stichting
mathematisch
centrum

MC

AFDELING NUMERIEKE WISKUNDE
(DEPARTMENT OF NUMERICAL MATHEMATICS)

NW 56/78 MEI

K. DEKKER

SEMI-DISCRETIZATION METHODS FOR PARTIAL DIFFERENTIAL EQUATIONS ON NON-RECTANGULAR GRIDS

Preprint

2e boerhaavestraat 49 amsterdam

Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O).

AMS(MOS) subject classification scheme (1970): 65M05, 65N05

Semi-discretization methods for partial differential equations on non-rectangular grids *)

by

K. Dekker

SUMMARY

This paper presents a semi-discretization method for two-dimensional partial differential equations, applicable to curvilinear meshes. The method yields nine-point approximation formulas for the first and second derivatives of a function. Error estimates for another discretization method are given, and both methods are compared in numerical examples. The new method turns out to be more accurate in our examples, whereas the calculation of the weights of the formulas is more time-consuming.

KEY WORDS & PHRASES: *numerical analysis, partial differential equations, finite differences, curvilinear grids.*

---

*) This report will be submitted for publication elsewhere.

INTRODUCTION

The problem of approximating the solution of time-dependent partial differential equations (PDE) is often solved by using direct grid methods, e.g. the alternating direction, the locally one-dimensional, and the hop-scotch methods (Gourlay[2], Yanenko[10]). An alternative approach consists in splitting the problem into two subproblems: firstly, transforming the PDE into a system of ordinary differential equations (ODE) by discretizing the space variables, and secondly, solving the resulting system of ODE's with a suitable integrator. At the moment several packages based on this idea are available (Schryer[6], Sincovec[7]). At the Mathematical Centre some investigation is done in this area, too. Several generalized splitting methods have been constructed (Van der Houwen[3]), which can integrate ODE's with five - and nine-point coupling, thereby setting a need for semi-discretization methods which yield a nine-point coupling.

On a square mesh, the approximation of the first and second derivatives of the dependent variable may be obvious. However, on non-rectangular grids it is not at all clear which finite difference formula is the best. Therefore, we will compare three semi-discretization methods in this paper. The first one is a special case of a method published recently by (Frey[1]), for which we give an alternative formulation admitting strict error bounds. The second method is developed by Kok e.a.[4], originally intended for grids with an explicitly given mesh, and the third one is a new method, which tries to minimize the truncation error of the derivative-approximations.

In the next sections we will derive and summarize the formulas on which the three methods are based. Moreover, we calculate error bounds for Frey's method.

In the last section we give some numerical results of the three methods. First we compute the error terms for a variety of different grids, then we calculate the errors in the derivatives of a set of analytically given functions on a fixed grid, and finally we compute the solution of an elliptic PDE using the three methods.

THE TRANSFORMATION METHOD OF FREY

Let a curvilinear grid $R \subset \mathbb{R}^2$ be given, together with the function values of a sufficiently differentiable function $u(x,y)$ at the grid points. The problem of approximating the first and second derivatives of u in an interior grid point, can be solved by considering a transformation T from an element E of R to a square element E' (see figure 1). This method was proposed by Frey; we will formulate it now in a slightly different notation.

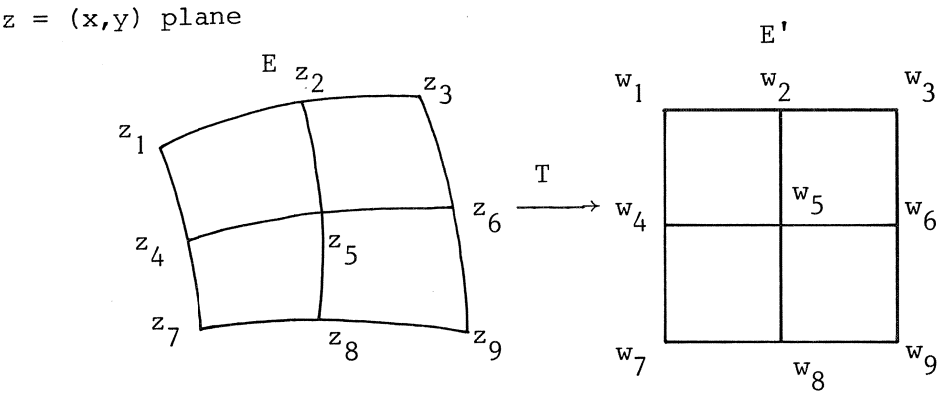

Figure 1. Curvilinear and square element

As T is an only locally defined transformation, we may assume without loss of generality that z_5 and w_5 are the origins of their respective planes. Then, we define the grid distance Δ and the element E' as follows.

DEFINITION 1.

(1) $\Delta = \max_{i=1,\ldots,9} \|z_i\|_1$; here $\|\ \|_1$ denotes the maximum norm, and z denotes the vector $(x,y)^T$.

The grid points of E' are given by

(2) $w_{i-3j+5} = \begin{pmatrix} i\Delta \\ j\Delta \end{pmatrix}$; $i = -1,0,1$; $j = -1,0,1$.

Now, we can express the transformation T as a Taylor-series about z_5, thus introducing a linear operator A and a bilinear operator B:

(3) $Tz = Az + \frac{1}{2}Bz^2 + O(z^3)$.

When the operators A and B are known, we can express the derivatives of u in z_5 by (using the chain-rule for differentiation)

$$\begin{Bmatrix} u_x \\ u_y \end{Bmatrix} = A^T \begin{Bmatrix} u_X \\ u_Y \end{Bmatrix}, \tag{4}$$

$$\begin{bmatrix} u_{xx} & u_{xy} \\ u_{yx} & u_{yy} \end{bmatrix} = A^T \begin{bmatrix} u_{XX} & u_{XY} \\ u_{YX} & u_{YY} \end{bmatrix} A + \langle u_X u_Y \rangle B. \tag{5}$$

Here, we assume u to be the function on E' defined by $u(w) = u(z)$ if $w = Tz$, and using central difference formulas on E', we can calculate the approximations for u_x, u_y, u_{xx}, u_{xy} and u_{yy} on E.

