JOURNAL OF COMPUTATIONAL PHYSICS **44**, 1–19 (1981)

Iterative Solution Methods for Certain Sparse Linear Systems with a Non-Symmetric Matrix Arising from PDE-Problems*

HENK A. VAN DER VORST

Academic Computer Centre, Budapestlaan 6, de Uithof, Utrecht, the Netherlands

Received September 13, 1979; revised June 5, 1981

In this paper methods are described for the solution of certain sparse linear systems with a non-symmetric matrix. The power of these methods is demonstrated by extensive numerical experiments. Application of the methods is limited to problems where the matrix has only eigenvalues with positive real part. An important class of this type of matrix arises from discretisation of second order partial differential equations with first order derivative terms.

1. INTRODUCTION

When an elliptic selfadjoint partial differential equation is discretised over some region, this results in a linear system $Ax = b$, where A is a symmetric matrix.

Efficient algorithms to solve this type of equation iteratively have been described by Axelsson [1], Concus *et al.* [3], Meijerink and van der Vorst [15, 16] and many others.

Another class of problems arises when the partial differential equation includes first order derivatives. Discretisation in this case yields a linear system with a non-symmetric matrix. These problems are much more difficult to solve and much research has been done and is currently going on to develop efficient algorithms. Algorithms have been published by Varga [20], Stone [18], Kershaw [11], Manteuffel [12, 13], Paige [17], Concus and Golub [4], Widlund [21], a.o.

Manteuffel has compared his version of Tchebycheff Iteration to the bidiagonalisation method of Paige [17] and the conjugate gradient method in a form as proposed by Kershaw [11]. From this comparison Manteuffel's method seems to be the most promising one.

In this paper we consider the use of preconditioning techniques in order to improve the efficiency of the Tchebycheff Iteration, which is briefly outlined in Section 3.

Some specific preconditionings, including a Fast Poisson Solver and incomplete Crout and Choleski decompositions, are discussed in Section 4. Certain instability

* The research described in this paper has been supported in part by the European Research Office, London, through Grant DAJA 37-80-C-0243.

0021-9991/81/110001-19$02.00/0

problems in the incomplete Crout decomposition can be overcome avoiding the necessity of partial pivoting.

From the numerical experiments described in Section 5 it appears that the use of an incomplete Crout decomposition as a preconditioning (eventually with modification) for the Tchebycheff Iteration leads to highly competitive iterative methods. This is even more the case if we make use of implementation ideas of Eisenstat [5] for preconditioned conjugate gradients, which can be easily adapted to the preconditioned conjugate gradients, which can be easily adapted to the preconditioned Tchebycheff Iteration.

2. Description of a Class of Non-Symmetric Matrices

In the following sections we often refer specifically to the important class of non-symmetric matrices that arises from 5-point difference discretisation of second order PDEs like:

$$-(Du'_x)'_x - (Eu'_y)'_y + Gu'_x + Hu'_y + Cu = F, \tag{2.1}$$

defined on a rectangular region R in the (x, y)-plane, with $D(x, y) > 0$, $E(x, y) > 0$ and $C(x, y) \geqslant 0$ for $x, y \in R$.

Equation (2.1), except for the part $Gu'_x + Hu'_y$, can be discretised in such a way that the resulting matrix is a symmetric positive definite M-matrix (see Varga [20]). The first derivative terms $Gu'_x + Hu'_y$ can be approximated by either central differences or backward/forward differences. If $G = G(y)$ and $H = H(x)$ then they contribute to the final discretisation matrix by a skew-symmetric matrix when central differences are used.

The final linear system resulting from the discretisation of (2.1) is denoted by $Ax = b$, where A has order nm and appears as in Fig. 1.

$$A = \left\{ \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \right. \quad \begin{array}{c} \overleftarrow{\quad m \quad} \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \quad \left. \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \right\}$$

$$a_{i1} \quad a_{i2} \quad a_{i3} \quad a_{i4} \quad a_{i5} \tag{2.2}$$

FIGURE 1

3. The Tchebycheff Iteration

In this section a short overview of the Tchebycheff Iteration for non-symmetric linear systems is presented. For more detailed information one is referred to Manteuffel [12].

The basic iteration formula is given by

$$x_{k+1} = -\alpha_k A x_k + (1+\beta_k) x_k - \beta_k x_{k-1} + \alpha_k b. \tag{3.1}$$

Manteuffel [12] shows that this iteration converges to the solution of $Ax = b$ if the spectrum of A can be enclosed in an ellipse in the right half plane.

The iteration parameters are defined in terms of parameters of this ellipse. Let $d-c$ and $d+c$ be the foci of the ellipse, c eventually complex; then α_k and β_k are defined by

$$\alpha_k = \frac{2}{c} \frac{T_k\left(\frac{d}{c}\right)}{T_{k+1}\left(\frac{d}{c}\right)}, \qquad \beta_k = \frac{T_{k-1}\left(\frac{d}{c}\right)}{T_{k+1}\left(\frac{d}{c}\right)}, \tag{3.2}$$

where $T_k(z) = \cos(k \arccos(z))$, the kth Tchebycheff polynomial. The constants d and c should be chosen to define a family of ellipses containing the ellipse that encloses the spectrum of A, for which the rate of convergence r_c (see formula (3.4)) is minimal. This leads to the following computational scheme.

Given x_0, define

$$r_0 = b - A x_0,$$

$$p_0 = \frac{1}{d} r_0,$$

$$\alpha_0 = \frac{2}{d},$$

Then

$$\left.\begin{aligned}
x_i &= x_{i-1} + p_{i-1}, && (3.3.1)\\
r_i &= b - A x_{i-1}, && (3.3.2)\\
\alpha_i &= \left(d - \left(\frac{c}{2}\right)^2 \alpha_{i-1}\right)^{-1}, && (3.3.3)\\
\beta_i &= d\alpha_i - 1, && (3.3.4)\\
p_i &= \alpha_i r_i + \beta_i p_{i-1}, && (3.3.5)
\end{aligned}\right\} \quad i = 1, 2, 3, \ldots .$$

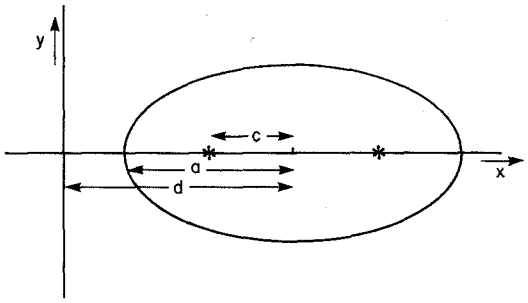

FIG. 2. Iteration parameters.

