This dissertation has been
microfilmed exactly as received 68-15,614

DUPONT, Todd, 1942-
A FACTORIZATION PROCEDURE FOR THE SOLUTION
OF ELLIPTIC DIFFERENCE EQUATIONS.

Rice University, Ph.D., 1968
Mathematics

University Microfilms, Inc., Ann Arbor, Michigan

RICE UNIVERSITY

A FACTORIZATION PROCEDURE FOR THE SOLUTION
OF ELLIPTIC DIFFERENCE EQUATIONS

by

Todd Dupont

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Thesis Director's signature:

A handwritten signature in cursive script, reading "James Douglas Jr.", is written over a horizontal line.

Houston, Texas

April, 1968

A Factorization Procedure for the Solution of Elliptic Difference Equations

Todd Dupont

1. Introduction

We are concerned here with a technique for the iterative solution of the systems of algebraic equations which arise from the discretization of elliptic boundary value problems. In particular, we consider here those types of systems which usually result from the discretization of the Dirichlet and Neumann problems for a bounded domain in $\mathbb{R}^n$. The procedure which is analyzed here gives an upper bound of

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

on the number of arithmetic operations necessary to reduce the norm of the error by a factor ϵ , where h is the mesh size.

In section 2 we define the iterations to be used and establish rates of convergence in an abstract setting. In section 3 we develop an approximate factorization of the matrix which comes from the discretization of a self-adjoint Dirichlet problem and make the estimations necessary to apply the results of section 2. We first obtain the work estimate above for an operator with smooth coefficients and then use this result to extend the

estimate to uniformly elliptic self-adjoint operators. In section 4 we treat the Neumann problem by using the approximate factorization. Again, we do the case for smooth coefficient operators first and use it to extend to more general operators. In section 5 we use the results of sections 2 and 3 to develop a method for the solution of the equations which result from a mildly non-linear Dirichlet problem. In section 6 we write out the version of the approximate factorization which one would use in practice.

The basic tool which we use here to get rates of convergence is that there is an L^2 comparability between the operator which we want to invert and its approximate factorization. Similar ideas have been employed by D'jakanov [2] and Gunn [4].

This work is an extension and generalization of some joint work of H. H. Rachford, Jr., R. P. Kendall, and the author [1]. The joint work was undertaken to improve and analyze a procedure due to H. L. Stone [6].

I wish to thank Jim Douglas, Jr., for his advice, encouragement, and continual dissatisfaction with what I had been able to accomplish.

2. Iterations and convergence

The first iteration which we shall consider will be used for the approximate solution of

$$(2.1) \quad Aw = f,$$

where A is a self-adjoint, positive-definite operator on a unitary space $\mathfrak{H}$. The basic idea which we shall use is that although we may not be able to solve (2.1) easily there may be an operator B on $\mathfrak{H}$ such that we can easily solve

$$(2.2) \quad (A+B)w = f,$$

and if $(A+B)$ is not too different from A , in a sense to be made precise, then we can use the fact that we can solve (2.2) to obtain an approximate solution of (2.1).

The iteration which we shall use is

$$(2.3) \quad (A+B)w_{m+1} = (A+B)w_m - \omega(Aw_m - f),$$

where $m=0,1,\dots$, and w_0 is any element of $\mathfrak{H}$. For suitable hypotheses on A , B and ω the convergence of this procedure is established in Lemma 1.

Let x and y be in the unitary space $\mathfrak{H}$ and denote the inner product of x and y by $\langle x, y \rangle$. Let $\|x\| = \langle x, x \rangle^{\frac{1}{2}}$. If A is a positive-definite operator on $\mathfrak{H}$, let $\|x\|_A = \langle Ax, x \rangle^{\frac{1}{2}}$. If G is an operator on $\mathfrak{H}$ let $\|G\| = \sup_{\|x\|=1} \|Gx\|$ and

$\|G\|_A = \sup_{\|x\|_A=1} \|Gx\|_A$. We shall call the quantity $\|x\|_A$ the

A-norm of x . Let G^* denote the adjoint of G .

Lemma 1: Let A and B be self-adjoint operators on $\mathcal{H}$. 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

$$(2.4) \quad \langle Ax, x \rangle / \langle (A+B)x, x \rangle \in [e_1, e_2].$$

Then for $\omega \in (0, 2e_2^{-1})$ the sequence $\{w_m\}$ defined by (2.3) converges to the solution of $Aw=f$. Further, for $\omega = 2(e_1 + e_2)^{-1}$ the A-norm of the error is reduced by at least a factor $(e_2 - e_1)(e_2 + e_1)^{-1}$ in each iteration.

Proof: Let w be the solution of (2.1) and $v_m = w_m - w$. Then v_m satisfies

$$(2.5) \quad v_{m+1} = (I - \omega(A+B)^{-1}A)v_m.$$

Multiplying (2.5) by $A^{\frac{1}{2}}$ we see that

$$(2.6) \quad A^{\frac{1}{2}}v_{m+1} = (I - \omega A^{\frac{1}{2}}(A+B)^{-1}A^{\frac{1}{2}})A^{\frac{1}{2}}v_m.$$

Noting that $\|x\|_A = \|A^{\frac{1}{2}}x\|$ we see that

$$(2.7) \quad \|v_{m+1}\|_A \leq \|I - \omega A^{\frac{1}{2}}(A+B)^{-1}A^{\frac{1}{2}}\| \|v_m\|_A.$$

The operator $I - \omega A^{\frac{1}{2}}(A+B)^{-1}A^{\frac{1}{2}}$ is self-adjoint and thus its norm is given by the maximum of the absolute value of its eigenvalues. We know that the eigenvalues of $A^{\frac{1}{2}}(A+B)^{-1}A^{\frac{1}{2}}$ are contained in the interval $[d_1, d_2]$, where

$$d_1 = \min \langle A^{\frac{1}{2}}(A+B)^{-1}A^{\frac{1}{2}}x, x \rangle / \langle x, x \rangle$$

$$d_2 = \max \langle A^{\frac{1}{2}}(A+B)^{-1}A^{\frac{1}{2}}x, x \rangle / \langle x, x \rangle.$$

Writing $x = (A^{\frac{1}{2}}(A+B)^{-1}A^{\frac{1}{2}})^{-\frac{1}{2}}A^{\frac{1}{2}}y$ we see that

$$\langle A^{\frac{1}{2}}(A+B)^{-1}A^{\frac{1}{2}}x, x \rangle = \langle Ay, y \rangle,$$

and

$$\begin{aligned} \langle x, x \rangle &= \langle A^{\frac{1}{2}}(A^{\frac{1}{2}}(A+B)^{-1}A^{\frac{1}{2}})^{-1}A^{\frac{1}{2}}y, y \rangle \\ &= \langle (A+B)y, y \rangle. \end{aligned}$$

Thus

$$\langle A^{\frac{1}{2}}(A+B)^{-1}A^{\frac{1}{2}}x, x \rangle / \langle x, x \rangle = \langle Ay, y \rangle / \langle (A+B)y, y \rangle.$$

Hence $[d_1, d_2] \subset [e_1, e_2]$. From this we see that the eigenvalues of $I - \omega A^{\frac{1}{2}}(A+B)^{-1}A^{\frac{1}{2}}$ are contained in the interval

$$[1 - \omega e_2, 1 - \omega e_1].$$

The conclusion follows.