In order to compute approximations to A and B, we consider an operator $\tilde{S}$: E' → E, such that $\tilde{S}w_i = z_i$ for i = 2,4,5,6,8, and $\tilde{S}w_i = z_i + 0(\Delta^3)$ for the other gridpoints. It is easily verified, that $\tilde{S}$ given by (6) satisfies this conditions:

$$\tilde{S}w = \tilde{C}w + \tfrac{1}{2}\tilde{D}w^2, \quad \text{with} \tag{6}$$

$$\tilde{C} = \frac{1}{2\Delta} \begin{bmatrix} x_6 - x_4 & x_2 - x_8 \\ y_6 - y_4 & y_2 - y_8 \end{bmatrix}, \tag{7}$$

$$\tilde{D} = \frac{1}{\Delta^2} \begin{bmatrix} x_6 + x_4 & \dfrac{-x_1 + x_3 + x_7 - x_9}{4} & \dfrac{-x_1 + x_3 + x_7 - x_9}{4} & x_2 + x_8 \\ y_6 + y_4 & \dfrac{-y_1 + y_3 + y_7 - y_9}{4} & \dfrac{-y_1 + y_3 + y_7 - y_9}{4} & y_2 + y_8 \end{bmatrix}. \tag{8}$$

Now, recalling that $Tz_i = w_i$, we see that $\tilde{S}$ is an approximation to T^{-1}, so that we are able to compute approximations $\tilde{A}$ and $\tilde{B}$ to A and B, using

$$z = \tilde{S}(\tilde{A}z + \tfrac{1}{2}\tilde{B}z^2 + 0(z^3)). \tag{9}$$

Comparing terms of the same order in (9), we finally find

(10) $\tilde{A} = \tilde{C}^{-1}$ and

(11) $\tilde{B} = -\tilde{A}\tilde{D}\tilde{A}\tilde{A}$.

Substituting $\tilde{A}$ for A and $\tilde{B}$ for B in (4) and (5), we obtain approximations $\tilde{u}_x$ to u_x etc., which are exactly the same as the equations (11)-(14) given by Frey, as some lengthy calculations may reveal. However, we did choose for the above formulation because it enables us to derive error bounds for the approximations.

In order to perform the error analysis we first notice that the inverse of T can be written as (regarding x and y as functions of X and Y)

$$(12)\qquad T^{-1}w = Sw = \begin{bmatrix} x_X & x_Y \\ y_X & y_Y \end{bmatrix} w + \tfrac{1}{2}\begin{bmatrix} x_{XX} & x_{XY} & x_{YX} & x_{YY} \\ y_{XX} & y_{XY} & y_{YX} & y_{YY} \end{bmatrix} w^2 + O(w^3)$$

$$= Cw + \tfrac{1}{2}Dw^2 + O(w^3).$$

Expanding x and y in a Taylor series w_5, we obtain (for small Δ)

$$(13)\qquad \delta C = \tilde{C}-C = \frac{\Delta^2}{6}\begin{bmatrix} x_{XXX} & x_{YYY} \\ y_{XXX} & y_{YYY} \end{bmatrix} + O(\Delta^4),$$

$$\delta D = \tilde{D}-D = \frac{\Delta^2}{12}\begin{bmatrix} x_{XXXX} & 2(x_{XXXY}+x_{XYYY}) & 2(x_{XXXY}+x_{XYYY}) & x_{YYYY} \\ y_{XXXX} & 2(y_{XXXY}+y_{XYYY}) & 2(y_{XXXY}+y_{XYYY}) & y_{YYYY} \end{bmatrix} + O(\Delta^4).$$

Hence, we have $\|\delta C\| \le k_c \Delta^2$ and $\|\delta D\| \le k_c \Delta^2$. Using $A = C^{-1}$, $B = -ADAA$, we find after some calculation (cf. Wilkinson[8])

$$(14)\qquad \delta A = \tilde{A}-A = -A\delta C(I-\tilde{A}\delta C)^{-1}\tilde{A}, \text{ or } \|\delta A\|_2 \le \frac{\|\tilde{A}\|_2^2}{1-\|\tilde{A}\|_2 k_c \Delta^2} k_c \Delta^2 = k_a \Delta^2,$$

$$(15)\qquad \|\delta B\|_2 = \|\tilde{B}-B\|_2 \le \|\tilde{A}\|_2^2 \left\{ \frac{3k_a}{\{1-\|A\|_2 k_c \Delta^2\}^2} \|\tilde{D}\|_2 + \|\tilde{A}\|_2 k_d \right\} \Delta^2 = k_b \Delta^2.$$

Obviously, the errors δC and δD depend on the curvature of the gridlines, and the variation in the distance between the gridpoints; δA and δB are influenced, too, by the condition of the transformation, and in consequence on the angle between the gridlines.

Now, let $\tilde{u}_X$ etc. denote the central difference approximation to u; using the error bounds given above, and the formulas (4), (5), we obtain the error estimates

$$(16) \qquad \left\| \begin{Bmatrix} \tilde{u}_x \\ \tilde{u}_y \end{Bmatrix} - \begin{Bmatrix} u_x \\ u_y \end{Bmatrix} \right\|_2 \le \left\{ k_a \left\| \begin{matrix} \tilde{u}_x \\ \tilde{u}_y \end{matrix} \right\|_2 + \frac{1}{6} \|\tilde{A}\|_2 \left\| \begin{matrix} u_{XXX}(\xi) \\ u_{YYY}(\eta) \end{matrix} \right\|_2 \right\} \Delta^2,$$

where ξ and η are certain points between w_4 and w_6, w_2 and w_7 respectively. In order to compute the error in the second derivative, we observe that

$$(17) \qquad \tilde{S}w_i = z_i + \delta z_i, \quad i = 1,3,7,9,$$

with e.g. $\delta z_3 = \langle -\tfrac{1}{2}x_{XXY} - \tfrac{1}{2}x_{XYY}, -\tfrac{1}{2}y_{XXY} - \tfrac{1}{2}y_{XYY} \rangle^T \Delta^3 + O(\Delta^4)$

and $$\sum_i \delta z_i = O(\Delta^4).$$

Thus, the error in $u_{XY} - \tilde{u}_{XY}$ is determined not only by the discretization but also by the evaluation of u in the points z_i instead of $\tilde{S}w_i$. Using $\tilde{u}_{XY} = \frac{1}{4\Delta^2}\{u_3+u_7-u_1-u_9\}$, this last error is given by