The asymptotic convergence factor for this iterative Tchebycheff method is given by

$$r_c = \frac{a + \sqrt{a^2 - c^2}}{d + \sqrt{d^2 - c^2}}, \tag{3.4}$$

where d is the center of the ellipse, c is the focal distance and a is the length of the axis in the x-direction (see Fig. 2). An adaptive procedure in which the values of d and c are estimated dynamically has been proposed by Manteuffel [12, 13].

4. PRECONDITIONINGS FOR THE TCHEBYCHEFF ITERATION

In order to improve the efficiency of the algorithm given in Section 3 one may consider the use of a non-singular preconditioning matrix K and solve the equation $KAx = Kb$. One hopes to be able to construct a preconditioning matrix so that the rate of convergence improves so much, that the total amount of work decreases taking into account the additional work required for the preconditioning.

Note that the use of a preconditioning only affects the scheme (3.3) insofar as the residual r_i now has to be computed from $r_i = K(b - Ax_i)$. For the algorithm to converge it is necessary that all eigenvalues are in the right half plane, i.e., all eigenvalues of KA should have positive real part. We will now present a number of succcessful preconditionings.

4.1 *Inverse of Symmetric Part*

If $A = M + N$, where M is symmetric positive definite and $N = -N^T$, then it is well known that all eigenvalues of $M^{-1}A$ have positive real part 1.0. If for given vectors x, the matrix vector product $M^{-1}x$ can be generated efficiently, e.g., by a Fast Poisson Solver, then the operator M^{-1} can be used as a preconditioning.

4.2. *Approximate Inverse of Symmetric Part*

If computation of $M^{-1}x$ becomes too inefficient, one might consider the use of an

incomplete Choleski decomposition [15] of the symmetric part, for which we give the following result.

THEOREM 1. *If $A = M + N$, M a symmetric M-matrix, $N = -N^T$, and K is an incomplete Choleski decomposition of M, then all eigenvalues of $K^{-1}A$ have positive real part.*

Proof. We write the incomplete Choleski decomposition as $K = LL^T$. The matrix $K^{-1}A$ has the same set of eigenvalues as $L^{-1}AL^{-T}$. The latter matrix can be written as

$$L^{-1}AL^{-T} = L^{-1}ML^{-T} + L^{-1}NL^{-T}.$$

Since $L^{-1}ML^{-T}$ is symmetric positive definite, all of its eigenvalues are in $[a, b]$, $a > 0$. The matrix $L^{-1}NL^{-T}$ is skew-symmetric and all of its eigenvalues are on the imaginary axis. Due to Householder [9, p. 79], the eigenvalues of $L^{-1}AL^{-T}$, and thus those of $K^{-1}A$, are in $[a, b] * [-ic, ic]$. ∎

4.3. *Approximate Inverse of A*

It is also possible to use some kind of incomplete Crout decomposition [15] as a preconditioning. Before we treat this important case in detail, we give the following result.

THEOREM 2. *If A is an M-matrix and K is an incomplete Crout decomposition of A, then all eigenvalues of $K^{-1}A$ have positive real part.*

Proof. The existence of an incomplete Crout decomposition K is guaranteed by a theorem of Meijerink and van der Vorst [15], who also prove that $A = K - R$ defines a regular splitting. From Varga [20, Theorem 3.12], it follows that $Kx_{i+1} = Rx_i + b$ converges for all x_0, or $\rho(K^{-1}R) < 1$. From $\rho(K^{-1}R) = \rho(K^{-1}(K - A)) = \rho(I - K^{-1}A) < 1$ it then follows that all eigenvalues of $K^{-1}A$ have positive real part. ∎

Except for the first order derivative terms, Eq. (2.1) can be discretised by 5-point central differences in such a way that the resulting matrix M is an irreducibly diagonally dominant symmetric M-matrix (see Varga [20]).

THEOREM 3. *If the first order derivative terms of* (2.1) *are discretised by central differences and if the resulting contributions to the discretisation matrix are in absolute value smaller than the respective elements of M, then the final discretisation matrix A is an M-matrix.*

Proof. From Varga [20, Theorem 3.4], it follows that $I - D^{-1}A$, where $D = \text{diag}(A)$, is a convergent matrix. Applying Varga [20, Theorem 3.10], gives that A is non-singular, $A^{-1} \geqslant 0$ and thus A is an M-matrix. ∎

The contributions of the first order derivative terms can be made smaller than the

other elements by choosing the mesh sizes Δx and Δy sufficiently small. When the first order terms are discretised by either backward or forward differences, such that they add to the main diagonal of A, A again is an M-matrix, independently of the choice of Δx or Δy. The proof is similar to the proof of Theorem 3.

If we write the incomplete Crout decomposition of A as

$$A = LD^{-1}U - R \tag{4.1}$$

then a very simple decomposition is defined by the following rules:

(a) $\text{diag}(L) = \text{diag}(U) = D^{-1}$;

(b) The off-diagonal parts of L and U are equal to the corresponding parts of A; (4.2)

(c) $\text{diag}(LD^{-1}U) = \text{diag}(A)$.

We will denote the decomposition defined in scheme (4.2) by the subscript 1: $L_1 D_1^{-1} U_1$. For the pentadiagonal blockmatrix of Section 2 this leads to the following recurrence relation for the elements d_i of D_1:

$$d_i = a_{i,3} - a_{i-1,4} a_{i,2} d_{i-1}^{-1} - a_{i-m,5} a_{i,1} d_{i-m}^{-1}. \tag{4.3}$$

For the case that $A = M + N$, where M is a symmetric M-matrix and $N = -N^T$, scheme (4.2) can be used to compute all elements of D_1; it can be proven that all of the d_i will be positive. Note that A itself is not required to be an M-matrix. However, when the elements of N are large then the factors L_1 and U_1 may be very ill conditioned, while A can be reasonably well conditioned. This ill-conditioning can eventually be prevented by a partial pivoting technique, which has the obvious disadvantage of destroying the sparsity structure.

