SIAM J. NUMER. ANAL.
Vol. 5, No. 3, September 1968
Printed in U.S.A.

AN APPROXIMATE FACTORIZATION PROCEDURE FOR SOLVING SELF-ADJOINT ELLIPTIC DIFFERENCE EQUATIONS*

TODD DUPONT†, RICHARD P. KENDALL AND H. H. RACHFORD, JR.‡

1. Introduction. Consider the Dirichlet problem for the self-adjoint differential equation

$$\frac{\partial}{\partial x_1}\left(a_1(x)\frac{\partial}{\partial x_1}u\right)+\frac{\partial}{\partial x_2}\left(a_2(x)\frac{\partial}{\partial x_2}u\right)+q(x)u=f(x), \tag{1.1}$$

where x lies in a bounded plane region, $a_i(x) > \eta_i > 0$, $q(x) \leqq 0$ and $a_i(x)$, $q(x)$, $f(x)$ are sufficiently smooth. Many iterative procedures for solving numerically the usual five-point discrete analogue

$$Aw = r \tag{1.2}$$

associated with (1.1) have been proposed. In particular, Stone [6], Buleev [1] and others have suggested that implicit procedures based on approxi mately factoring the discrete operator A into easily invertible factors might facilitate the solution of (1.2). In this vein, we consider the implicit procedure

$$(A + B)w_{n+1} = (A + B)w_n - \omega(Aw_n - r), \tag{1.3}$$

where B is chosen so that

$$A + B = LL^*, \tag{1.4}$$

where L is a lower triangular matrix with no more than three entries per row.

For an appropriate fixed choice of B and the parameter ω, an estimate of $O(h^{-3}\log\epsilon^{-1})$ is obtained for the computational work required to reduce the A-norm of the error by a factor of ϵ. A Chebyshev sequence of parameters, $\{\omega_n\}$, is shown to give a work estimate of $O(h^{-5/2}\log\epsilon^{-1})$. Here h is the step-size in a uniform grid over which (1.2) is solved.

2. The difference equations. Suppose we wish to approximate the solution u of the Dirichlet problem associated with (1.1) on a bounded plane region Ω. Let $\bar{\Omega}_h$ be the points (jh, kh) which lie in $\bar{\Omega}$, where h is a small positive number and j and k are integers. Let Ω_h be the N points (jh, kh) in $\bar{\Omega}_h$ such that $((j + 1)h, kh)$, $((j - 1)h, kh)$, $(jh, (k + 1)h)$ and

* Received by the editors March 8, 1968. This research was supported by Esso Production Research Company, Houston, Texas.

† Department of Mathematics, University of Chicago, Chicago, Illinois 60637.
‡ Department of Mathematics, Rice University, Houston, Texas 77001.

$(jh, (k-1)h)$ are in $\bar{\Omega}_h$. Let $\partial\Omega_h$ be the points in $\bar{\Omega}_h$ which are not in Ω_h. We shall assume that the value of u is known at the points in $\partial\Omega_h$. Let $w_{j,k}$ denote the value of the grid function w at the point (jh, kh) of Ω_h. At each point of Ω_h we shall approximate the differential operator in (1.1) by a self-adjoint difference operator A which, when applied to a grid function w at the point (jh, kh), assumes the form

$$(Aw)_{j,k} = b_{j,k}w_{j,k} + c_{j,k}w_{j+1,k} + f_{j,k}w_{j,k+1} + c_{j-1,k}w_{j-1,k} + f_{j,k-1}w_{j,k-1}. \tag{2.1}$$

The coefficients will be taken as

$$\begin{aligned} c_{j,k} &= -h^{-2}a_1((j+\tfrac{1}{2})h, kh), \\ f_{j,k} &= -h^{-2}a_2(jh, (k+\tfrac{1}{2})h), \\ b_{j,k} &= h^{-2}[a_1((j+\tfrac{1}{2})h, kh) + a_1((j-\tfrac{1}{2})h, kh) + a_2(jh, (k+\tfrac{1}{2})h) \\ &\quad + a_2(jh, (k-\tfrac{1}{2})h)] - q(jh, kh). \end{aligned} \tag{2.2}$$

This approximation of the difference operator is second order correct (for smooth a_1 and a_2) and is one that is frequently used in practice. We choose this specific difference operator only for definiteness, however, since the procedure given below for solving the difference equations works as well for several similar five-point difference operators.

When A is written as a matrix, the coefficients which multiply $w_{j,k}$ do not appear for (jh, kh) not in Ω_h, since such terms are incorporated into r of (1.2). We adopt the convention that these coefficients are zero; with this convention, (2.1) gives the value of the matrix A operating on the grid function w. This convention can be stated as

$$\begin{aligned} c_{j,k} &= 0 \quad \text{if} \quad (jh, kh) \notin \Omega_h \quad \text{or} \quad ((j+1)h, kh) \notin \Omega_h; \\ f_{j,k} &= 0 \quad \text{if} \quad (jh, kh) \notin \Omega_h \quad \text{or} \quad (jh, (k+1)h) \notin \Omega_h. \end{aligned} \tag{2.3}$$

We remark that A so defined is symmetric and positive definite.