$$(18) \qquad \left| \frac{1}{4\Delta^2} \left\{ \langle u_x\, u_y \rangle (\delta z_3 + \delta z_7 - \delta z_1 - \delta z_9) + \begin{bmatrix} u_{xx} & u_{xy} \\ u_{yx} & u_{yy} \end{bmatrix} O(\Delta^4) \right\} \right| = c\Delta^2$$

because $\delta z_3 + \delta z_7 - \delta z_1 - \delta z_9 = 0$. Finally we get

$$(19) \qquad \left\| \begin{bmatrix} \tilde{u}_{xx} & \tilde{u}_{xy} \\ \tilde{u}_{yx} & u_{yy} \end{bmatrix} - \begin{bmatrix} u_{xx} & u_{xy} \\ u_{yx} & u_{yy} \end{bmatrix} \right\|_2 \le \left[\|A\|_2^2 \left\{ \frac{1}{12} \left\| \begin{bmatrix} u_{XXXX} & 2(u_{XXXY}+u_{XYYY}) \\ 2(u_{XXXY}+u_{XYYY}) & u_{YYYY} \end{bmatrix} \right\|_2 + c \right\} \right.$$

$$\left. + 2 \left\| \begin{bmatrix} \tilde{u}_{xx} & \tilde{u}_{xy} \\ \tilde{u}_{yx} & \tilde{u}_{yy} \end{bmatrix} \right\|_2 \|A\|_2 k_a + \|\tilde{B}\|_2 \left\| \begin{matrix} u_{XXX} \\ u_{YYY} \end{matrix} \right\|_2 + \left\| \begin{matrix} \tilde{u}_X \\ \tilde{u}_Y \end{matrix} \right\|_2 k_b \right] \Delta^2.$$

The estimates (16) and (19) show that the finite difference approximations are of second order, just as is the case for square elements. However,

for curvilinear elements the error constants will be larger, as more terms occur in the error bounds.

A TRANSFORMATION METHOD WITH EXPLICITLY GIVEN GRIDLINES

The method of Kok[4] is based on the idea, that the gridlines are known functions of the coordinates z and y, such that their derivatives can be obtained easily. When the gridlines are not known, they are locally approximated; to that end the lines in the x-direction are considered to be functions of x, and a three-point interpolation formula yields

$$(20) \qquad y = f_-(x) = y_7 + \frac{y_8-y_7}{x_8-x_7}(x-x_7) + \frac{\frac{y_9-y_8}{x_9-x_8} - \frac{y_8-y_7}{x_8-x_7}}{x_9-x_7}(x-x_7)(x-x_8),$$

and similarly for f_+, f, g_+, g and g_- (see figure 3).

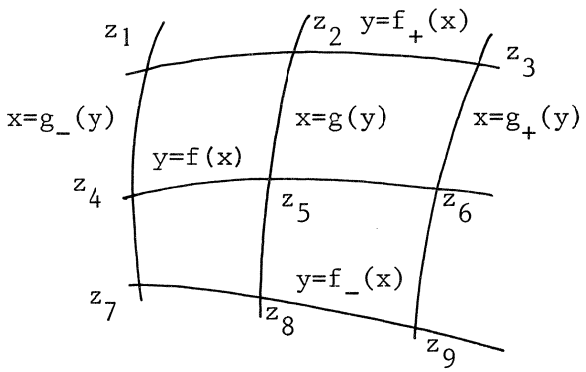

Figure 2. The gridlines as second order interpolation polynomials

Now, we will denote by f,f' etc. the function and derivative values at the point x_5, and similarly by g,g' the values at the point y_5. Using the same notation as in the previous section, the derivatives of a function u at the point $z_5 = (x_5, y_5)$ are given by the formulas (4) and (5), with matrix A defined by

$$(21) \qquad A_{11} = \frac{2\Delta}{g_+-g_-}, \qquad A_{22} = \frac{2\Delta}{f_+-f_-},$$

$$A_{12} = -g'A_{11}, \qquad A_{21} = -f'A_{22},$$

and B defined by

$$B_{111} = -8\,\frac{g_+-2g+g_-}{(g_+-g_-)^3}\,\Delta, \qquad B_{222} = -8\,\frac{f_+-2f+f_-}{(f_+-f_-)^3}\,\Delta,$$

$$(22)\qquad B_{112} = B_{121} = B_{111}\,\frac{A_{12}}{A_{11}} - A_{11}^2\,\frac{g_+'-g_-'}{2\Delta}, \quad B_{212}=B_{221}=B_{222}\,\frac{A_{21}}{A_{22}} - A_{22}\,\frac{f_+'-f_-'}{2\Delta},$$

$$B_{122} = -\{(g')^2B_{111}+2g'B_{112}+g''A_{11}\}, \quad B_{211} = -\{(f')^2B_{222}+2f'B_{221}+f''B_{22}\}.$$

For a detailed derivation of these formulas we refer to Kok[4].

THE MINIMIZATION METHOD

In the previous sections we have described two discretization method based on a transformation of the elements. Here, we will follow an alternative approach. An approximation formula might be regarded as a function of nine parameters, the weights in the points z_i, $i = 1,\ldots,9$. Expanding each function value in z_i as a Taylor-series, we obtain for the approximation formula a Taylor-series about z_5. Each term of this series has a coefficient depending on several of the weights used in the formula. Now, we may wish these coefficients to have a predescribed value, for example one for the derivative to be approximated, and zero for the other coefficients up to and including third order. However, we obtain ten equations with nine unknowns in this way, and a solution generally does not exist. (Here, we note that in the rectangular case, a (not unique) solution exists). Restricting ourselves to the lower order terms, we get only six equations, and we have a wide variety of solutions to them. In the following we will compute that solution to the latter system, which minimizes the error constants of the third order terms, and hope that this solution yields a good approximation to the required derivative.