Ill-conditioning can also be prevented by constructing an incomplete decomposition of $A + (\sigma - 1) * \text{diag}(A)$, with the constant σ chosen large enough. This is similar to what Manteuffel [14] proposes to do for the decomposition of symmetric positive definite matrices.

In this case formula (4.3) changes to

$$d_i = \sigma a_{i,3} - a_{i-1,4} a_{i,2} d_{i-1}^{-1} - a_{i-m,5} a_{i,1} d_{i-m}^{-1}. \tag{4.4}$$

We will denote this "stabilized" decomposition by $L_\sigma D_\sigma^{-1} U_\sigma$. If the parameter σ is chosen so that the resulting d_i compare in magnitude with the sum of the off-diagonal elements in absolute value of L_σ and U_σ (those elements do not depend on σ), then the resulting L_σ and U_σ are well conditioned. For the modelequation $-u''_{xx} - u''_{yy} + \beta(u'_x + u'_y) + cu = f$ one can show that the d_i, for increasing i, rapidly tend to the largest root of

$$d = \sigma a_{i,3} - a_{i-1,4} a_{i,2} d^{-1} - a_{i-m,5} a_{i,1} d^{-1}, \tag{4.5}$$

which makes it easy to determine the proper value of σ. We will call this value σ_{opt}.

When A is a diagonally dominant M-matrix, then it is not necessary to incorporate a σ for stabilization. Meijerink and van der Vorst [15] show that the decomposition in this case is stable. Manteuffel [14] shows that optimal convergence in the symmetric case (for the preconditioned *cg*-process) is achieved for $\sigma \leqslant 1.0$. A similar result for the symmetric case is published by Gustafsson [7], he, in fact, proposes different σ_i for each diagonal element.

Now the next question is how the convergence behaviour is affected by the parameter σ. Numerical experiments indicate that for the modelequation the rate of convergence of the Tchebycheff iteration with the preconditioning $K = (L_\sigma D_\sigma^{-1} U_\sigma)^{-1}$ is minimal for σ in the neighbourhood of σ_{opt}.

This is an embarassing result since σ was only introduced in order to have well-conditioned factors. Apparently we should strive for decompositions where the diagonalelements are at least comparable to the sum of off-diagonal elements in the factors. That suggests the following parameterless incomplete decomposition scheme. The off-diagonal elements of L and U are set equal to the corresponding elements in A, and $\text{diag}(L) = \text{diag}(U) = D^{-1}$. Let $\sum_{L,i} = \sum_{i>j} |a_{ij}|$ and $\sum_{U,i} = \sum_{i<j} |a_{ij}|$.

Then the kth step in the decomposition construction process is defined by

(a) compute the kth diagonal element of D, using the relation $\text{diag}(A) = \text{diag}(LD^{-1}U)$; this element is denoted by Δ_k;

(b) compute $\sum_{L,k}$ and $\sum_{U,k}$; (4.6)

(c) the kth diagonal element of D is now replaced by $d_k = \max\{\Delta_k, \sum_{L,k}, \sum_{U,k}\}$.

The above-defined decomposition is denoted by $L_{EQ} D_{EQ}^{-1} U_{EQ}$.

Since $L_\sigma D_\sigma^{-1} U_\sigma$ includes $L_1 D_1^{-1} U_1$ we will in the numerical examples only consider $L_\sigma D_\sigma^{-1} U_\sigma$ and $L_{EQ} D_{EQ}^{-1} U_{EQ}$, besides the preconditionings mentioned in Sections 4.1 and 4.2.

We were not able to prove that in general all the eigenvalues of the preconditioned matrices are in the right half plane for the stabilized preconditionings. For a large number of problems we computed all the eigenvalues and they always happened to have positive real part.

4.4. *Efficient Implementation of the Unsymmetric Preconditionings*

If we consider the computational cost per iterationstep of the preconditioned iterationprocess then we recognize that this is significantly higher than without the use of any preconditioning. As we know, we have to compute matrix vector products $K^{-1}x$ besides Ax for each iterationstep.

Eisenstat [5] shows that the preconditioned conjugate gradient iteration for certain preconditionings can be implemented in such a way that the preconditioned iterationsteps are almost as cheap as the unpreconditioned ones. Eisenstat's ideas can easily be adapted to the preconditioned Tchebycheff Iteration for the preconditionings as described in Section 4.3. We will give here a brief outline of how this can be done.

Let $LD^{-1}U$ be either the $L_\sigma D_\sigma^{-1} U_\sigma$ or the $L_{EQ} D_{EQ}^{-1} U_{EQ}$ incomplete decomposition

of A. We assume that A has been scaled such that $D = I$, so we omit D in the formulas.

The linear system $Ax = b$ has the same solution as

$$L^{-1}AU^{-1}Ux = L^{-1}b. \tag{4.7}$$

Now we define $\hat{A} = L^{-1}AU^{-1}$, $\hat{x} = Ux$ and $\hat{b} = L^{-1}b$, giving $\hat{A}\hat{x} = \hat{b}$. The Tchebycheff iteration for this equation, where the residual is now computed by a recurrence relation, looks like

$$\hat{x}_i = \hat{x}_{i-1} + \hat{p}_{i-1}, \tag{4.8.1}$$

$$\hat{r}_i = \hat{r}_{i-1} - \hat{A}\hat{p}_{i-1}, \tag{4.8.2}$$

$$\hat{\alpha}_i = \left(\hat{d} - \left(\frac{\hat{c}}{2}\right)^2 \hat{\alpha}_{i-1}\right)^{-1}, \tag{4.8.3}$$

$$\hat{\beta}_i = \hat{d}\hat{\alpha}_i - 1, \tag{4.8.4}$$

$$\hat{p}_i = \hat{\alpha}_i \hat{r}_i + \hat{\beta}_i \hat{p}_{i-1}. \tag{4.8.5}$$