Figure 1 is useful in visualizing which grid points are involved in the equation associated with (jh, kh) and what the coefficients are in that equation. This type of graph-theoretic tool will be useful to us several times in the manipulation of matrices.

3. The factorization algorithm. Ideally, since A is self-adjoint, one might like to factor the discrete operator A above into

$$A = LL^*,$$

where L is given by

$$(Lw)_{j,k} = v_{j,k}w_{j,k} + t_{j-1,k}w_{j-1,k} + g_{j,k-1}w_{j,k-1},$$

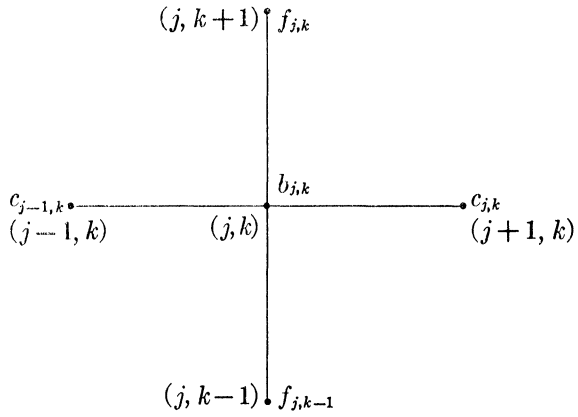

FIG. 1

for such an operator is easily invertible (in matrix form, L can be made lower triangular and has no more than three entries per row). This is not possible. However, we shall show that it is possible to factor A approximately in the form (1.4), with B defined by

$$
\begin{aligned}
(Bw)_{j,k} &= (\tilde{B}w)_{j,k} + (Dw)_{j,k}, \\
(\tilde{B}w)_{j,k} &= h_{j,k}w_{j-1,k+1} + h_{j+1,k-1}w_{j+1,k-1} + (-h_{j,k} - h_{j+1,k-1})w_{j,k}, \\
(Dw)_{j,k} &= \alpha_{j,k}b_{j,k}w_{j,k},
\end{aligned} \tag{3.1}
$$

where $\alpha_{j,k}$ is an iteration parameter associated with the point (jh, kh). Figures 2–4 may be of assistance in visualizing the points involved in the equations associated with the point (jh, kh); in each, the center point is (jh, kh). If we form the product LL^*, we see that it involves the points and coefficients shown in Fig. 5.

Thus, (1.4) follows if v, t, g and h satisfy

$$
\begin{aligned}
v_{j,k} &= [b_{j,k}(1 + \alpha_{j,k}) - h_{j,k} - h_{j+1,k-1} - t_{j-1,k}^2 - g_{j,k-1}^2]^{1/2}, \\
g_{j,k} &= \frac{f_{j,k}}{v_{j,k}}, \\
t_{j,k} &= \frac{c_{j,k}}{v_{j,k}}, \\
h_{j,k} &= t_{j-1,k}g_{j-1,k},
\end{aligned} \tag{3.2}
$$

where the convention that the $c_{j,k}$'s and $f_{j,k}$'s are zero if these are coefficients corresponding to points outside the region allows us to use (3.2) even at the boundaries. From the recursiveness of these relations, we can

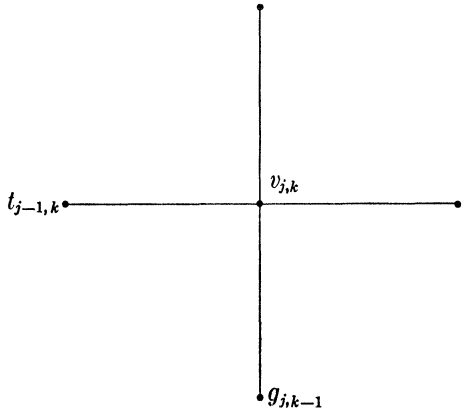

FIG. 2. L

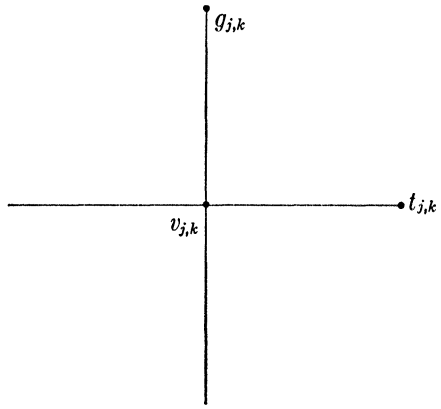

FIG. 3. L^*

see that the factorization is possible provided $v_{j,k}^2$ is positive. Let us show that this is the case for $\alpha_{j,k} \geqq 0$.