First, we introduce the following notations

$$u^T = \langle u_1-u_5, u_2-u_5, \ldots, u_4-u_5, u_6-u_5, \ldots, u_9-u_5\rangle,$$

$$(23)\qquad d^T = \langle u_x, u_y, \frac{u_{xx}}{\sqrt{2}}, u_{xy}, \frac{u_{yy}}{\sqrt{2}}, \frac{u_{xxx}}{\sqrt{6}}, \frac{u_{xxy}}{\sqrt{2}}, \frac{u_{xyy}}{\sqrt{2}}, \frac{u_{yyy}}{\sqrt{6}}\rangle,$$

$$\tilde{d}^T = \langle \tilde{u}_x, \tilde{u}_y, \frac{\tilde{u}_{xx}}{\sqrt{2}}, \tilde{u}_{xy}, \frac{\tilde{u}_{yy}}{\sqrt{2}} \rangle,$$

$$d_s^T = \langle u_x, u_y, \frac{u_{xx}}{\sqrt{2}}, u_{xy}, \frac{u_{yy}}{\sqrt{2}} \rangle.$$

Here, u_i denotes the function value in z_i (cf. figure 1), u_x, etc. the derivatives in z_5, and $\tilde{u}_x$ etc. approximations to these derivatives. Further we note that we have added the factors $\sqrt{2}$ and $\sqrt{6}$, in order to have a rotation independent Euclidean norm $\| \ \|_E$. This is illustrated by the following example.

EXAMPLE 1. Consider the function $u(x,y) = x^2-y^2$. The second derivatives are $\langle u_{xx}, u_{xy}, u_{yx}, u_{yy} \rangle = \langle 2,0,0,2 \rangle$ and their Euclidean norm is $2\sqrt{2}$. A rotation of $\frac{\pi}{4}$ yields the function $u(\xi,\eta) = 2\xi\eta$, with second derivatives given by $\langle 0,2,2,0 \rangle$, again with Euclidean norm $2\sqrt{2}$. As u_{xy} equals u_{yx}, the latter term can be deleted to shorten the notation. However, the norm is influenced by this deletion and $\| \langle u_{xx}, u_{xy}, u_{yy} \rangle \|_E$ is no longer rotation independent. Thus, we have to divide u_{xx} and u_{yy} by a factor $\sqrt{2}$ in order to preserve this property.

Now, a nine-point approximation method is defined by a matrix of weights W(5×8)

$$\tilde{d} = Wu. \tag{24}$$

Furthermore, expanding u_i in a Taylor-series around z_5, we have

$$u = Md + O(\Delta^4), \tag{24}$$

where M is a 8×9 matrix, only depending on the size and shape of the element E (see figure 1). The error in the approximation can be expressed by

$$\tilde{d} - d_s = WMd + \text{higher order terms} - d_s, \tag{26}$$

and we call F defined by

$$F_{ij} = (WM)_{ij} - \delta_{ij}, \quad i = 1,\ldots,5, \quad j = 1,\ldots,9, \tag{27}$$

the error matrix. Obviously, F depends only on the given element, and the approximation method chosen, and the minimizing problem is nothing else than constructing a kind of inverse of M. To be more precise, we want to minimize $\|F\|_E$ under the constraints $F_{ij} = 0$, $i,j = 1,\ldots,5$ (these are the first and second order terms), as is illustrated in figure 3.

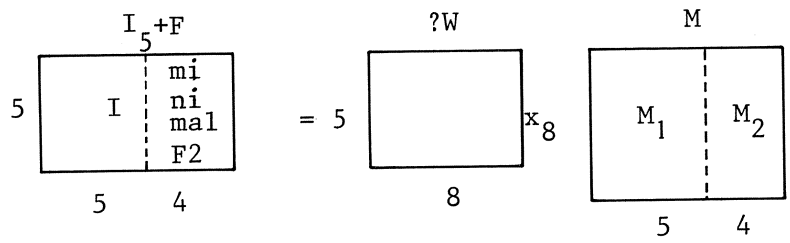

Figure 3. The minimizing problem

The solution may be found by numerical algebra techniques. First, we compute the pseudo inverse M_1^+ of M_1 (Wilkinson & Reinsch[9]), such that $M_1^+M_1 = I$. Now, let $M_1^\perp$ be the matrix consisting of the three vectors which are orthogonal to M_1. Then we solve the overdetermined systems

$$(28) \qquad (M_2^T M_1^\perp)\ X = (M_1^+ M_2)^T, \quad X \text{ a 3 by 5 matrix,}$$

in the least squares sense, and the desired optimal solution W_{opt} is given by

$$(29) \qquad W_{opt} = M_1^+ - (M_1^\perp X)^T.$$

It is easily verified that this solution is optimal, by considering

$$(30) \qquad F_2 = (M_1^+ - (M_1^\perp X)^T)M_2.$$

In the next sections, the computations of W_{opt} were made by using the NAG-library[5] routine F01BHF, which performs singular value decompositions.

REMARKS.

1. We note that the matrix M_1 has rank 5, if and only if no quadratic form exists, which is satisfied by the nine points z_i. As a quadratic form is defined by five points, M_1 will always have full rank, unless the grid is chosen very awkwardly (e.g. all points on a circle, on two straight lines). Thus M_1^+ indeed exists.

2. It may be advisable to scale the matrix M before executing the formulas (28) and (29), in order to avoid ill-conditioning. Division of the first two columns of M by Δ, the next three by Δ^2 and the last four by Δ^3 (Δ as defined in (1)) will be appropriate. Afterwards, the first two rows of W_{opt} should be divided by Δ, the next ones by Δ^2.

3. In the derivation of W_{opt}, we did not use any specific property of the nine points given, except that they did not lie on a quadratic form. Thus, the method can be used for boundary points, too, when we select 8 points in the neighbourhood of the boundary point, all lying within the domain or on the boundary (see figure 4). (Note that the numbering is irrelevant).

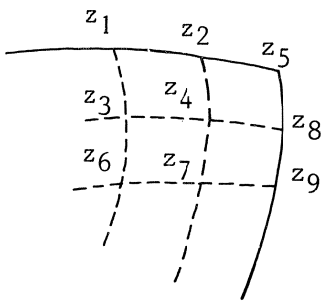

Figure 4. An element on the boundary

However, as the distances z_i-z_5 are larger for boundary points than for interior points, the approximations will be less accurate.

4. It may happen that the matrix $(M_2^T M_1^\perp)$ is not of full rank. This case occurs if the three-dimensional space $M_1^\perp$ is not spanned by the columns of M_2, or equivalently, if the nine columns of M do not span the whole space $\mathbb{R}^8$. For example, when the element is uniform and rectangular, M spans only a 7-dimensional subspace of $\mathbb{R}^8$.