We now simply replace (4.8.1) by

$$x_i = x_{i-1} + U^{-1}\hat{p}_{i-1}. \tag{4.8.1a}$$

Now we consider $\hat{A}\hat{p}_{i-1}$:

$$\hat{A}\hat{p}_{i-1} = L^{-1}(L + A - L - U + U)\, U^{-1}\hat{p}_{i-1}.$$

Since the off-diagonal elements of L and U are equal to the corresponding elements of A and since $\operatorname{diag}(L) = \operatorname{diag}(U) = I$, this yields, if we write $\hat{t}_i = U^{-1}\hat{p}_{i-1}$:

$$\begin{aligned}\hat{A}\hat{p}_{i-1} &= U^{-1}\hat{p}_{i-1} + L^{-1}(\hat{p}_{i-1} + (D - 2I)\, U^{-1}\hat{p}_{i-1}) \\ &= \hat{t}_{i-1} + L^{-1}(\hat{p}_{i-1} + (D - 2I)\, \hat{t}_{i-1}),\end{aligned} \tag{4.9}$$

where $D = \operatorname{diag}(A)$.

Expression (4.9) can be inserted in (4.8.2) and we see that the cost of the preconditioned Tchebycheff iterationsteps has been reduced to almost the cost of the unpreconditioned iterationsteps. Note that the same reduction in computational cost cannot be achieved for the preconditionings described in Sections 4.1 and 4.2.

5. Numerical Experiments

The numerical experiments described in this section have all been carried out on a CDC-Cyber 175/100 of the Academic Computer Centre, Utrecht, in 48 bits relative working precision. The residuals in all experiments, as far as listed, have been

TABLE I

Iteration Results for 5.1 with $\beta = 4.0$

Method	Number of iterations	Final residual / Initial residual
Tchebycheff iteration without preconditioning	92	1.5_{10-8}
id. with $L_\sigma D_\sigma^{-1} U_\sigma$- preconditioning, $\sigma = 1.0$	25	9.1_{10-9}
id. with $L_{EQ} D_{EQ}^{-1} U_{EQ}$- preconditioning	25	9.0_{10-9}

computed as $\|Ax_i - b\|_2$, where x_i is the ith iterand in the iterative solution process for $Ax = b$. No efforts have been made to start the Tchebycheff iteration with good parameters d and c. This means that in most cases the first 20 iterationsteps were only used to get estimates for d and c, while sometimes the iterationprocess even diverged during these first 20 steps. Therefore the number of iterations gives mostly a pessimistic impression of the actual convergence behaviour. We have chosen this crude strategy because we believe that in most practical situations it is the only possible choice. However, if one has to solve a set of similar problems a proper determination of d and c beforehand can improve the efficiency significantly.

5.1. $-u''_{xx} - u''_{yy} + \beta(u'_x + u'_y) + u = 1$

This simple problem has been chosen since it has been described extensively by Manteuffel [13] and since a number of properties can easily be verified (e.g., solution, eigenvalues, stability).

The equation is discretised over a square region with gridspacing 1.0 in both

TABLE II

Iteration Results for 5.1 with $\beta = 20.0$

Method	Number of iterations	Final residual / Initial residual
Tchebycheff iteration without preconditioning	200	6.5_{10-7}
id. with $L_\sigma D_\sigma^{-1} U_\sigma$- preconditioning, $\sigma = 2.6$	40	6.4_{10-9}
id. with $L_{EQ} D_{EQ}^{-1} U_{EQ}$- preconditioning	37	8.3_{10-9}

directions. The boundary conditions are of Dirichlet type: $u = 1.0$ along the boundary. The first order derivative terms are discretised by central differences. The starting vector is always chosen to be $x_0 = 0.0$. The initial parameters for the Tchebycheff iteration are $d = 1$, $c = 0$ for the preconditioned iteration and $d = 5$, $c = 0$ for the unpreconditioned iteration. Using formula (4.5) we observe that the diagonal elements in the incomplete $L_\sigma D_\sigma^{-1} U_\sigma$ decomposition rapidly tend to the largest root of the equation $d = 5\sigma + (2/d)(\beta/2 + 1)(\beta/2 - 1)$. Since the factors L_σ and U_σ are well conditioned when $d = \beta + 2$, it follows that $\sigma_{\text{opt}} = (3 + \beta/2)/5$. It should be mentioned here that in many cases it is possible to find by trial and error a value for σ for which the total number of iterations is smaller than for σ_{opt}. When the number of unknowns equals 841 ($=29^2$) the iteration results as given in Tables I and II are obtained.

For $\sigma = 20.0$ we have checked how the convergence of the Tchebycheff iteration with the $L_\sigma D_\sigma^{-1} U_\sigma$-preconditioning depends on the choice of σ (formula (4.4)). From the same straightforward computation as above it follows that the sum of the off-diagonal elements in a typical row of L_σ in absolute value is equal to the corresponding diagonal element if $\sigma = 2.6$. In Fig. 3 it is shown how many iterations actually are required to obtain a final residual $\|Au_n - b\|_2 < 10^{-6}$ for different values of σ.

It is interesting to compare the ellipses for the case $\beta = 20.0$ that contain the eigenvalues, or more precisely the field of values, of A and KA for some preconditioning K. This gives an impression of the effect of the preconditioning. In Fig. 4 the big ellipse is estimated by Manteuffel's algorithm to contain the field of values of A, while the small ellipse (indicated by an arrow) contains the field of values of $(L_\sigma D_\sigma^{-1} U_\sigma)^{-1} A$, with $\sigma = 2.6$.