Note that for (jh, kh) in Ω_h,

$$b_{j,k} = -(\tilde{c}_{j,k} + \tilde{f}_{j,k} + \tilde{c}_{j-1,k} + \tilde{f}_{j,k-1} + q_{j,k}),$$

where

$$\tilde{c}_{l,m} = -h^{-2}a_1((l + \tfrac{1}{2})h, mh),$$

$$\tilde{f}_{l,m} = -h^{-2}a_2(lh, (m + \tfrac{1}{2})h),$$

$$q_{j,k} = q(x_{j,k}).$$

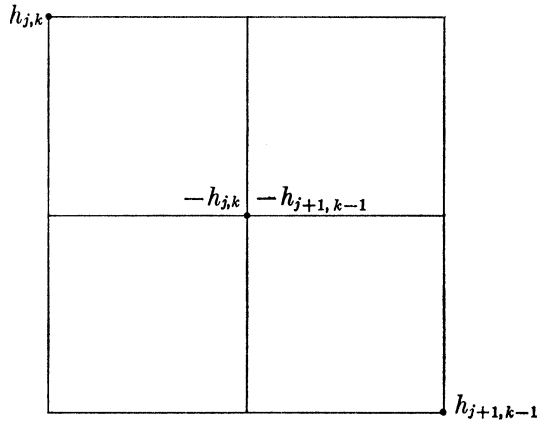

FIG. 4. $\tilde{B}$

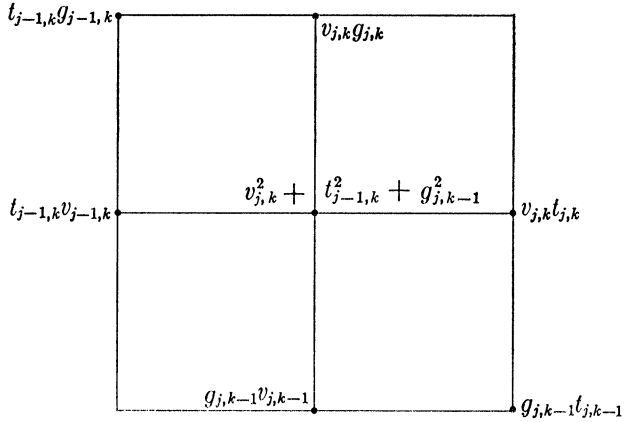

FIG. 5. LL^*

Then, for all l, m,

$$\tilde{c}_{l,m} + \tilde{f}_{l,m} < -(\eta_1 + \eta_2)h^{-2}.$$

LEMMA 1. *For the factorization above and* $\alpha_{j,k} \geqq 0$, *we have*

$$\frac{v_{j,k}^2}{-(\tilde{c}_{j,k} + \tilde{f}_{j,k})} \geqq \beta_{j,k} > 0, \tag{3.3}$$

$$\begin{aligned} 0 \leqq h_{j+1,k} &\leqq -\frac{c_{j,k}f_{j,k}}{\beta_{j,k}(\tilde{c}_{j,k} + \tilde{f}_{j,k})} \\ &\leqq -\frac{c_{j,k}f_{j,k}}{\beta(\tilde{c}_{j,k} + \tilde{f}_{j,k})}, \end{aligned} \tag{3.4}$$

where

$$\rho_{j,k} = \frac{\tilde{c}_{j-1,k} + \tilde{f}_{j,k-1}}{\tilde{c}_{j,k} + \tilde{f}_{j,k}},$$

$$\tilde{\beta}_{j,k} = \frac{(1 + \alpha_{j,k})(1 + \rho_{j,k}) + [((1 + \alpha_{j,k})(1 + \rho_{j,k}))^2 - 4\rho_{j,k}]^{1/2}}{2},$$

$$\beta_{j,k} = \min_{l \leqq j, m \leqq k} \tilde{\beta}_{l,m},$$

$$\beta = \min \beta_{j,k}.$$

Proof. We have by (3.2),

$$v_{j,k}^2 = -(1 + \alpha_{j,k})(\tilde{c}_{j,k} + \tilde{f}_{j,k} + \tilde{c}_{j-1,k} + \tilde{f}_{j,k-1} + q_{j,k}) - \frac{f_{j-1,k} c_{j-1,k}}{v_{j-1,k}^2} - \frac{f_{j,k-1} c_{j,k-1}}{v_{j,k-1}^2} - \left(\frac{c_{j-1,k}}{v_{j-1,k}}\right)^2 - \left(\frac{f_{j,k-1}}{v_{j,k-1}}\right)^2.$$

This formula is correct even near the boundary, if we define $v_{j,k} = 1$ for (jh, kh) not in Ω_h. If we drop the nonnegative term $-(1 + \alpha_{j,k})q_{j,k}$ and then divide by $-\tilde{c}_{j,k} - \tilde{f}_{j,k}$, we obtain

$$\begin{aligned} \frac{v_{j,k}^2}{-\tilde{c}_{j,k} - \tilde{f}_{j,k}} \geqq (1 + \alpha_{j,k}) + (1 + \alpha_{j,k}) \frac{\tilde{c}_{j-1,k} + \tilde{f}_{j,k-1}}{\tilde{c}_{j,k} + \tilde{f}_{j,k}} \\ + \frac{c_{j-1,k}}{\tilde{c}_{j,k} + \tilde{f}_{j,k}} \left[\frac{c_{j-1,k} + f_{j-1,k}}{v_{j-1,k}^2}\right] \\ + \frac{f_{j,k-1}}{\tilde{c}_{j,k} + \tilde{f}_{j,k}} \left[\frac{c_{j,k-1} + f_{j,k-1}}{v_{j,k-1}^2}\right]. \end{aligned} \tag{3.5}$$

