respect to the canonical basis of $\mathbf{C}^{n\times n}$ so that each weight $a_{i-1/2}$ contributes in the matrix $X_n(a)$ to $E_{i-1,i-1}$, $E_{i,i-1}$, $E_{i-1,i}, E_{i,i}$, where $E_{s,t} = \mathbf{e}_s\mathbf{e}_t^T$. Furthermore, this notion of "locality" is geometrical as well since vectors of the canonical basis that are close ($\mathbf{e}_s$ and $\mathbf{e}_t$ are close if $|s-t|/n = o(1)$) correspond to dyads $Q_n(s-1), Q_n(s)$, $Q_n(t-1)$, $Q_n(t)$ such that the related weights come from close points in the interval $[0, 1]$.

Therefore we can say that the matrices $\{X_n(a)\}_n$ have a local decomposition with respect to the Toeplitz matrices $\{X_n(1)\}_n$. In [43] the locally Toeplitz matrices have been deeply studied by making heavy use of the underlying Toeplitz structure. Here we analyze general locally X matrices by emphasizing spectral results that do not depend on the specific choice of the basic sequence $X = \{X_n(1)\}_n$. This is why we choose a very generic "name" such as X: because we want to stress that X will have only a very marginal role in the subsequent results. Indeed, as shown in the next section we will make only very mild assumptions on the sequence X.

2. Notations and definitions. Let $X = \{X_n\}_n$ be a sequence of $n \times n$ Hermitian nonnegative definite matrices. Let us consider a sequence $\{\mathbf{v}_j^{(n)} \in \mathbf{C}^n : \ 1 \le j \le n+k\}_{n\in\hat{\mathbf{N}}\subset\mathbf{N}}$ ($\#\hat{\mathbf{N}} = \infty$) of sets of $n+k$ nonzero vectors of $\mathbf{C}^n$ so that

- $k = k(n)$ and
- $\sum_{j=1}^{n+k} \mathbf{v}_j^{(n)}[\mathbf{v}_j^{(n)}]^H = X_n$.

We first assume that $V_n = \{\mathbf{v}_j^{(n)} \in \mathbf{C}^n : \ 1 \le j \le n+k\}$ generates $\mathbf{C}^n$, so that X_n is positive definite; i.e.,

$$k \ge 0, \quad \operatorname{span}_{1\le j\le n+k}(\mathbf{v}_j^{(n)}) = \mathbf{C}^n. \tag{7}$$

Remark 2.1. Concerning the specific example given in (6), we have that $\{X_n(1)\}_n$ is a special instance of $X = \{X_n\}_n$ satisfying (7) with $k = 1$. However, we notice that if we use high precision FD formulas in problems such as (3) the value of k is bigger.

For our construction we need to characterize some grids of points belonging to a set $\Omega \subset \mathbf{R}^d$.

DEFINITION 2.1. *Let $\Omega \subset \mathbf{R}^d$ be a bounded domain and $K = \bar{\Omega}$ be its closure. A sequence of grids $\mathcal{G} = \{\hat{\mathcal{G}}_n\}_n$ with $\hat{\mathcal{G}}_n = \{x_i^{(n)}\}$, $x_i^{(n)} \in \Omega$ and $\#(\hat{\mathcal{G}}_n) = n$ is called quasi-uniform with respect to K if for any ball $B(x_0, \rho) = \{x \in \mathbf{R}^d : \ \|x - x_0\|_\infty \le \rho\} \subset K$ with positive radius ρ we have*

$$\#\left\{x_i^{(n)} \in B(x_0, \rho)\right\} = n \cdot \frac{\operatorname{meas}(B(x_0, \rho))}{\operatorname{meas}(K)} + o(n).$$

The sequence of grids is uniform if for any ball $B(x_0,\rho)\subset K$ we find that

$$\#\left\{x_i^{(n)}\in B(x_0,\rho)\right\}=n\cdot\frac{\mathrm{meas}(B(x_0,\rho))}{\mathrm{meas}(K)}+O(n^{\frac{d-1}{d}}).$$

The sequence of grids is $*$-uniform if for any connected set $S\subset K$ with Lipschitz continuous boundaries we find that

$$\#\left\{x_i^{(n)}\in S\right\}=n\cdot\frac{\mathrm{meas}(S)}{\mathrm{meas}(K)}+O(n^{\frac{d-1}{d}}).$$

THEOREM 2.1. *Let $S\subset\Omega$ be a measurable set in the Peano–Jordan sense* [18, pp. 28–29] *(i.e., its characteristic function is integrable in the multidimensional Riemann sense) and let $\Omega\subset\mathbf{R}^d$ be a bounded domain. Let $\{\hat{\mathcal{G}}_n\}_n$ with $\hat{\mathcal{G}}_n=\{x_i^{(n)}\}$, $x_i^{(n)}\in\Omega$, and $\#(\hat{\mathcal{G}}_n)=n$ be quasi-uniform with respect to $K=\bar{\Omega}$. Then*

$$\#\left\{x_i^{(n)}\in S\right\}=n\cdot\frac{\mathrm{meas}(S)}{\mathrm{meas}(K)}+o(n).$$

Proof. Since S is Peano–Jordan measurable, then it is Lebesgue measurable so that it makes sense to write $\mathrm{meas}(S)$. In addition, for any $\epsilon>0$, there exist two plurirectangles $P_1\subset P_2$ so that

$$\begin{aligned}
&P_1\subset S\subset P_2, && \mathrm{meas}(P_2)-\mathrm{meas}(P_1)<\epsilon,\\
&P_1=\bigcup_{j=1}^{k_1(\epsilon)}B(x_j,\rho_j), && P_2=\bigcup_{j=1}^{k_2(\epsilon)}B(y_j,r_j),\\
&B^\circ(x_j,\rho_j)\cap B(x_k,\rho_k)=\emptyset, && B^\circ(y_j,r_j)\cap B(y_k,r_k)=\emptyset,\quad j\neq k,
\end{aligned}$$

where X° denotes the interior part of a set $X\subset\mathbf{R}^d$. By set inclusion it easily follows that

$$\#\left\{x_i^{(n)}\in P_1\right\}\le\#\left\{x_i^{(n)}\in S\right\}\le\#\left\{x_i^{(n)}\in P_2\right\}.$$

By Definition 2.1 we have

$$\#\left\{x_i^{(n)}\in P_1\right\}=n\cdot\frac{\mathrm{meas}(P_1)}{\mathrm{meas}(K)}+o(n)\ge n\cdot\frac{\mathrm{meas}(S)-\epsilon}{\mathrm{meas}(K)}+o(n)$$

and

$$\#\left\{x_i^{(n)}\in P_2\right\}=n\cdot\frac{\mathrm{meas}(P_2)}{\mathrm{meas}(K)}+o(n)\le n\cdot\frac{\mathrm{meas}(S)+\epsilon}{\mathrm{meas}(K)}+o(n).$$

Since ϵ is arbitrary the proof is over. ∎

Remark 2.2. The quasi uniformity is equivalent by Theorem 2.1 to the fact that $\mathcal{G}$ can be used to approximate the Lebesgue measure of Peano–Jordan sets. The uniformity is clearly implied by the $*$-uniformity. There is the open problem of understanding whether the uniformity implies the $*$-uniformity.

Sometimes in the applications we require some extra points not generally belonging to Ω.

DEFINITION 2.2. *Let $\{\hat{\mathcal{G}}_n\}_n$ be defined as in the previous definition. Then*

$$\mathcal{G}_n=\hat{\mathcal{G}}_n\cup\{x_i^{(n)}\}_{n+1\le i\le n+k(n)}.$$

Here $k(n) = o(n)$ if $\{\hat{\mathcal{G}}_n\}_n$ is quasi-uniform and $k(n) = O(n^{\frac{d-1}{d}})$ if $\{\hat{\mathcal{G}}_n\}_n$ is uniform/$*$-uniform. In this way, the new grids sequence $\{\mathcal{G}_n\}_n$ inherits the quasi uniformity or the uniformity/$*$-uniformity from $\{\hat{\mathcal{G}}_n\}_n$.

At this point we are ready to introduce a property that is the "key" for the "locality" in a sense that will be clear in the following few lines. Let us consider a sequence $Z = \{\{\mathbf{z}_j^{(n)}\}_{1\le j\le n}\}_n$ of orthonormal bases of $\{\mathbf{C}^n\}_n$ (orthogonal and unitary with respect to the inner product inducing the $\|\cdot\|_2$ vector norm). Let $\mathcal{N}_n$ be a bijection from $\hat{\mathcal{G}}_n$ to $\{\mathbf{z}_j^{(n)}\}$ and let $\mathcal{M}_n$ be a bijection from $\mathcal{G}_n$ to $\{\mathbf{v}_j^{(n)}\}$ with $\mathcal{M}_n^{-1}(\mathbf{v}_i^{(n)}) \equiv \mathcal{N}_n^{-1}(\mathbf{z}_i^{(n)})$ for $i = 1, \dots, n$. In this way, to each point x of $\hat{\mathcal{G}}_n \subset \Omega$ we associate a unique vector $\mathbf{z}_{j(x)}^{(n)}$ of the basis $\{\mathbf{z}_j^{(n)}\}_{1\le j\le n}$ and a unique vector $\mathbf{v}_{j(x)}^{(n)}$ of the dyadic decomposition of X_n. Let us call Z_n the unitary matrix constructed as $[\mathbf{z}_1^{(n)}, \dots, \mathbf{z}_n^{(n)}]$ and let us suppose that the sequence of sets of vectors

$$V = \left\{\{\mathbf{v}_j^{(n)}\}_{1\le j\le n+k(n)}\right\}_n$$

is "locally" weakly distributed with regard to Z in the following sense. For any $i \le n + k(n)$, it holds that

$$\#\{j : \ [\mathbf{v}_i^{(n)}]^H \mathbf{z}_j^{(n)} \neq 0\} = O(1).$$

Moreover, for any $j \le n$ and any $i \le n + k(n)$, the relation

$$[\mathbf{v}_i^{(n)}]^H \mathbf{z}_j^{(n)} \neq 0 \quad \text{implies that} \quad \|x_i^{(n)} - x_j^{(n)}\|_\infty = o(1). \tag{8}$$

In other words each $\mathbf{v}_i^{(n)}$ is made up of a *constant* number independent of n of vectors $\{\mathbf{z}_j^{(n)}\}$; moreover each grid point $x_j^{(n)}$ associated via $\mathcal{N}_n$ with $\mathbf{z}_j^{(n)}$ fulfilling (8) belongs to a neighborhood of $x_i^{(n)} \in \mathcal{G}_n$ whose radius is infinitesimal as n tends to infinity. The "locality assumption is strong" when (8) holds with $o(1) = O(n^{-1/d})$; i.e.,

$$[\mathbf{v}_i^{(n)}]^H \mathbf{z}_j^{(n)} \neq 0 \quad \text{implies that} \quad \|x_i^{(n)} - x_j^{(n)}\|_\infty = O(n^{-1/d}). \tag{9}$$

Now we are ready for the main definition.

DEFINITION 2.3. *Let K, $\{\hat{\mathcal{G}}_n\}_n$, and $\{\mathcal{G}_n\}_n$ be quasi-uniform as in Definition 2.1 and let $a : K \to \mathbf{R}$ be a continuous function. We define the operator $X_n(\cdot)$ from $C(K)$ into the space of the Hermitian matrices as follows:*

$$X_n(a) = \sum_{j=1}^{n+k(n)} a_j^{(n)} \mathbf{v}_j^{(n)} [\mathbf{v}_j^{(n)}]^H, \tag{10}$$

where $a_j^{(n)} = a(x_j^{(n)})$. Suppose that (8) holds; then the real space $\mathcal{LL}(V_n) = \mathcal{LL}(X_n)$ of matrices is defined as $\bigcup_{a\in C(K)} X_n(a)$. The class $\mathcal{LL}(X)$ of "locally X" matrices is defined as the sequence $\{\mathcal{LL}(X_n)\}_n$. The "locality" is strong if (9) holds and both $\{\hat{\mathcal{G}}_n\}_n$ and $\{\mathcal{G}_n\}_n$ are $*$-uniform.

As previously noted, it may happen that some of the points $\{x_j^{(n)}\}$ are "extra points" in the sense that they do not belong to Ω. Nevertheless, we impose that

$$\operatorname{dist}(\Omega, \mathcal{G}_n) = \sup_{x\in\Omega} \inf_{y\in\mathcal{G}_n} \|x - y\|_\infty = o(1)$$

while, when the assumption of "locality" holds in the strong sense, it is natural to require that

$$\operatorname{dist}(\Omega, \mathcal{G}_n) = O(n^{-1/d}).$$

In the case where some extra points occur (this is the case of high-precision formulae [36, 38]), we can extend the function a by using Taylor expansions (or other asymptotic expansions) with the crucial constraint given by the following relation:

$$\min_{y\in K} a(y) \le a(x) \le \max_{y\in K} a(y) \quad \forall x \in K^* = K \cup \Big(\bigcup_n \mathcal{G}_n\Big). \tag{11}$$

Here the assumption made in (11) is added in order to guarantee that the extension of a to K^* is positive if and only if $a|_K$ is positive since this property is crucial for the results stated in Theorems 3.1 and 3.2.

2.1. LPOs: Definitions and preliminary results. Let us begin with some theoretical results on linear positive operators (LPOs) proved in [14] and based on simple theorems discussed in [33].

We take two function sets $\mathcal{F} = \mathcal{F}(A, \mathbf{R})$ and $\mathcal{G} = \mathcal{G}(B, \mathbf{R})$, where A and B are sets and $\mathcal{F}$ and $\mathcal{G}$ are (real) subspaces of the space of all the real-valued functions defined over A and B, respectively.

Now let us take an LPO $\Phi : \mathcal{F} \to \mathcal{G}$, where the concept of positivity is with respect to the partial ordering induced in $\mathcal{F}$ and $\mathcal{G}$ by the ordering of $\mathbf{R}$: if $a \in \mathcal{F}$ is nonnegative, then $\Phi(a)$ is nonnegative in $\mathcal{G}$. The concept of linearity expressed by $\Phi(\alpha a + \beta b) = \alpha\Phi(a) + \beta\Phi(b)$ is the natural one.

DEFINITION 2.4. *A linear positive operator Φ is called strong (SLPO) if* $\inf \Phi(a) > 0$ *when a is nonnegative and not identically zero. Φ is called normally positive (NLPO) if* $\infty > \sup \Phi(1) \ge \inf \Phi(1) > 0$ *when the constant function 1 belongs to $\mathcal{F}$.*

We denote by $BILH(\mathbf{S}^n, \mathbf{C})$ the linear space of the homogeneous quadratic forms from $\mathbf{S}^n$ to $\mathbf{R}$, with $\mathbf{S}^n$ being the Euclidean unit sphere of $\mathbf{C}^n$, and by $BILH(\mathbf{S}^n, \mathbf{R})$ the real subspace of $BILH(\mathbf{S}^n, \mathbf{C})$ made up of the self-adjoint forms. It is evident that to each functional in $BILH(\mathbf{S}^n, \mathbf{C})$ corresponds a unique $n \times n$ complex matrix representing the form with respect to the canonical basis of $\mathbf{C}^n$ and vice versa. It is also clear that any $g \in BILH(\mathbf{S}^n, \mathbf{R})$ is related uniquely to a Hermitian matrix.

Moreover, if V is a complex matrix, then V^+ denotes the pseudoinverse in the sense of Moore and Penrose [24, 26]; this notion will be useful to treat numerically the case of singular linear systems by using singular preconditioners.

THEOREM 2.2 (see [14]). *Let us suppose that $\{\Phi_n(\cdot)\}_n$ is a sequence of LPOs from $\mathcal{F}$ to $\mathcal{G} = BILH(\mathbf{S}^n, \mathbf{R})$ representing homogeneous quadratic forms related to the Hermitian matrix-valued operators $\{F_n(\cdot)\}_n$ (i.e., $F_n(a)$ is the Hermitian matrix so that $\mathbf{x}^H F_n(a)\mathbf{x} = (\Phi_n(a))(\mathbf{x})$ for any $\mathbf{x} \in \mathbf{S}^n$). We assume that $a \in \mathcal{F}$ and $b \in \mathcal{F}$ with $b \ge 0$. Then*

1. *each nonzero eigenvalue of $F_n^+(b)F_n(a)$ belongs to $[r, R]$, where*
$$r = \inf_C \frac{a}{b} \quad \text{and} \quad R = \sup_C \frac{a}{b}$$
with $C = \{x \in A : \ a(x) \ne 0 \ \text{ or } \ b(x) \ne 0\}$.
2. *In addition if r and R are finite, then the kernel of the matrices $F_n^+(b)$ and $F_n(b)$ is contained in the kernel of $F_n(a)$.*
3. *Finally if $0 < r \le R < \infty$ (a and b are equivalent) then the kernels of $F_n(a)$ and $F_n(b)$ coincide.*

THEOREM 2.3 (see [14]). *Let us suppose that $\{\Phi_n(\cdot)\}_n$ is a sequence of SLPOs from $\mathcal{F}$ to $\mathcal{G} = BILH(\mathbf{S}^n, \mathbf{R})$ representing homogeneous quadratic forms related to the Hermitian matrix-valued operators $\{F_n(\cdot)\}_n$. We assume that $a \in \mathcal{F}$ and $b \in \mathcal{F}$ with $b \geq 0$. Then*

(1) *each eigenvalue of $F_n^{-1}(b)F_n(a)$ belongs to (r, R), where*

$$r = \inf_C \frac{a}{b} \quad \text{and} \quad R = \sup_C \frac{a}{b}$$

with $C = \{x \in A : \ a(x) \neq 0 \ \ or \ \ b(x) \neq 0\}$ and $r < R$.

(2) *If $r = R$, then $F_n^{-1}(b)F_n(a) = rI_n$.*

Now we link the properties of the "locally X" matrices with the concept of linear and positive operators.