When the above described situation arises, the solution of equation (28) is not unique; using the pseudo-inverse of $(M_2^T M^\perp)$, we obtain the unique solution vector with minimal norm in this solution space, for each of the five columns of the right hand side. However, on square elements we do not obtain the usual central difference formulas by this method, although these formulas obviously lie in the solution space of (28).

This difficulty can be overcome by adding the vector m_{10} with elements $x_i^2 y_i^2$ to the matrix M. When this vector is not linearly dependent on the other columns of M, we will obtain within the solution space of (28) the vector which is perpendicular to m_{10}.

For the proof, that the new solution lies in the original solution space, we refer to the appendix.

NUMERICAL EXAMPLES

In this section we will give some numerical results, produced by the methods described in the previous sections. These methods will be denoted by their generating weight-matrices W_2 (Frey's method), W_3 (Kok's method) and W_4 (the minimization method).

In the first subsection we will compute the entries of the error matrix F defined by (27) for various elements. Here, it turns out that W_2, W_3 and W_4 produce identical error matrices on rectangular elements, whereas the F generated by W_4 has the smallest norm on curvilinear elements.

In the next subsection we approximate the derivatives of a set of analytic functions with the various methods on several non-rectangular grids. Again, W_4 gives the best results, as it shows the highest order of convergence when the functions are smoothed.

Finally, we solved an elliptic problem on a curvilinear grid, and tabulated the error between the analytic solution and the solution of the discrete system for various numbers of gridlines.

A. The error matrices F for the methods W_2, W_3 and W_4

In this series of tests we computed the entries of F for elements given

by the formulas (see also figure 5)

$$\begin{Bmatrix} x_6 \\ y_6 \end{Bmatrix}, \begin{Bmatrix} x_4 \\ y_4 \end{Bmatrix} = dx \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \begin{Bmatrix} 1 + c_1 d_x \\ c_2 dx \end{Bmatrix}, \begin{Bmatrix} -1 + c_1 dx \\ c_2 dx \end{Bmatrix},$$

$$\begin{Bmatrix} x_2 \\ y_2 \end{Bmatrix}, \begin{Bmatrix} x_8 \\ y_8 \end{Bmatrix} = dy \begin{bmatrix} \cos(\alpha+\beta) & -\sin(\alpha+\beta) \\ \sin(\alpha+\beta) & \cos(\alpha+\beta) \end{bmatrix} \begin{Bmatrix} 1 + c_3 dy \\ c_4 dy \end{Bmatrix}, \begin{Bmatrix} -1 + c_3 dy \\ c_4 dy \end{Bmatrix}.$$

(31)

$$z_1 = z_2 + z_4 + c_5\ dxdy \begin{Bmatrix} dx \\ dy \end{Bmatrix}, \quad z_3 = z_2 + z_6 + c_5\ dxdy \begin{Bmatrix} dx \\ -dy \end{Bmatrix},$$

$$z_7 = z_4 + z_8 + c_5\ dxdy \begin{Bmatrix} -dx \\ dy \end{Bmatrix}, \quad z_9 = z_6 + z_8 + c_5\ dxdy \begin{Bmatrix} -dx \\ -dy \end{Bmatrix}.$$

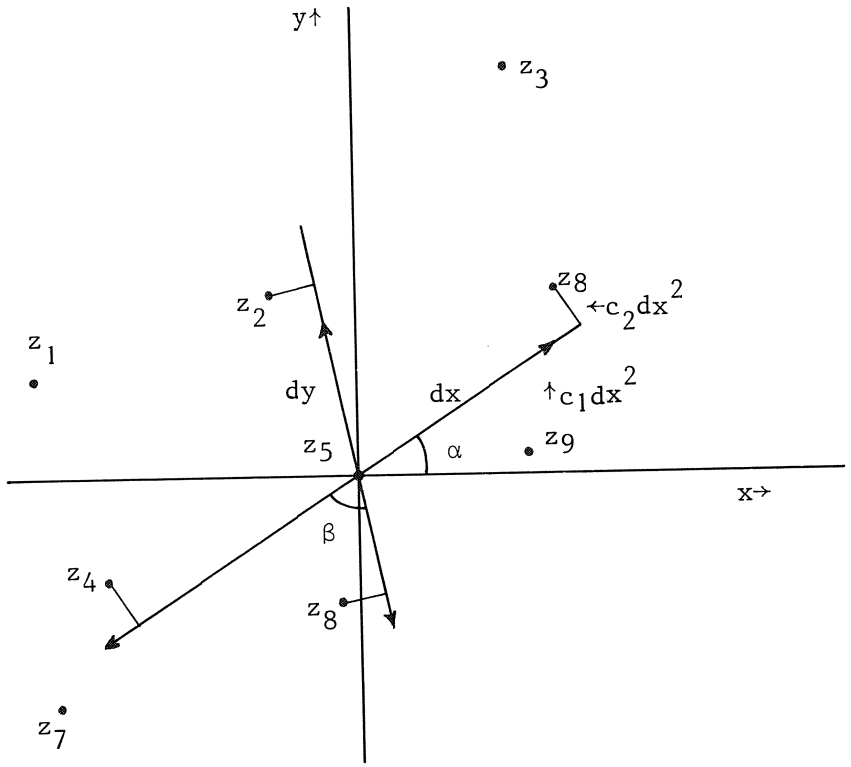

Figure 5. The element defined by (31)

For the parameters in (31) we have chosen the default values

$$
\begin{aligned}
&\alpha = 0,\\
&\beta = \pi/2,\\
&dx = dy = 1,\\
&c_1 = c_2 = c_3 = c_4 = c_5 = 0,
\end{aligned}
\tag{32}
$$

and in each test we varied some of these parameters, namely

(33)

a) $\beta = \frac{i\pi}{18}$, $i = 1,\ldots,9$, yielding a diamond,

b) $dy = \frac{i}{10}$, $i = 1,\ldots,10$, yielding a rectangle,

c) $dy = \frac{i}{10}$, $\beta = \frac{10+i}{40}\,\pi$, $i = 1,\ldots,10$, yielding a parallelogram,

d) $c_5 = \frac{i}{20}$, $i = 1,\ldots,10$, yielding a distorted square,

e) $c_1 = c_2 = c_3 = c_4 = \frac{i}{20}$, $i = 1,\ldots,10$, yielding a curvilinear element, in which each quadrilateral remains a parallelogram,

f) $dy = .5$, $\beta = \frac{\pi}{3}$, $c_1 = \frac{i}{30}$, $c_2 = -\frac{i}{20}$, $c_3 = \frac{i}{50}$, $c_4 = \frac{i}{40}$, $c_5 = \frac{i}{40}$, $i = 1,\ldots,10$, yielding a curvilinear element.