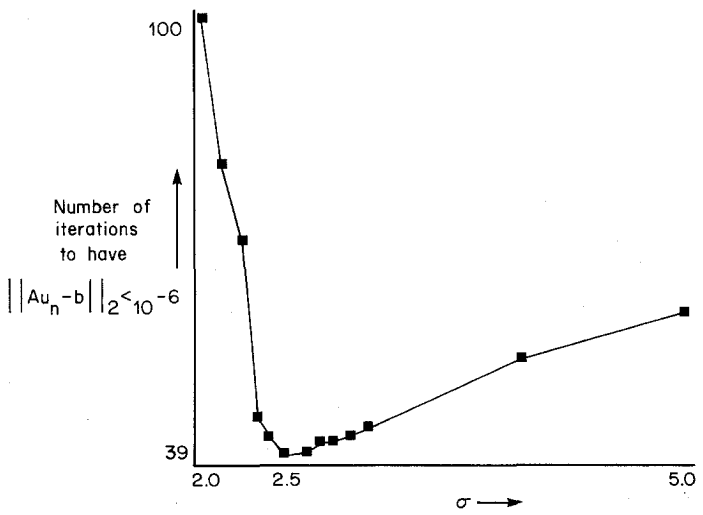

FIG 3. Number of iterations for $\beta = 20.0$.

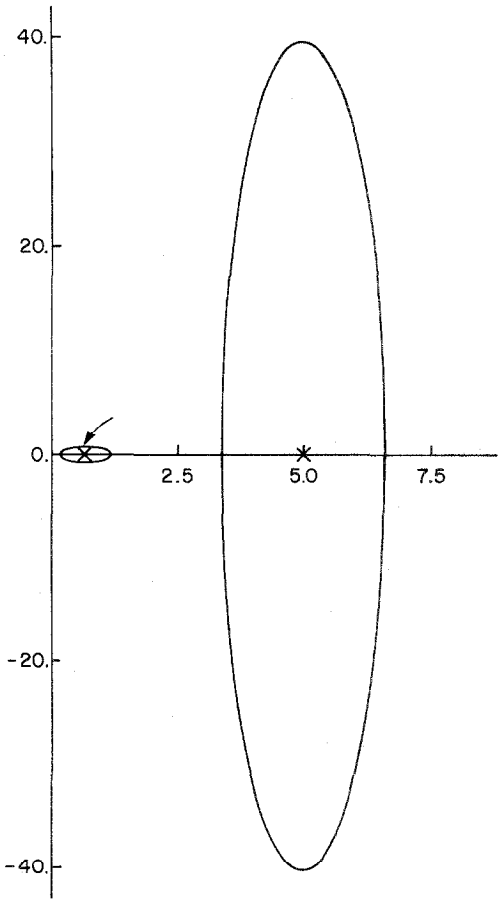

FIG. 4. Ellipses for A and KA, for $\beta = 20.0$.

The convergence factors r_c associated with each ellipse are:

1. for the A-ellipse: $r_c = 0.918$;
2. for the KA-ellipse: $r_c = 0.586$.

5.2. $-u''_{xx} - u''_{yy} + \beta(u'_x + u'_y) = 0$

This problem has also been considered extensively by Manteuffel [13]. Manteuffel compares for this case the Tchebycheff Iteration, with dynamically estimated parameters, with some other iterative methods:

(a) The bidiagonalisation method (Golub and Kahan [6], Paige [17]).

(b) The *cg*-method applied to A^TA (Hesteness and Stiefel [8], Kershaw [11]).

For this problem the Tchebycheff iteration appeared to be superior for a wide variety of values of β.

TABLE III

Results for the Unpreconditioned Tchebycheff Iteration

β	Number of iterations	Final residual / Initial residual	Number of iterations to gain one decimal*
0.1	198	7.5_{10-8}	27.8
0.4	200	1.1_{10-5}	40.5
4.0	192	2.6_{10-8}	25.3
20.0	200	4.1_{10-7}	31.3
40.0	200	2.3_{10-4}	55.0

* The number of iterations to gain one decimal is computed as $n_f/^{10}\log(r_0/r_f)$, where r_0, r_f are initial, final residual resp. and n_f is the number of iterations.

Again the equation has been discretised over a grid with gridspacing 1.0, using central differences. For the resulting linear system consisting of 1600 ($=40^2$) unknowns, the results of the *un*preconditioned Tchebycheff iteration with *exact* parameters d and c are listed in Table III (for exact values of d and c see Manteuffel [13, p. 35]).

In Table IV we present the results of the Tchebycheff iteration *with* preconditioning $(L_\sigma D_\sigma^{-1} U_\sigma)^{-1}$ with optimal values for σ determined as in Section 5.1. No efforts have been made to find optimal iteration parameters d and c, in each case they were initially chosen to be $d = 1.0$ and $c = 0.0$ and have been adjusted dynamically.

From the results in the Tables III and IV we conclude that also in this case, the use of a suitable preconditioning, even with sometimes poor initial approximations for the parameters d and c, leads to a big improvement. Note that for small β (e.g., $\beta = 0.1$ and 0.4) the value of the optimal σ is less than 1.0. In this case the discretisation matrix is an M-matrix and although we introduced σ only to prevent ill-conditioning, Manteuffel proved that optimal efficiency, in case of the *cg*-process with preconditioning, is reached for $\sigma \leqslant 1.0$ (see Section 4.3).

TABLE IV

Tchebycheff Iteration with $(L_\sigma D_\sigma^{-1} U_\sigma)^{-1}$-Preconditioning

β	σ_{opt}	Number of iterations	Final residual / Initial residual	Number of iterations to gain one decimal
0.1	0.7625	47	7.0_{10-8}	6.6
0.4	0.8	39	4.5_{10-8}	5.3
4.0	1.25	15	2.4_{10-8}	2.0
20.0	3.25	48	8.8_{10-9}	6.0
40.0	5.75	83	3.8_{10-9}	9.9

TABLE V

Tchebycheff Iteration with $(L_{EQ}D_{EQ}^{-1}U_{EQ})^{-1}$-Preconditioning

β	Number of iterations	Final residual / Initial residual	Number of iterations to gain one decimal
0.1	125	$7.2_{10^{-8}}$	17.5
0.4	46	$1.4_{10^{-8}}$	5.9
4.0	15	$1.7_{10^{-8}}$	1.9
20.0	50	$6.6_{10^{-9}}$	6.1
40.0	84	$3.3_{10^{-9}}$	9.9

In Table V we give the results achieved when the $(L_{EQ}D_{EQ}^{-1}U_{EQ})^{-1}$-preconditioning is used (again with initial parameters $d = 1.0$, $c = 0.0$). We see that for small β, when A is an M-matrix and no actions need to be undertaken to prevent ill-conditioning, the $(L_{EQ}D_{EQ}^{-1}U_{EQ})^{-1}$-preconditioning is less efficient, whilst competing with the $(L_\sigma D_\sigma^{-1}U_\sigma)^{-1}$-preconditioning for larger values of β.