PROPOSITION 2.1. *Let $Ra : \mathbf{C}^{n\times n} \to BILH(\mathbf{S}^n, \mathbf{C})$ be the Rayleigh operator such that $Ra(A)(\mathbf{x}) = \mathbf{x}^H A\mathbf{x}$ for $A \in \mathbf{C}^{n\times n}$ and $\mathbf{x} \in \mathbf{S}^n$. For any fixed dimension n, let $X_n(\cdot)$ be defined as in Definition 2.3. Then the operator*

$$\tilde{X}_n(\cdot) \equiv Ra \circ X_n(\cdot) : C(K) \to BILH(\mathbf{S}^n, \mathbf{R})$$

is linear and positive (LPO). If we assume that (7) holds, then the considered operators $\tilde{X}_n(\cdot)$ are also normally positive in the sense of Definition 2.4 (i.e., $X_n(1)$ is positive definite); in this case we write that the matrix-valued operator $X_n(\cdot)$ is normally positive.

Proof. Let $\tilde{X}_n(\cdot) = Ra(X_n(\cdot))$. The linearity of $\tilde{X}_n(\cdot)$ is trivial. Owing to the Hermitianity of $X_n(a)$ the minimum of the quadratic form $\mathbf{x}^H X_n(a)\mathbf{x}$, $\mathbf{x} \in \mathbf{S}^n$, is exactly the minimum eigenvalue of the matrix $X_n(a)$. By exploiting (10) and since $X_n(1) = X_n$, it follows that the minimal eigenvalue of $X_n(a)$ is not less than $(\min a)\lambda_{\min}(X_n)$. Therefore if a is nonnegative then $X_n(a)$ is nonnegative definite and this proves the positivity. The normal positivity is contained in the assumption (7) from which it follows that $X_n = X_n(1)$ is positive definite for any choice of the dimension n. ∎

3. Main results. In the first part of this section, we report the localization features of the spectra of the sequences $\{\mathcal{P}_n(a; b)\}_n$ by making heavy use of the notion of LPO. The same concept of LPO is still the key to proving asymptotics on the extreme eigenvalues and conditioning of $\{X_n(a)\}_n$ and $\{\mathcal{P}_n(a; b)\}_n$. Finally, in section 3.2, by exploiting the notion of "locality" of the operators $\{X_n(\cdot)\}_n$, we prove the distributional results concerning the matrices $\{\mathcal{P}_n(a; b)\}_n$.

3.1. Definitions and localization results.

DEFINITION 3.1. *Let $\{A_n\}_n$ and $\{B_n\}_n$ be two sequences of $n \times n$ nonnegative definite Hermitian matrices. $\{B_n\}_n$ are optimal preconditioners for $\{A_n\}_n$ if for any n*

- $\ker(B_n^+ A_n) = \ker(A_n)$,
- *all the nonzero eigenvalues of $B_n^+ A_n$ belong to a bounded interval $[r, R]$, with r and R positive constants independent of n.*

Note that the preconditioner needs only to work in the subspace where A_n is invertible; in the orthogonal subspace, where A_n is the null matrix, the problem of finding a solution of the related linear system is ill-posed, and therefore it is convenient that B_n also acts like the null matrix. This definition coincides with the one given by Axelsson and Lindskog [2] when each A_n is nonsingular.

DEFINITION 3.2. *Consider a sequence of n-dimensional real vectors $\{\mathbf{a}_n = \{a_i^{(n)}\}_{i=1}^n\}_{n\in\hat{\mathbf{N}}\subset\mathbf{N}}$ ($\#\hat{\mathbf{N}} = \infty$) and a set M in the real line. Denote by M_ϵ the ϵ-extension of M, which is the union of all balls of radius ϵ centered at points of M.*

For any n, let $\gamma_n(\epsilon)$ count those components $a_i^{(n)}$ of $\mathbf{a}_n$ that do not belong to M_ϵ. Assume that, for any $\epsilon > 0$,

$$\gamma_n(\epsilon) = o(n), \quad n \to \infty.$$

Then M is called a general cluster. If, for any $\epsilon > 0$ there exists a constant $c(\epsilon)$ such that

$$\gamma_n(\epsilon) \le c(\epsilon),$$

then M is called a proper cluster. If $M = \{p\}$, then we say that $\{\mathbf{a}_n\}_n$ is clustered at p.

The same notion applies to a sequence of Hermitian matrices $\{A_n\}_n \in \mathbf{C}^{n\times n}$: in this case the vector $\mathbf{a}_n$ is given by the complete set of the eigenvalues of A_n [44]. Furthermore the same notion applies to a sequence of (general) matrices $\{A_n\}_n \in \mathbf{C}^{n\times n}$: in this case the vector $\mathbf{a}_n$ is given by the complete set of the singular values of A_n [44].

Remark 3.1. If $\{B_n\}_n$ is an optimal preconditioning sequence for $\{A_n\}_n$ in the sense of Definition 3.1, then the number of PCG iterations in order to reach the solution (in the least squares sense in this case) within a preassigned accuracy is bounded by a constant independent of the size n [2]. In this case we say that the PCG method is optimal.

Remark 3.2. When $\{B_n\}_n$ is such that the preconditioned sequence for $\{B_n^+ A_n\}_n$ is spectrally clustered at $p > 0$ in the sense of Definition 3.2 and the minimal singular value does not go to zero too fast, then the corresponding PCG method is optimal and the "quality" of the convergence is superlinear [2].

DEFINITION 3.3. *Let a be a continuous function on K and b be continuous and nonnegative on K. The preconditioned matrix $\mathcal{P}_n(a,b)$ is defined as $X_n^+(b)X_n(a)$, where the $X_n(\cdot)$ is as in Definition* 2.3.

DEFINITION 3.4. *Let a be a continuous nonnegative function over K. The preconditioned matrix $\mathcal{P}_n^*(a)$ is defined as $\mathcal{P}_n^*(a) = [\hat{D}_n(a)X_n(1)\hat{D}_n(a)]^+ X_n(a)$, where $\hat{D}_n(a) = Z_n D_n^{1/2}(a) Z_n^H$, $D_n(a)$ is the diagonal matrix obtained by extracting the main diagonal of $Y_n(a) = Z_n^H X_n(a) Z_n$ and Z_n is the nth unitary matrix of Z. Here $X_n(\cdot)$ is as in Definition* 2.3 *and $Z = \{\{\mathbf{z}_j^{(n)}\}_{1\le j\le n}\}_n$ is the sequence of orthonormal bases of $\{\mathbf{C}^n\}_n$ used for defining the locality notion in* (8) *and* (9).

THEOREM 3.1. *Let a, $b \in C(K)$ and $\mathcal{P}_n(a,b)$ be as in Definition* 3.3. *Under the assumption that b is nonnegative, each nonzero eigenvalue of $\mathcal{P}_n(a;b)$ belongs to $[r,R]$. Here r and R are the inf and the sup of a/b taken over the set C where the function a/b is well defined (the same set C defined in Theorem* 2.2*). In addition, if r and R are finite, then the kernel of the matrices $X_n^+(b)$ and $X_n(b)$ is contained in the kernel of $X_n(a)$. Finally if $0 < r \le R < \infty$ (a and b are equivalent), then the kernels of $X_n(a)$ and $X_n(b)$ coincide.*

Proof. By Proposition 2.1, the operator $\tilde{X}_n(\cdot) = Ra(X_n(\cdot))$ fulfills the assumptions of Theorem 2.2 so that the theorem is proved. □

THEOREM 3.2. *Under the same hypotheses as in Theorem* 3.1 *and by supposing that* (7) *holds, an eigenvalue of $\mathcal{P}_n(a,b)$ lies in (r,R) if $r < R$. Otherwise $\mathcal{P}_n(a,b)$ coincides with r times the identity.*

Proof. By Proposition 2.1, the operator $\tilde{X}_n(\cdot) = Ra(X_n(\cdot))$ fulfills the assumptions of Theorem 2.3 so that the theorem is proved. □

We now give a result connecting the concepts of LPO and the asymptotic spectral behavior of matrix sequences related to linear positive matrix-valued operators.

THEOREM 3.3 (see [14]). *Let us suppose that $\{\Phi_n(\cdot)\}_n$ is a sequence of LPOs from $\mathcal{F}$ to $\mathcal{G} = BILH(\mathbf{S}^n, \mathbf{R})$ with $1 \in \mathcal{F}$ and $\Phi_n(1) = 1$, where $\{\Phi_n(\cdot)\}_n$ represent a sequence of homogeneous quadratic forms related to the Hermitian matrix-valued operators $\{F_n(\cdot)\}_n$. We assume that $a \in \mathcal{F}$ and $b \in \mathcal{F}$ with $b \geq 0$ and that they satisfy the condition $a - m_a \sim b - m_b$, where m_a and m_b are the* inf *of a and b, respectively. Then, setting $\lambda_{j,n}(\cdot)$ as the jth eigenvalue of $F_n(\cdot)$, we have*

$$\lambda_{j,n}(a) - m_a \sim \lambda_{j,n}(b) - m_b.$$

If $m_a = m_b = 0$ and $\kappa_{2,n}(X)$ denotes the spectral condition number of an $n \times n$ matrix X, then $\kappa_{2,n}(F_n(a)) \sim \kappa_{2,n}(F_n(b))$. If $\{\Phi_n(\cdot)\}_n$ does not reproduce the constants, i.e., if $F_n(1)$ is not the identity, then from $a \sim b$ we infer

$$\lambda_{j,n}(a) \sim \lambda_{j,n}(b) \quad \textit{and} \quad \kappa_{2,n}(F_n(a)) \sim \kappa_{2,n}(F_n(b)).$$

Remark 3.3. Since $X_n(\cdot)$ is a linear positive matrix-valued operator, we infer that the preceding theorem applies to $\{X_n(\cdot)\}_n$ as well.

3.2. Ergodic and distribution results. In order to prove the density and the distribution results for the preconditioned matrix sequences (in the spirit of the Szegő theorem for Toeplitz matrices [15]), we need preparatory results about the asymptotic inertia of the matrices $X_n(a)$. We stress that it is possible to obtain similar results by following other approaches as the one described in [37]; on the other hand, in [37] we assumed that the basic matrix sequences $\{X_n(a)\}_n$ fulfill an ergodic limit relation and this is not necessarily our case.

Notice that as a by-product of Theorem 3.1, we obtained that the inertia of the matrices $X_n(a)$ and $X_n(b)$ coincide when a and b are equivalent (i.e., when $r \cdot a \leq b \leq R \cdot a$ with $0 < r \leq R < \infty$). Here we are interested in linking the asymptotical inertia of $X_n(a)$ to the distribution of sign of the coefficient function $a(x)$ and in particular to the Lebesgue measure of the sets where a is positive or negative or zero.

We will use the following definitions and notations.

DEFINITION 3.5. *Fix $s < t$ and a, b as two continuous functions defined over K such that b is nonnegative. Then*

$$\begin{aligned}
N(s,t,a,n) &= \#\{i:\ \lambda_i(X_n(a)) \in (s,t)\},\\
N(s,t,(a,b),n) &= \#\{i:\ \lambda_i(X_n^+(b)X_n(a)) \in (s,t)\},\\
A_+(a) &= \{x \in K:\ a(x) > 0\},\\
A_-(a) &= \{x \in K:\ a(x) < 0\},\\
A_0(a) &= \{x \in K:\ a(x) = 0\}.
\end{aligned}$$

Moreover, when $(s,t) = (-\infty, 0)$ or $(s,t) = (0,\infty)$ we have in short

$$\begin{aligned}
N(0,\infty,a,n) &= N^+(X_n(a)),\\
N(0,\infty,(a,b),n) &= N^+(X_n^+(b)X_n(a)),\\
N(-\infty,0,a,n) &= N^-(X_n(a)),\\
N(-\infty,0,(a,b),n) &= N^-(X_n^+(b)X_n(a)).
\end{aligned}$$

Finally, by $\alpha_-(a)$, $\alpha_+(a)$, and $\alpha_0(a)$ we denote the Lebesgue measure meas$\{\cdot\}$ *of the sets $A_-(a)$, $A_+(a)$, and $A_0(a)$, respectively.*

THEOREM 3.4. *Let a be a continuous function on K. Under the preceding notations of Definition* 3.5*, the relationships*

$$\begin{aligned}
N^+(X_n(a)) &= n\tfrac{\alpha_+(a)}{\mathrm{meas}(K)} + O(n^{\frac{d-1}{d}}),\\
N^-(X_n(a)) &= n\tfrac{\alpha_-(a)}{\mathrm{meas}(K)} + O(n^{\frac{d-1}{d}}),\\
\dim(\mathrm{Ker}(X_n(a))) &= n\tfrac{\alpha_0(a)}{\mathrm{meas}(K)} + O(n^{\frac{d-1}{d}})
\end{aligned}$$

hold true when the following conditions are satisfied:

hp:a1: $A_0(a)$ is made up of a finite collection of connected compact sets with Lipschitz continuous boundaries;

hp:a2: the sequence $\{X_n(\cdot)\}_n$ is locally X in the strong sense.

Proof. Let $I_n^+(a) = \{i :\ a(x_i^{(n)}) > 0\}$, $I_n^-(a) = \{i :\ a(x_i^{(n)}) < 0\}$, and $I_n^0(a) = \{i : a(x_i^{(n)}) = 0\}$, where $\mathcal{G}_n = \{x_i^{(n)}\}$ is the nth grid of $\mathcal{G}$ where we sample the functions of $C(K)$.

By the assumption of continuity of the coefficient function $a(x)$ and since $\{\hat{\mathcal{G}}_n\}_n$ is $*$-uniformly distributed over K in the sense of Definition 2.1, it follows that

$$\begin{aligned} \#I_n^+(a) &= n\tfrac{\alpha_+(a)}{\mathrm{meas}(K)} + O(n^{\frac{d-1}{d}}), \\ \#I_n^-(a) &= n\tfrac{\alpha_-(a)}{\mathrm{meas}(K)} + O(n^{\frac{d-1}{d}}), \\ \#I_n^0(a) &= n\tfrac{\alpha_0(a)}{\mathrm{meas}(K)} + O(n^{\frac{d-1}{d}}). \end{aligned} \tag{12}$$

Now, by taking into account the dyadic decomposition of the matrix $X_n(a)$ with respect to the set $\{\mathbf{v}_j^{(n)} \in \mathbf{C}^n :\ 1 \le j \le n+k\}$, it is evident that each dyad $\mathbf{v}_j^{(n)}[\mathbf{v}_j^{(n)}]^H$ is made up of a weighted sum of at most $O(1)$ terms of the form $\mathbf{z}_s^{(n)} \cdot [\mathbf{z}_t^{(n)}]^H$, where $\mathbf{z}_k^{(n)}$ is the kth element of the basis Z_n of $\mathbf{C}^n$ and $\|x_s^{(n)} - x_t^{(n)}\|_\infty = O(n^{-1/d})$; more precisely,

$$\mathbf{v}_j^{(n)}[\mathbf{v}_j^{(n)}]^H = \sum\Big\{s,t : \|x_j^{(n)} - x_s^{(n)}\|_\infty + \|x_j^{(n)} - x_t^{(n)}\|_\infty = O(n^{-1/d})\Big\} \alpha_s\alpha_t \mathbf{z}_s^{(n)} \cdot [\mathbf{z}_t^{(n)}]^H,$$

with the convention that the terms corresponding to nonexistent vectors $\{\mathbf{z}_k^{(n)}\}$ are discarded. As a consequence we will partition the space $\mathbf{C}^n$ as a direct sum of orthogonal subspaces.

First define the set of indices

$$\begin{aligned} \hat{I}_n^+(a) &= \{j \in \{1,\dots,n\} :\ \forall i \in I_n^-(a) \cup I_n^0(a),\ [\mathbf{v}_i^{(n)}]^H \mathbf{z}_j^{(n)} = 0\}, \\ \hat{I}_n^-(a) &= \{j \in \{1,\dots,n\} :\ \forall i \in I_n^+(a) \cup I_n^0(a),\ [\mathbf{v}_i^{(n)}]^H \mathbf{z}_j^{(n)} = 0\}, \\ \hat{I}_n^0(a) &= \{j \in \{1,\dots,n\} :\ \forall i \in I_n^-(a) \cup I_n^+(a),\ [\mathbf{v}_i^{(n)}]^H \mathbf{z}_j^{(n)} = 0\}, \\ W(a) &= \{j : j \in \{1,\dots,n\} \backslash (\hat{I}_n^+(a) \cup \hat{I}_n^-(a) \cup \hat{I}_n^0(a))\}, \end{aligned}$$

where the definition of the set $W(a)$ is necessary to give evidence of vectors of $\{\mathbf{z}_k^{(n)}\}$ belonging to those "adjacent" dyads $\mathbf{v}_i^{(n)}[\mathbf{v}_i^{(n)}]^H$ at which a change in the sign of the coefficient function $a(x)$ occurs.

It is easy to check that

$$\begin{aligned} \hat{I}_n^+(a) &\subset I_n^+(a), \\ \hat{I}_n^-(a) &\subset I_n^-(a), \\ \hat{I}_n^0(a) &\subset I_n^0(a), \\ \hat{I}_n^0(a) &\cap \hat{I}_n^+(a) = \emptyset, \\ \hat{I}_n^-(a) &\cap \hat{I}_n^0(a) = \emptyset, \\ \hat{I}_n^-(a) &\cap \hat{I}_n^+(a) = \emptyset, \end{aligned}$$

and that

$$\begin{aligned} &\#W(a) = O(n^{\frac{d-1}{d}}), \\ &\#(\hat{I}_n^+(a)) = [\#(I_n^+(a)) - O(n^{\frac{d-1}{d}})]^+, \\ &\#(\hat{I}_n^-(a)) = [\#(I_n^-(a)) - O(n^{\frac{d-1}{d}})]^+, \\ &\#(\hat{I}_n^0(a)) = [\#(I_n^0(a)) - O(n^{\frac{d-1}{d}})]^+, \end{aligned}$$

where the symbol x^+ denotes $\max\{x, 0\}$ and where the relationships $\#(\hat{I}_n^*(a)) = [\#(I_n^*(a)) - O(n^{\frac{d-1}{d}})]^+$, $* \in \{+, -, 0\}$ follow from the assumption that a vanishes at most at a finite number of connected sets with Lipschitz continuous boundaries.