If we use a sequence $\{\omega_m\}$ instead of a fixed ω in (2.3) we obtain a sequence $\{w_m\}$ defined by

$$(2.8) \quad (A+B)w_{m+1} = (A+B)w_m - \omega_{m+1}(Aw_m - f).$$

Just as before, using v_m as the error vector, we get

$$\|v_m\|_A \leq \left\| \prod_{i=1}^m (I - \omega_i A^{\frac{1}{2}}(A+B)^{-1}A^{\frac{1}{2}}) \right\| \|v_0\|_A.$$

As $\prod_{i=1}^m (I - \omega_i A^{\frac{1}{2}}(A+B)^{-1}A^{\frac{1}{2}})$ is self-adjoint its norm is given

by the maximum of the absolute values of its eigenvalues.

Thus its norm is majorized by

$$(2.9) \quad \max_{x \in [e_1, e_2]} \left| \prod_{i=1}^m (1 - \omega_i x) \right|,$$

because the eigenvalues of $A^{\frac{1}{2}}(A+B)^{-1}A^{\frac{1}{2}}$ lie in the interval $[e_1, e_2]$. If we wish to pick the ω_i 's to minimize (2.9) we are in a position to use the classical Chebyshev analysis. The polynomial, $P_m(x)$, of degree m which has minimal maximum absolute value on $[e_1, e_2]$ subject to the constraint $P_m(0)=1$ is given by

$$T_m\left(\frac{e_1+e_2-2x}{e_2-e_1}\right) [T_m\left(\frac{e_1+e_2}{e_2-e_1}\right)]^{-1},$$

where T_m is the Chebyshev polynomial of degree m . For $|x| < 1$

$$T_m(x) = \cos(m \cos^{-1}(x));$$

using this expression we can see that the ω_i 's are given by

$$(2.10) \quad 2[e_1+e_2-(e_2-e_1)\cos(\frac{(2i-1)}{2m}\pi)]^{-1}, \quad i=1, \dots, m.$$

This analysis is standard and the reader is referred to Forsythe and Wasow [3, section 21.5] or Varga [7, section 5.1]. The result of this procedure is given in Lemma 2.

Lemma 2: Let A, B, e_1 , and e_2 be as in Lemma 1 and let $\epsilon > 0$. If we take m to be the smallest integer not less than $\cosh^{-1}(\epsilon^{-1}) / \cosh^{-1}(\frac{e_2+e_1}{e_2-e_1})$ and take ω_i to be given by (2.10) and let w_i be given by (2.8) and $w=A^{-1}f$

then

$$\|w_m - w\|_A \leq \epsilon \|w_0 - w\|_A.$$

The importance of Lemma 2 is that as e_2/e_1 becomes large (as is the case in our application) the number of iterations required to reduce the error by a factor ϵ using a sequence $\{w_i\}$ is approximately proportional to the square root of the number of iterations that would be required if we used a fixed w . It should be noted that if one uses the procedure of Lemma 2 one must decide on m in advance and then calculate w_m because the w_i 's for $i < m$ are not approximations to w . However, by a modification of a technique of Stiefel [5] we can calculate directly a sequence $\{w_i\}$ where each w_i is the result of a Chebyshev sequence of length i . The algorithm for this process is given below. The arguments which give this process can be found in Forsythe and Wasow [3] at the end of section 21.5.

$$\begin{aligned} w_{m+1} &= w_m + \delta w_m, \\ (2.11) \quad \delta w_m &= q_{1,m} (A+B)^{-1} (f - Aw_m) \\ &\quad + q_{2,m} \delta w_{m-1}, \\ \delta w_0 &= 2[e_1 + e_2]^{-1} (A+B)^{-1} (f - Aw_0), \end{aligned}$$

where

$$\begin{aligned} q_{1,m} &= 4 \cosh(m\theta) / [(e_2 - e_1) \cosh((m+1)\theta)], \\ q_{2,m} &= \cosh((m-1)\theta) / \cosh((m+1)\theta), \\ \theta &= \cosh^{-1}((e_1 + e_2) / (e_2 - e_1)). \end{aligned}$$

The second type of iteration we shall consider will be used for the approximate solution of

$$(2.12) \quad Dw = f,$$

where

$$(2.13) \quad \operatorname{Re} D = \frac{1}{2}(D+D^*)$$

is positive-definite. (Note that if (2.13) is positive-definite and $x \neq 0$ then $0 < \langle \operatorname{Re} D x, x \rangle = \operatorname{Re} \langle D x, x \rangle$. Thus $D x \neq 0$ if $x \neq 0$ and hence D is invertible.) The iteration we shall consider here is a nested one. The outer iteration is due to Gunn [4]; the inner iteration is the one we have just analyzed. This process will also be useful in obtaining approximate solutions to mildly nonlinear equations.

Gunn's iteration for operators $\mathfrak{L}$ and A , is defined as

$$(2.14) \quad A w_{m+1} = A w_m - \rho \mathfrak{L}(w_m).$$

Under suitable hypotheses on $A, \mathfrak{L}$, and ρ , w_m converges to a solution of

$$(2.15) \quad \mathfrak{L}(w) = 0.$$

The precise statement of Gunn's result is Lemma 3.

Lemma 3: Let A be a positive-definite operator on $\mathfrak{H}$. Let $\mathfrak{L}$ be a not-necessarily-linear operator on $\mathfrak{H}$. Let u be a solution of (2.15). Let γ_1, γ_2 be positive and γ_3 nonnegative. Suppose that for each $w \in \mathfrak{H}$ there exists a linear operator $D(w)$ such that

$$(2.16) \quad \mathcal{L}(w) = D(w)[w-u],$$

where $D(w)$ is such that for all nonzero x

$$(2.17) \quad \begin{aligned} \gamma_1 &\leq \operatorname{Re}\langle D(w)x, x \rangle / \langle Ax, x \rangle \leq \gamma_2 \\ |\operatorname{Im}\langle D(w)x, x \rangle / \langle Ax, x \rangle| &\leq \gamma_3. \end{aligned}$$

Then for

$$(2.18) \quad \rho = 2[\gamma_1 + \gamma_2]^{-1} [1 - \{\gamma_3^2 [\gamma_3^2 + \gamma_1(\gamma_1 + \gamma_2)]^{-1}\}^{\frac{1}{2}}]$$

the sequence $\{w_m\}$ defined by (2.14) converges to u .

Further the number of iterations required to reduce the A -norm of the error by a factor ϵ is no more than

$$(2.19) \quad O([4\gamma_3^2 + \gamma_1(\gamma_1 + \gamma_2)] / \gamma_1^2) \log \epsilon^{-1}.$$

It may be that the A which is a natural choice in light of relations (2.17) is not easily inverted. In this case we might try to approximately solve (2.14) at each step by using an iteration of the form (2.3). Obviously, $A+B$ could be inserted in (2.17) in place of A , but the asymptotic work estimates would suffer in the applications to follow. If we use k inner iterations at each step and take w_m as our first guess at w_{m+1} we obtain

$$(2.20) \quad \begin{aligned} w_{m+1}^{(0)} &= w_m, \\ (A+B)w_{m+1}^{(i+1)} &= (A+B)w_{m+1}^{(i)} - w_{i+1}(Aw_{m+1}^{(i)} - b_m), \\ w_{m+1} &= w_{m+1}^{(k)}, \end{aligned}$$

where

$$b_m = Aw_m - \rho \mathcal{L}(w_m).$$

If we let

$$v_{m+1}^{(i)} = w_{m+1}^{(i)} - A^{-1}b_m$$

then

$$(A+B)v_{m+1}^{(i+1)} = (A+B)v_{m+1}^{(i)} - \omega_{i+1}Av_{m+1}^{(i)}.$$

Thus

$$v_{m+1}^{(k)} = \prod_{i=1}^k (I - \omega_i(A+B)^{-1}A)v_{m+1}^{(0)}.$$

Hence

$$w_{m+1} - A^{-1}b_m = \left[\prod_{i=1}^k (I - \omega_i(A+B)^{-1}A) \right] [w_m - A^{-1}b_m].$$

Hence

$$\begin{aligned} w_{m+1} - w_m + \rho A^{-1}\mathcal{L}(w_m) \\ = \left[\prod_{k=1}^k (I - \omega_i(A+B)^{-1}A) \right] [w_m - w_m + \rho A^{-1}\mathcal{L}(w_m)]. \end{aligned}$$

Hence

$$w_{m+1} = w_m - \rho \left[I - \prod_{i=1}^k (I - \omega_i(A+B)^{-1}A) \right] A^{-1}\mathcal{L}(w_m).$$

Now letting

$$(2.21) \quad E = A \left[I - \prod_{i=1}^k (I - \omega_i(A+B)^{-1}A) \right]^{-1}$$

we get

$$(2.22) \quad Ew_{m+1} = Ew_m - \rho \mathcal{L}(w_m).$$

We shall show that if the ω_i 's are chosen properly then the inverse taken in (2.21) exists and that $\langle Ax, x \rangle / \langle Ex, x \rangle$ can be confined to a small interval about 1. Thus we

shall show that (2.22) is an iteration of the type (2.14).