For this problem we have also compared the effect of different preconditionings as described in Section 4. The gridspacing is still 1.0, but the number of unknowns has been reduced to 961 $(=31^2)$. This particular choice was necessary to be able to include the Fast Poisson Solver in the comparison. The following preconditionings have been chosen.

(a) Fast Poisson Solver [2].

(b) The incomplete Choleski $K_{1,3}$ on the symmetric part [16, 19].

(c) $L_\sigma D_\sigma^{-1}U_\sigma$ with $\sigma = \sigma_{\text{opt}}$.

The eigenvalues of the preconditioned matrix for (a) and (b) have always real parts, whereas we were only able to prove this property for (c) when $\beta \leqslant 2.0$ (see Section 4). The $L_\sigma D_\sigma^{-1}U_\sigma$-preconditioning can be implemented very efficiently, in contrast to both other types of preconditionings. We therefore have included the CPU-times in the tables in order to give an impression of the actual efficiency of the three different choices. The computational results are given in Tables VI, VII, and VIII.

TABLE VI

Iteration Results for $\beta = 0.1$

Preconditioning	Number of iterations	Final residual / Initial residual	CPU-time
FAST POISSON	17	$6.4_{10^{-8}}$	0.88
INC. CHOLESKI	34	$7.4_{10^{-8}}$	1.02
$L_\sigma D_\sigma^{-1}U_\sigma$, $\sigma = 0.7625$	46	$6.7_{10^{-8}}$	0.75

TABLE VII

Iteration Results for $\beta = 0.4$

Preconditioning	Number of iterations	Final residual / Initial residual	CPU-time
FAST POISSON	58	5.7_{10-8}	2.8
INC. CHOLESKI	35	4.8_{10-8}	1.0
$L_\sigma D_\sigma^{-1} U_\sigma$, $\sigma = 0.8$	38	3.3_{10-8}	0.63

For the case $\beta = 0.4$ we have also plotted the ellipses containing the field of values for A, $P^{-1}A$, $K_{1,3}^{-1}A$ and $(L_\sigma D_\sigma^{-1} U_\sigma)^{-1}A$, where P represents the inversion by the Fast Poisson Solver. The ellipses displayed in Fig. 5 (two of them are degenerated to straight lines) are derived from the results in the last dynamic estimationstep of the Manteuffel algorithm, usually performed after each 20th iterationstep. The respective values for r_c are:

1. A $\quad r_c = 0.798;$
2. $P^{-1}A$ $\quad r_c = 0.624;$
3. $K_{1,3}^{-1}A$ $\quad r_c = 0.538;$
4. $(L_\sigma D_\sigma^{-1} U_\sigma)^{-1}A$ $\quad r_c = 0.343.$

5.3. $-u''_{xx} - u''_{yy} + \left(\frac{\partial}{\partial x}(au) + a\frac{\partial u}{\partial x}\right)\Big/2 = f$

This equation has been taken from Widlund [21]. Since only a first derivative in one direction (the x-direction) is present, it can be shown that the incomplete decomposition yields well-conditioned factors which implies that the $L_1 D_1^{-1} U_1$-decomposition can be used as a preconditioning. The equation is discretised over a rectangular grid with equal gridspacings in both directions over the region $[0, 1] * [0, 1]$, and along the boundaries a Dirichlet boundary condition is imposed.

TABLE VIII

Iteration Results for $\beta = 4.0$

Preconditioning	Number of iterations	Final residual / Initial residual	CPU-time
FAST POISSON	200	2.7_{10-1}	9.5
INC. CHOLESKI	200	4.7_{10-5}	3.8
$L_\sigma D_\sigma^{-1} U_\sigma$, $\sigma = 1.25$	15	2.3_{10-8}	0.26

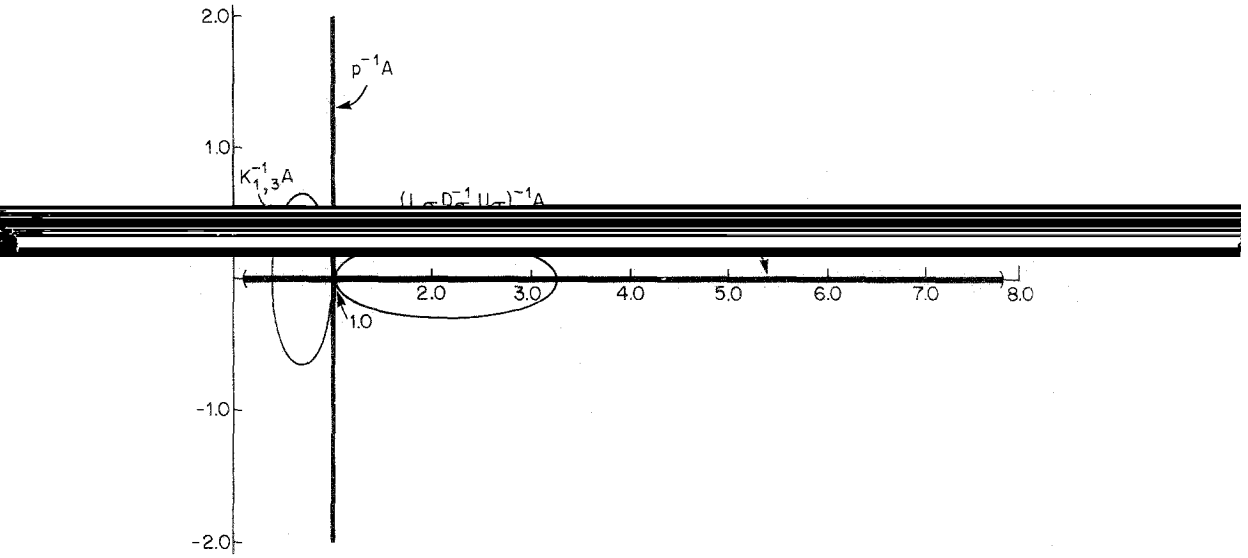

FIG. 5. Eigenvalue ellipses for $\beta = 0.4$.