By defining

$$\begin{aligned}
\mathcal{S}^+(a) &= \text{span}\{\mathbf{z}_j^{(n)} : \ j \in \hat{I}_n^+(a)\},\\
\mathcal{S}^-(a) &= \text{span}\{\mathbf{z}_j^{(n)} : \ j \in \hat{I}_n^-(a)\},\\
\mathcal{S}^0(a) &= \text{span}\{\mathbf{z}_j^{(n)} : \ j \in \hat{I}_n^0(a)\},\\
\mathcal{W}(a) &= \text{span}\{\mathbf{z}_j^{(n)} : \ j \in W(a)\},
\end{aligned}$$

we have that $\mathbf{C}^n$ is the direct sum of the four orthogonal subspaces $\mathcal{S}^+(a)$, $\mathcal{S}^-(a)$, $\mathcal{S}^0(a)$, and $\mathcal{W}(a)$.

It is crucial to observe that for any choice of the vectors $\mathbf{x} \in \mathcal{S}^+(a)$, $\mathbf{y} \in \mathcal{S}^-(a)$, and $\mathbf{z} \in \mathcal{S}^0(a)$ with $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = \|\mathbf{z}\|_2 = 1$, we find

$$\begin{aligned}
&\mathbf{x}^H X_n(a)\mathbf{x} > 0, \quad X_n(a)\mathbf{x} \in \mathcal{S}^+(a) \oplus \mathcal{W}(a),\\
&\mathbf{y}^H X_n(a)\mathbf{y} < 0, \quad X_n(a)\mathbf{y} \in \mathcal{S}^-(a) \oplus \mathcal{W}(a),\\
&\quad X_n(a)\mathbf{z} = 0.
\end{aligned}$$

In fact, since $\mathbf{x} \in \mathcal{S}^+(a)$, by the dyadic decomposition of $X_n(a)$, we have

$$\mathbf{x}^H X_n(a)\mathbf{x} = \sum_{i \in I_n^+(a)} a(x_i^{(n)}) \mathbf{x}^H \mathbf{v}_i^{(n)} [\mathbf{v}_i^{(n)}]^H \mathbf{x}. \tag{13}$$

Now let us consider a strictly positive continuous function $\hat{a}$ which coincides with a over $A_+(a) \cap \mathcal{G}_n$. By definition of $\mathcal{S}^+(a)$ and $I_n^+(a)$ and from the relation (13), we can check that

$$\mathbf{x}^H X_n(a)\mathbf{x} = \mathbf{x}^H X_n(\hat{a})\mathbf{x},$$

where

$$\mathbf{x}^H X_n(\hat{a})\mathbf{x} > 0$$

since, from Proposition 2.1, we know that the LPO $\tilde{X}_n(\cdot)$ is also *normally positive*. The other two relationships are obtained in a similar manner.

Now, by the relation $X_n(a)\mathbf{z} = 0$ and by virtue of the third relation displayed in (12), we find that

$$\dim(\text{Ker}(X_n(a))) \geq \#(\hat{I}_n^0(a)) = n\frac{\alpha_0(a)}{\text{meas}(K)} + O(n^{\frac{d-1}{d}}).$$

At this point call $\mathbf{v}_1, \ldots, \mathbf{v}_{N^+}$, $\mathbf{w}_1, \ldots, \mathbf{w}_{N^-}$, and $\mathbf{s}_1, \ldots, \mathbf{s}_{N^0}$ the eigenvectors of $X_n(a)$, respectively, related to its positive, negative, and null eigenvalues. Call $\mathcal{Z}^+(a)$, $\mathcal{Z}^-(a)$, and $\mathcal{Z}^0(a)$ the subspaces of $\mathbf{C}^n$ spanned by these three sets of vectors. Here, in short, the symbols N^+, N^-, and N^0 denote $N^+(X_n(a))$, $N^-(X_n(a))$, and $\dim(\text{Ker}(X_n(a)))$, respectively.

In light of their definitions the set-theoretic relationships

$$\begin{aligned}
&\mathcal{S}^+(a) \cap (\mathcal{Z}^-(a) \oplus \mathcal{Z}^0(a)) = \{\mathbf{0}\},\\
&\mathcal{S}^-(a) \cap (\mathcal{Z}^+(a) \oplus \mathcal{Z}^0(a)) = \{\mathbf{0}\},\\
&(\mathcal{S}^+(a) \oplus \mathcal{S}^0(a)) \cap \mathcal{Z}^-(a) = \{\mathbf{0}\},\\
&(\mathcal{S}^-(a) \oplus \mathcal{S}^0(a)) \cap \mathcal{Z}^+(a) = \{\mathbf{0}\}
\end{aligned}$$

hold true. Moreover, by exploiting the well-known inequalities

$$[\dim(\mathcal{X})+\dim(\mathcal{Y})-n]^+ \le \dim(\mathcal{X}\cap\mathcal{Y}) \le \min\{\dim(\mathcal{X}),\dim(\mathcal{Y})\},$$

we deduce that

$$\begin{aligned}
0=\dim(\mathcal{S}^+(a)\cap(\mathcal{Z}^-(a)\oplus\mathcal{Z}^0(a))) &\ge [\#(\hat{I}_n^+(a))+(n-N^+)-n]^+\\
&= [\#(\hat{I}_n^+(a))-N^+]^+,\\
0=\dim(\mathcal{S}^-(a)\cap(\mathcal{Z}^+(a)\oplus\mathcal{Z}^0(a))) &\ge [\#(\hat{I}_n^-(a))+(n-N^-)-n]^+\\
&= [\#(\hat{I}_n^-(a))-N^-]^+,\\
0=\dim((\mathcal{S}^+(a)\oplus\mathcal{S}^0(a))\cap\mathcal{Z}^-(a)) &\ge [n-[\#(\hat{I}_n^-(a))+\dim(\mathcal{W}(a))]+N^--n]^+\\
&= [N^--[\#(\hat{I}_n^-(a))+\dim(\mathcal{W}(a))]]^+,\\
0=\dim((\mathcal{S}^-(a)\oplus\mathcal{S}^0(a))\cap\mathcal{Z}^+(a)) &\ge [n-[\#(\hat{I}_n^+(a))+\dim(\mathcal{W}(a))]+N^+-n]^+\\
&= [N^+-[\#(\hat{I}_n^+(a))+\dim(\mathcal{W}(a))]]^+
\end{aligned}$$

and finally

$$\begin{array}{lllll}
\#(\hat{I}_n^+(a)) & \le & N^+ & \le & \#(\hat{I}_n^+(a))+\dim(\mathcal{W}(a)),\\
\#(\hat{I}_n^-(a)) & \le & N^- & \le & \#(\hat{I}_n^-(a))+\dim(\mathcal{W}(a)),\\
\#(\hat{I}_n^0(a)) & \le & \dim(\mathrm{Ker}(X_n(a)) & \le & \#(\hat{I}_n^0(a))+\dim(\mathcal{W}(a)).
\end{array}$$

By recalling that $\dim(\mathcal{W}(a))=O(n^{\frac{d-1}{d}})$ and the relations linking the sizes of the sets $\hat{I}_n^*(a)$ and $I_n^*(a)$ with $*\in\{+,-,0\}$, the claimed thesis directly follows. □

The assumption [*hp*:*a*1] in the preceding result is necessary and cannot be removed, as shown in the following example.

EXAMPLE 3.1. Let $d=1$ and let $\Omega=(0,1)$. Let us consider the grid sequence $\mathcal{G}=\{G_n\}_n$ with

$$G_n=\{x_i^{(n)}=i/n, i=1,\dots,n\}\subset K.$$

It is trivial to check that $\mathcal{G}$ is $*$-uniform with respect to $K=[0,1]$.

Suppose that [*hp*:*a*1] is violated, that is, that $A_0(a)$ is made up of an infinite sequence of disjoint connected compact sets with Lipschitz boundaries. In particular take $a(x)=\min_{y\in Q}|y-x|=\mathrm{dist}(x,Q)$ so that

$$Q\equiv A_0(a)=\bigcup_{j=0}^{\infty}\left[3^{-j-1},3^{-j}/2\right]\cup\{0\}.$$

It is a direct computation to verify that

(1) $\mathrm{meas}(A_0(a))=1/4$,

(2) $\#\{i: x_i^{(n)}\in A_0(a)\}=n/4+r_n$ and $r_n\ne O(1)$.

Let us prove (2) (part (1) is trivial). Since the basic intervals $\left[3^{-j-1},3^{-j}/2\right]$ are pairwise disjoint, it follows that

$$\#\{i: x_i^{(n)}\in A_0(a)\}=\sum_{j=1}^{\infty}\#\{i: x_i^{(n)}\in\left[3^{-j-1},3^{-j}/2\right]\}.$$

Now take $n=3^p$ and observe that $3^{-j}/2<1/n$ for $j>p-1$. Therefore $\#\{i: x_i^{(n)}\in\left[3^{-j-1},3^{-j}/2\right]\}=0$ for any $j>p-1$ while

$$\begin{aligned}
\#\{i: x_i^{(n)}\in\left[3^{-j-1},3^{-j}/2\right]\} &= \#\{i: 3^{p-j-1}\le i\le 3^{p-j}/2\}\\
&= 3^{p-j}/2-3^{p-j-1}+1/2 \qquad \text{for } \ j\le p-1.
\end{aligned}$$

Consequently we obtain that

$$\begin{aligned}\#\{i : x_i^{(n)} \in A_0(a)\} &= \sum_{j=1}^{p-1} 3^{p-j}/2 - 3^{p-j-1} + 1/2 \\ &= p/2 + 3^p \sum_{j=1}^{p-1} 3^{-j}(1/2 - 1/3) \\ &= p/2 + 3^p/6 \sum_{j=1}^{p-1} 3^{-j} \\ &= p/2 + n/4(1 - n^{-1}) \\ &= n/4 + p/2 - 1/4,\end{aligned}$$

and finally $r_n = p/2 - 1/4$ is asymptotic to $\log(n)$.

Now define $X_n(a) = \sum_{j=1}^n a(x_i^{(n)})\mathbf{e}_j^{(n)}[\mathbf{e}_j^{(n)}]^T$. It is evident that $\{X_n(\cdot)\}_n$ is locally X in the strong sense with $X = \{X_n = I_n\}_n$ and

$$\dim(\mathrm{Ker}(X_n(a))) = \#\{i : x_i^{(n)} \in A_0(a)\} = n\frac{\alpha_0(a)}{\mathrm{meas}(K)} + r_n,$$

so that the third relation of Theorem 3.4 is violated since $d = 1$ and $r_n = p/2 - 1/4 \sim \log(n) \neq O(n^{\frac{d-1}{d}}) \equiv O(1)$. □

THEOREM 3.5. *Let a be a continuous function on K. Under the preceding notation of Definition* 3.5 *the relationships*

$$\begin{array}{lcl} N^+(X_n(a)) & = & n\frac{\alpha_+(a)}{\mathrm{meas}(K)} + o(n), \\ N^-(X_n(a)) & = & n\frac{\alpha_-(a)}{\mathrm{meas}(K)} + o(n), \\ \dim(\mathrm{Ker}(X_n(a))) & = & n\frac{\alpha_0(a)}{\mathrm{meas}(K)} + o(n) \end{array}$$

hold true when the following conditions are satisfied:

hp:b1: $A_0(a)$ is measurable in the sense of Peano and Jordan;

hp:b2: the sequence $\{X_n(\cdot)\}_n$ is locally X.

Proof. It is enough to follow the same steps as in Theorem 3.4 by taking into account the quasi uniformity of $\{\mathcal{G}_n\}_n$ and Theorem 2.1. □

The assumption [*hp*:*b*1] in the preceding result is necessary and cannot be removed as shown in the following example.

EXAMPLE 3.2. Let $d = 1$ and let $\Omega = (0, 1)$. Let us consider the grid sequence $\mathcal{G} = \{G_n\}_n$ as in the preceding example.

Suppose that [*hp*:*b*1] is violated, i.e., that $A_0(a)$ is not measurable with respect to the Peano–Jordan procedure.

We construct a compact set Q in the following way. Let $\epsilon < 10^{-1}$ and let $\{q_i\}_{i\in\mathbf{N}}$ be the sequence of all the rational numbers belonging to the real interval $[0, 1]$ with $q_0 = 1/2$. Let $S = \cup_{i\in\mathbf{N}} B^\circ(q_i, \epsilon 2^{-1})$ and $Q = [0, 1]\backslash S$. The following facts hold:

(1) Q is compact (since S is open),

(2) S is dense in $[0, 1]$ and Q is totally unconnected,

(3) the Lebesgue measure meas(Q) of Q is t and $1 - 2/5 \leq t \leq 1 - 1/5$.

In particular take the continuous $a(x) = \min_{y\in Q} |y - x| = \mathrm{dist}(x, Q)$ so that $A_0(a) = Q$ and $\mathrm{meas}(A_0(a)) = t \geq 3/5$.

However, since each $x_i^{(n)}$ is rational it follows that

$$\#\{i : x_i^{(n)} \in A_0(a)\} = 0.$$

Therefore $X_n(a)$ is positive definite and

$$\dim(\mathrm{Ker}(X_n(a))) = 0 \neq n\frac{\alpha_0(a)}{\mathrm{meas}(K)} + o(n) \geq 3n/5 + o(n). \qquad \square$$

Now, we are ready to state the first Szegő-style theorem concerning the asymptotic distribution of the eigenvalues of the preconditioned matrices $\{\mathcal{P}_n(a;b) = X_n^+(b)X_n(a)\}_n$ whose proof follows the same idea as in Theorem 4.6 in [36].

THEOREM 3.6. *Let a and b be two equivalent functions belonging to $C(K)$. Let (s,∞) be the generic half-line and let us denote by F_s the characteristic function of (s,∞). Suppose that $A_0(a-sb)$ and $A_0(a)$ are measurable in the Peano–Jordan sense and that [hp:b2] is satisfied. Then, by calling $\{\lambda_i^{(n)}\}$ the eigenvalues of the matrices $\{\mathcal{P}_n(a;b)\}_n$, we find that*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} F_s\left(\lambda_i^{(n)}\right) = \int_{A_+(b)} F_s(a/b)dx + \mathrm{meas}\{A_0(b)\}F_s(0). \tag{14}$$

If in addition $A_0(a-sb)$ and $A_0(a)$ fulfill [hp:a1] and [hp:a2] holds true, then we have

$$\frac{1}{n}\sum_{i=1}^{n} F_s\left(\lambda_i^{(n)}\right) - \int_{A_+(b)} F_s(a/b)dx - \mathrm{meas}\{A_0(b)\}F_s(0) = O(n^{\frac{d-1}{d}}). \tag{15}$$

Proof. Let us consider the Hermitian matrices

$$\Psi(s) = X_n(a) - sX_n(b)$$

and

$$\Phi_n(a;b) = (X_n^+(b))^{1/2}X_n(a)(X_n^+(b))^{1/2}.$$

Then by following the same argument as in Theorem 4.6 in [36] we know that $P_n(a;b)$ and $\Phi_n(a;b)$ have the same spectra and there exist a Hermitian matrix Θ_s, a positive definite matrix L, and a unitary matrix U such that

$$\Phi_n(a;b) - sI = U \begin{bmatrix} -s & & & \\ & \ddots & & 0 \\ & & -s & \\ & 0 & & \Theta_s \end{bmatrix} U^H$$

and

$$\Psi(s) = U \begin{bmatrix} 0 & & & \\ & \ddots & & 0 \\ & & 0 & \\ & 0 & & L\Theta_s L \end{bmatrix} U^H.$$

Therefore in the subspace orthogonal to the kernel of $A_n(b)$ ($\mathcal{Z}^0(b)$), the matrices $\Psi(s)$ and $\Phi_n(a,b) - sI$ have the same inertia, and as a consequence we have

$$\begin{aligned} N(s,\infty,(a,b),n) &= \#\{i:\ \lambda_i(\mathcal{P}_n(a,b)) > s\} \\ &= \#\{i:\ \lambda_i(\Phi_n(a,b) - sI) > 0\} \\ &= \#\{i:\ \lambda_i(\Phi_n(a,b)) = 0\}F_s(0) + \#\{i:\ \lambda_i(\Theta_s) > 0\} \\ &= \dim(\ker(\mathcal{P}_n(a;b)))F_s(0) + \#\{i:\ \lambda_i(L\Theta_s L) > 0\}. \end{aligned}$$

Now, through Theorems 3.4 and 3.5, and by comparison with the splitting of $\Psi(s)$, we have that

$$\begin{aligned} N(s,\infty,(a,b),n) &= (n\cdot \mathrm{meas}\{A_0(b)\} + O(r_n))F_s(0) + \#\{i:\ \lambda_i(L\Theta_s L) > 0\} \\ &= (n\cdot \mathrm{meas}\{A_0(b)\} + O(r_n))F_s(0) + \#\{i:\ \lambda_i(\Psi(s)) > 0\}, \end{aligned}$$

where, again in light of Theorems 3.4 and 3.5,

$$\begin{aligned} \#\{i:\ \lambda_i(\Psi(s)) > 0\} &= N(0,\infty,a-sb,n) \\ &= n\cdot \mathrm{meas}\{x\in K:\ a - sb > 0\} + O(r_n) \\ &= n\cdot \mathrm{meas}\{x\in A_+(b):\ a/b > s\} + O(r_n) \\ &= n\int_{A_+(b)} F_s(a/b) + O(r_n). \end{aligned}$$

This ends the proof if we recall that $r_n = O(n^{\frac{d-1}{d}})$ if [*hp*:a1] is fulfilled for $a - sb$ and the sequence $\{X_n(\cdot)\}_n$ is locally X in the strong sense or $o(n)$ if [*hp*:b1] is fulfilled for $a - sb$ and the sequence $\{X_n(\cdot)\}_n$ is locally X in the weak sense. ∎

The following is the main Szegő-type result.