Lemma 4: Let A and $A+B$ be positive-definite operators on $\mathbb{H}$, and let e_1 and e_2 be positive numbers such that for $x \neq 0$, $x \in \mathbb{H}$,

$$(2.23) \quad \langle Ax, x \rangle / \langle (A+B)x, x \rangle \in [e_1, e_2].$$

Then given $\delta \in (0, 1)$ there exists a sequence $\{\omega_i\}_{i=1}^k$ such that for all nonzero $x \in \mathbb{H}$

$$(2.24) \quad \langle Ax, x \rangle / \langle Ex, x \rangle \in [1-\delta, 1+\delta],$$

where E is given by (2.21). Further k can be chosen to be the smallest integer not less than

$$\cosh^{-1}(\delta^{-1}) / \cosh^{-1}\left(\frac{e_1+e_2}{e_2-e_1}\right).$$

Proof: Let

$$F = \prod_{i=1}^k (I - \omega_i (A+B)^{-1} A).$$

Then F is the error propagator in the iteration (2.8).

If k and $\{\omega_i\}_{i=1}^k$ are taken as directed in Lemma 2 then we can obtain

$$\|A^{\frac{1}{2}} F A^{-\frac{1}{2}}\| \leq \delta < 1.$$

Thus $(I-F)^{-1}$ exists. Now E is self-adjoint if and only if $A^{-\frac{1}{2}} E A^{-\frac{1}{2}}$ is. Then,

$$\begin{aligned}
A^{-\frac{1}{2}}EA^{-\frac{1}{2}} &= A^{-\frac{1}{2}}A(I-F)^{-1}A^{-\frac{1}{2}} \\
&= A^{\frac{1}{2}}(I-F)^{-1}A^{-\frac{1}{2}} \\
&= (I-A^{\frac{1}{2}}FA^{-\frac{1}{2}})^{-1}, \\
A^{\frac{1}{2}}FA^{-\frac{1}{2}} &= \prod_{i=1}^k (I-\omega_i A^{\frac{1}{2}}(A+B)^{-1}A^{\frac{1}{2}}).
\end{aligned}$$

Thus as $A^{\frac{1}{2}}(A+B)^{-1}A^{\frac{1}{2}}$ is self-adjoint, $A^{\frac{1}{2}}FA^{-\frac{1}{2}}$ is self-adjoint; thus $A^{-\frac{1}{2}}EA^{-\frac{1}{2}}$ is self-adjoint: thus E is self-adjoint.

In order to establish (2.24) we write

$$x = A^{-\frac{1}{2}}[A^{-\frac{1}{2}}EA^{-\frac{1}{2}}]^{-\frac{1}{2}}z.$$

Then

$$\begin{aligned}
\langle Ax, x \rangle / \langle Ex, x \rangle &= \frac{\langle [A^{-\frac{1}{2}}EA^{-\frac{1}{2}}]^{-1}z, z \rangle}{\langle [A^{-\frac{1}{2}}EA^{-\frac{1}{2}}]^{-\frac{1}{2}}(A^{-\frac{1}{2}}EA^{-\frac{1}{2}})[A^{-\frac{1}{2}}EA^{-\frac{1}{2}}]^{-\frac{1}{2}}z, z \rangle} \\
&= \langle A^{\frac{1}{2}}E^{-1}A^{\frac{1}{2}}z, z \rangle / \langle z, z \rangle.
\end{aligned}$$

Now

$$\begin{aligned}
A^{\frac{1}{2}}E^{-1}A^{\frac{1}{2}} &= A^{\frac{1}{2}}[I-F]A^{-\frac{1}{2}} \\
&= I-A^{\frac{1}{2}}FA^{-\frac{1}{2}}.
\end{aligned}$$

Thus

$$\langle Ax, x \rangle / \langle Ex, x \rangle = 1 - \langle A^{\frac{1}{2}}FA^{-\frac{1}{2}}z, z \rangle / \|z\|^2.$$

Hence (2.24) holds because

$$\|A^{\frac{1}{2}}FA^{-\frac{1}{2}}\| \leq \delta.$$

Lemma 4 shows that we get convergence of the iteration (2.20) if we take k and $\{\omega_i\}_{i=1}^k$ as directed in Lemma 2 and ρ as directed in Lemma 3. In the formula for ρ we use $(1-\delta)\gamma_1$, $(1+\delta)\gamma_2$, and $(1+\delta)\gamma_3$ instead of γ_1, γ_2 , and γ_3 respectively; the same change is made in the work estimate (2.19). The norm reduction will be in the E-norm, but the A-norm and the E-norm are comparable by (2.24). Thus if we take $0 < \delta \leq \frac{1}{2}$ and $0 < \epsilon \leq \frac{1}{2}$ we can get an estimate of the form of (2.19) for reduction of the A-norm of the error by a factor ϵ .

3. A factorization procedure for solving a self-adjoint Dirichlet problem.

The procedure to be examined in this section will give us a means of solving the difference equations associated with the problem

$$\begin{aligned} \mathfrak{L}(u) &= - \sum_{i=1}^n \frac{\partial}{\partial x_i} (a_i(x) \frac{\partial}{\partial x_i} u) + q(x)u \\ (3.1) \quad &= f(x) \quad \text{in } D \\ u &= g \quad \text{on } \partial D. \end{aligned}$$

We shall take D to be open, bounded in $\mathbb{R}^n$, $n \geq 2$, and

$$\begin{aligned} 0 < \eta_i &\leq a_i(x) \leq \mu_i, \\ q(x) &\geq 0. \end{aligned}$$

The method we shall use is, if the difference equations associated with (3.1) are written

$$Au = \tilde{f},$$

where A is a matrix, then we shall produce a matrix B such that $A+B = LL^*$, where L is lower triangular and sparse, and $A+B$ is comparable with A in the sense of Lemma 1 and Lemma 2. After we have this procedure we can, of course, use Lemma 3 and Lemma 4 to solve more general types of difference equations.