Suppose that for all (i, l) such that $(ih, lh) \in \Omega_h$ and $i < j$ and $l \leqq k$ or $i = j$ and $l < k$, we have

$$\frac{v_{i,l}^2}{-(\tilde{c}_{i,l} + \tilde{f}_{i,l})} \geqq \beta > 0.$$

Then the right-hand side of (3.5) is decreased by replacing the terms in brackets by $-1/\beta$ (if there are some c's or f's equal to zero, this is still true). Thus we get

$$\frac{v_{j,k}^2}{-(\tilde{c}_{j,k} + \tilde{f}_{j,k})} \geqq (1 + \alpha_{j,k}) + \rho_{j,k}\left(1 + \alpha_{j,k} - \frac{1}{\beta}\right). \tag{3.6}$$

Hence, if

$$\beta \leqq (1 + \alpha_{j,k}) + \rho_{j,k}\left(1 + \alpha_{j,k} - \frac{1}{\beta}\right), \tag{3.7}$$

we have

$$\frac{v_{j,k}^2}{-(\tilde{c}_{j,k} + \tilde{f}_{j,k})} \geqq \beta.$$

In order to establish (3.7) we first note that for $\beta > 0$, (3.7) is equivalent to

$$\psi(\beta) = \beta^2 - (1 + \alpha_{j,k})(1 + \rho_{j,k})\beta + \rho_{j,k} \leqq 0. \tag{3.8}$$

The polynomial $\psi(\beta)$ has real roots because

$$\begin{aligned}[(1 + \alpha_{j,k})(1 + \rho_{j,k})]^2 - 4\rho_{j,k} \\ = (1 + \rho_{j,k})^2(2\alpha_{j,k} + \alpha_{j,k}^2) + (1 - \rho_{j,k})^2 \geqq 0.\end{aligned} \tag{3.9}$$

Hence (3.8) holds if and only if β is between the roots of $\psi(\beta)$. The smaller root is

$$\tilde{r} = \tfrac{1}{2}(1 + \alpha_{j,k})(1 + \rho_{j,k}) - \tfrac{1}{2}[(1 + \alpha_{j,k})^2(1 + \rho_{j,k})^2 - 4\rho_{j,k}]^{1/2}. \tag{3.10}$$

If

$$[(1 + \alpha_{j,k})(1 + \rho_{j,k}) - 2]^2 \leqq (1 + \alpha_{j,k})^2(1 + \rho_{j,k})^2 - 4\rho_{j,k}, \tag{3.11}$$

then $\tilde{r} \leqq 1$. To see that (3.11) is valid, we expand the left-hand side and subtract $(1 + \alpha_{j,k})^2(1 + \rho_{j,k})^2$ from both sides; (3.11) then becomes

$$4[1 - (1 + \alpha_{j,k})(1 + \rho_{j,k})] \leqq -4\rho_{j,k}. \tag{3.12}$$

Expanding the left-hand side again and dropping the nonpositive term $-4\alpha_{j,k}(1 + \rho_{j,k})$, we see that (3.12) holds. Thus $\tilde{r}$ is less than or equal to 1. Similarly, it is easy to see that the larger root $\tilde{\beta}_{j,k}$ is greater than 1.

Since (3.8), and hence (3.7), holds for any $\beta \in [1, \beta_{j,k}]$, (3.3) will follow by induction if we can show it is satisfied for the first points, i.e., the points for which one or both of the multipliers of the terms in brackets in (3.5) are zero. We note that in the event both multipliers are zero, (3.6) holds immediately for any positive β, and $\beta \in [1, \beta_{j,k}]$ suffices for (3.7) so that (3.3) follows for this case. In the event only one multiplier is nonzero, the value in its bracket is associated with a point below or to the left of the point in question. As the induction hypothesis applies for such points, the value of the bracket does not exceed $-1/\beta$. Thus, in this case also, (3.6), and hence (3.3), must hold.

As a consequence of the conclusion above and the definition of $h_{j,k}$, we get

$$h_{j+1,k} \leqq \frac{-c_{j,k}f_{j,k}}{\beta_{j,k}(\tilde{c}_{j,k} + \tilde{f}_{j,k})}.$$

Note that whereas factorization into LL^* is preferable from the standpoint of analysis, it may be computationally more desirable to factor A into LU, where L is lower triangular and U is upper triangular with 1's on the diagonal, for this avoids the taking of square roots which is relatively costly in computer time. That the former factorization and the latter are equivalent follows from the observation that if $S = \text{diag}\ (L)$,

$$LL^* = (LS)(S^{-1}L^*) = \tilde{L}U.$$

4. Convergence and work estimate. Our convergence analysis consists of making precise the intuitively reasonable idea that if $A + B$ is "not too different" from A and we can solve $(A + B)w = f$, then we can approximately solve $Aw = f$. We shall take $A + B$ to be close to A in the sense that there is an L^2 comparability between them. This same type of idea has been used by D'Yakanov [2] and Gunn [4].