THEOREM 3.7. *Let a and b be two equivalent functions. Suppose that $A_0(a - sb)$ is Peano–Jordan measurable for $s \in T$ with T dense in $\mathbf{R}$ and $0 \in T$. Suppose that* [*hp*:b2] *holds true. Then the eigenvalues $\{\lambda_i^{(n)}\}$ of the matrices $\{\mathcal{P}_n(a;b)\}_n$ verify the ergodic formula* (16); *i.e., for any continuous function F with bounded support, it follows that*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} F\left(\lambda_i^{(n)}\right) = \int_{A_+(b)} F(a/b)dx + \mathrm{meas}\{A_0(b)\}F(0). \tag{16}$$

Proof. Call S the space of the linear combinations of functions of the form F_s with $s \in T$. Owing to the hypothesis $\bar{T} = \mathbf{R}$, we deduce that the topological closure with respect to the infinity norm of the space S contains all the continuous functions vanishing at infinity. Now the general statement given in (16) is a consequence of Theorem 3.6 and of the linearity of both the sides of (16) with respect to F. ∎

Therefore we write that $\{\mathcal{P}_n(a;b)\}_n$ is distributed as $\eta = a/b$, where b is positive and zero otherwise.

As a direct consequence of Theorem 3.7 and of Theorem 3.1, we find the following.

COROLLARY 3.1. *Let a and b be two equivalent functions belonging to $C(K)$. Suppose that the assumptions of Theorem* 3.7 *are satisfied. Then the topological closure Λ of the set of the eigenvalues of all the matrices $\{\mathcal{P}_n(a;b)\}_n$ is contained in $[r,R]\bigcup\{0\}$ and the set Λ contains the usual range of the function a/b. Moreover, the eigenvectors related to the null eigenvalue belong, for any dimension, to the null space of $X_n(a)$.*

When the two functions are not equivalent (with nonnegative b) then the matrices $\{X_n(b)\}_n$ are not good preconditioners for $\{X_n(a)\}_n$; indeed for any positive ϵ there

exists $\phi_\epsilon > 0$ such that for n large enough at least $\phi_\epsilon n$ eigenvalues of the preconditioned matrices not related to the null space of $X_n(a)$ have modulus belonging either to $[0, \epsilon)$ or to (ϵ^{-1}, ∞).

Therefore, if the equivalence condition does not hold, then the n-sized preconditioned matrix is characterized by a linear quantity of very small and/or very big eigenvalues and, in the light of the Axelsson–Lindskog analysis [2], this fact leads to a substantial deterioration of the performances of the related PCG method.

Some further remarks are useful.

Remark 3.4. When $d = 1$ and $a - sb$ has only a finite number of isolated zeros and a, b, and F are Lipschitz continuous, then the quoted result tells us that the error

$$\left| \frac{1}{n} \sum_{i=1}^{n} F\left(\lambda_i^{(n)}\right) - \int_K F(\eta) dx \right|$$

is bounded by c/n with c an absolute constant and $\eta(x) = 0$ if $b(x) = 0$ and $a(x)/b(x)$ otherwise. Therefore the given statement is also in the style of the second order results due to Widom [47] in the Toeplitz context.

Remark 3.5. The assumption $k(n) = O(n^{\frac{d-1}{d}})$ is completely natural for a "good" grid sequence $\{\mathcal{G}_n\}_n$. Indeed, for a domain Ω good enough (e.g., connected with Lipschitz continuous boundaries), we count the number of points of $\hat{\mathcal{G}}_n$ whose distance from $\partial\Omega$ is less than $K \cdot n^{-1/d}$ with K an absolute constant. Since $\Omega \subset \mathbf{R}^d$, the set $\partial\Omega$ is a $(d-1)$-dimensional manifold and therefore if any point $x_i^{(n)}$ has $O(1)$ associated extra points, we have to expect that

$$\#\{\mathcal{G}_n \backslash \hat{\mathcal{G}}_n\} \equiv k(n) \leq O(1) \cdot O(n^{\frac{d-1}{d}}) = O(n^{\frac{d-1}{d}}).$$

Remark 3.6. If $\{\mathcal{G}_n\}_n$ is uniformly distributed over K with regard to another measure μ (different from the Lebesgue one), then the distributional results will hold with all the statements involving the μ-measure instead of the Lebesgue measure (compare [37] and [32]). From a technical point of view, the proofs will only change when dealing with measures and integrals. In particular a delicate change will concern the conditions [*hp*:*a*1] and [*hp*:*b*1].

3.3. Further ergodic results. Concerning the main ergodic result, i.e., Theorem 3.7, it should be observed that the assumption that "$A_0(a - sb)$ is Peano–Jordan measurable" is generally fulfilled in the applications for any s. On the other hand we have constructed examples ad hoc in order to violate this assumption with $s = 0$. Therefore the result is not theoretically satisfactory, so that in this short subsection we discuss this point.

The following facts will be used:

(1) For any number s, and for any pair of polynomials p and q the set $A_0(p - sq)$ is Peano–Jordan measurable.
(2) Any continuous function over a compact set $K \subset \mathbf{R}^d$ can be approximated by polynomials: use the theorem of Weierstrass over an hypercube containing K and the Tietze extension theorem (see Theorem 3.2 in [10, pp. 10–11] and [27, pp. 422–423]).

Let a and b be two equivalent functions of $C(K)$. By item (2), for any positive ϵ there exist polynomials $p_{a,\epsilon}$ and $p_{b,\epsilon}$ such that

$$\|\sqrt{a} - p_{a,\epsilon}\|_\infty < \epsilon, \quad \|\sqrt{b} - p_{b,\epsilon}\|_\infty < \epsilon.$$

It follows that

$$\|a - p_{a,\epsilon}^2\|_\infty < \epsilon(2\|a\|_\infty + \epsilon), \quad \|b - p_{b,\epsilon}^2\|_\infty < \epsilon(2\|b\|_\infty + \epsilon).$$

Set $q_{a,\epsilon} = p_{a,\epsilon}^2$ and $q_{b,\epsilon} = p_{b,\epsilon}^2$ and assume that a is strictly positive over K. Then we have

$$\begin{aligned} X_n^{-1}(b)X_n(a) &\sim X_n^{-1/2}(b)X_n(a)X_n^{-1/2}(b) \\ &= X_n^{-1/2}(b)X_n(q_{a,\epsilon})X_n^{-1/2}(b) \\ &\quad + X_n^{-1/2}(b)X_n(a - q_{a,\epsilon})X_n^{-1/2}(b), \end{aligned}$$

where $\sim$ denotes the similarity relation. Since $X_n^{-1/2}(b)X_n(a-q_{a,\epsilon})X_n^{-1/2}(b)$ is similar to $X_n^{-1}(b)X_n(a-q_{a,\epsilon})$, by Theorem 3.2 it follows that $X_n^{-1/2}(b)X_n(a-q_{a,\epsilon})X_n^{-1/2}(b)$ has spectral norm bounded by $q_1\epsilon$ with q_1 absolute constant. Now take $T_{\epsilon,n,1} = X_n^{-1/2}(b)X_n(q_{a,\epsilon})X_n^{-1/2}(b)$ and consider

$$T_{\epsilon,n,1}^{-1} \sim X_n(b)X_n^{-1}(q_{a,\epsilon}) \sim X_n^{-1/2}(q_{a,\epsilon})X_n(b)X_n^{-1/2}(q_{a,\epsilon}).$$

Define $T_{\epsilon,n,2}^{-1} = X_n^{-1/2}(q_{a,\epsilon})X_n(q_{b,\epsilon})X_n^{-1/2}(q_{a,\epsilon})$ and set

$$T_{\epsilon,n,1}^{-1} = T_{\epsilon,n,2}^{-1} + X_n^{-1/2}(q_{a,\epsilon})X_n(b - q_{b,\epsilon})X_n^{-1/2}(q_{a,\epsilon}).$$

Therefore by using Theorem 3.1 we obtain that $X_n^{-1/2}(q_{a,\epsilon})X_n(b - q_{b,\epsilon})X_n^{-1/2}(q_{a,\epsilon})$ has spectral norm bounded by $q_2\epsilon$ and by using Theorem 3.7 we infer that $\{T_{\epsilon,n,2}^{-1} \sim X_n^{-1}(q_{a,\epsilon})X_n(q_{b,\epsilon})\}_n$ is distributed as $\frac{q_{b,\epsilon}}{q_{a,\epsilon}}$ and $\{T_{\epsilon,n,2}\}_n$ distributes as $\frac{q_{a,\epsilon}}{q_{b,\epsilon}}$. Consequently it holds that

$$T_{\epsilon,n,1} = T_{\epsilon,n,2} - T_{\epsilon,n,1}X_n^{-1/2}(q_{a,\epsilon})X_n(b - q_{b,\epsilon})X_n^{-1/2}(q_{a,\epsilon})T_{\epsilon,n,2}$$

with $T_{\epsilon,n,1}X_n^{-1/2}(q_{a,\epsilon})X_n(b-q_{b,\epsilon})X_n^{-1/2}(q_{a,\epsilon})T_{\epsilon,n,2}$ having spectral norm bounded by $q_3\epsilon$ with q_3 absolute constant. Finally, recalling that

$$\begin{aligned} X_n^{-1}(b)X_n(a) &\sim T_{\epsilon,n,1} + O(\epsilon)\text{−norm} \\ &= T_{\epsilon,n,2} + O(\epsilon)\text{−norm} \end{aligned}$$

and that $\{T_{\epsilon,n,2}\}_n$ distributes as $\frac{q_{a,\epsilon}}{q_{b,\epsilon}}$, by the arbitrariness of ϵ, we conclude that $\{X_n^{-1}(b)X_n(a)\}_n$ is distributed as a/b; i.e., for any continuous function F with bounded support, it follows that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} F\left(\lambda_i^{(n)}\right) = \int_{A_+(b)} F(a/b)dx + \text{meas}\{A_0(b)\}F(0) \tag{17}$$

with $A_0(b) = \emptyset$. With the same trick we can treat the case where a has isolated zeros or the more general case where $A_0(a)$ is Peano–Jordan.

4. The analysis of the preconditioners $\{\mathcal{P}_n^*(a)\}_n$. Let us introduce the following technical tools.

LEMMA 4.1 (see [35]). *Suppose that $\{A_n\}_n$ and $\{B_n\}_n$ are two sequences of $n \times n$ matrices and $\{B_n\}_n$ are nonsingular for any n large enough, and let us denote by $\{\sigma_i^{(n)}\}$ the singular values of B_n. Let us assume that $\{B_n\}_n$ is "sparsely vanishing"*

[39]; *i.e., we assume that there exists a nonnegative function* $x(s)$ *independent of* n *so that,* $\forall \varepsilon > 0$ *and for* n *large enough,*

$$\frac{1}{n}\#\{i : \sigma_i^{(n)} \le \varepsilon\} \le x(\varepsilon), \qquad \lim_{s\to 0} x(s) = 0,$$

and that there exist matrices $\{D_n = D_n(\varepsilon)\}_n$ *such that*

$$\lim_{n\to\infty} \|A_n - B_n - D_n\|_2 = 0$$

with $\operatorname{rank}(D_n) \le \varepsilon n$. *Then the matrices* $\{B_n^{-1}A_n\}_n$ *are spectrally clustered at* 1 *(weak clustering property).*

LEMMA 4.2. *Under the assumptions as in Lemma* 4.1, *if* $\operatorname{rank}(D_n) = O(1)$ *and* $\|B_n^{-1}\|_2\|A_n - B_n - D_n\|_2 = o(1)$, *then the matrices* $\{B_n^{-1}A_n\}_n$ *are spectrally properly clustered at* 1 *(strong clustering property).*

Proof. It is enough to make a slight modification of the proof of the preceding lemma for which we refer the reader to [35] (see also [31]). □

THEOREM 4.1. *Let* a *be a nonnegative function with* $A_0(a)$ *fulfilling* [*hp:b*1], *let* [*hp:b*2] *hold, and let* $\{X_n(1)\}_n$ *be sparsely vanishing. Let us suppose that*

$$\sum_{j=1}^{n+k(n)} |(\mathbf{s}_j^{(n)})_i|^2 = c^*, \quad (\mathbf{s}_j^{(n)} = Z_n^H \mathbf{v}_j^{(n)}),\ i = 1, \dots, n, \tag{18}$$

holds with c^* *independent of* i. *Then the eigenvalues* $\{\lambda_i^{(n)}\}$ *of the matrices* $\{c^* \cdot \mathcal{P}_n^*(a)\}_n$ *are such that, for any continuous function* F *with bounded support, it holds that*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} F\left(\lambda_i^{(n)}\right) = \int_{A_+(a)} F(1)dx + \operatorname{meas}\{A_0(a)\}F(0). \tag{19}$$

If (18) *is not fulfilled, then we set* $\mathcal{P}_n^{**}(a) = [D_n^*(a)X_n(1)D_n^*(a)]^{-1}X_n(a)$ *with* $D_n^*(a) = Z_n D_n^{1/2}(a) D_p^{-1/2} Z_n^H$, *where* D_p *is the diagonal matrix whose element* (i,i) *is the value*

$$\sum_{j=1}^{n+k(n)} |(\mathbf{s}_j^{(n)})_i|^2.$$

In this case a formula like (19) *holds for the sequence* $\{\mathcal{P}_n^{**}(a)\}_n$. *Moreover, if the function* a *is strictly positive or has isolated zeros, then the sequence* $\{\mathcal{P}_n^{**}(a)\}_n$ *is spectrally clustered at one.*

Proof. We give the proof under the assumption that a is strictly positive; the case where a is nonnegative with $A_0(a)$ being Peano–Jordan can be simply handled by observing that

$$\#\{i :\ (D_n(a))_{i,i} = 0\} \sim n \cdot \frac{\operatorname{meas}\{A_0(a)\}}{\operatorname{meas}\{K\}}.$$

Let $Y_n(a) = Z_n^H X_n(a) Z_n$. From the locality assumptions we have

$$\begin{aligned} Y_n(a) &= \sum_{j=1}^{n+k(n)} a_j Z_n^H \mathbf{v}_j^{(n)} [\mathbf{v}_j^{(n)}]^H Z_n \\ &= \sum_{j=1}^{n+k(n)} a_j \mathbf{s}_j^{(n)} [\mathbf{s}_j^{(n)}]^H, \end{aligned}$$

where $a_j = a(x_j^{(n)})$ and $\mathbf{s}_j^{(n)}$ is a linear combination of $O(1)$ vectors of the canonical basis of $\mathbf{C}^n$ corresponding to grid points close to $x_j^{(n)}$.

Now let us consider $\mathcal{P}_n^*(a) = [\hat{D}_n(a)X_n(1)\hat{D}_n(a)]^{-1}X_n(a)$. By the definition of $\hat{D}_n(a)$ it is clear that $\mathcal{P}_n^*(a)$ is similar to $[D_n^{1/2}(a)Y_n(1)D_n^{1/2}(a)]^{-1}Y_n(a)$, and the latter is similar to

$$[Y_n(1)]^{-1}D_n^{-1/2}(a)Y_n(a)D_n^{-1/2}(a).$$

Call $V_n = D_n^{-1/2}(a)Y_n(a)$. It is evident that $(V_n)_{s,t}$ is nonzero if and only if $(Y_n(a))_{s,t}$ is nonzero and this is true if and only if $[\mathbf{s}_j^{(n)}]_s \neq 0$ and $[\mathbf{s}_j^{(n)}]_t \neq 0$ for some j of the decomposition of $Y_n(\cdot)$. But, by invoking the strong (weak) locality assumption, the latter is true only when $\mathbf{e}_s$ and $\mathbf{e}_t$ belong to the same dyad, i.e., when

$$\|x_s^{(n)} - x_t^{(n)}\|_\infty = \begin{cases} O(n^{-1/d}) & \text{if the "locality" is strong,} \\ o(1) & \text{if the "locality" is weak.} \end{cases}$$

Therefore by applying the same argument to $W_n = V_n D_n^{-1/2}(a)$, by virtue of (18), and since a is continuous and positive, we obtain that the preconditioned matrix is similar to

$$[Y_n(1)]^{-1}((c^*)^{-1}Y_n(1) + E_n(a)),$$

where $E_n(a)$ is a matrix having the same "sparsity" pattern as $Y_n(1)$ and having spectral norm infinitesimal as n if a is strictly positive. More precisely, it holds that $\|E_n(a)\|_2 \leq c\ \omega_a(n^{-1/d})$, where c is an absolute constant and $\omega_a(\cdot)$ denotes the modulus of continuity of a. By using the fact that $\{X_n(1)\}_n$ has sparsely vanishing spectra, we deduce that $\{Y_n(1)\}_n$ has sparsely vanishing spectra, and therefore by Lemma 4.1 the claimed thesis follows. If $\|[Y_n(1)]^{-1}\|_2\|E_n(a)\|_2 = o(1)$ then we can apply Lemma 4.2 by deducing the strong clustering property. Finally, if (18) does not hold it is enough to repeat the same proof with $\mathcal{P}_n^{**}(a)$ in place of $\mathcal{P}_n^*(a)$. □

Therefore, the matrices are spectrally clustered at one whenever the assumption of locality is fulfilled and a is positive or with isolated zeros.