Before we can carry out the above program we must develop some notation and specify what types of difference

equations we shall discuss. Let h be a small positive number. Let $\bar{D}_h$ be the points of the grid $\{x : x=(x_1, \dots, x_n), x_i=kh, k \text{ integer}\}$ which are in $\bar{D}$ (the closure of D). Let D_h be the points x of $\bar{D}_h$ which are such that $x+he_i$ and $x-he_i$ are in $\bar{D}_h$ for $i=1, \dots, n$, where e_i is the unit vector in the x_i direction. Let $\partial D_h = \bar{D}_h \setminus D_h = \{x \in \bar{D}_h : x \notin D_h\}$. We shall be attempting to find or approximate a grid function u_h defined on $\bar{D}_h$ which satisfies a difference analogue of $\mathcal{L}(u) = f$ at each point of D_h and whose value is given on ∂D_h . Let the value of u_h at the point αh be denoted u_α , where α is an n -tuple with integral components. We shall call such α 's multi-indices and shall write $\alpha \geq 0$ if $\alpha = (\alpha_1, \dots, \alpha_n)$, $\alpha_i \geq 0$; $\alpha \geq \beta$ if $\alpha - \beta \geq 0$; and

$|\alpha| = \sum_{i=1}^n |\alpha_i|$. We wish to find u_h which satisfies

$$(3.2) \quad \begin{aligned} u_\alpha &= \tilde{g}(\alpha h) & \alpha h \in \partial D_h \\ \mathcal{L}_h u_\alpha &= f_\alpha & \alpha h \in D_h, \end{aligned}$$

where

$$(3.3) \quad \mathcal{L}_h u_\alpha = b_\alpha u_\alpha + \sum_{i=1}^n [c_{\alpha, e_i} u_{\alpha+e_i} + c_{\alpha-e_i, e_i} u_{\alpha-e_i}].$$

The coefficients in the operator $\mathcal{L}_h$ are obtained from $\mathcal{L}$ by replacing $-\frac{\partial}{\partial x_i}(a_i(x)\frac{\partial}{\partial x_i} u)$ by

$$(3.4) \quad h^{-2} \{ a_i(x + \frac{h}{2}e_i)[u(x) - u(x+he_i)] \\ + a_i(x - \frac{h}{2}e_i)[u(x) - u(x-he_i)] \},$$

or

$$(3.5) \quad h^{-2} \{ [a_i(x)a_i(x+he_i)]^{\frac{1}{2}}[u(x) - u(x+he_i)] \\ + [a_i(x)a_i(x-he_i)]^{\frac{1}{2}}[u(x) - u(x-he_i)] \},$$

or

$$(3.6) \quad h^{-2} \{ \frac{1}{2} [a_i(x) + a_i(x+he_i)][u(x) - u(x+he_i)] \\ + [a_i(x) + a_i(x-he_i)][u(x) - u(x-he_i)] \},$$

and the term $q(x)u$ is simply evaluated.

We shall write (3.2) as

$$(3.7) \quad Au = \tilde{f},$$

where A is a matrix, u is a vector whose components are the values u_α for $\alpha h \in D_h$ and $\tilde{f}$ comes from f and the boundary values $\tilde{g}$. Now if we adopt the convention that the coefficients c_{α, e_i} which multiply values of u_h not in D_h are zero (i.e., $c_{\alpha, e_i} = 0$ if $h(\alpha + e_i)$ or $h\alpha \notin D_h$) then the formula (3.3) gives the value of Au at the α component. This convention will be useful several times in that it allows us to use formulas which would otherwise not hold at the boundary.

We shall find a B such that we may write

$$(3.8) \quad (A+B) = LL^*,$$

where L is such that

$$(3.9) \quad Lu_\alpha = v_\alpha u_\alpha + \sum_{i=1}^n t_{\alpha-e_i, e_i} u_{\alpha-e_i}.$$

L^* is the adjoint of L considered as an operator on $\mathbb{C}^N$, where N is the number of grid points in D_h , with the usual innerproduct (if u and w are grid functions then

$$\langle u, w \rangle = \sum_{\alpha \in D_h} u_\alpha \bar{w}_\alpha). \quad \text{Thus}$$

$$(3.10) \quad L^*u_\alpha = v_\alpha u_\alpha + \sum_{i=1}^n t_{\alpha, e_i} u_{\alpha+e_i}.$$

In (3.9) and (3.10) we are using the convention that $t_{\alpha, e_i} = 0$ if $h\alpha \notin D_h$ or $h(\alpha+e_i) \notin D_h$. The factorization LL^* is useful for theoretical purposes, but in practice we would use $A+B = LU$, where U is upper triangular and $\text{diag } U = I$. We will need to study the LL^* factorization in detail so a complete algorithm will be given for it. In section 6 we will give the algorithm for the LU factorization.

In order to specify completely the factorization we need more constraints. Given a nonnegative grid function ϕ_h we can specify $LL^* = A+B$ as follows.

$$B = \tilde{B} + \tilde{D},$$

where

$$(3.11) \quad \begin{aligned} \tilde{B}u_\alpha = & \sum_{n \geq i > j \geq 1} [h_{\alpha, e_i - e_j} (u_{\alpha+e_i-e_j} - u_\alpha) \\ & + h_{\alpha-e_i+e_j, e_i-e_j} (u_{\alpha-e_i+e_j} - u_\alpha)], \end{aligned}$$

and

$$(3.12) \quad \tilde{D}u_\alpha = \varphi_\alpha b_\alpha u_\alpha.$$

If we simply multiply out LL^* and set corresponding components equal we obtain the following set of relations

$$(3.13) \quad \begin{aligned} h_{\alpha, e_i - e_j} &= t_{\alpha - e_j, e_j} t_{\alpha - e_j, e_i}, \quad i > j \\ v_\alpha &= [b_\alpha (1 + \varphi_\alpha) - \sum_{n \geq i > j \geq 1} (h_{\alpha, e_i - e_j} + h_{\alpha - e_i + e_j, e_i - e_j}) \\ &\quad - \sum_{i=1}^n t_{\alpha - e_i, e_i}^2]^{1/2} \\ t_{\alpha, e_i} &= c_{\alpha, e_i} / v_\alpha. \end{aligned}$$

The convention that the coefficients corresponding to points outside D_h are zero allows us to use (3.13) for all $\alpha h \in D_h$.

If we order the points of the grid in such a fashion that $\alpha - e_i$ precedes α for $i=1, \dots, n$ then (3.13) gives an easily computable set of relations for $v_\alpha, h_{\alpha, e_i - e_j}$, and t_{α, e_i} so long as v_α^2 is not zero. We shall show that v_α^2 remains positive if $\varphi_\alpha \geq 0$ for all α , and we shall obtain an estimate for the coefficients $h_{\alpha, e_i - e_j}$.

Let

$$(3.14) \quad b_\alpha = - \sum_{i=1}^n (\tilde{c}_{\alpha, e_i} + \tilde{c}_{\alpha - e_i, e_i}) + q(\alpha h),$$

where

$$\tilde{c}_{\alpha, e_i} = c_{\alpha, e_i} \quad \text{if } c_{\alpha, e_i} \neq 0.$$

These relations define $\tilde{c}_{\alpha, e_i}$ uniquely only away

from ∂D_h but this will not be important.

Lemma 5: For the factorization (3.13) with $\varphi_\alpha \geq 0$ we have

$$(3.15) \quad 1^\circ \quad v_\alpha^2 / \sum_{i=1}^n \tilde{c}_{\alpha, e_i} \geq \theta_\alpha > 0,$$

$$(3.16) \quad 2^\circ \quad 0 \leq h_{\alpha+e_j, e_i-e_j} \leq c_{\alpha, e_i} c_{\alpha, e_j} [\theta_\alpha \sum_{i=1}^n \tilde{c}_{\alpha, e_i}]^{-1} \\ \leq c_{\alpha, e_i} c_{\alpha, e_j} [\theta \sum_{i=1}^n \tilde{c}_{\alpha, e_i}]^{-1},$$

where

$$(3.17) \quad \rho_\alpha = \left(\sum_{i=1}^n \tilde{c}_{\alpha-e_i, e_i} \right) \left(\sum_{i=1}^n \tilde{c}_{\alpha, e_i} \right)^{-1},$$

$$(3.18) \quad \tilde{\theta}_\alpha = \frac{1}{2} [(1+\varphi_\alpha)(1+\rho_\alpha) + \{[(1+\varphi_\alpha)(1+\rho_\alpha)]^2 - 4\rho_\alpha\}^{\frac{1}{2}}],$$

$$(3.19) \quad \theta_\alpha = \min_{\alpha \geq \beta} \tilde{\theta}_\beta,$$

$$(3.20) \quad \theta = \min \theta_\alpha.$$

Proof: Note that

$$\sum_{j=1}^n [t_{\alpha-e_j, e_j}^2 + \sum_{i>j} (h_{\alpha, e_j-e_j} + h_{\alpha-e_i+e_j, e_i-e_j})] \\ = \sum_{j=1}^n [t_{\alpha-e_j, e_j}^2 + \sum_{i>j} (t_{\alpha-e_j, e_j} t_{\alpha-e_j, e_i} + t_{\alpha-e_i, e_i} t_{\alpha-e_i, e_j})] \\ = \sum_{ij} t_{\alpha-e_j, e_j} t_{\alpha-e_j, e_i}$$

$$= \sum_j (c_{\alpha-e_j, e_j} [\sum_j c_{\alpha-e_j, e_i}] / v_{\alpha-e_j}^2).$$