If $x = (x_1, \cdots, x_N)$ and $y = (y_1, \cdots, y_N)$, are N-component vectors, then $\langle x, y\rangle = \sum_{i=1}^{N} x_i y_i^*$, where y_i^* denotes the complex conjugate of y_i. We let $\| x \| = \| x \|_{L^2} = \langle x, x\rangle^{1/2}$. If C is an $N \times N$ complex matrix, then $\| C \| = \| C \|_{L^2} = \max_{x \neq 0} \| Cx \|/\| x \|$. If A is a positive definite $N \times N$ matrix, then $\| x \|_A = \langle Ax, x\rangle^{1/2}$ and $\| C \|_A = \max_{x \neq 0} \| Cx \|_A/\| x \|_A$.

LEMMA 2. *Let A and $A + B$ be positive definite. Suppose there exist positive numbers e_1 and e_2 such that for all nonzero x we have*

$$\frac{\langle Ax, x\rangle}{\langle (A + B)x, x\rangle} \in [e_1, e_2]. \tag{4.1}$$

Then for $0 < \omega < 2e_2^{-1}$ the sequence $\{w_n\}$ defined by (1.3) converges to $w = A^{-1}r$. Further, for $\omega = 2(e_1 + e_2)^{-1}$ the A-norm of the error is reduced by a factor of at least $(e_2 - e_1)/(e_2 + e_1)$ in each iteration.

Proof. Let $v_n = w_n - w$; then v_n satisfies

$$(A + B)v_{n+1} = (A + B)v_n - \omega A v_n. \tag{4.2}$$

If we multiply (4.2) by $A^{1/2}(A + B)^{-1}$, we see that

$$A^{1/2}v_{n+1} = (I - \omega A^{1/2}(A + B)^{-1}A^{1/2})A^{1/2}v_n. \tag{4.3}$$

Since

$$\| x \|_A = \| A^{1/2}x \|_{L^2},$$

we see from (4.3) that

$$\| v_{n+1} \|_A \leqq \| I - \omega A^{1/2}(A + B)^{-1}A^{1/2} \|_{L^2} \| v_n \|_A.$$

Note that $I - \omega A^{1/2}(A + B)^{-1}A^{1/2}$ is Hermitian and thus its L^2-norm is its spectral radius. Clearly, the eigenvalues of $A^{1/2}(A + B)^{-1}A^{1/2}$ are

between the minimum and the maximum of

$$\frac{\langle A^{1/2}(A + B)^{-1}A^{1/2}x, x\rangle}{\| x \|^2}. \tag{4.4}$$

If we let

$$x = [A^{1/2}(A + B)^{-1}A^{1/2}]^{-1/2}A^{1/2}y,$$

then (4.4) becomes

$$\frac{\langle Ay, y\rangle}{\langle (A + B)y, y\rangle}.$$

Thus by (4.1) the eigenvalues of $A^{1/2}(A + B)^{-1}A^{1/2}$ are in the interval $[e_1, e_2]$. Hence

$$\| I - \omega A^{1/2}(A + B)^{-1}A^{1/2} \| \leqq \max\{1 - \omega e_1, \omega e_2 - 1\}$$

and the result follows.

In order to get bounds for e_1 and e_2, we shall use the following lemma.

LEMMA 3. *Let c and f be positive and let a, b, e be complex; then*

$$\frac{cf}{c + f}|a - b|^2 \leqq f|a - e|^2 + c|e - b|^2. \tag{4.5}$$

Proof. Equation (4.5) holds if and only if

$$|a - b|^2 \leqq \left(\frac{cf}{c + f}\right)^{-1} f|a - e|^2 + \left(\frac{cf}{c + f}\right)^{-1} c|e - b|^2. \tag{4.6}$$

The right-hand side of (4.6) can be written as

$$(1 + \epsilon)|a - e|^2 + (1 + \epsilon^{-1})|e - b|^2,$$

where $\epsilon = fc^{-1}$. But it is well known that

$$|a - b|^2 \leqq (1 + \epsilon)|a - e|^2 + (1 + \epsilon^{-1})|e - b|^2$$

for any $\epsilon > 0$.

By an easy computation (summation by parts) we obtain

$$\langle \tilde{B}x, x\rangle = -\sum_{S_h} h_{j+1,k}|x_{j+1,k} - x_{j,k+1}|^2 \tag{4.7}$$

and

$$\begin{aligned}\langle Ax, x\rangle = &-\sum_{\Omega_h}[c_{j,k}|x_{j,k} - x_{j+1,k}|^2 + f_{j,k}|x_{j,k} - x_{j,k+1}|^2] \\ &+ \sum_{\Omega_h}(b_{j,k} + c_{j,k} + f_{j,k} + c_{j-1,k} + f_{j,k-1})|x_{j,k}|^2,\end{aligned} \tag{4.8}$$

where $S_h = \{(j, k); c_{j,k}f_{j,k} \neq 0\}$.