Remark 4.1. We stress that the proof of Theorem 4.1 does not exploit explicitly the symmetry (in the Hermitian sense) of the given structure $X_n(a)$. This means that the same result holds with the singular values in place of the eigenvalues if the considered matrices are not Hermitian. The latter remark has particular interest when considering the discretization of partial differential equations (PDEs) over irregular domains by FD approximation schemes (see section 5.1.3).

Remark 4.2. Theorem 4.1 suggests a unified definition of preconditioners as $\mathcal{P}_n^*(a)$ in the cases where (18) holds and where it does not hold. Define $\mathcal{P}_n^{***}(a)$ as $\mathcal{P}_n^*(a)$, where $D_n(a)$ must be replaced by $\Delta_n(a)$. The matrix $\Delta_n(a)$ is diagonal and its ith diagonal entry is given by $a(x_k^{(n)})$, $x_k^{(n)} \in \hat{\mathcal{G}}_n$, and $\mathcal{N}_n(x_k^{(n)}) = \mathbf{z}_i^{(n)}$.

5. Applications. In this section we briefly report some applications of the theoretical results given in the preceding sections. In particular, we remark the importance of Theorem 4.1: it states that diagonal matrix sequences and $\{X_n(1)\}_n$ can be combined to obtain superlinear preconditioners for $\{X_n(a)\}_n$. If we consider, for instance, nonlinear elliptic BVPs, then the resulting nonlinear algebraic systems must be solved by an iterative method. By considering the Newton method we obtain that at the kth step we can solve a linear system (related to the Jacobian) whose coefficient matrix

is of the form $X_n(a_k)$. Therefore we factorize $X_n(1)$ once one time and we have a superlinear preconditioner for all the sequences $\{X_n(a_k)\}_n$ of matrices. (Compare [31], where this idea was exploited in the FD case, and [33, 39] for some related results regarding matrix sequences arising in FE and collocation methods.) In addition, in the case where the domain has a rectangular form we observe that $X_n(1)$ has a (multilevel) band Toeplitz structure. In this specific case the quoted literature furnishes a wide choice of optimal solvers [5, 13], and therefore the complexity of solving systems with matrices $\{X_n(a)\}_n$ is reduced to the simpler and well-studied Toeplitz system problem.

Of course the same remark applies to Theorem 3.1 when $b \equiv 1$ and a is positive: in that case the sequence $\{X_n(1)\}_n$ is an optimal preconditioning sequence for $\{X_n(a)\}_n$.

5.1. FD matrices.

5.1.1. Unidimensional semielliptic differential operators. Let us consider the function $a \in C(K)$, $K = [0,1]$, and the differential problem

$$\begin{cases} (-)^t \frac{d^t}{dx^t}\left(a(x)\frac{d^t}{dx^t}u(x)\right) = b(x) & \text{on } \Omega = (0,1), \quad a \geq 0, \\ u^{(j)}(x) = 0, & j = 0, \ldots, t-1, \quad \text{on } \partial\,\Omega. \end{cases} \tag{20}$$

Although the FD discretization process could be performed in a more general way (refer to [36]), here we limit ourselves to the case where both the inner and the outer differential operator $\frac{d^t}{dx^t}$ are discretized by means of the FD formula of maximal precision order $\nu(q)$, defined through the coefficient vector $\mathbf{c} \in \mathbf{R}^q$, where q is the number of contiguous grid points involved by the discretization. It is worthwhile stressing that, owing to the comparison between the computational cost and the precision order, we will always consider q odd when t is even and vice versa.

More precisely, by setting $v(x) = \frac{d^t}{dx^t}u(x)$ and $x_{r+j} = x_r + jh$, we have

$$v(x_r) = \frac{1}{h^t}\sum_j c_j u(x_{r+j}) + O(h^\nu),$$

so that the expression

$$z_r = \frac{(-)^t}{h^{2t}} \sum_i c_i a(x_{r+i}) \left(\sum_j c_j u(x_{r+i+j}) \right) \tag{21}$$

represents the discrete approximation of

$$(-)^t \left(\frac{d^t}{dx^t}\left(a(x)\frac{d^t}{dx^t}u(x)\right)\right)_{|x=x_r},$$

where the global precision is still ν under the assumption that a is regular enough (see [36] for details concerning (21)).

Such an expression has a useful matrix interpretation. Calling $\mathbf{a}_r$ the q-dimensional vector of the evaluations of $a(x)$ as in (21), $\mathbf{a}_r * \mathbf{c}$ the q-dimensional vector obtained by Hadamard [4] (componentwise) product between $\mathbf{a}_r$ and $\mathbf{c}$, and T the $q \times (2q-1)$ Toeplitz matrix having on its first row the vector $(\mathbf{c}, 0, \ldots, 0)$, the quantity z_r can be represented in a compact way as a weighted convolution

$$z_r = \frac{(-)^t}{h^{2t}} (\mathbf{a}_r * \mathbf{c})^T T \mathbf{u}_r, \tag{22}$$

where the vector $\mathbf{u}_r \in \mathbf{R}^{2q-1}$ is constructed by the equispaced evaluations of $u(x)$ displayed in (21).

Clearly, the quantity z_r represents the left-hand side of the rth row of the linear system arising from the discretization of the template problem (20), with the convention that the factor h^{2t} is discarded on its right-hand side. More precisely, we have the following definition.

DEFINITION 5.1. *By setting $h = (n+1)^{-1}$, the symbol $X_n(a, m, t)$, with $m = \lfloor \frac{q}{2} \rfloor$, denotes the $n \times n$ matrix discretizing the problem* (20) *through the maximal precision FD formula related to the coefficient vector $\mathbf{c} \in \mathbf{R}^q$ for both the inner and the outer derivative and according to* (22) *for the rth row.*

The subsequent theorem guarantees that $X_n(\cdot, m, t)$ is a linear normally positive operator so that the localization results given in section 3.1 plainly hold.

THEOREM 5.1 (see [36]). *The matrices $X_n(a, m, t)$ can be expressed as*

$$\sum_{i=1}^{n+k(m,\nu,t)} a(x_i) Q_n(\mathbf{c}, \mathbf{c}, i), \tag{23}$$

where the matrices $Q_n(\mathbf{c}, \mathbf{c}, i)$ are symmetric nonnegative definite dyads given by

$$Q_n(\mathbf{c}, \mathbf{c}, i) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & D_i & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad D_i = (-)^t \mathrm{rev}(\mathbf{c}) \mathbf{c}^T \in \mathbf{R}^{q \times q}, \quad \mathrm{rev}(\mathbf{c}) = (-)^t \mathbf{c}.$$

As a consequence, we have that, for any n, t, and m, the matrix-valued operator $X_n(\cdot, m, t) : C(K) \to \mathcal{S}^{n \times n}$ is linear and normally positive. Here $\mathcal{S}^{n \times n}$ denotes the real subspace of $\mathbf{C}^{n \times n}$ of the symmetric real $n \times n$ matrices.

For instance, for $t = 1$ and $q = 2$ (that is, the minimal number of points that insures the convergence of the FD formula), we meet again the example considered in the introduction so that $\mathbf{c}^T = (1, -1)$, $D_i = \left[\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right]$, and the ith row of the matrix $X_n(a, 1, 1)$ is given by $\big(\mathbf{0}, -a_{z_i}, a_{z_i} + a_{z_{i+1}}, -a_{z_{i+1}}, \mathbf{0}\big)$, where $a_{z_i} = a(z_i)$, $z_i = (i - 1/2)h$, and $h = (n+1)^{-1}$. In this case we have $\hat{\mathcal{G}}_n = \{z_i\}_{1 \le i \le n}$ and $\mathcal{G}_n = \hat{\mathcal{G}}_n \cup \{z_{n+1}\}$. It is evident that both the sets $\{\mathcal{G}_n\}$ and $\{\hat{\mathcal{G}}_n\}$ are $*$-uniformly distributed in the sense of Definition 2.1. In addition we notice that the $k(n) = 1$ and the $(n+1)$th point belong to Ω. To have some real extra points we have to increase the precision order, that is, the number q.

For $t = 1$ and $q = 4$ ($m = 2$), we have a formula whose error in the approximation of the elliptic operator is $O(h^4)$ with $h = (n+1)^{-1}$. In this case, up to some scaling factor, we have $\mathbf{c}^T = (1, -27, 27, -1)$ and therefore the ith row of the associated matrix $X_n(a, 1, 2)$ is defined as

$$\begin{aligned} &(\mathbf{0}, -a_{z_{i-1}}, 27(a_{z_{i-1}} + a_{z_i}), -27(a_{z_{i-1}} + 27a_{z_i} + a_{z_{i-1}}), a_{z_{i-1}} + 27^2 a_{z_i} + 27^2 a_{z_{i+1}} + a_{z_{i+2}}, \\ &\quad -27(a_{z_i} + 27a_{z_{i+1}} + a_{z_{i+2}}), 27(a_{z_{i+1}} + a_{z_{i+2}}), -a_{z_{i+2}}, \mathbf{0}). \end{aligned}$$

Hence $\hat{\mathcal{G}}_n = \{z_i\}_{1 \le i \le n}$ and $\#\{\mathcal{G}_n \backslash \hat{\mathcal{G}}_n\} = 3$, where one of the additional points z_{n+1} is internal to Ω and the other two are extra points, namely, z_0 and z_{n+2}.

In the general case all the assumptions regarding the "locality" are fulfilled so that $\{X_n(\cdot, m, t)\}_n$ is "locally X" in the strong sense where $X = \{X_n\}_n$ and $X_n = X_n(1, m, t)$. In particular we have $Z_n = I_n$; the constant c^* considered in Theorem 4.1 is positive, finite, and independent of n; and (18) holds. Finally the operator

$X_n(\cdot, m, t)$ does not give the identity when applied to 1 so that only the assumptions of the second part of Theorem 3.3 are fulfilled with $F_n(\cdot) = X_n(\cdot, m, t)$.

Therefore all the theorems given in sections 3 and 4 are valid for these structured matrices (see also [31, 35, 36]).

Remark 5.1. *Since for any n, m, and t the matrix $X_n(1, m, t)$ is Toeplitz, it follows that each sequence $\{X_n(a, m, t)\}_n$ is locally Toeplitz too* [43].

5.1.2. Multidimensional semielliptic differential operators. Let $\Omega \subset \mathbf{R}^2$ be a bounded domain with piecewise Lipschitz continuous boundaries. We consider the semielliptic problem over Ω:

$$-\frac{\partial}{\partial x}\left(a(x,y)\frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\left(a(x,y)\frac{\partial u}{\partial y}\right) = b(x,y) \tag{24}$$

with Dirichlet boundary conditions and $a(x, y) \geq 0$. Preliminarily we analyze in detail the simpler case in which $\Omega = (0, 1)^2$.

As a case study we consider the well-known five-point formula so that, by ordering the unknowns in the classic manner, we have to solve an $n \times n$ linear system whose coefficient matrix $X_n(a)$, $a > 0$, is a symmetric positive definite block tridiagonal matrix with tridiagonal blocks and is semidefinite positive if $a \geq 0$. By referring to the notations of the preceding sections, we may represent the nontrivial entries of a generic row of $X_n(a)$ in the following way:

$$\left(-a_{i,z_j}, \mathbf{0}, -a_{z_i,j}, Z_{i,j}, -a_{z_{i+1},j}, \mathbf{0}, -a_{i,z_{j+1}}\right),$$

with $Z_{i,j}$ equal to the sum of the absolute values of the nondiagonal entries of the related row and

$$a_{s,t} = a\left(x_s, x_t\right), \quad x_v = vh, \; z_v = x_{v-1/2}, \quad h = \frac{1}{n^{1/2}+1}.$$

In this way, setting $\mathbf{c} = (1 - 1)$ as the vector of coefficients of the unidimensional formula used to construct the x and y second derivatives we have

$$X_n(a) = \sum_{i,j=1}^{\sqrt{n}+1} a_{i,z_j} Q_n^x(\mathbf{c}, \mathbf{c}, i, j) + a_{z_i,j} Q_n^y(\mathbf{c}, \mathbf{c}, i, j),$$

where

$$Q_n^x(\mathbf{c}, \mathbf{c}, i, j) = \mathbf{e}_j \mathbf{e}_j^T \otimes Q_n(\mathbf{c}, \mathbf{c}, i)$$

and

$$Q_n^y(\mathbf{c}, \mathbf{c}, i, j) = Q_n(\mathbf{c} \otimes \mathbf{e}_j, \mathbf{c} \otimes \mathbf{e}_j, i).$$

It is evident that $X_n(\cdot)$ is linear normally positive. In particular we have $Z_n = I_n$; the constant c^* considered in Theorem 4.1 is positive, finite, and independent of n, and (18) holds. Furthermore the operator $X_n(\cdot)$ does not reproduce the identity when applied to 1 and consequently only the assumptions of the second part of Theorem 3.3 are fulfilled with $F_n(\cdot) = X_n(\cdot)$.

However, the number of nodes of the nth grid is not n but is asymptotical to $2n$ and therefore the proof of the ergodic and distributional theorems in section 3.2 fails. On the other hand, the FD methods are local schemes and therefore the "locality

principle" given in (8) is fulfilled with the remark that $\mathcal{N}_n$ is not a bijection but a function onto so that the important Theorem 4.1 stands. For a specific analysis of the considered two-dimensional problem see [31], and for a general extension of the concept of locality that takes into account the case where $\#\hat{\mathcal{G}}_n \sim \theta n$, $\theta \geq 1$ see section 7.

Remark 5.2. Since for any n, the matrix $X_n(1)$ has a two-level Toeplitz structure whose nonseparable generating function is the sum of two separable functions, it follows that each sequence $\{X_n(a)\}_n$ can be expressed as a sum of two multilevel locally Toeplitz sequences [40].

5.1.3. Irregular domains. When considering nonrectangular domains, we generally lose the symmetry of the resulting matrices $X_n(a)$. In this case the only property that is maintained is the "locality principle" stated in (8) (see [40] for the spectral analysis of $\{X_n(a)\}$ from the point of view of (generalized) multilevel locally Toeplitz sequences). In this case Theorem 4.1 still holds (see the associated Remark 4.1). However, in this case, since the matrices are not symmetric, it is useless to consider a symmetric scaling: indeed the preconditioner $\mathcal{P}_n^*(a)$ should be replaced by $[\hat{D}_n^2(a)X_n(1)]^{-1}X_n(a)$, where the chosen preconditioning matrix is not symmetric. Therefore, with regard to Theorem 4.1 we have that the distributional results concern the singular values instead of the eigenvalues.

5.2. Graph theory. Let $H = (\mathcal{U}, \mathcal{V})$ be a graph with n nodes $\mathcal{U} = \{u_1, \ldots, u_n\}$ and m arcs $\mathcal{V} = \{v_1, \ldots, v_m\}$.

Let us consider the $(0, 1, -1)$-incidence matrix $E = E(H)$ defined as $E_{i,j} = 0$ if the arc v_j does not involve the node u_i, $E_{i,j} = 1$ if v_j emanates from u_i and $E_{i,j} = -1$ if v_j terminates at u_i. It is transparent that $E \in \{0, 1, -1\}^{n\times m}$.

Now set θ to be an m-dimensional vector of positive weights and let Θ be the $m \times m$ diagonal matrix whose diagonal elements are the components of the vector θ. In some optimization problems involving interior point methods and graph theory, it is of crucial interest to solve linear systems of the form

$$E\Theta E^T\mathbf{x} = \mathbf{b}.$$

If $\{m = m_n\}_n$ is a fixed increasing sequence of positive integers, it is evident that $\{\Phi_n(\cdot)\}_n$ with $\Phi_n(\theta) = E\Theta E^T$ is a sequence of LPOs in a matrix sense because each nonnegative vector θ is mapped to a nonnegative definite matrix. The operators are normally positive if E is full rank and this is generally the case in the applications.

Therefore all the statements concerning the localization of the eigenvalues of $\{\mathcal{P}_n(\theta_1; \theta_2)\}_n$ and all the statements regarding the spectral extremal behavior of $\{\Phi_n(\cdot)\}_n$ hold.

On the other hand, the "locality" requirements are generally violated. First, m_n can grow as n^2 in the case of quasi-complete graphs and this implies that $k(n) \neq O(n^{\frac{d-1}{d}})$ for any positive d.

Second, there is not "locality" in the sense that the diagonal elements $[\Phi_n(\theta)]_{i,i}$ can have contributions from all the θ_j (again this situation occurs in the case of quasi-complete graphs).

However, it should be stressed that there exist important subclasses of graphs for which we can easily identify a "local structure." Let $\{\mathcal{V} = \mathcal{V}_{m_n}\}_n$ be such that $\exists k > 0$ and $\forall n$, $\forall v \in \mathcal{V}_{m_n}$ we have that $v = (u_i, u_j)$ implies that $|j-i| \leq k$. This kind of "local graphs" where the locality is in a graph sense is associated with sequences of matrices $\{\Phi_n(\theta)\}_n$ that are "locally X" with $X = \{X_n\}_n$ and $X_n = EE^T$, $E = E(H_n)$.

Notice that in the case of local graphs, the matrix $E\Theta E^T$ shows a pattern which is very similar to those characterizing finite differences for one-dimensional problems. (For a more general notion of "local graphs" containing multidimensional structures and for a more specific analysis of these problems see [14].)

Remark 5.3. Notice that the matrix sequences $\{X_n(a)\}_n$ considered in this section furnish an example of locally X matrix sequences that are not locally Toeplitz except for few very special cases [14].

6. Comparison with the literature. In this section we compare an instance of our technique with the iterative solvers used in the relevant literature. The case under study is the one of PDEs such as (24) over a rectangle of $\mathbf{R}^2$ discretized by uniformly spaced finite differences. We analyze the following three situations:

(a.1) $a > 0$ and $a \in C^2([0,1]^2)$,

(a.2) $a > 0$ and $\infty > \sup a \geq \inf a > 0$,

(a.3) $a \geq 0$ with at most isolated zeros and a piecewise continuous.