If we substitute the above into the expression for v_{α}^2 in (3.13), replace b_{α} by its representation (3.14), drop the $q(\alpha h)$ term, and divide by $\sum_{i=1}^n \tilde{c}_{\alpha, e_i}$ we obtain

$$v_{\alpha}^2 / \sum \tilde{c}_{\alpha, e_i} \geq (1+\varphi_{\alpha})(1+\rho_{\alpha}) + [\sum \tilde{c}_{\alpha, e_i}]^{-1} \sum_{j=1}^n (c_{\alpha-e_j, e_j} [\sum_{i=1}^n c_{\alpha-e_j, e_i}] / v_{\alpha-e_j}^2).$$

Now if

$$0 < \theta \leq v_{\beta}^2 / \sum_{i=1}^n \tilde{c}_{\beta, e_i} \quad \text{for } \beta \leq \alpha, \beta \neq \alpha$$

then

$$\frac{-1}{\theta} \leq \left(\sum_{i=1}^n c_{\beta, e_i} \right) / v_{\beta}^2 \quad \text{for } \beta \leq \alpha, \beta \neq \alpha.$$

Thus

$$v_{\alpha}^2 / \sum_{i=1}^n \tilde{c}_{\alpha, e_i} \geq (1+\varphi_{\alpha})(1+\rho_{\alpha}) - \rho_{\alpha} / \theta.$$

Hence

$$(3.21) \quad v_{\alpha}^2 / \sum_{i=1}^n \tilde{c}_{\alpha, e_i} \geq 1 + \varphi_{\alpha} + \rho_{\alpha} (\varphi_{\alpha} + 1 - \frac{1}{\theta}).$$

It is immediate from (3.21) and an easy induction that

$$(3.22) \quad v_{\alpha}^2 / \sum_{i=1}^n \tilde{c}_{\alpha, e_i} \geq 1.$$

If

$$(3.23) \quad \theta \leq (1+\varphi_{\alpha})(1+\rho_{\alpha}) - \rho_{\alpha} / \theta$$

then

$$v_{\alpha}^2 / \sum_{i=1}^n \tilde{c}_{\alpha, e_i} \geq \theta.$$

Now by (3.22) we can take $\theta \geq 1$ (and thus ≥ 0) so (3.23) holds if and only if

$$(3.24) \quad \theta^2 - (1+\varphi_{\alpha})(1+\rho_{\alpha})\theta + \rho_{\alpha} \leq 0.$$

This polynomial in θ has positive discriminant, since

$$\begin{aligned} (1+\varphi_{\alpha})^2(1+\rho_{\alpha})^2 - 4\rho_{\alpha} &= (2\varphi_{\alpha} + \varphi_{\alpha}^2)(1+\rho_{\alpha})^2 + (1-\rho_{\alpha})^2 \\ &\geq 0; \end{aligned}$$

thus (3.24) holds if and only if θ is in the interval $[r, \tilde{\theta}_{\alpha}]$, where r is the smaller root of the polynomial. Note that r is less than or equal to one; this is a straightforward estimation. Thus for $\theta \geq 1$ (3.24) holds if θ is less than $\tilde{\theta}_{\alpha}$. Hence if $\theta \geq 1$ is a lower bound for

$$v_{\beta}^2 / \sum_{i=1}^n \tilde{c}_{\beta, e_i} \quad \text{for } \beta \leq \alpha, \beta \neq \alpha,$$

and $\theta \leq \tilde{\theta}_{\alpha}$ then θ is a lower bound for

$$v_{\alpha}^2 / \sum_{i=1}^n \tilde{c}_{\alpha, e_i}.$$

Thus (3.15) holds. It should be noted that the argument given above is valid when some (or all) of the points $(\alpha - e_i)h$ are not in D_h . In this case some of the non-positive terms in the expression for v_{α}^2 which we estimated by $\frac{-1}{\theta}$ are zero; hence we can estimate them

by $-\frac{1}{\theta}$ for any $\theta > 0$. Finally (3.16) is an easy consequence of (3.15) and the proof is complete.

In order to establish the comparability of A and $A+B$ we shall use the following inequality.

Lemma 6: Let c_i be nonnegative for $i=1, \dots, n$.

Let $\sum_{i=1}^n c_i$ be positive. Let e and a_i , $i=1, \dots, n$, be complex. Then

$$(3.25) \quad \left[\sum_{i=1}^n c_i \right]^{-1} \sum_{1 \leq i < \ell \leq n} c_i c_\ell |a_i - a_\ell|^2 \leq \sum_{i=1}^n c_i |a_i - e|^2.$$

Proof: It is clear that we may assume $\sum_{i=1}^n c_i = 1$.

Assume that the a_i 's and e are real. Now the right-hand side of (3.25) is minimized with respect to e by

$$e = \sum_{i=1}^n c_i a_i. \quad \text{Thus}$$

$$\begin{aligned} \sum_{i=1}^n c_i |a_i - e|^2 &\geq \sum_{i=1}^n c_i (a_i - \sum_{\ell} c_\ell a_\ell)^2 \\ &= \sum_{i=1}^n c_i a_i^2 - 2 \sum_{i, \ell=1}^n c_i c_\ell a_i a_\ell + \sum_{i=1}^n c_i \left[\sum_{\ell=1}^n c_\ell a_\ell \right]^2 \\ &= \sum_{i=1}^n c_i a_i^2 - 2 \sum_{i, \ell=1}^n c_i c_\ell a_i a_\ell + \left(\sum_{\ell=1}^n c_\ell a_\ell \right)^2 \\ &= \sum_{i=1}^n c_i a_i^2 - \sum_{i, \ell=1}^n c_i c_\ell a_i a_\ell \\ &= \sum_{i, \ell=1}^n c_i c_\ell [a_i^2 - a_i a_\ell] \end{aligned}$$

$$\begin{aligned}
&= \sum_{1 \leq i < \ell \leq n} c_i c_\ell [a_i^2 - a_i a_\ell] + \sum_{1 \leq \ell < i \leq n} c_i c_\ell [a_i^2 - a_i a_\ell] \\
&= \sum_{1 \leq i < \ell \leq n} c_i c_\ell [a_i^2 - 2a_i a_\ell + a_\ell^2] \\
&= \sum_{1 \leq i < \ell \leq n} c_i c_\ell [a_i - a_\ell]^2.
\end{aligned}$$

Thus the lemma holds for real numbers. It follows for complex numbers by applying it to their real and imaginary parts. (It is clear that the lemma holds for a_i and e in a Hilbert space and absolute value replaced by norms.)