In the tables 1-6- we listed the non-zero values of f_{11}, f_{12}, f_{13}, f_{21}, f_{22} and f_{23} which are defined by the matrix F in the following way:

(34)

F a 5×9 matrix

	2	3	4
2	F_{11}	F_{12}	F_{13}
3	F_{21}	F_{22}	F_{23}

, $f_{ij} = \|F_{ij}\|_E$.

As the values of f_{ij} turned out to be independent of α, we dropped this parameter from the tests.

β	W_2	W_3	W_4
π/18	.810	.810	.014
2π/18	.792	.792	.057
3π/18	.764	.764	.126
4π/18	.727	.727	.216
5π/18	.686	.686	.320
6π/18	.646	.646	.423
7π/18	.610	.610	.507
8π/18	.586	.586	.560
π/2	.577	.577	.577

Table 1. f_{13} element for (33^a)

dy	W_2, W_3, W_4
.1	.408
.2	.409
.3	.410
.4	.413
.5	.421
.6	.434
.7	.455
.8	.485
.9	.525
10	.577

Table 2. f_{13} element (33^b)

dy	β	W_2	W_3	W_4
.1	11π/40	.536	.536	.245
.2	12π/40	.503	.503	.272
.3	13π/40	.478	.478	.299
.4	14π/40	.461	.461	.327
.5	15π/40	.453	.453	.358
.6	16π/40	.455	.455	.391
.7	17π/40	.467	.467	.429
.8	18π/40	.491	.491	.473
.9	19π/40	.527	.527	.522
10	π/2	.577	.577	.577

Table 3. f_{13} for element (33^c)

c_5	f_{13}			f_{22}		
	W_2	W_3	W_4	W_2	W_3	W_4
.05	.577	.577	.576	.100	.100	0
.1	.577	.577	.573	.200	.200	0
.15	.577	.577	.568	.301	.301	0
.2	.577	.577	.560	.402	.402	0
.25	.577	.577	.549	.504	.405	0
.3	.577	.577	.533	.607	.607	0
.35	.577	.577	.513	.711	.711	0
.4	.577	.577	.488	.816	.816	0
.45	.577	.577	.458	.922	.922	0
.5	.577	.577	.423	1.031	1.031	0

Table 4. f_{13} and f_{22} for element (33^d)

c_1	f_{12}			f_{13}			f_{22}			f_{23}		
	W_2	W_3	W_4	W_2	W_3	W_4	W_2	W_3	W_4	W_2	W_3	W_4
.05	.122	.122	0	.582	.582	.407	.010	.017	0	.182	.183	.064
.1	.245	.245	0	.595	.595	.404	.040	.070	0	.364	.365	.128
.15	.367	.368	0	.618	.619	.399	.090	.159	0	.543	.548	.191
.2	.490	.492	0	.653	.655	.392	.160	.286	0	.719	.734	.252
.25	.612	.618	0	.700	.706	.383	.250	.458	0	.893	.931	.310
.3	.735	.749	0	.760	.775	.371	.360	.683	0	1.063	1.151	.364
.35	.857	.890	0	.836	.868	.356	.490	.977	0	1.23	1.42	.414
.4	.980	1.05	0	.926	.991	.337	.640	1.37	0	1.40	1.77	.459
.45	1.10	1.24	0	1.03	1.16	.312	.810	1.92	0	1.56	2.28	.497
.5	1.22	1.47	0	1.15	1.39	.280	1.00	2.74	0	1.73	3.09	.530

Table 5. f_{12}, f_{13}, f_{22} and f_{23} for element (33^e)

c_2	f_{12}			f_{13}			f_{21}			f_{22}			f_{23}		
	W_2	W_3	W_4	W_2	W_3	W_4	W_2	W_3	W_4	W_2	W_3	W_4	W_2	W_3	W_4
-.05	.081	.081	0	.480	.480	.274	0	.004	0	.056	.056	0	.115	.115	.053
-.1	.163	.163	0	.485	.485	.268	0	.018	0	.118	.126	0	.228	.229	.107
-.15	.244	.245	0	.493	.494	.261	0	.043	0	.187	.219	0	.339	.344	.161
-.2	.326	.327	0	.506	.507	.252	0	.086	0	.265	.340	0	.449	.462	.217
-.25	.407	.410	0	.523	.525	.241	0	.156	0	.354	.495	0	.557	.586	.273
-.3	.488	.494	0	.547	.548	.229	0	.268	0	.454	.692	0	.664	.720	.329
-.35	.570	.578	0	.577	.579	.216	0	.444	0	.567	.940	0	.769	.873	.386
-.4	.651	.662	0	.615	.616	.201	0	.719	0	.692	1.25	0	.875	1.06	.443
-.45	.733	.748	0	.661	.661	.186	0	1.15	0	.830	1.65	0	.983	1.29	.500
-.5	.814	.833	0	.716	.714	.171	0	1.84	0	.981	2.16	0	1.09	1.60	.555

Table 6. f_{12}, f_{13}, f_{21} and f_{23} for element (33^f)

Finally we tested in which way the f_{ij} were influenced by a change of the scaling factors dx and dy. The results for the element defined by (33f) with i = 5 are given in table 7. Obviously, all error constants are $O(\Delta^2)$

dx=dy	$-^{10}\log f_{12}$			$-^{10}\log f_{13}$			$-^{10}\log f_{21}$			$-^{10}\log f_{22}$			$-^{10}\log f_{23}$			$-^{10}\log f_{11}$
	w_2	w_3	w_4	w_2	w_3	w_4	w_2	w_3	w_4	w_2	w_3	w_4	w_2	w_3	w_4	w_3
1	.4	.4	-	.3	.3	.6	-	.8	-	.5	.3	-	.3	.2	.6	1.0
$10^{-\frac{1}{2}}$	1.4	1.4	-	1.3	1.3	1.6	-	1.9	-	1.5	1.3	-	1.2	1.2	1.6	2.0
10^{-1}	2.4	2.4	-	2.3	2.3	2.6	-	2.9	-	2.5	2.3	-	2.2	2.2	2.6	3.0
$10^{-\frac{1}{2}}$	3.4	3.4	-	3.3	3.3	3.6	-	3.9	-	3.5	3.3	-	3.2	3.2	3.6	4.0
10^{-2}	4.4	4.4	-	4.3	4.3	4.6	-	4.9	-	4.5	4.3	-	4.2	4.2	4.6	5.0