The PCG methods based on preconditioners from incomplete LU factorizations [22, 8, 16] and from the circulant algebra [6, 19, 21] are sublinear, i.e., require a number of iterations $O(n^\beta)$ with positive β. This is true even in the case (a.1), where a is positive and smooth. (For other numerical techniques, see, e.g., [41].)

On the other hand the PCG methods defined by using separable preconditioners [12] and the multigrid algorithms [17, 3] are optimal in the sense of Definition 3.1 in the cases (a.1) and (a.2) but not in the case (a.3). On the contrary our technique is superlinear in the case (a.1) (therefore also optimal) and ensures a "weak" clustering in the cases (a.2) and (a.3); this property does not guarantee theoretically the optimality but the numerous numerical experiments performed in [35, 36] suggest a convergence rate independent of the size n. Finally we recall that the "weak" clustering property also holds in the case where $a \in L^1([0,1]^2)$ [31] with the convention that $a_{i,j}$ is not $a(x_i, x_j)$ but $\frac{1}{\mathrm{meas}(I_{i,j})} \int_{I_{i,j}} a(s,t)dsdt$, with $I_{i,j} = [x_i, x_{i+1}] \times [x_j, x_{j+1}]$.

7. Further extension. What section 5.1.2 revealed is that the requirement that $\#\hat{\mathcal{G}}_n = n$ so that it is possible to associate a bijection from $\hat{\mathcal{G}}_n$ to $\{\mathbf{z}_j^{(n)}\}$ can be too restrictive. Actually the simple five-point formula used in the two-dimensional problem considered in section 5.1.2 leads to a number of internal grid points which is asymptotic to $2n$. The same remark holds in the case of FE methods. In fact if n is the cardinality of the triangulation [7] associated with the method, that is the size of the matrix $X_n(a)$, then the number of grid points (the degrees of freedom [7] of the triangulation) is asymptotic to θn with some $\theta \geq 1$. This is true in general for any Faedo–Galerkin–Ritz (FGR) methods with locally supported basis because the number of the degrees of freedom is generally bigger than the number of the elements of the triangulation; more precisely, if n is the cardinality of the triangulation then the number of degrees of freedom is asymptotical to θn with $\theta \geq 1$ an absolute constant depending on d and on the geometry of the triangulation. Finally we notice that the same relation stands between the number of the nodes and the arcs when local graphs are considered.

Concerning the FGR methods it should be observed that the related matrices $X_n(a)$ are by definition symmetric and positive definite since $X_n(a)$ represents a positive scalar product on an n-dimensional vector space of functions defined over Ω. In addition, by the monotonicity of the integral, for any n we infer that the matrix-valued operator $X_n(\cdot)$ is linear and positive; therefore all the localization results given in section 3.1 are valid.

However, the locality in the sense stated in Definition 2.3 such that

$$X_n(a) = \sum_{j=1}^{n+k(n)} a_j^{(n)} \mathbf{v}_j^{(n)} [\mathbf{v}_j^{(n)}]^H, \quad a_j^{(n)} = a(x_j^{(n)}),$$

cannot hold for these matrices whose coefficients are defined through integrals and not through some sampling of $a(x)$.

7.1. A more general notion of "locality." First we simply adapt Definition 2.1 to deal with the case where we have $\theta n(1+o(1))$ points with $\theta > 1$: as previously observed this situation occurs frequently in the case of discretization of differential operators over multidimensional domains as well as in graph theory.

DEFINITION 7.1. *Let $\Omega \subset \mathbf{R}^d$ be a bounded domain and $K = \bar{\Omega}$ be its closure. A sequence of grids $\mathcal{G} = \{\hat{\mathcal{G}}_n\}_n$ with $\hat{\mathcal{G}}_n = \{x_i^{(n)}\}$, $x_i^{(n)} \in \Omega$, and $\#(\hat{\mathcal{G}}_n) \sim \theta n$, $\theta \geq 1$, is called quasi-uniform if for any ball $B(x_0, \rho) = \{x \in \mathbf{R}^d : \ \|x - x_0\|_\infty \leq \rho\} \subset K$ with positive radius ρ we have*

$$\# \left\{ x_i^{(n)} \in B(x_0, \rho) \right\} = \theta n \cdot \frac{\text{meas}(B(x_0, \rho))}{\text{meas}(K)} + o(n).$$

The sequence of grids is uniform if for any ball set $B(x_0, \rho) \subset K$ we find that

$$\# \left\{ x_i^{(n)} \in B(x_0, \rho) \right\} = \theta n \cdot \frac{\text{meas}(B(x_0, \rho))}{\text{meas}(K)} + O(n^{\frac{d-1}{d}}).$$

The sequence of grids is $*$-uniform if for any connected set $S \subset K$ with Lipschitz continuous boundaries we find that

$$\# \left\{ x_i^{(n)} \in S \right\} = \theta n \cdot \frac{\text{meas}(S)}{\text{meas}(K)} + O(n^{\frac{d-1}{d}}).$$

We now consider some partitionings $\mathcal{P}$ of the set $K = \bar{\Omega}$.

DEFINITION 7.2. *A partitioning $\mathcal{P}_n$ of K having size n is a collection of Peano–Jordan connected subsets $\{\Omega_i^{(n)}\}_{i=1}^n$ with the following features:*

- $K = \bigcup_{i=1}^n \Omega_i^{(n)}$,
- $\Omega_i^{(n)} \bigcap \Omega_j^{(n)} = \emptyset$ *for* $i \neq j$.

DEFINITION 7.3. *Let $\Omega \subset \mathbf{R}^d$ be a bounded domain and $K = \bar{\Omega}$ be its closure. A sequence of partitionings $\mathcal{P} = \{\mathcal{P}_n\}_n$ of K having size n is called quasi-uniform if for any ball $B(x_0, \rho) = \{x \in \mathbf{R}^d : \ \|x - x_0\|_\infty \leq \rho\} \subset K$ with positive radius ρ we have*

$$\# \left\{ \Omega_i^{(n)} \subset X \right\} = n \cdot \frac{\text{meas}(X)}{\text{meas}(K)} + o(n), \quad X \in \{B(x_0, \rho), K \backslash B(x_0, \rho)\}.$$

The sequence of partitionings is uniform if for any ball $B(x_0, \rho) \subset K$ we find that

$$\# \left\{ \Omega_i^{(n)} \subset X \right\} = n \cdot \frac{\text{meas}(X)}{\text{meas}(K)} + O(n^{\frac{d-1}{d}}), \quad X \in \{B(x_0, \rho), K \backslash B(x_0, \rho)\}.$$

The sequence of partitionings is $*$-uniform if for any connected set $S \subset K$ with Lipschitz continuous boundaries we find that

$$\# \left\{ \Omega_i^{(n)} \subset X \right\} = n \cdot \frac{\text{meas}(X)}{\text{meas}(K)} + O(n^{\frac{d-1}{d}}), \quad X \in \{S, K \backslash S\}.$$

PROPOSITION 7.1. *Let $\mathcal{P} = \{\mathcal{P}_n\}_n$, with $\mathcal{P}_n$ having size n, be quasi-uniform with respect to $K = \bar{\Omega}$ with Ω bounded domain. Then for any ball $B(x_0, \rho) \subset K$ with positive radius ρ we have*

$$\#\left\{\Omega_i^{(n)} \cap \partial B(x_0,\rho) \neq \emptyset\right\} = o(n).$$

Proof. Let $\epsilon < \rho$. Set $B_\epsilon = B(x_0, \rho+\epsilon)$, $B_{-\epsilon} = B(x_0, \rho-\epsilon)$, and $B = B(x_0,\rho)$. Then by the quasi uniformity we have

$$\#\left\{\Omega_i^{(n)} \subset B_\epsilon\right\} = n \cdot \frac{\text{meas}(B_\epsilon)}{\text{meas}(K)} + o(n), \quad \#\left\{\Omega_i^{(n)} \subset B_{-\epsilon}\right\} = n \cdot \frac{\text{meas}(B_{-\epsilon})}{\text{meas}(K)} + o(n).$$

Since

$$\left\{\Omega_i^{(n)} \cap \partial B\right\} \subset \mathcal{P}_n \backslash \left\{\Omega_i^{(n)} \subset B_{-\epsilon} \text{ or } \Omega_i^{(n)} \subset K\backslash B_\epsilon\right\}$$

it follows that

$$\#\left\{\Omega_i^{(n)} \cap \partial B(x_0,\rho) \neq \emptyset\right\} \leq n - n \cdot \frac{\text{meas}(B_\epsilon)}{\text{meas}(K)} - n \cdot \frac{\text{meas}(K\backslash B_{-\epsilon})}{\text{meas}(K)} + o(n) \leq S\epsilon n$$

with S absolute constant and arbitrary ϵ. □

PROPOSITION 7.2. *Let $R = B(x_1,r_1) \cup B(x_2,r_2)$ with $B^\circ(x_1,r_1) \cap B(x_2,r_2) = \emptyset$ and $B(x_i,r_i) \subset K$, $r_i > 0$, $i = 1,2$. Let $\mathcal{P} = \{\mathcal{P}_n\}_n$, with $\mathcal{P}_n$ having size n, be quasi-uniform with respect to $K = \bar{\Omega}$ with Ω bounded domain. Then*

$$\#\left\{\Omega_i^{(n)} \subset X\right\} = n \cdot \frac{\text{meas}(X)}{\text{meas}(K)} + o(n)$$

with $X \in \{R, K\backslash R\}$.

Proof. If $B(x_1,r_1) \cap B(x_2,r_2) = \emptyset$, then the claim is trivial so that we assume that $B(x_1,r_1)$ and $B(x_2,r_2)$ have a line in common. Since each $\Omega_i^{(n)}$ is connected, it follows that

$$\#\left\{\Omega_i^{(n)} \subset R\right\} \geq \#\left\{\Omega_i^{(n)} \subset B(x_1,r_1)\right\} + \#\left\{\Omega_i^{(n)} \subset B(x_2,r_2)\right\}$$

because we have to count those $\Omega_i^{(n)}$ that intersect the line in common with $B(x_1,r_1)$ and $B(x_2,r_2)$ and are contained in R. Therefore it follows that

$$\#\left\{\Omega_i^{(n)} \subset R\right\} \leq \#\left\{\Omega_i^{(n)} \subset B(x_1,r_1)\right\} + \#\left\{\Omega_i^{(n)} \subset B(x_2,r_2)\right\} + f_n,$$

where f_n is the totality of the sets $\Omega_i^{(n)}$ that have nonempty intersection with the frontier of $B(x_1,r_1)$ or $B(x_2,r_2)$. Consequently the proposition is proved owing to $f_n = o(n)$, the latter being implied by Proposition 7.1 and by the quasi uniformity of $\mathcal{P}$. The part concerning $K\backslash R$ is analogous. □

As done in the second section, we can prove a result which is analogous to Theorem 2.1 when a quasi-uniform partitioning sequence is considered in place of a quasi-uniform grid sequence.

THEOREM 7.1. *Let $S \subset \Omega$ be a measurable set in the Peano–Jordan sense and let $\Omega \subset \mathbf{R}^d$ be a bounded domain. Let $\mathcal{P} = \{\mathcal{P}_n\}_n$, with $\mathcal{P}_n$ having size n, be quasi-uniform with respect to $K = \bar{\Omega}$. Then*

$$\#\left\{\Omega_i^{(n)} \subset S\right\} = n \cdot \frac{\text{meas}(S)}{\text{meas}(K)} + o(n).$$

Proof. For any $\epsilon > 0$, let two plurirectangles $P_1 \subset P_2$ be as in Theorem 2.1. By set inclusion it easily follows that

$$\#\left\{\Omega_i^{(n)} \subset P_1\right\} \leq \#\left\{\Omega_i^{(n)} \subset S\right\} \leq \#\left\{\Omega_i^{(n)} \subset P_2\right\}.$$

By repeated applications of Proposition 7.2 we infer that

$$\#\left\{\Omega_i^{(n)} \subset P_1\right\} = n \cdot \frac{\text{meas}(P_1)}{\text{meas}(K)} + o(n) \geq n \cdot \frac{\text{meas}(S) - \epsilon}{\text{meas}(K)} + o(n),$$

$$\#\left\{\Omega_i^{(n)} \subset P_2\right\} = n \cdot \frac{\text{meas}(P_2)}{\text{meas}(K)} + o(n) \leq n \cdot \frac{\text{meas}(S) + \epsilon}{\text{meas}(K)} + o(n),$$

and consequently the proof is concluded owing to the arbitrariness of ϵ. ∎

The two notions of uniform distribution can have relationships. More specifically, define $I^{(n)}(j)$ as $\hat{\mathcal{G}}_n \cap \Omega_j^{(n)}$. Therefore if we make some assumption on the cardinality of each $I^{(n)}(j)$, then we deduce the uniform distribution of $\{\hat{\mathcal{G}}_n\}_n$ from the uniform distribution of $\{\mathcal{P}_n\}_n$. This fact is established in the following theorem.

THEOREM 7.2. *Suppose that $\mathcal{G} = \{\hat{\mathcal{G}}_n\}_n$ is a sequence of grids with $\#(\hat{\mathcal{G}}_n) = d_n$ and suppose that $\mathcal{P} = \{\mathcal{P}_n\}_n$ is a sequence of partitions with $\#(\mathcal{P}_n) = n$. Let $I^{(n)}(j) = \hat{\mathcal{G}}_n \cap \Omega_j^{(n)}$ and suppose that $\{\#(I^{(n)}(j))\}_n$ is clustered at θ (in the sense of Definition 3.2), where θ is a positive integer and that $\#(I^{(n)}(j))$ is uniformly bounded by a constant z independent of n. If $\mathcal{P}$ is uniformly distributed on K, then*

(1) *$\#(\hat{\mathcal{G}}_n) \sim \theta n$, so that $\lim_{n\to\infty} d_n/n = \theta$ and*
(2) *$\{\hat{\mathcal{G}}_n\}_n$ is uniformly distributed on $K = \bar{\Omega}$.*

Proof.

(1) By the definition of cluster (Definition 3.2) it follows that

$$\gamma_n(\theta, 1/2) = \#\{j : |\#(I^{(n)}(j)) - \theta| > 1/2\} = o(n).$$

Since $\#(I^{(n)}(j)) \in \mathbf{N}$, it holds that

$$\gamma_n(\theta, 1/2) = \#\{j : \#(I^{(n)}(j)) \geq \theta + 1\} + \#\{j : \#(I^{(n)}(j)) \leq \theta - 1\} = o(n).$$

Since $\mathcal{P}_n$ is a partitioning we have

$$\begin{aligned}\#(\hat{\mathcal{G}}_n) &= \sum_{j=1}^{n} \#(I^{(n)}(j)) \\ &= \theta(n - \gamma_n(\theta, 1/2)) + \epsilon_n(\theta),\end{aligned}$$

where

$$\epsilon_n(\theta) = \sum_{j:\#(I^{(n)}(j))\neq\theta} \#(I^{(n)}(j)).$$

Finally, owing to the hypothesis $0 \leq \#(I^{(n)}(j)) \leq z$, it follows that

$$0 \leq \epsilon_n(\theta) \leq z\ \gamma_n(\theta, 1/2) = o(n)$$

and the proof of the first part is concluded.

(2) By a set theoretic argument and by part (1) it easily follows that

$$\begin{aligned}\#\left\{x_i^{(n)} \in B(x_0,\rho)\right\} &\geq \sum_{j=1}^{n} \#\left\{x_i^{(n)} \in \Omega_j^{(n)} : \Omega_j^{(n)} \subset B(x_0,\rho)\right\} \\ &\geq \theta\left(\#\left\{\Omega_j^{(n)} \subset B(x_0,\rho)\right\} - \gamma_n(\theta,1/2)\right) \\ &= \theta n \frac{\mathrm{meas}(B(x_0,\rho))}{\mathrm{meas}(K)} + o(n).\end{aligned}$$

The application of the same argument to $K\backslash B(x_0,\rho)$ yields

$$\#\left\{x_i^{(n)} \in K\backslash B(x_0,\rho)\right\} \geq \theta n \frac{\mathrm{meas}(K\backslash B(x_0,\rho))}{\mathrm{meas}(K)} + o(n)$$

so that

$$\#\left\{x_i^{(n)} \in B(x_0,\rho)\right\} \leq \theta n \frac{\mathrm{meas}(B(x_0,\rho))}{\mathrm{meas}(K)} + o(n)$$

and the theorem is proved. □

Now we give an alternative notion of locality.

Let $\{X_n(\cdot)\}_n$ be a sequence of $n \times n$ Hermitian matrix-valued LPOs defined over $PC(K)$ with K being the closure of Ω and $PC(K)$ denoting the space of the piecewise continuous and bounded functions over K: here $a \in PC(K)$ if there exists a finite partitioning of K in Peano–Jordan connected measurable sets T such that a restricted to T is continuous.