We shall also use the following representations of $\langle Au, u \rangle$ and $\langle \tilde{B}u, u \rangle$:

$$\begin{aligned}
(3.26) \quad \langle Au, u \rangle &= \sum_{\alpha h \in D_h} \left[\sum_{i=1}^n -c_{\alpha, e_i} |u_\alpha - u_{\alpha+e_i}|^2 \right. \\
&\quad \left. + (b_\alpha + \sum_{i=1}^n c_{\alpha, e_i}) |u_\alpha|^2 \right],
\end{aligned}$$

$$(3.27) \quad \langle \tilde{B}u, u \rangle = \left[\sum_{\alpha h \in D_h} \sum_{i < \ell} -h_{\alpha+e_i, e_\ell - e_i} |u_{\alpha+e_\ell} - u_{\alpha+e_i}|^2 \right].$$

These representations result from an application of summation by parts.

Using the representations (3.26) and (3.27) in connection with Lemma 6 we can see that

$$(3.28) \quad 0 \leq -\langle \tilde{B}u, u \rangle \leq \frac{1}{\theta} \langle Au, u \rangle,$$

where θ is given in Lemma 5. In order to get (3.28) we replace $h_{\alpha+e_i, e_\ell - e_i}$ in (3.27) by its upper bound given in

(3.16) and apply Lemma 6 for each α . In the application of Lemma 6, $u_{\alpha+e_i}$ plays the role of a_i , u_α the role of e , and $-c_{\alpha,e_i}$ the role of c_i .

Note that

$$(3.29) \quad \langle Au, u \rangle / \langle (A+B)u, u \rangle = \frac{1}{1 + \frac{\langle Bu, u \rangle}{\langle Au, u \rangle} + \frac{\langle \tilde{D}u, u \rangle}{\langle Au, u \rangle}}.$$

Now use the facts that

$$\langle \tilde{D}u, u \rangle \geq 0$$

and

$$\langle \tilde{B}u, u \rangle \leq 0$$

to see that

$$(3.30) \quad [1 + \langle \tilde{D}u, u \rangle / \langle Au, u \rangle]^{-1} \leq \langle Au, u \rangle / \langle (A+B)u, u \rangle \leq [1 - \frac{1}{\theta}]^{-1}.$$

Theorem 1: Let A be the matrix which arises in the discrete Dirichlet problem (3.2). Suppose that there exists $k_1 \geq 0$ such that ρ_α (defined by 3.17) satisfies

$$(3.31) \quad \rho_\alpha \geq 1 - k_1 h$$

for all $\alpha h \in D_h$. Let $\varphi_\alpha = c_0 h^2$, and let $A+B = LL^*$, where L is given by (3.9) and (3.13). Then given $\epsilon > 0$ we can find a sequence $\{w_i\}_{i=1}^k$ such that the iteration (2.8) yields w_k , where

$$(3.32) \quad \|w_k - A^{-1}f\|_A \leq \epsilon \|w_0 - A^{-1}f\|_A.$$

Further k can be taken to be

$$(3.33) \quad O(h^{-\frac{1}{2}} \log \epsilon^{-1}).$$

Thus the number of arithmetic operations is

$$(3.34) \quad O(h^{-n-\frac{1}{2}} \log \epsilon^{-1}).$$

(Note that if the coefficients $a_i(x)$ are uniformly Lipschitz in the x_i direction and the operator $\mathcal{L}_h$ is chosen by any of the formulas (3.4), (3.5), or (3.6) then the hypothesis on ρ_α is satisfied.)

Proof: The proof is a direct application of Lemma 2. To find e_1 of the interval $[e_1, e_2]$ of Lemma 2 we observe that

$$1 + \langle \tilde{D}u, u \rangle / \langle Au, u \rangle \leq 1 + c_0 h^2 (\max b_\alpha) \|A^{-1}\| \leq k_2,$$

where k_2 is independent of h . To find e_2 we note that for some α

$$\begin{aligned} \theta &= \frac{1}{2}[(1+\varphi_\alpha)(1+\rho_\alpha) + \{[(1+\varphi_\alpha)(1+\rho_\alpha)]^2 - 4\rho_\alpha\}^{\frac{1}{2}}] \\ &\geq 1 + \frac{1}{2}[\rho_\alpha - 1 + \{2\varphi_\alpha(1+\rho_\alpha)^2 + (1-\rho_\alpha)^2\}^{\frac{1}{2}}] \\ &\geq 1 + \frac{1}{2}\varphi_\alpha(1+\rho_\alpha)^2 \{2\varphi_\alpha(1+\rho_\alpha)^2 + (1-\rho_\alpha)^2\}^{-\frac{1}{2}} \\ &= 1 + \frac{1}{2}\varphi_\alpha^{\frac{1}{2}}(1+\rho_\alpha)^2 \{2(1+\rho_\alpha)^2 + \varphi_\alpha^{-1}(1-\rho_\alpha)^2\}^{-\frac{1}{2}} \\ (3.35) \quad &\geq 1 + k_3 h, \end{aligned}$$

where k_3 is independent of h . Thus if we take e_1 and e_2

to be

$$(3.36) \quad e_1 = k_2^{-1}$$

$$(3.37) \quad e_2 = k_3^{-1} h^{-1} + 1$$

we see that for all $u \neq 0$

$$\langle Au, u \rangle / \langle (A+B)u, u \rangle \in [e_1, e_2].$$

Thus the conclusion of the theorem follows from Lemma 2 with k the smallest integer not less than

$$\cosh^{-1}(\epsilon^{-1}) / \cosh^{-1}\left(\frac{e_2 + e_1}{e_2 - e_1}\right).$$

In order to obtain (3.33) we use

$$\cosh^{-1}(x) = \log(x + (x^2 - 1)^{\frac{1}{2}}),$$

for $x \geq 1$.

Theorem 1 is the basic result for this factorization procedure. Using it and a few tricks we shall be able to extend the work estimate (3.34) to cases in which (3.31) does not hold, to Dirichlet problems that are not self-adjoint, and to problems involving the nine-point Laplacian in the plane.

Consider the problem

$$(3.38) \quad Gu = f,$$

where G is a matrix which comes from discretizing a problem of the type (3.1) in which we do not know that the coefficients are smooth. It may well be the case that (3.31) does not hold (with k_1 independent of h).

If the coefficients are sufficiently bad we might wish to use integral averages over cells of the a_i 's instead of just evaluating them at one or two points as in (3.4), (3.5), and (3.6). However, this poses no new difficulties so long as

$$0 < \eta_i \leq a_i(x) \leq M_i.$$

We shall suppose that $\mathcal{L}(u)$ is discretized by

$$(3.39) \quad \mathcal{L}_h u_\alpha = q(\alpha h) u_\alpha + h^{-2} \sum_{i=1}^n \{ I_{\alpha, e_i, h}(a_i(x)) [u_\alpha - u_{\alpha+e_i}] \\ + I_{\alpha-e_i, e_i, h}(a_i(x)) [u_\alpha - u_{\alpha-e_i}] \},$$

where $\min_{\bar{D}} a_i(x) \leq I_{\alpha, e_i, h}(a_i(x)) \leq \max_{\bar{D}} a_i(x)$.