The function $a(x, y)$ is chosen as $2 * \exp(3.5(x^2 + y^2))$, $20 * \exp(3.5(x^2 + y^2))$, resp. The right-hand side $f(x, y)$ is chosen in such a way that $u(x, y) = \sin \pi x \sin \pi y \exp((x/2 + y)^3)$ satisfies the equation. No efforts have been made to estimate good Manteuffel-parameters, in each case they were initialised as $d = 1.0$ and $c = 0.0$. The results are given in Table IX. If we compare these results with those published by Widlund [21, Table 1], then it appears that our method is highly competitive (remember that in each iterationstep of Widlunds algorithm a symmetric linear system $(A + A^T)/2\, x = b'$ has to be solved).

TABLE IX

Iteration Results for Discretised Linear System of (5.3)

$a(x, y)$	Method	Number of unknowns	Number of iterations	Final residual / Initial residual
$2e^{3 \cdot 5(x^2+y^2)}$	no precond.	225	200	4.9_{10-2}
$2e^{3 \cdot 5(x^2+y^2)}$	$L_1 D_1^{-1} U_1$-prec	225	29	2.1_{10-8}
$2e^{3 \cdot 5(x^2+y^2)}$	no precond.	961	200	5.2_{10-2}
$2e^{3 \cdot 5(x^2+y^2)}$	$L_1 D_1^{-1} U_1$-prec	961	47	3.1_{10-8}
$20e^{3 \cdot 5(x^2+y^2)}$	no precond.	225	200	2.9_{10-1}
$20e^{3 \cdot 5(x^2+y^2)}$	$L_1 D_1^{-1} U_1$-prec	225	14	2.0_{10-9}
$20e^{3 \cdot 5(x^2+y^2)}$	no precond.	961	200	6.5_{10-1}
$20e^{3 \cdot 5(x^2+y^2)}$	$L_1 D_1^{-1} U_1$-prec	961	26	2.2_{10-9}

5.4. $-u''_{xx} - (1+y^2)\,u''_{yy} + u'_x + (1+y^2)u'_y = f$

This example has been taken from Houstis *et al.* [10] and is chosen since the various derivative terms appear with unequal coefficients. The equation is discretised with an equidistant grid over the region $[0, 1] * [0, 1]$, along the boundaries of which a Dirichlet boundary condition holds. The right-hand side, as well as the solution along the boundary, is chosen such that $u = e^{x+y} + (x^2 - x)^2 \ln(1 + y^2)$ is the solution of the equation. In our experiments we have chosen no special initial values for the Manteuffel-parameters d and c ($d = 1.0$, $c = 0.0$). This implies that the first 20 iterationsteps in most cases are required only for computing better approximations. This influences the number of iterations to gain one decimal as we have computed it. In order to get a better impression of the asymptotic speed of convergence we have also computed the number of iterations to gain 1 decimal in the following way:

$$N^* = (n_2 - n_1)/^{10}\log(r_1/r_2),$$

where n_2 = the number of iterations to reach a residual less than $_{10} - 6$, r_2 = the value of the n_2th residual, and n_1 and r_1 analogously corresponding to a residual less than $_{10} - 3$. The results are summarized in Table X.

5.5. *Discretisation of First Order Terms by Backward/Forward Differences*

The matrix that arises from standard 5-point discretisation of the second order and linear terms in the partial differential equation is a diagonally dominant symmetric M-matrix. If the first order derivative terms are discretised by backward or forward differences depending on the sign of the functions $G(x, y)$ and $H(x, y)$ (see Section 2) in such a way that the contribution to the diagonal elements in the discretisation matrix is positive, then the resulting matrix A is a diagonally dominant non-symmetric M-matrix (see Section 4). According to Meijerink and van der Vorst [15] an incomplete Crout decomposition K exists, the factors are well conditioned and due to Theorem 3 all the eigenvalues of the preconditioned matrix $K^{-1}A$ have positive real part. In this case we can safely use Tchebycheff iteration with $L_1 D_1^{-1} U_1$-preconditioning. A disadvantage of the forward/backward differences is that the

TABLE X

Iteration Results for the Discretised Linear System of (5.4)

Method	Number of unknowns	Number of iterations	Final residual / Initial residual	N*
no precond.	225	80	2.3_{10-4}	—
INC. CHOLESKI	225	25	1.1_{10-8}	2.18
$L_\sigma D_\sigma^{-1} U_\sigma$, $\sigma = 1.0$	225	38	2.0_{10-8}	3.14
no precond.	841	299	1.6_{10-8}	40.2
INC. CHOLESKI	841	36	9.2_{10-9}	2.92
$L_\sigma D_\sigma^{-1} U_\sigma$, $\sigma = 1.0$	841	45	9.1_{10-9}	3.05

TABLE XI

Iteration Results for the Discretised Linear System of (5.5)

		Backward differences			Central differences		
a	No. of unknowns	No. of iterations	Final res. / Init. res.	Discr. error	No of iterations	Final. res. / Init. res.	Discr. error
$2e^{3\cdot 5(x^2+y^2)}$	225	31	$1.1_{10^{-8}}$	0.252	29	$2.1_{10^{-8}}$	0.219
$2e^{3\cdot 5(x^2+y^2)}$	961	49	$3.0_{10^{-8}}$	0.169	47	$3.1_{10^{-8}}$	0.059
$20e^{3\cdot 5(x^2+y^2)}$	225	19	$1.2_{10^{-9}}$	0.238	14	$2.0_{10^{-9}}$	0.512
$20e^{3\cdot 5(x^2+y^2)}$	961	30	$2.2_{10^{-9}}$	0.167	26	$2.2_{10^{-9}}$	0.066

discretisation error is of lower order than when central differences are applied. We will not discuss here the arguments for a choice between central or backward/forward differences, we only consider the effects of the preconditioning for both cases.