Let us consider a sequence $Z = \{\{\mathbf{z}_j^{(n)}\}_{1\leq j\leq n}\}_n$ of orthonormal bases of $\{\mathbf{C}^n\}_n$. Let $\mathcal{N}_n$ be a bijection from $\{\Omega_i^{(n)}\}$ onto $\{\mathbf{z}_i^{(n)}\}$. In this way, to each subset $\Omega_j^{(n)}$ we associate a unique vector $\mathbf{z}_{j(x)}^{(n)}$ of the basis $\{\mathbf{z}_j^{(n)}\}_{1\leq j\leq n}$ and to any vector we associate a unique subset of points $I^{(n)}(j)$ defined as $\hat{\mathcal{G}}_n \cap \Omega_j^{(n)}$. Let us call Z_n the unitary matrix constructed as $[\mathbf{z}_1^{(n)}, \ldots, \mathbf{z}_n^{(n)}]$, let us call I_A the characteristic function of a set A, and let us suppose that the sequence of matrices

$$V = \left\{\{X_n(I_{\Omega_j^{(n)}})\}_{1\leq j\leq n}\right\}_n$$

is "locally" weakly distributed with regard to Z in the following sense. For any $i \leq n$, there holds

$$\#\{j :\ [\mathbf{z}_j^{(n)}]^H X_n(I_{\Omega_i^{(n)}})\mathbf{z}_j^{(n)} \neq 0\} = O(1) \quad (\mathrm{rank}(X_n(I_{\Omega_i^{(n)}})) = O(1)). \tag{25}$$

Moreover, for any $i, j \leq n$ the relation

$$\begin{aligned}&[\mathbf{z}_j^{(n)}]^H X_n(I_{\Omega_i^{(n)}})\mathbf{z}_j^{(n)} \neq 0 \quad \text{implies that} \\ &\quad \mathrm{dist}(\Omega_i^{(n)}, \Omega_j^{(n)}) = o(1) \quad \text{so that} \\ &(\|x_l^{(n)} - x_k^{(n)}\|_\infty = o(1) \quad \forall x_l^{(n)} \in I^{(n)}(i)\ \forall x_k^{(n)} \in I^{(n)}(j)).\end{aligned} \tag{26}$$

In other words each $X_n(I_{\Omega_i^{(n)}})$ is made up of linear combinations of a *constant* number (independent of n) of dyads $\{\mathbf{z}_j^{(n)}[\mathbf{z}_j^{(n)}]^H\}$: more precisely each grid point $x_k^{(n)}$ associated via $\mathcal{N}_n$ with a $\mathbf{z}_j^{(n)}$ that satisfies (26) belongs to a neighborhood of $x_l^{(n)} \in \Omega_i^{(n)}$

whose radius is infinitesimal as n tends to infinity. The "locality assumption is strong" when (26) holds with $o(1) = O(n^{-1/d})$; i.e.,

$$\begin{aligned}
&[\mathbf{z}_j^{(n)}]^H X_n(I_{\Omega_i^{(n)}})\mathbf{z}_j^{(n)} \neq 0 \quad \text{implies that}\\
&\operatorname{dist}(\Omega_i^{(n)}, \Omega_j^{(n)}) = O(n^{-1/d}) \quad \text{so that}\\
&(\|x_l^{(n)} - x_k^{(n)}\|_\infty = O(n^{-1/d}) \ \ \forall x_l^{(n)} \in I^{(n)}(i) \ \forall x_k^{(n)} \in I^{(n)}(j)).
\end{aligned}$$

According to previous notation, the following definition is introduced.

DEFINITION 7.4. *Let $\mathcal{P}$ be a sequence of quasi-uniform partitionings. Let $\{X_n(\cdot)\}$ be a sequence of matrix-valued LPOs acting on $PC(K)$ and satisfying* (25) *and* (26). *The space $\mathcal{LL}(X_n)$ of matrices is defined as $\bigcup_{a\in PC(K)} X_n(a)$. The class $\mathcal{LL}(X)$ of the "locally X" matrices is then given by $\{\mathcal{LL}(X_n)\}_n$. If* (26) *holds with $O(n^{-1/d})$ in place of $o(1)$ and $\mathcal{P}$ is $*$-uniform, then we say that the "locality" is strong.*

We notice that with this new extended definition the distribution results stated in section 3.2 hold if we require this further property.

THEOREM 7.3. *For any set S measurable in the Peano–Jordan sense, it holds that*

$$\operatorname{rank}(X_n(I_S)) = n \cdot \frac{\operatorname{meas}(S)}{\operatorname{meas}(K)} + o(n).$$

Proof. Consider the "generalized plurirectangles"

$$P_1(\mathcal{P}_n) \equiv \bigcup_{\Omega_j^{(n)} \subset S} \Omega_j^{(n)} \subset P_2(\mathcal{P}_n) \equiv \bigcup_{\Omega_j^{(n)} \cap S \neq \emptyset} \Omega_j^{(n)}.$$

Then for any positive ϵ we have

$$P_1(\mathcal{P}_n) \subset S \subset P_2(\mathcal{P}_n)$$

and $\operatorname{meas}(P_2(\mathcal{P}_n) \backslash P_1(\mathcal{P}_n)) < \epsilon$ for n large enough (the existence of "generalized plurirectangles" having these properties is guaranteed by the uniformity of $\mathcal{P}$). From the positivity of the operator $X_n(\cdot)$ it follows that

$$0 \leq \tilde{X}_n(I_{P_1(\mathcal{P}_n)}) \leq \tilde{X}_n(I_S) \leq \tilde{X}_n(I_{P_2(\mathcal{P}_n)})$$

with $\tilde{X}_n(\cdot)$ being the associated Rayleigh operator, and consequently

$$\operatorname{rank}\left(X_n(I_{P_1(\mathcal{P}_n)})\right) \leq \operatorname{rank}\left(X_n(I_S)\right) \leq \operatorname{rank}\left(X_n(I_{P_2(\mathcal{P}_n)})\right).$$

The idea is to evaluate $\operatorname{rank}(X_n(I_P))$ with $P = P_j \equiv P_j(\mathcal{P}_n)$, $j = 1, 2$. Let us consider $j = 2$, the other case being similar: we define a set of indices $T = \{k : \mathbf{z}_k^H X_n(I_{P_2})\mathbf{z}_k = 0\}$ and we observe that therefore we have

$$\begin{aligned}
T &= \{k : \mathbf{z}_k^H X_n(I_{\Omega_j^{(n)}})\mathbf{z}_k = 0 \ \forall \Omega_j^{(n)} \cap S \neq \emptyset\}\\
&\subset \{k : \Omega_k^{(n)} \cap P_2 = \emptyset\}\\
&= \{k : \Omega_k^{(n)} \cap S = \emptyset\}\\
&= \{k : \Omega_k^{(n)} \subset K \backslash S\}.
\end{aligned}$$

Consequently, by the quasi uniformity of $\mathcal{P}$, we deduce that

$$\#(T) \le n \cdot \frac{\text{meas}(K\backslash S)}{\text{meas}(K)} + o(n).$$

Since $\dim(\ker(X_n(I_{P_2}))) = \#(T) + o(n)$ we have

$$\text{rank}(X_n(I_{P_2})) \ge n \cdot \frac{\text{meas}(S)}{\text{meas}(K)} + o(n). \tag{27}$$

On the other hand we observe that

$$\begin{aligned} T &= \{k : \mathbf{z}_k^H X_n(I_{\Omega_j^{(n)}})\mathbf{z}_k = 0 \ \forall \Omega_j^{(n)} \cap S \neq \emptyset\} \\ &\supset \{k : \Omega_k^{(n)} \cap S_\epsilon = \emptyset\}, \end{aligned}$$

and so by the quasi uniformity of $\mathcal{P}$ we deduce that

$$\#(T) \ge n \cdot \frac{\text{meas}(K\backslash S_\epsilon)}{\text{meas}(K)} + o(n).$$

Consequently it follows that

$$\text{rank}(X_n(I_{P_2}) \le n \cdot \frac{\text{meas}(S_\epsilon)}{\text{meas}(K)} + o(n). \tag{28}$$

The use of the same argument for $P_1 = P_1(\mathcal{P}_n)$ and the arbitrariness of ϵ end the proof. □

THEOREM 7.4. *Let $\mathcal{P}$ be a $*$-uniform grid sequence and $\{X_n(\cdot)\}_n$ be a sequence of operators as in Definition* 2.3. *Then for any connected set S with Lipschitz continuous boundaries it holds that*

$$\text{rank}(X_n(I_S)) = n \cdot \frac{\text{meas}(S)}{\text{meas}(K)} + O(n^{\frac{d-1}{d}}).$$

Proof. It is enough to observe that

$$\text{rank}\left(X_n\left(I_{\bigcup_{\Omega_i^{(n)} \subset S} \Omega_i^{(n)}}\right)\right) \le \text{rank}(X_n(I_S)) \le \text{rank}\left(X_n\left(I_{\bigcup_{\Omega_i^{(n)} \cap S \neq \emptyset} \Omega_i^{(n)}}\right)\right)$$

and that

$$\#\{i : \ \Omega_i^{(n)} \cap S \neq \emptyset \text{ and } \ \Omega_i^{(n)} \cap K\backslash S \neq \emptyset\} = O(n^{\frac{d-1}{d}}). \qquad \square$$

Under the assumption of $*$-uniformity of $\mathcal{P}$, an analog of Theorem 3.4, i.e., Theorem 7.5, holds true. We remark that the result stated in Theorem 3.4 and in its analog is very important because all the ergodic results concerning the sequences $\{\mathcal{P}_n(a;b)\}_n$ for b nonnegative depend on it.

Here for the sake of completeness we give a sketch of the proof of this crucial result.

THEOREM 7.5. *Let a be a continuous function over K. Under the notations given in Definition* 3.5*, the following relationships*

$$\begin{aligned} N^+(X_n(a)) &= n\tfrac{\alpha_+(a)}{\text{meas}(K)} + O(n^{\frac{d-1}{d}}), \\ N^-(X_n(a)) &= n\tfrac{\alpha_-(a)}{\text{meas}(K)} + O(n^{\frac{d-1}{d}}), \\ \dim(\text{Ker}(X_n(a))) &= n\tfrac{\alpha_0(a)}{\text{meas}(K)} + O(n^{\frac{d-1}{d}}) \end{aligned}$$

are true when the following conditions are satisfied:

hp:a1: $A_0(a)$ is made up of a finite collection of connected compact sets with Lipschitz continuous boundaries;

hp:a2: the sequence $\{X_n(\cdot)\}_n$ is locally X in the strong sense with respect to the partitionings $\mathcal{P}$.

Proof. From *hp:a1*, it follows that $A_-(a)$ and $A_+(a)$ are open in K with Lipschitz continuous boundaries that coincide with those of $A_0(a)$. Moreover the sets $A_-(a)$ and $A_+(a)$ are made up of a finite collection of connected sets. Now by Theorem 7.4, we see that

$$\operatorname{rank}(X_n(I_{A_+(a)})) = n \cdot \frac{\alpha_+(a)}{\operatorname{meas}(K)} + O(n^{\frac{d-1}{d}})$$

and therefore $X_n(I_{A_+(a)})$ has exactly $\frac{\alpha_+(a)}{\operatorname{meas}(K)} + O(n^{\frac{d-1}{d}})$ positive eigenvalues and $n - \operatorname{rank}(X_n(I_{A_+(a)}))$ null eigenvalues. Now consider the new set

$$A_{+,*}(a) = A_+(a) \backslash \bigcup_j \left\{ \Omega_j^{(n)} : \inf_{\Omega_j^{(n)}} a \le 0 \right\}.$$

The hypotheses *hp:a2* and *hp:a1* imply that

$$\# \left\{ \Omega_j^{(n)} : \inf_{\Omega_j^{(n)}} a \le 0 \ \text{ and } \ \Omega_j^{(n)} \cap A_+(a) \neq \emptyset \right\} = O(n^{\frac{d-1}{d}}).$$

By (25) we have $\operatorname{rank}(X_n(I_{\Omega_j^{(n)}})) = O(1)$, and then it is evident that

$$\operatorname{rank}(X_n(I_{A_{+,*}(a)})) = \operatorname{rank}(X_n(I_{A_+(a)})) - O(n^{\frac{d-1}{d}}).$$

From the construction, we have $\inf_{A_{+,*}(a)} a = \theta_n > 0$ and $\sup_{A_{+,*}(a)} a = \|a\|_\infty < \infty$ so that $I_{A_{+,*}(a)} \sim a|_{A_{+,*}(a)}$ and therefore by the equivalence Theorem 3.1, we deduce that $X_n(I_{A_{+,*}(a)})$ and $X_n(a|_{A_{+,*}(a)})$ have the same inertia and therefore the same rank.

By means of the same argument and by the definition of "locally X" matrices, it is easy to see that

$$\# \left\{ j : \left([\mathbf{z}_j^{(n)}]^H X_n(I_{A_{+,*}(a)}) \mathbf{z}_j^{(n)} \right) \cdot \left([\mathbf{z}_j^{(n)}]^H X_n(I_{A_-(a)} + I_{A_0(a)}) \mathbf{z}_j^{(n)} \right) \neq 0 \right\} = O(n^{\frac{d-1}{d}}).$$

Therefore we infer that $X_n(a)$ is positive definite over a subspace of dimension $\frac{\alpha_+(a)}{\operatorname{meas}(K)} + O(n^{\frac{d-1}{d}})$, is negative definite over a space of dimension $\frac{\alpha_-(a)}{\operatorname{meas}(K)} + O(n^{\frac{d-1}{d}})$, and is singular over a subspace of dimension $\frac{\alpha_0(a)}{\operatorname{meas}(K)} + O(n^{\frac{d-1}{d}})$. The claimed thesis follows from this as in Theorem 3.4. ∎

THEOREM 7.6. *Let a be a continuous function over K. Under the notation given in Definition* 3.5 *the relationships*

$$\begin{aligned} N^+(X_n(a)) &= n\tfrac{\alpha_+(a)}{\operatorname{meas}(K)} + o(n), \\ N^-(X_n(a)) &= n\tfrac{\alpha_-(a)}{\operatorname{meas}(K)} + o(n), \\ \dim(\operatorname{Ker}(X_n(a))) &= n\tfrac{\alpha_0(a)}{\operatorname{meas}(K)} + o(n) \end{aligned}$$

hold true when the following conditions are satisfied:

hp:b1: $A_0(a)$ is Peano–Jordan;

hp:b2: the sequence $\{X_n(\cdot)\}$ is locally X with respect to the partitionings $\mathcal{P}$.

Proof. It is enough to follow the same proof as in Theorem 7.5. □

Concerning the matrices $\mathcal{P}_n^*(a)$ and $\mathcal{P}_n^{**}(a)$ considered in Theorem 4.1, it should be observed that in their definition we use some quantities (see relation (18)) related to the vectors $\{\mathbf{s}_j^{(n)} = Z_n^H \mathbf{v}_j^{(n)}\}$ which arise in the case of "sample" matrices. In the case of FGR matrices this kind of definition makes no sense. Therefore, as suggested in Remark 4.2, we can consider the sequence $\{\mathcal{P}_n^{***}(a)\}_n$ whose construction depends only on the knowledge of function a over the grid points (in some sense this approach corresponds to approximating an integral over a local domain by a certain mean value). More specifically, $\mathcal{P}_n^{***}(a)$ is defined as $\mathcal{P}_n^*(a)$, where $D_n(a)$ is replaced by the diagonal matrix $\Delta_n(a)$ and $(\Delta_n(a))_{i,i}$ is given by $a(x_s^{(n)})$ or by $(\int_{\Omega_k^{(n)}} a)/\mathrm{meas}(\Omega_k^{(n)})$, $x_s^{(n)} \in \Omega_k^{(n)}$, and $\mathcal{N}_n(\Omega_k^{(n)}) = \mathbf{z}_i^{(n)}$.

Remark 7.1. It seems redundant to define the "locally X" matrices by using at the same time a sequence $\mathcal{G} = \{\mathcal{G}_n\}_n$ of grids and a sequence $\mathcal{P} = \{\mathcal{P}_n\}_n$ of partitionings of $K = \bar{\Omega}$ which are somehow related via the *uniformity* of the distributions. However, this "double" nature of the definition is useful to treat in a unified way both "sample" matrices and "local integral" matrices. The first kind of matrices that contain the FD and the graph matrices are defined through some samplings of a function (therefore the notion of grid sequence is necessary). The second kind of matrices that contain the FE matrices are defined through some weighted integrals over "local domains" so that the notion of a partitioning of the domain K is also required.

What the new definition reveals is that the two approaches to the discretization of differential problems (FD and FE methods) are in some sense similar, and the common thing is the locality of the approximation schemes. This fact is now put in evidence from a linear algebra point of view since the associated types of matrices are both "locally X" with regard to a certain sequence $X = \{X_n\}_n$ of matrices.

Acknowledgments. Hearty thanks go to Giuseppe Buttazzo and Paolo Tilli for discussing with me these and other things and to the referees for very helpful suggestions and remarks.

REFERENCES

[1] F. Avram, *On bilinear forms on Gaussian random variables and Toeplitz matrices*, Probab. Theory Related Fields, 79 (1988), pp. 37–45.

[2] O. Axelsson and G. Lindskog, *The rate of convergence of the preconditioned conjugate gradient method*, Numer. Math., 52 (1986), pp. 499–523.

[3] O. Axelsson and M. Neytcheva, *The algebraic multilevel iteration methods—theory and applications*, Proceedings of the Second Colloquium on Numerical Analysis, D. Bainov, ed., Plovdiv, Bulgaria, VSP, Utrecht, 1993, pp. 13–23.

[4] R. Bhatia, *Matrix Analysis*, Springer-Verlag, New York, 1997.

[5] D. A. Bini and B. Meini, *Effective methods for solving banded Toeplitz systems*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 700–719.

[6] R. Chan and T. Chan, *Circulant preconditioners for elliptic problems*, J. Numer. Linear Algebra Appl., 1 (1992), pp. 77–101.

[7] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[8] P. Concus, G. Golub, and G. Meurant, *Block Preconditioning for the Conjugate Gradient Method*, Tech. Rep. LBL-14856, Lawrence Berkeley Laboratory, University of California, Los Angeles, CA, 1982.

[9] D. Cvetkovic, M. Doob, and H. Sachs, *Spectra of Graphs*, Academic Press, New York, 1979.