LEMMA 4. *If* $\alpha_{j,k} = C_0 h^2$ *for* $C_0 > 0$ *independent of* h, $a_1(x)$ *and* $a_2(x)$ *are sufficiently smooth, and* $A + B = LL^*$ *is obtained from* (3.2), *then there exist positive numbers* e_1 *and* e_2 *such that*

$$\frac{\langle Aw, w\rangle}{\langle (A+B)w, w\rangle} \in [e_1, e_2], \tag{4.9}$$

where e_1 *is independent of* h *and* $e_2 = O(h^{-1})$. (*We remark that sufficiently smooth could be replaced by continuously differentiable in* $\bar{\Omega}$, *but all that is actually needed is a one-sided Lipschitz condition on* a_i.)

Proof. We replace $h_{j+1,k}$ in (4.7) by its upper bound from (3.4):

$$\frac{-c_{j,k} f_{j,k}}{\beta(c_{j,k} + f_{j,k})}.$$

(Note that $(j, k) \in S_h$ implies that $c_{j,k} = \tilde{c}_{j,k}$ and $f_{j,k} = \tilde{f}_{j,k}$.) We next apply Lemma 3 to each point (j, k) in S_h to obtain the inequality

$$\frac{1}{\beta} \sum_{S_h} [c_{j,k} \mid x_{j,k} - x_{j+1,k} \mid^2 + f_{j,k} \mid x_{j,k} - x_{j,k+1} \mid^2] \leqq \langle \tilde{B}x, x\rangle. \tag{4.10}$$

The left-hand side is then decreased to $-\langle Ax, x\rangle/\beta$ to give

$$-\frac{1}{\beta}\langle Ax, x\rangle \leqq \langle \tilde{B}x, x\rangle \leqq 0. \tag{4.11}$$

Note that

$$0 \leqq \frac{\langle Dx, x\rangle}{\langle Ax, x\rangle} \leqq k_1, \tag{4.12}$$

where k_1 is independent of h. We write $\langle Ax, x\rangle/\langle (A + B)x, x\rangle$ as

$$\frac{1}{1 + \langle \tilde{B}x, x\rangle/\langle Ax, x\rangle + \langle Dx, x\rangle/\langle Ax, x\rangle}, \tag{4.13}$$

and use (4.13), (4.12) and (4.11) to yield

$$\frac{1}{1+k_1} \leqq \frac{\langle Ax, x\rangle}{\langle (A+B)x, x\rangle} \leqq \frac{1}{1 - 1/\beta}. \tag{4.14}$$

Thus the lemma will hold if

$$\beta \geqq 1 + k_2 h \tag{4.15}$$

for some $k_2 > 0$ independent of h. If the a_i's are smooth, we shall have a k_3 independent of h such that for all (j, k),

$$| \rho_{j,k} - 1 | \leqq k_3 h. \tag{4.16}$$

Let (j, k) be a point such that $\beta = \tilde{\beta}_{j,k}$; then

$$\beta = 1 + \tfrac{1}{2}[\alpha_{j,k} + \alpha_{j,k}\rho_{j,k} + (\rho_{j,k} - 1) + \{(\alpha_{j,k}^2 + 2\alpha_{j,k})(\rho_{j,k} + 1)^2 + (\rho_{j,k} - 1)^2\}^{1/2}].$$

If we drop the first and second terms in the brackets and the $\alpha_{j,k}^2$ terms in the braces, then estimate the square root by a first order expansion about $(\rho_{j,k} - 1)^2$, we get

$$\begin{aligned}\beta &\geqq 1 + \frac{1}{2}\left[(\rho_{j,k} - 1) + |1 - \rho_{j,k}| + \frac{\alpha_{j,k}(\rho_{j,k} + 1)^2}{\{2\alpha_{j,k}(\rho_{j,k} + 1)^2 + (\rho_{j,k} - 1)^2\}^{1/2}}\right] \\ &\geqq 1 + \frac{1}{2}\frac{\alpha_{j,k}^{1/2}(\rho_{j,k} + 1)^2}{\{2(\rho_{j,k} + 1)^2 + (\rho_{j,k} - 1)^2\alpha_{j,k}^{-1}\}^{1/2}}.\end{aligned} \tag{4.17}$$

If we now use $\alpha_{j,k} = C_0 h^2$ and (4.16), we see that (4.15) holds. Thus the proof is complete.