Theorem 2: Let G be the matrix obtained from (3.1) by some process of the type (3.39). Let A be the matrix obtained from a problem of the type (3.1) but with the a_i 's replaced by smooth (e.g. constant) functions. Assume A is obtained by (3.4), (3.5) or (3.6). Let $A+B = LL^*$ be obtained from the factorization (3.11) using A with $\varphi_\alpha = c_0 h^2$, then the procedure

$$LL^* w_{m+1} = LL^* w_m - \omega_{m+1} (G w_m - f)$$

gives a sequence $\{w_m\}$ which converges to the solution of

$$(3.40) \quad Gw = f,$$

provided the ω_i 's are properly chosen. Further in order

to reduce the G -norm of the error by a factor ϵ it suffices to calculate $\{w_i\}_{i=1}^k$, where

$$k = O(h^{-\frac{1}{2}} \log \epsilon^{-1}).$$

Proof: It is immediate from (3.26) that there exist positive v_1 and v_2 , independent of h , such that

$$(3.41) \quad \langle Gu, u \rangle / \langle Au, u \rangle \in [v_1 v_2],$$

for all nonzero u . Thus

$$(3.42) \quad \langle Gu, u \rangle / \langle (A+B)u, u \rangle \in [v_1 e_1, v_2 e_2],$$

where

$$\langle Au, u \rangle / \langle (A+B)u, u \rangle \in [e_1, e_2].$$

Hence the theorem follows from Lemma 2 using the estimates which we obtained for e_1 and e_2 in the proof of Theorem 1.

We remark that even if G comes from a problem in which the a_i 's are discontinuous we may be able to use the factorization associated directly with G in order to obtain an estimate $O(h^{-\frac{1}{2}} \log \epsilon^{-1})$ on the number of iterations. If the a_i 's are such that there exist M_1 and M_2 , independent of h , such that on each line ℓ through D in the x_i direction there are open intervals

$\{I_k\}_{k=1}^{M_1}$ such that $\ell = \bigcup_{k=1}^{M_1} \overline{I_k}$ and a_i is Lipschitz continuous on I_k with Lipschitz constant M_2 then we can obtain the above work estimate if we take $G+B=LL^*$ to be given by the factorization (3.13) with $\varphi_\alpha = c_0 h^2$ at those points α such that all the a_i 's are Lipschitz continuous

near αh and $\varphi_\alpha = c_1 h$ at the other α 's. If we use the above method for choosing φ_α then it is clear from the proof of Theorem 1 that at those points α where $\varphi_\alpha = c_0 h^2$, $\tilde{\theta}_\alpha \geq 1 + k_1 h$, k_1 independent of h and α . Also if in the derivation of (3.35) we do not drop the $\varphi_\alpha(1 + \rho_\alpha)$ term in the first line then it is clear that

$\tilde{\theta}_\alpha \geq 1 + \frac{c_1 h}{2}$ for those α such that $\varphi_\alpha = c_1 h$. Hence to get the estimate claimed all we need to do is show that

$$\langle \tilde{D}x, x \rangle / \langle Gx, x \rangle$$

is bounded independent of h . This result follows easily from the fact that if A_1 corresponds to $-\frac{\partial^2}{\partial x^2}$ on the interval $[0, K]$ and $\tilde{D}_1$ corresponds to $h^{-1} \delta_{x_0}$ then $\langle \tilde{D}_1 x, x \rangle / \langle A_1 x, x \rangle$ is bounded independent of h .

This fact is an easy consequence of the fact that $\max |x_{kh}| \leq h^{-\frac{1}{2}} \langle Ax, x \rangle^{\frac{1}{2}}$.

We conclude this section with a remark about the nine point discrete Laplacian in the plane. The nine point discrete Laplacian is defined as

$$\begin{aligned} \Delta_h^{(9)} u_\alpha &= h^{-2} \left[\frac{1}{6} \{ u_{\alpha+e_1+e_2} + u_{\alpha-e_1-e_2} + u_{\alpha+e_1-e_2} + u_{\alpha-e_1+e_2} \} \right. \\ &\quad \left. + \frac{2}{3} \{ u_{\alpha+e_1} + u_{\alpha+e_2} + u_{\alpha-e_1} + u_{\alpha-e_2} \} \right. \\ &\quad \left. - \frac{10}{3} u_\alpha \right] \\ (3.43) \quad &= \frac{2}{3} \Delta_h u_\alpha + \frac{1}{3} \Delta_h^{(x)} u_\alpha, \end{aligned}$$

where Δ_h is the usual five point approximation to the Laplacian and $\Delta_h^{(x)}$ is a rotation of Δ_h .

If we let G be the matrix associated with $-\Delta_h^{(9)}$ and A be the matrix associated with $-\Delta_h$ then

$$(3.44) \quad \frac{2}{3} \leq \langle Gu, u \rangle / \langle Au, u \rangle \leq \frac{4}{3}.$$

The lower bound comes from the fact that the part of G which corresponds to $-\Delta_h^{(x)}$ is positive-definite and the upper bound follows from the argument used to bound $-\langle \tilde{B}u, u \rangle$ in terms of $\langle Au, u \rangle$.

Now if we get $LL^* = A + B$ from A using $\varphi = c_0 h^2$ then

$$\langle Gu, u \rangle / \langle LL^*u, u \rangle \in [\frac{2}{3}e_1, \frac{4}{3}e_2]$$

where

$$\langle Au, u \rangle / \langle LL^*u, u \rangle \in [e_1, e_2].$$

Thus

$$LL^*w_{m+1} = LL^*w_m - w_{m+1}(Gw_m - f)$$

can be used to approximate the solution to

$$Gw = f.$$

The work estimate is of course the same expression (3.34).

It should be noted that we can obtain a much better work estimate in case the region D is a rectangle by using alternating direction methods approximately to invert A , $(A \sim \Delta_h)$. That is, if we use the outer iteration

$$(3.45) \quad Aw_{m+1} = Aw_m - \rho(Gw_m - f)$$

and approximately invert A by alternating direction techniques then $w_m \rightarrow w = G^{-1}f$, for ρ properly chosen. See Gunn [4] pages 28 and 29 for the comparability between A and $A[I-\Lambda]^{-1}$ where Λ is the error propagator for an alternating direction method.

4. A factorization procedure for solving a self-adjoint Neumann Problem.

As an illustration of the ability of the procedure given in section 3 to be adapted to boundary conditions other than Dirichlet we shall consider the self-adjoint analogue of

$$(4.1) \quad \sum_{i=1}^n \frac{\partial}{\partial x_i} (a_i(x) \frac{\partial u}{\partial x_i}) = f \quad \text{in } D,$$

$$\frac{\partial u}{\partial n} = g \quad \text{on } \partial D.$$

Here $\frac{\partial}{\partial n}$ denotes differentiation in the direction normal to the boundary. We assume D is a bounded domain in $\mathbb{R}^n$ and that each $a_i(x)$ is bounded above and below by positive numbers. In order to get a self-adjoint matrix for the discrete problem and to simplify certain estimates we shall take D to be $\bigcup_{i=1}^k D_i$ where D_i is an open rectangular parallelopiped with edges parallel to the coordinate axes. I.e., D_i is of the form $\{x: x=(x_1, \dots, x_n), \eta_{i,l} < x_l < \eta_{i,l}^{i,l}\}$. We further assume that $\eta_{i,l}$ and $\eta_{i,l}^{i,l}$ are integer multiples of $h_0 > 0$. We shall take h to be h_0 divided by a positive integer. Put a grid on $\mathbb{R}^n$ of the form $G_h = \{x: x=(x_1, \dots, x_n) \ x_l = (k_l + \frac{1}{2})h, k_l \text{ integer}\}$.

Let $\bar{D}_h = \bar{D} \cap G_h$. Let $D_h = \{x: x \in \bar{D}_h, x + he_i \in \bar{D}_h \text{ and } x - he_i \in \bar{D}_h \text{ for } i=1, \dots, n\}$.