[10] R. De Vore and G. Lorentz, *Constructive Approximation*, Springer-Verlag, Berlin, 1991.
[11] F. Di Benedetto, G. Fiorentino, and S. Serra, *C.G. preconditioning for Toeplitz matrices*, Comput. Math. Appl., 25 (1993), pp. 33–45.
[12] H. C. Elman and M. H. Shultz, *Preconditioning by fast direct methods for nonselfadjoint nonseparable elliptic equations*, SIAM J. Numer. Anal., 23 (1986), pp. 44–57.
[13] G. Fiorentino and S. Serra, *Multigrid methods for Toeplitz matrices*, Calcolo, 28 (1991), pp. 283–305.
[14] A. Frangioni and S. Serra Capizzano, *Matrix-Valued Linear Positive Operators and Applications to Graph Optimization*, Tech. Rep. 04-99, Dept. of Informatica, University of Pisa, Italy, 1999.
[15] U. Grenander and G. Szegő, *Toeplitz Forms and Their Applications*, 2nd ed., Chelsea, New York, 1984.
[16] I. Gustafsson, *Stability and Rate of Covergence of Modified Incomplete Cholesky Factorization Methods*, Tech. Rep. 79.02R, Chalmers University of Technology, Sweden, 1979.
[17] W. Hackbush, *Multigrid Methods and Applications*, Springer-Verlag, Berlin, 1985.
[18] C. Jordan, *Cours d'Analyse de l'Ecole Polytecnique: Vol.* I, Gauthier-Villars, Paris, 1909.
[19] S. Holmgren and K. Otto, *Iterative methods for block-tridiagonal systems of equations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 863–886.
[20] E. Isaacson and H. Keller, *Analysis of Numerical Methods*, John Wiley, New York, 1966.
[21] I. Lirkov, S. Margenov, and P. Vassilevsky, *Circulant block factorization for elliptic problems*, Computing, 53 (1994), pp. 59–74.
[22] J. Meijerink and H. Van Der Vorst, *An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix*, Math. Comp., 31 (1977), pp. 148–162.
[23] J. Meijerink and H. Van Der Vorst, *Guidelines for the usage of incomplete decompositions in solving sets of linear equations as they occur in practical problems*, J. Comput. Phys., 14 (1981), pp. 134–155.
[24] E. Moore, *General Analysis. Part* I, Amer. Phil. Society, Philadelphia, 1935.
[25] S. Parter, *On the distribution of singular values of Toeplitz matrices*, Linear Algebra Appl., 80 (1986), pp. 115–130.
[26] R. Penrose, *A generalized inverse for matrices*, Proc. Cambridge Philos. Soc., 51 (1955), pp. 406–413.
[27] W. Rudin, *Real and Complex Analysis*, McGraw-Hill, Singapore, 1986.
[28] S. Serra, *Preconditioning strategies for asymptotically ill-conditioned block Toeplitz systems*, BIT, 34 (1994), pp. 579–594.
[29] S. Serra, *The extension of the concept of generating function to a class of preconditioned Toeplitz matrices*, Linear Algebra Appl., 267 (1997), pp. 139–161.
[30] S. Serra, *Asymptotic results on the spectra of block Toeplitz preconditioned matrices*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 31–44.
[31] S. Serra, *The rate of convergence of Toeplitz based PCG methods for second order nonlinear boundary value problems*, Numer. Math., 81 (1999), pp. 461–495.
[32] S. Serra Capizzano, *Korovkin tests, approximation, and ergodic theory*, Math. Comp., in press.
[33] S. Serra Capizzano, *Some theorems on linear positive operators and functionals and their applications*, Comput. Math. Appl., 39 (2000), pp. 139–167.
[34] S. Serra Capizzano, *How bad can positive definite Toeplitz matrices be?*, Numer. Funct. Anal. Optim., 21 (2000), pp. 255–261.
[35] S. Serra Capizzano, *Spectral Behavior of Matrix-Sequences and Discretized Boundary Value Problems*, manuscript, 1998; a preliminary version appeared as Tech. Rep. 31, LAN, Dept. Math., Univ. of Calabria, Italy, 1998.
[36] S. Serra Capizzano and C. Tablino Possio, *Spectral and structural analysis of high order finite difference matrices for elliptic operators*, Linear Algebra Appl., 293 (1999), pp. 85–131.
[37] S. Serra Capizzano, *An ergodic theorem for classes of preconditioned matrices*, Linear Algebra Appl., 282 (1998), pp. 161–183.
[38] S. Serra Capizzano and C. Tablino Possio, *High Order Finite Difference Schemes and Toeplitz Based Preconditioners for Elliptic Problems*, manuscript, 1997.
[39] S. Serra Capizzano and C. Tablino Possio, *Superlinear Preconditioning of Optimal Preconditioners for Collocation Linear Systems*, manuscript, 1998.
[40] S. Serra Capizzano and P. Tilli, *From Partial Differential Equations to (Generalized) Multilevel Locally Toeplitz Sequences*, Tech. Rep. 12-99, Dept. of Informatica, Univ. of Pisa, Italy, 1999.
[41] P. N. Swarztrauber, *A direct method for the discrete solution of separable elliptic equations*,

SIAM J. Numer. Anal., 11 (1974), pp. 1136–1150.
[42] P. Tilli, *Singular values and eigenvalues of non Hermitian block Toeplitz matrices*, Linear Algebra Appl., 272 (1998), pp. 59–89.
[43] P. Tilli, *Locally Toeplitz sequences: Spectral properties and applications*, Linear Algebra Appl., 278 (1998), pp. 91–120.
[44] E. Tyrtyshnikov, *A unifying approach to some old and new theorems on distribution and clustering*, Linear Algebra Appl., 232 (1996), pp. 1–43.
[45] E. Tyrtyshnikov and N. Zamarashkin, *Spectra of multilevel Toeplitz matrices: Advanced theory via simple matrix relationships*, Linear Algebra Appl., 270 (1998), pp. 15–27.
[46] H. Widom, *Toeplitz matrices*, in Studies in Real and Complex Analysis, I. Hirshman, Jr., ed., Mathematical Association of America, Washington, DC, 1965.
[47] H. Widom, *On the singular values of Toeplitz matrices*, Z. Anal. Anwendungen, 8 (1989), pp. 221–229.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 4, pp. 1389–1391
© 2000 Society for Industrial and Applied Mathematics

A COUNTEREXAMPLE OF THE CONVEXITY OF SUMS OF SINGULAR VALUES*

LIU HONGWEI† AND LIU SANYANG†

Abstract. In this note we present a counterexample to Seeger's conjecture on the convexity of sums of singular values of rectangular matrices. Thus we answer the related open question in the negative.

Key words. convex function, conjecture, singular values

AMS subject classifications. 15A18, 49R05

PII. S0895479899361801

Let $K(x) = (k_{ij}(x))_{s\times r}$ be a rectangular matrix depending on a parameter $x \in E$ (E is a linear space) and denote by

$$g(x) := \sum_{i=1}^{m} \sigma_i(K(x))$$

the sum of m (with $1 \leqslant m \leqslant \min\{s, r\}$) greatest singular values

$$\sigma_1(K(x)) \geqslant \sigma_2(K(x)) \geqslant \cdots \geqslant \sigma_m(K(x))$$

of $K(x)$.

The following conjecture on the convexity of $g(x)$ is due to Seeger [1].

CONJECTURE. *Suppose that for all $i = 1, \ldots, s$ and $j = 1, \ldots, r$, $x \in E \to |k_{ij}(x)|$ is a (finite) convex function. Then*

$$x \in E \to \quad g(x) := \sum_{i=1}^{m} \sigma_i(K(x))$$

is also convex.

The conjecture is true only if $\min\{s, r\} = 1$. The proof follows immediately by Proposition 4.1 in [1] and the equality

$$\begin{aligned} g(x) &= \sigma_1(K(x)) \\ &= \lambda_1 \left(\begin{pmatrix} 0 & K(x) \\ K(x)^T & 0 \end{pmatrix} \right) \\ &= \sqrt{\sum_{j=1}^{r} k_{1j}^2(x)} \qquad \left(\text{or } \sqrt{\sum_{i=1}^{s} k_{i1}^2(x)} \right) \\ &= \sigma_1 \left((|k_{11}(x)|, |k_{12}(x)|, \ldots, |k_{1r}(x)|)_{1\times r} \right) \\ &\quad \left(\text{or } \sigma_1 \left((|k_{11}(x)|, |k_{21}(x)|, \ldots, |k_{s1}(x)|)_{1\times s}^T \right) \right). \end{aligned}$$

*Received by the editors October 1, 1999; accepted for publication (in revised form) by P. Van Dooren December 9, 1999; published electronically May 11, 2000.

http://www.siam.org/journals/simax/21-4/36180.html

†Department of Applied Mathematics, Xidian University, Xi'an 710071, People's Republic of China (liusanyang@263.net or syliu@mail.xidian.edu.cn).

When $\min\{s,r\} \geqslant 2$, we give a 2×2 counterexample which can trivially be embedded in larger ones to also disprove the result for larger values of m.

A counterexample. Let $K(x) = \left(\begin{smallmatrix} x^2 & x \\ x & x \end{smallmatrix}\right)$ with $x \in R$. Since $K(x)^T K(x) = \left(\begin{smallmatrix} x^4+x^2 & x^3+x^2 \\ x^3+x^2 & 2x^2 \end{smallmatrix}\right)$

$$\begin{aligned} |\lambda I - K(x)^T K(x)| &= \left|\lambda I - \begin{pmatrix} x^4+x^2 & x^3+x^2 \\ x^3+x^2 & 2x^2 \end{pmatrix}\right| \\ &= \begin{vmatrix} \lambda I-(x^4+x^2) & -(x^3+x^2) \\ -(x^3+x^2) & \lambda - 2x^2 \end{vmatrix} \\ &= \lambda^2 - (x^4+3x^2)\lambda + 2x^2(x^4+x^2) - (x^3+x^2)^2 \\ &= \lambda^2 - (x^4+3x^2)\lambda + x^6 + x^4 - 2x^5, \end{aligned}$$

then

$$\begin{aligned} \sigma_1(K(x)) &= \sqrt{\frac{x^4+3x^2+\sqrt{(x^4+3x^2)^2-4(x^6+x^4-2x^5)}}{2}} \\ &= \sqrt{\frac{(x^4+3x^2)+x^2\sqrt{x^4+2x^2+8x+5}}{2}}, \\ \sigma_2(K(x)) &= \sqrt{\frac{(x^4+3x^2)-x^2\sqrt{x^4+2x^2+8x+5}}{2}}, \\ \sigma_1(K(x)) &+\sigma_2(K(x)) = \sqrt{x^4+3x^2+2\sqrt{x^4+x^6-2x^5}}. \end{aligned}$$

When $m=1$

$$\begin{aligned} g(x) &= \sigma_1(K(x)), \\ g(0) &= \sigma_1(K(0)) = 0, \\ g(-1) &= \sigma_1(K(-1)) = \sqrt{2}, \\ g\left(-\frac{1}{2}\right) &= \sigma_1\left(K\left(-\frac{1}{2}\right)\right) = \frac{3}{4}, \end{aligned}$$

then

$$\frac{g(0)+g(-1)}{2} = \frac{0+\sqrt{2}}{2} = \frac{\sqrt{2}}{2} \leq \frac{3}{4} = g\left(-\frac{1}{2}\right) = g\left(0\times\frac{1}{2}+(-1)\times\frac{1}{2}\right).$$

Hence $g(x)$ is not convex for $m=1$.

When $m=2$

$$\begin{aligned} g(x) &= \sigma_1(K(x)) + \sigma_2(K(x)) \\ &= \sqrt{x^4+3x^2+2\sqrt{x^4+x^6-2x^5}}, \\ g(0) &= 0, \\ g(1) &= 2, \\ g\left(\frac{1}{2}\right) &= \frac{\sqrt{17}}{4}, \end{aligned}$$

then

$$\frac{g(0)+g(1)}{2}=\frac{0+2}{2}=1<\frac{\sqrt{17}}{4}=g\left(\frac{1}{2}\right)=g\left(0\times\frac{1}{2}+1\times\frac{1}{2}\right),$$

so $g(x)$ is not convex for $m=2$.

The above counterexample shows that Seeger's conjecture is not true.

REFERENCES

[1] A. Seeger, *Sensitivity analysis of nondifferentiable sums of singular values of rectangular matrices*, Numer. Funct. Anal. Optim., 16 (1995), pp. 247–260.

[2] D. Ye, *When is the greatest eigenvalue of a parameterized symmetric matrix a convex function of the parameter?*, Linear Algebra Appl., 169 (1992), pp. 103–109.

SIAM J. MATRIX ANAL. APPL.
Vol. 21, No. 4, pp. 1392–1393
© 2000 Society for Industrial and Applied Mathematics

A NOTE ON P-REGULAR SPLITTING OF HERMITIAN MATRIX*

ZHI-HAO CAO†

Abstract. In this note we examine the spectral property of the iteration matrix induced by the P-regular splitting of an invertible Hermitian matrix and point out that a theorem given by Weissler (cf. Theorem A2 in [*SIAM J. Matrix Anal. Appl.*, 16 (1995), pp. 448–461]) is incorrect.

Key words. P-regular splitting, spectral radius, positive definite matrix

AMS subject classification. 65F10

PII. S089547989935469X

First we recall the definition of the P-regular splitting (cf. [2]).

DEFINITION. *A splitting $A = M - N$ of a matrix A is called P-regular if matrix $M + N$ is positive definite.*

Obviously, matrix $M + N$ is positive definite if and only if matrix $M^* + N$ is positive definite. In particular, if A is a Hermitian matrix, then the splitting $A = M - N$ is P-regular if and only if $M + M^* - A$ is a Hermitian positive definite matrix.

In [3], first some historical remarks are given on the spectral property of the iteration matrix $T \equiv M^{-1}N$ induced by the P-regular splitting $A = M - N$ of matrix A, where A is an invertible Hermitian matrix, i.e., the so-called Householder–John theorem.

THEOREM A1. *Let A be an invertible Hermitian matrix, and let $A = M - N$ be a P-regular splitting of A. Then $\rho(M^{-1}N) < 1$ if and only if A is positive definite.*

The following slightly different result is given in [3].

THEOREM A2. *Let A be a Hermitian positive definite matrix, and let $A = M - N$ be a splitting of A. Then $\rho(M^{-1}N) < 1$ if and only if the splitting $A = M - N$ is P-regular.*

However, Theorem A2 is incorrect, as the following example shows.

Example. Let $A = I_2$, and $A = M - N$ is a splitting of A, where

$$M = \begin{pmatrix} \frac{1}{\epsilon} & -\frac{m}{\epsilon^2} \\ & \frac{1}{\epsilon} \end{pmatrix}, \quad N = \begin{pmatrix} \frac{1}{\epsilon} - 1 & -\frac{m}{\epsilon^2} \\ & \frac{1}{\epsilon} - 1 \end{pmatrix};$$

then

$$M^{-1} = \begin{pmatrix} \epsilon & m \\ & \epsilon \end{pmatrix} \quad \text{and} \quad M^{-1}N = \begin{pmatrix} 1-\epsilon & -m \\ & 1-\epsilon \end{pmatrix}.$$

Let ϵ be such that $0 < \epsilon < 1$: then $\rho(M^{-1}N) = 1 - \epsilon < 1$.

On the other hand, we have

$$M + M^* - A = \begin{pmatrix} \frac{2}{\epsilon} - 1 & -\frac{m}{\epsilon^2} \\ -\frac{m}{\epsilon^2} & \frac{2}{\epsilon} - 1 \end{pmatrix}.$$

*Received by the editors April 12, 1999; accepted for publication by P. Van Dooren August 12, 1999; published electronically May 11, 2000.

http://www.siam.org/journals/simax/21-4/35469.html

†Department of Mathematics, Fudan University, Shanghai 200433, People's Republic of China (zcao@fudan.edu.cn). This work was supported by China State Major Key Project for Basic Research, the Experimental Foundation of the Laboratory of Computational Physics, and the Doctoral Point Foundation of China.

Let m satisfy $m \geq \epsilon(2-\epsilon)$, then matrix $M + M^* - A$ is not positive definite, i.e., $A = M - N$ is not a P-regular splitting. ∎

In fact, the correct result should be the following theorem which should credit Young (cf. [4, Theorem 5.3]). We give a proof for completeness and also because of its simplicity.

THEOREM A3. *Let A be a Hermitian positive definite matrix, and let $A = M - N$ be a splitting of A. Then $\|M^{-1}N\|_{A^{\frac{1}{2}}} < 1$ if and only if the splitting $A = M - N$ of A is P-regular.*

Proof. Let $T = M^{-1}N$; then we have

$$A - T^*AT = (M^{-1}A)^*(M + M^* - A)(M^{-1}A). \tag{1}$$

Multiplying (1) on the left and right by $A^{-\frac{1}{2}}$ results in the following

$$I - (A^{\frac{1}{2}}TA^{-\frac{1}{2}})^*(A^{\frac{1}{2}}TA^{-\frac{1}{2}}) = (M^{-1}A^{\frac{1}{2}})^*(M + M^* - A)(M^{-1}A^{\frac{1}{2}}). \tag{2}$$

Since $\|A^{\frac{1}{2}}TA^{-\frac{1}{2}}\|_2 = \|M^{-1}N\|_{A^{\frac{1}{2}}}$, from (2) we deduce immediately the conclusion of the theorem. ∎

Finally, we point out that a related result is given by Ortega (cf. [1, Theorem A.2.6]).

THEOREM A4. *Let A be an invertible Hermitian matrix, and let $A = M - N$ be a splitting of A. If M is Hermitian positive definite and $\rho(M^{-1}N) < 1$, then A is positive definite and the splitting $A = M - N$ is P-regular.*

REFERENCES

[1] J. M. ORTEGA, *Introduction to Parallel and Vector Solution of Linear Systems*, Plenum Press, New York, 1988.

[2] J. M. ORTEGA, *Numerical Analysis, A Second Course*, Academic Press, New York, 1972.

[3] F. B. WEISSLER, *Some remarks concerning iterative methods for linear systems*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 448–461.

[4] D. M. YOUNG, *Iterative Solution of Large Linear Systems*, Academic Press, New York, 1971.