Table 7. The dependence of f_{ij} on the gridsize for element (^{f}) with i = 5

B. The error in the approximation of known functions

The results of the previous subsection indicates that we may expect method w_4 to give the most accurate results in the approximation of the derivatives; using this method, the third derivatives of the function to be approximated occur in the truncation error with minimal constants. Here, we will investigate the actual error in the approximation, and for that end we have chosen the following set of testfunctions

(35) $$f_{i,j,\alpha}(x,y) = g_i(\alpha h_j(x,y)), \quad i = 1,\ldots,4,\ j = 1,2,\ \alpha\in[10^{-3},1],$$

with $$h_1=(x,y)=x+y,\ h_2(x,y)=x-y/2,\ f_1 = \sin,\ f_2 = \cos,\ f_3 = \exp,\ f_4(x) = (1+x)^3.$$

For each testfunction f, a given element and a given approximation method, we computed a mixed error $\varepsilon(f)$ defined by

(36) $$\varepsilon(f) = \frac{1}{5}\sqrt{\{(f_x-\tilde{f}_x)^2+(f_y-\tilde{f}_y)^2+\tfrac{1}{2}(f_{xx}-\tilde{f}_{xx})^2+\tfrac{1}{2}(f_{yy}-\tilde{f}_{yy})^2+(f_{xy}-\tilde{f}_{xy})^2\}},$$

where $\sim$ denotes the calculated approximation. To eliminate the influence of the function considered, we formed a mean value over the set of test-functions given by

$$(37) \qquad sd(\alpha) = \frac{1}{8} \sum_{i=1}^{4} \sum_{j=1}^{2} - \log(\varepsilon(f_{i,j,\alpha})).$$

The values of $sd(\alpha)$ are listed in table 8; the approximations were made on an element given by Frey' (see table 9) and on the square element defined by (31) and (32). In the latter case the three methods gave identical results as could be expected.

α	element of table 9			square element
	w_2	w_3	w_4	$w_2=w_3=w_4$
1	.97	.95	1.02	.82
$10^{-1/2}$	2.32	2.09	2.71	2.47
10^{-1}	3.49	2.81	4.36	4.10
$10^{-3/2}$	4.63	3.46	5.99	5.73
10^{-2}	5.57	4.09	7.62	7.35
$10^{-5/2}$	6.88	4.72	9.24	8.98
10^{-3}	8.01	5.34	10.88	10.60

Table 8. The values of $sd(\alpha)$ for the element given by Frey, and a square element

i	x_i	y_i
1	-.7500	.2813
2	-.1875	.6563
3	.5313	1.0000
4	-.6250	-.2813
5	0	0
6	.7138	.1875
7	-.5625	-.8433
8	.0938	-.7188
9	.8125	-.6875

Table 9. The coordinates of the element given by Frey

C. Application to an elliptic problem

We considered the equation

(38) $\Delta u = 2e^{x+y}$ on Ω,

$u(x,y) = e^{x+y}$ on $\partial\Omega$,

with the domain Ω given by (see figure 6): $\{(x,y)\,|\,|x|\le 1, -1\le y\le 0\}\cup\{(x,y)\,|\,x^2+y^2\le 1\}$

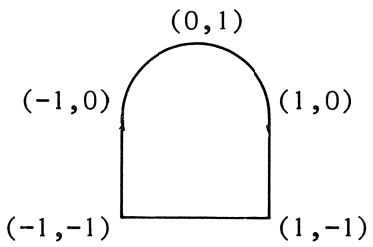

Figure 6. The domain Ω

The analytic solution is given by $u(x,y) = e^{x+y}$. We regarded the three straight lines and the arc as the four boundaries, and divided each one in N equal parts. Then we connected the corresponding points on the upper and lower boundary and defined the interior nodes to be

$$(39)\qquad z_{i,j} = \left(\frac{j}{N}\,\frac{2i-N}{N} - \frac{N-j}{N}\cos\frac{i\pi}{N},\ -\frac{j}{N} + \frac{N-j}{N}\sin\frac{i\pi}{N}\right)^T.$$

Of course we could have used a more sophisticated mesh generation scheme, but the subdivision given above is quite suitable for our testing purposes. Since the boundary conditions determine the values of u at the boundary nodes, we can set up a system of $(N-1)^2$ linear algebraic equations, by using the discretization methods from the previous sections. The linear systems were solved by successive overrelaxation. In table 10 we tabulated the differences between the analytic solution and the solution of the discrete problem, using the maximum norm.

N	$-^{10}\log$ max-error		
	w_2	w_3	w_4
2	1.5	1.5	1.4
3	1.6	0.7	1.6
4	1.3	0.7	1.7
5	1.3	0.8	1.9
6	1.4	0.9	2.0
8	1.6	1.0	2.3
10	1.8	1.2	2.5
12	1.9	1.3	2.6
16	2.1	1.4	2.9
20	2.3	1.5	3.1

Table 10. The maximum error in the approximation of the solution of (3.8)

Finally, we solved the same problem on domains with sizes $\frac{1}{\sqrt{10}}$ and $\frac{1}{10}$ of the original ones. We list the results in table 11, together with the computation time needed to set up the linear system, as measured on the Cyber 72 computer. Note, however, that these timings serve merely as an indication of the complexity of the methods; they were implemented only for testing purposes, and were not optimalized with respect to efficiency.