THEOREM 1. *If we take $\alpha_{i,j} = C_0 h^2$, C_0 positive and independent of h, and if the a_i's are sufficiently smooth and if $A + B = LL^*$ is obtained from* (3.2), *then for an appropriate choice of ω the iteration* (1.3) *gives a sequence w_n which converges to $A^{-1}r$, and the number of arithmetic operations necessary to reduce the A-norm of the error by a factor ϵ is no more than*

$$O(h^{-3} \log \epsilon^{-1}).$$

Proof. Lemma 2 shows that for ω properly chosen, each iteration reduces the A-norm by a factor

$$\frac{e_2 - e_1}{e_2 + e_1} = 1 - \frac{2}{1 + e_2/e_1}.$$

Lemma 4 says that $e_2/e_1 = O(h^{-1})$. Thus for some k_4, positive and independent of h, the A-norm is reduced by a factor

$$1 - k_4 h$$

in each iteration. Thus it requires $O(h^{-1} \log \epsilon^{-1})$ iterations to reduce the error by a factor ϵ. The conclusion follows as it requires $O(h^{-2})$ arithmetic operations to carry out each iteration.

5. Extension to more general equations. In this section we shall extend the work estimate obtained in Theorem 1 to cases in which the coefficients are not smooth. We shall assume that $0 < \eta_i \leqq a_i(x) \leqq M_i$, and that we

pick the coefficients $c_{i,j}$ and $f_{i,j}$ in such a fashion that

$$h^{-2}\eta_1 \leqq -c_{i,j} \leqq h^{-2}M_1\,, \qquad h^{-2}\eta_2 \leqq -f_{i,j} \leqq h^{-2}M_2\,. \tag{5.1}$$

Let Γ denote the matrix associated with the above difference approximation of (1.1). Let A denote the matrix associated with a difference approximation of a problem of the same form as (1.1) with the functions a_i replaced by smooth (e.g., constant) functions which are bounded above and below by positive numbers. Under these assumptions there exist positive numbers v_1 and v_2, independent of h, such that for nonzero vectors x,

$$\frac{\langle \Gamma x, x\rangle}{\langle Ax, x\rangle} \in [v_1\,, v_2]. \tag{5.2}$$

In what is to follow, (5.2) is the important property of Γ.

THEOREM 2. *Let A and Γ be as above. Let $LL^* = A + B$ be obtained from* (3.2) *using A and $\alpha_{j,k} = C_0h^2$. Then for ω properly chosen, the sequence $\{w_n\}$ defined by*

$$LL^*w_{n+1} = LL^*w_n - \omega(\Gamma w_n - r) \tag{5.3}$$

converges to $\Gamma^{-1}r$ and the number of arithmetic operations required to reduce the Γ-norm of the error by a factor ϵ is

$$O(h^{-3}\log \epsilon^{-1}).$$

Proof. If we write

$$\frac{\langle \Gamma x, x\rangle}{\langle LL^*x, x\rangle} = \frac{\langle \Gamma x, x\rangle}{\langle Ax, x\rangle}\,\frac{\langle Ax, x\rangle}{\langle LL^*x, x\rangle}\,,$$

it follows from (5.2) and Lemma 4 that

$$\frac{\langle \Gamma x, x\rangle}{\langle LL^*x, x\rangle} \in [v_1e_1\,, v_2e_2],$$

where v_1, v_2, e_1 are independent of h and $e_2 = O(h^{-1})$. The conclusion follows exactly as in Theorem 1 from Lemma 2.

6. Chebyshev sequence of parameters. In this section we shall use a sequence of ω_n's instead of a single ω, and in this fashion we are able to obtain a work estimate of

$$O(h^{-5/2}\log \epsilon^{-1})$$

for the reduction of the error by a factor ϵ. We shall consider the iteration

$$LL^*w_{m+1} = LL^*w_m - \omega_{m+1}(Aw_m - r). \tag{6.1}$$

Just as in Lemma 2, we see that the error v_m satisfies

$$\| v_m \|_A \leqq \left\| \prod_{i=1}^{m} (I - \omega_i A^{1/2}(LL^*)^{-1}A^{1/2}) \right\|_{L^2} \| v_0 \|_A . \tag{6.2}$$

The norm of the matrix in (6.2) is less than

$$\max_{x \in [e_1, e_2]} \left| \prod_{i=1}^{m} (1 - \omega_i x) \right| , \tag{6.3}$$

where the interval $[e_1, e_2]$ is such that

$$\frac{\langle Ax, x \rangle}{\langle LL^*x, x \rangle} \in [e_1, e_2]. \tag{6.4}$$

By classical Chebyshev analysis, the polynomial $P_m(x)$ of degree m with the smallest maximum absolute value on $[e_1, e_2]$ which has the property that $P_m(0) = 1$ is given by

$$P_m(x) = \frac{T_m((e_2 + e_1 - 2x)/(e_2 - e_1))}{T_m((e_2 + e_1)/(e_2 - e_1))} , \tag{6.5}$$

where T_m is the Chebyshev polynomial of degree m. Its maximum absolute value on $[e_1, e_2]$ is

$$\left(T_m \left[\frac{e_1 + e_2}{e_2 - e_1} \right] \right)^{-1} . \tag{6.6}$$

In §4 and §5 we showed how to get LL^* from A such that e_1 and e_2 of (6.4) are positive, e_1 is independent of h, and e_2 is $O(h^{-1})$. (LL^* comes by directly factoring, using (3.2) or by smoothing the a_i's and then factoring.) Thus

$$\frac{e_1 + e_2}{e_2 - e_1} = 1 + O(h).$$

Hence we need take only

$$m = O(h^{-1/2} \log \epsilon^{-1}) \tag{6.7}$$

to make (6.6) less than ϵ. (In order to get (6.7), use $T_m(x) = \cosh (m \cosh^{-1} x)$ for $x \geqq 1$, and $\cosh^{-1} x = \log (x + (x^2 - 1)^{1/2})$.)