We present here the results achieved for the equation $-u''_{xx} - u''_{yy} + ((\partial/\partial x)(au) + a(\partial u/\partial x))/2 = f$, which has been discussed in Section 5.3.

For given function a, the function f and the Dirichlet boundary conditions have been chosen in such a way that the solution is

$$u = \sin \pi x \sin \pi y \, e^{(x/2+y)^3}.$$

For both backward/forward and central differences we have computed the discretisation error as the maximum of the absolute differences of the discretised solution and the exact solution on the gridpoints.

The results are given in Table XI. In all cases the $L_1 D_1^{-1} U_1$-preconditioning has been used.

6. Conclusions

The major conclusion is that the Tchebycheff iteration with suitable preconditioning can be an efficient solutionmethod for a class of problems. From the theory and the experiments we conclude that for linear systems coming from pde's where the first order terms contribute only little, a preconditioning constructed by incomplete Crout decomposition is a good choice. If the matrix is nearly symmetric then a Fast Poisson Solver or an incomplete Choleski decomposition for the symmetric part can be helpful.

If the first order terms are dominant, e.g., preventing the matrix to be an M-matrix, then incomplete Crout decomposition by itself fails due to ill-conditioning of the factors. However, this can be repaired by introducing a parameter in order to stabilize the decomposition (Section 4.3). In practical situations the parameterless

decomposition $L_{EQ}D_{EQ}^{-1}U_{EQ}$, see Section 4.3, scheme (4.6), may be a very good alternative, although sometimes some efficiency might be lost as compared to the parameter version. The Eisenstat implementation as described in Section 4.4 deserves special attention, making essentially the preconditioned Tchebycheff iteration almost as cheap as the unpreconditioned one per iteration for the preconditionings described in Section 4.3. This possibility is lost for the Fast Poisson Solver and the incomplete Choleski decompositions mentioned above. Of course it is also possible to construct more elaborate incomplete Crout-decompositions of a given matrix, as described by Meijerink and van der Vorst [16]. Their use decreases in general the number of iterations, but since they take more memory and more computational cost per iteration and since the Eisenstat implementation cannot be used for them, it is uncertain whether they really can compete (see also van der Vorst and Van Kats [19]).

ACKNOWLEDGMENTS

The algorithms presented in this report are all based on Manteuffel's adaptive Tchebycheff iteration. I wish to express my thanks to Tom Manteuffel for making available the FORTRAN-code of his program. Moreover personal communications with him proved very helpful. Also I wish to acknowledge valuable discussions with professor Van der Sluis and stimulating discussions with Gene Golub. All three referees have helped me very much in trying to clarify some rather vague points in the first version of this paper.

REFERENCES

1. O. AXELSSON, *BIT* **13** (1972), 443–467.
2. B. L. BUZBEE, "Program Description of TBPSDN—Fast Direct Poisson Solver," LASL, 1973.
3. P. CONCUS, G. H. GOLUB, AND D. P. O'LEARY, A generalized conjugate gradient method for the numerical solution of elliptic partial differential equations, *in* "Proceedings, Symp. on Sparse Matrix Computation" J. R. Bunch and D. J. Rose, Eds.), New York, 1975.
4. P. CONCUS AND G. H. GOLUB, A generalized conjugate gradient method for non-symmetric systems of linear equations, *in* "Proceedings, Second Internat. Symp. on Computing Methods in Applied Sciences and Engineering, IRIA (Paris, Dec. 1975)," (R. Glowinski and J. L. Lions, Eds.), Lecture Notes in Economics and Mathematical Systems No. 134, Springer-Verlag, Berlin, 1976.
5. S. C. EISENSTAT, "Efficient Implementation of a Class of Preconditioned Conjugate Gradient Methods," Research Report No. 185, Yale University, 1980.
6. G. H. GOLUB AND W. KAHAN, *Siam J. Numer. Anal.* **2** (1965), 205–224.
7. I. GUSTAFSSON, *BIT* **18** (1977), 142–156.
8. M. R. HESTENES AND E. L. STIEFEL, *N.B.S.J. of Res.* **49** (1952) 409–436.
9. A. S. HOUSEHOLDER, "The Theory of Matrices in Numerical Analysis," Blaisdell, New York, 1964.
10. E. N. HOUSTIS, R. E. LYNCH, J. R. RICE, AND T. S. PAPATHEODOROU, *J. Comput. Phys.* **27**(3) (1978), 323–350.
11. D. S. KERSHAW, *J. Comput. Phys.* **26**(1) (1978), 43–65.
12. T. A. MANTEUFFEL, *Numer. Math.* **28** (1977), 307–327.
13. T. A. MANTEUFFEL, "Adaptive Procedure for Estimating Parameters for the Non-Symmetric Tchebycheff Iteration," Sandia Laboratories Report SAND 77-8239, Livermore, 1977.

14. T. A. MANTEUFFEL, "The Shifted Incomplete Cholesky Factorization," Sandia Laboratories Report SAND 78-8226, Livermore, 1978.
15. J. A. MEIJERINK AND H. A. VAN DER VORST, *Math. Comput.* **31**(137) (1977), 148–162.
16. J. A. MEIJERINK AND H. A. VAN DER VORST, *J. Comput. Phys.*, in press.
17. C. C. PAIGE, *Siam J. Numer. Anal.* **11** (1974), 197–209.
18. H. L. STONE, *Siam J. Numer. Anal.* **5** (1968), 530–558.
19. H. A. VAN DER VORST AND J. M. VAN KATS, "Manteuffel's Algorithm with Preconditioning for the Iterative Solution of Certain Sparse Linear Systems with a Non-Symmetric Matrix," Technical Report, TR-11, ACCU, Utrecht, 1979.
20. R. S. VARGA, "Matrix Iterative Analysis," Prentice–Hall, Englewood Cliffs N. J., 1962.
21. O. WIDLUND, *Siam J. Numer. Anal.* **15**(4) (1978), 801–812.