N	size $\frac{1}{\sqrt{10}}$ $-^{10}\log$ max-error			size $\frac{1}{10}$ $-^{10}\log$ max-error			time in seconds		
	w_2	w_3	w_4	w_2	w_3	w_4	w_2	w_3	w_4
2	2.4	2.4	3.4	3.3	3.5	5.4	.02	.02	.12
4	2.5	1.7	3.6	3.6	2.4	5.1	.17	.05	1.09
8	2.9	2.0	4.1	4.0	2.7	5.7	.91	.27	6.05
16	3.4	2.3	4.7	4.5	2.9	6.3	4.24	1.18	27.7

Table 11. The maximum error and timing on domains with size $\frac{1}{\sqrt{10}}$ and $\frac{1}{10}$.

APPENDIX:

Here we give the proof of the assertion in remark 4.

Let $m_1,\ldots,m_9$ denote the columns of M, defined by (25). Suppose that these vectors do not span $\mathbb{R}^8$, and let m_{10} be a vector which is linearly independent of $m_1,\ldots,m_9$.
Now, we consider three linear subspaces of $\mathbb{R}^8$, V_1 with dimension 5 spanned by $m_1,\ldots,m_5$, V_2 with dimension 2 and V_3 with dimension 1 such that $V_2 \perp V_1$, $V_3 \perp V_1 \oplus V_2$ and $V_1 \oplus V_2$ is spanned by $m_1,\ldots,m_9$. Let $v_1,\ldots,v_5$ form a basis for V_1, v_6 and v_7 a basis for V_2, and v_8 for V_3.
Then, a row w_i of w_{opt}, satisfying the minimization property, can be written as

$$(1) \qquad w_i = \sum_{j=1}^{7} \alpha_{ij} v_j, \quad i = 1,\ldots,5,$$

with the properties

$$(2) \qquad (w_i,m_j) = \delta_{ij}, \quad j = 1,\ldots,5,$$

$$(3) \qquad \sum_{j=6}^{9} (w_i,m_j)^2 = f(\alpha_{i1},\ldots,\alpha_{i7}) \quad \text{minimal}$$

or equivalently (α_{ij}, $j = 1,\ldots,5$ are fixed by (2))

$$(4) \qquad \frac{\partial}{\partial \alpha_{ij}} f(\alpha_{i1},\ldots,\alpha_{i7}) = 0, \quad j = 6,7.$$

Now, we will show that the solution w_i^* to the minimization problem under the constraint (2)

$$(5) \qquad \sum_{j=6}^{10} (w_i^*,m_j)^2 \quad \text{minimal},$$

satisfies the relations (4) too, and thus lies in the solution space of the original minimization problem.

Let us write $w_i^* = \sum_{j=1}^{7} \alpha_{ij}^* v_j + \gamma_i v_8$. Rewriting of (5) yields

$$\sum_{j=6}^{10} (w_i^*, m_j)^2 = \sum_{j=6}^{9} \left\{ \sum_{k=1}^{7} \alpha_{ik}^* (v_k, m_j) + \gamma_i (v_8, m_j) \right\}^2 +$$

$$+ \left\{ \sum_{j=1}^{7} \alpha_{ij}^* (v_j, m_{10}) + \gamma_i (v_8, m_{10}) \right\}^2 =$$

$$= f(\alpha_{i1}^*, \ldots, \alpha_{i7}^*) + \left\{ \sum_{j=1}^{7} \alpha_{ij}^* (v_j, m_{10}) + \gamma_i (v_8, m_{10}) \right\}^2,$$

because $(v_8, m_j) = 0, \quad j = 1, \ldots, 9.$

As m_{10} does not lie in $V_1 \oplus V_2$, it is not orthogonal to V_3, and therefore (v_8, m_{10}) does not vanish. Thus, minimality of (5) implies

$$\gamma_i = - \frac{\sum_{j=1}^{7} \alpha_{ij}^* (v_j, m_{10})}{(v_8, m_{10})}, \tag{6}$$

and

$$\frac{\partial}{\partial \alpha_{ij}} f(\alpha_{i1}^*, \ldots, \alpha_{i7}^*) = 0, \quad j = 6,7, \tag{7}$$

whereas the solution of

$$(w_i^*, m_j) = \sum_{k=1}^{5} \alpha_{ik}^* (v_k, m_j) = \delta_{ij}, \quad j = 1, \ldots, 5, \tag{8}$$

uniquely determines $\alpha_{ik}^* = \alpha_{ik}, \quad k = 1, \ldots, 5.$
As $(\alpha_{i6}, \alpha_{i7})$ is the solution to eq. (4), it solves (7) too, and we find that w_i^* can be written as

$$w_i^* = w_i + \gamma_i v_8.$$

From (7) and (8) it follows that w_i^* satisfies the original minimization problem given by (2) and (3), and it is the unique solution within the solution space of (2) and (3), with the property

$$(w_i^*, m_{10}) = 0,$$

which property is a consequence of (6).

REFERENCES

1. W.H. FREY, 'Flexible finite-difference stencils from isoparametric finite elements'. *Int. J. num. Meth. Engng*, 11, 1653-1665 (1977).

2. A.R. GOURLAY, 'Splitting methods for time-dependent partial differential equations', in the preceedings of the 1976 conference: *The state of art in numerical analysis*, ed. D.A.H. Jacobs, Academic Press (to appear).

3. P.J. VAN DER HOUWEN & J. VERWER, 'Non-linear splitting methods for semi-discretized parabolic diffrential equations', *Report NW 51/77*, Mathematisch Centrum, Amsterdam (1977).

4. J. KOK, P.J. VAN DER HOUWEN & P.H.M. WOLKENFELT, 'A semi-discretization algorithm for two-dimensional partial differential equations', *Report NW*, Mathematisch Centrum, Amsterdam (to appear).

5. *NAG Library Manual*, Mark 5, Oxford (1977).

6. N.L. SCHRYER, 'Numerical solution of time-varying partial differential equation in one space variable', *Comp. Science Techn. Rep. No. 53*, Bell Laboratories, Murray Hill (1975).

7. R.F. SINCOVEC & N.K. MADSEN, 'Software for nonlinear partial differential equations', *ACM Transactions on Mathematical Software* 1, 232-260 (1975).

8. J.H. WILKINSON, *The algebraic eigenvalue problem*, Clarendon Press (1965).

9. J.H. WILKINSON & C. REINSCH, *Handbook for automatic computation*, vol. 2, Linear Algebra, Springer Verlag (1971).

10. N.N. YANENKO, *The method of fractional steps*, Springer Verlag, Berlin (1971).