THEOREM 3. *Let A be the matrix associated with a difference approximation of* (1.1) *and let LL^* be obtained as directed in* §4 *and* §5 *such that* (6.3) *holds with e_1 and e_2 positive and $e_2/e_1 = O(h^{-1})$. Then given $\epsilon > 0$ we can take $\{\omega_i\}$, $i = 1, \cdots, m$, such that if w_m is defined by* (6.1) *then*

$$\| w_m - A^{-1}r \|_A \leqq \epsilon \| w_0 - A^{-1}r \| . \tag{6.8}$$

Further, we may take m to be

$$O(h^{-1/2} \log \epsilon^{-1})$$

so that the total number of arithmetic operations involved in finding w_m is

$$O(h^{-5/2} \log \epsilon^{-1}). \tag{6.9}$$

The actual Chebyshev iteration may be carried out as follows. Let U_k denote the result of using a Chebyshev sequence of length k. By a modification of the technique of Stiefel [5] we can calculate U_{k+1} by

$$\begin{aligned} U_{k+1} &= U_k + \delta U_k, \\ \delta U_k &= q_{1,k}(LL^*)^{-1}(r - AU_k) + q_{2,k}\delta U_{k-1}, \\ \delta U_0 &= \left(\frac{2}{e_1 + e_2}\right)(LL^*)^{-1}(r - AU_0), \end{aligned} \tag{6.10}$$

where

$$\begin{aligned} q_{1,k} &= \frac{4\cosh(k\theta)}{(b-a)\cosh((k+1)\theta)}, \\ q_{2,k} &= \frac{\cosh((k-1)\theta)}{\cosh((k+1)\theta)}, \\ \theta &= \cosh^{-1}\left(\frac{e_1 + e_2}{e_2 - e_1}\right), \end{aligned}$$

and the eigenvalues of $(LL^*)^{-1}A$ are contained in the interval $[e_1, e_2]$. An account of this process can be found in Forsythe and Wasow [3] at the end of Section 21.5.

Remark 1. As the reader may suspect, the analysis above is not sharp. The crude estimates used to get bounds for e_1 and e_2 in §4 are the sources of this lack of exactness. Experiments indicate that for the Dirichlet problem with $c_{j,k} = f_{j,k} = \text{const.}$, the best value of k_4 is about 2.5 times the best value the analysis gives. Also, it is clear that as

$$\frac{\max |\, c_{j,k} \,|}{\min |\, f_{j,k} \,|} \to 0, \tag{7.1}$$

the procedure (1.3) approaches exact inversion of A if the operator D is taken sufficiently small. In the case of $c_{j,k} = \text{const.}$, $f_{j,k} = \text{const.}$, the analysis gives an improved bound in the case (7.1), but it is far from sharp.

Remark 2. Equation (1.3) can be normalized as follows for efficient computation:

$$LUw_{n+1} = LUw_n - \omega(LUw_n - \tilde{B}\, w_n - \tilde{r}), \tag{7.2}$$

where diag (L) = diag (U) = identity. After initialization and factorization, (7.2) can be carried out with eight multiplications and nine additions per point per iteration.

Remark 3. In the discussion of any procedure which uses lower triangular matrices, the manner in which the points of the grid are ordered becomes important. We remark that the analysis above can be done as long as the difference equation at each point involves no more than two preceding points and two following points. Further, the factorization is identical for any two such orderings which have the property that for each difference equation, the preceding points involved are the same for either ordering. Stone [6] suggests using different orderings (which give different factorizations) in alternate iterations.

REFERENCES

[1] N. I. Buleev, *A numerical method for the solution of two-dimensional and three-dimensional equations of diffusion*, Mat. Sb., 51 (1960), pp. 227–238.

[2] E. G. D'Yakanov, *On an iterative method for the solution of finite difference equations*, Dokl. Akad. Nauk SSSR, 138 (1961), pp. 522–525.

[3] G. E. Forsythe and W. R. Wasow, *Finite Difference Methods for Partial Differential Equations*, John Wiley, New York, 1960.

[4] J. E. Gunn, *The solution of elliptic difference equations by semi-explicit iterative techniques*, this Journal, 2 (1965), pp. 24–45.

[5] E. L. Stiefel, *Kernel polynomials in linear algebra and their numerical applications*, Nat. Bur. Standards Appl. Math. Ser., 49 (1958), pp. 1–22.

[6] H. L. Stone, *Iterative solution of implicit approximations of multidimensional partial differential equations*, this Journal, 5 (1968), pp. 530–558.