We obtain a discrete analogue

$$(4.2) \quad Aw = \tilde{f},$$

of (4.1) by taking

$$(4.3) \quad Aw_\alpha = b_\alpha w_\alpha + \sum_{i=1}^n [c_{\alpha, e_i} w_{\alpha+e_i} + c_{\alpha-e_i, e_i} w_{\alpha-e_i}],$$

where

$$(4.4) \quad b_\alpha + \sum_{i=1}^n [c_{\alpha, e_i} + c_{\alpha-e_i, e_i}] = 0.$$

We take

$$c_{\alpha, e_i} = \begin{cases} -h^{-2} a_i ((\alpha + \frac{1}{2} e_i)h), & \text{if } (\alpha + \frac{1}{2} e_i)h \in D, \\ = 0 & \text{otherwise.} \end{cases}$$

The data f and g give rise to $\tilde{f}$. A is self-adjoint and the sum of the elements in any row or column is zero.

We shall make the additional assumption that A is irreducible: this simply means that we cannot write D_h as a disjoint

union of subregions and solve the problem independently on each. Note that (4.4) implies that if ξ is the constant grid function equal everywhere to one, then

$$(4.5) \quad A\xi = 0.$$

The assumption of irreducibility assures us that the null space of A , which we shall denote $n(A)$, consists solely of multiples of ξ . As A is self-adjoint its range, $\mathcal{R}(A)$, is the orthogonal complement of its null space. I.e., $\mathcal{R}(A)$ is the set of grid functions with zero average value. In order to solve (4.2) we assume that $\tilde{f}$ is in the range of A .

The factorization given in section 3 is still possible for $\varphi_\alpha \geq 0$; however, L is singular if $\varphi_\alpha \equiv 0$ and we would not be able to carry out the iteration. If we take $\varphi_\alpha > 0$ then L is nonsingular and we obtain almost the same lemma as before for estimation of $h_{\alpha, e_i - e_j}$.

Lemma 7: For the factorization (3.13) and

$\varphi_\alpha > 0$ we have:

$$(4.6) \quad 1^\circ \quad v_\alpha^2 / \sum_{i=1}^n -c_{\alpha, e_i} \geq \theta_\alpha > 0$$

$$\text{if } \sum_{i=1}^n c_{\alpha, e_i} \neq 0,$$

$$(4.7) \quad 2^\circ \quad v_\alpha^2 > 0, \text{ for all } \alpha,$$

$$3^\circ \quad 0 \leq h_{\alpha + e_i, e_i - e_j} \leq c_{\alpha, e_i} c_{\alpha, e_j} [\theta_\alpha \sum_{i=1}^n -c_{\alpha, e_i}]^{-1}$$

$$(4.8) \quad \leq c_{\alpha, e_i} c_{\alpha, e_j} [\theta_\alpha \sum_{i=1}^n -c_{\alpha, e_i}]^{-1},$$

$$\text{if } \sum_{i=1}^n -c_{\alpha-e_i} \neq 0.$$

where

$$(4.9) \quad \rho_{\alpha} = \left(\sum_{i=1}^n c_{\alpha-e_i, e_i} \right) \left(\sum_{i=1}^n c_{\alpha, e_i} \right)^{-1},$$

$$(4.10) \quad \tilde{\theta}_{\alpha} = \frac{1}{2}[(1+\varphi_{\alpha})(1+\rho_{\alpha}) + \{[(1+\varphi_{\alpha})(1+\rho_{\alpha})]^2 - 4\rho_{\alpha}\}^{\frac{1}{2}}],$$

$$(4.11) \quad \theta_{\alpha} = \min\{\tilde{\theta}_{\beta} : \beta \leq \alpha \text{ and } \sum_{i=1}^n c_{\beta, e_i} \neq 0\},$$

$$(4.12) \quad \theta = \min\{\tilde{\theta}_{\beta} : \sum_{i=1}^n c_{\beta, e_i} \neq 0\}$$

The proof of Lemma 7 is quite similar to that of Lemma 5 and will be omitted.

Since $\tilde{f}$ is in $\mathcal{R}(A)$ and A is a one-to-one map of $\mathcal{R}(A)$ onto $\mathcal{R}(A)$ we know that there is a unique solution of (4.2) in $\mathcal{R}(A)$. It is natural therefore to require the successive approximations w_m to be in $\mathcal{R}(A)$. Let P be the orthogonal projection onto $\mathcal{R}(A)$. As an operator on $\mathcal{R}(A)$, PLL^* is self-adjoint and positive-definite. Thus the following iteration defines uniquely a sequence $\{w_m\} \subset \mathcal{R}(A)$:

$$(4.13) \quad w_0 \in \mathcal{R}(A)$$

$$PLL^*w_{m+1} = PLL^*w_m - w_{m+1}(Aw_m - \tilde{f}).$$

In order to establish the usefulness of (4.13) we must show that as an operator on $\mathcal{R}(A)$, PLL^* is easily invertible and we must establish bounds for $\langle Au, u \rangle / \langle PLL^*u, u \rangle$, for u any nonzero element of $\mathcal{R}(A)$. Before we do this we

remark that the iteration (2.8) with $A+B = LL^*$ produces a sequence $\{w_m\}$ which, for w_m 's chosen properly, converges in the A -semi-norm to a solution of (4.2). However, we have no bounds for the comparability which needs to be estimated (between A and the inverse on $\mathcal{R}(A)$ of $P(LL^*)^{-1}$).

In order to establish that PLL^* is easily inverted on $\mathcal{R}(A)$ we note that for $f \in \mathcal{R}(A)$

$$(4.14) \quad PLL^*u = f$$

if and only if

$$(4.15) \quad LL^*u = f + c\xi,$$

where ξ is the vector each of whose components is one.

Now (4.15) holds if and only if

$$(4.16) \quad u = (LL^*)^{-1}f + c(LL^*)^{-1}\xi,$$

where the inverse is taken on the whole space. But in order for u to be in $\mathcal{R}(A)$ it is only necessary that

$$(4.17) \quad \langle u, \xi \rangle = 0.$$

Thus we need

$$c = -\langle f, (LL^*)^{-1}\xi \rangle / \langle \xi, (LL^*)^{-1}\xi \rangle.$$

Hence

$$(4.18) \quad u = (LL^*)^{-1}f - \frac{\langle f, (LL^*)^{-1}\xi \rangle}{\langle \xi, (LL^*)^{-1}\xi \rangle} (LL^*)^{-1}\xi$$

is in $\mathcal{R}(A)$ and is a solution of (4.14). As LL^* is easily inverted we see that PLL^* is.

Hence,

$$(4.19) \quad w_{m+1} = w_m - w_{m+1} (LL^*)^{-1} [Aw_m - f - \frac{\langle Aw_m - f, (LL^*)^{-1} \xi \rangle}{\langle \xi, (LL^*)^{-1} \xi \rangle} \xi].$$

If we proceed as in section 3 we see that for $u \in \mathcal{R}(A)$

$$(4.19) \quad [1 + \langle Du, u \rangle / \langle Au, u \rangle]^{-1} \leq \langle Au, u \rangle / \langle PLL^*u, u \rangle \leq [1 - \frac{1}{\theta}]^{-1}.$$

Theorem 3: Let A be given by (4.3) and satisfy (4.4). Suppose that for $c_{\alpha, e_i} \neq 0$

$$(4.21) \quad -M_i h^{-2} \leq c_{\alpha, e_i} \leq -\eta_i h^{-2} < 0.$$

Suppose that there exists $k_1 \geq 0$ such that for all α ,

such that $\sum_{i=1}^n c_{\alpha, e_i} \neq 0$ and $\prod_{i=1}^n c_{\alpha - e_i, e_i} \neq 0$, we have

$$(4.22) \quad \rho_\alpha \geq 1 - k_1 h.$$

Let $LL^* = A + B$ be given by the factorization (3.13) with φ_α given by

$$(4.23) \quad \varphi_\alpha = c_0 h^2 \text{ if for all } i=1, \dots, n \quad c_{\alpha - e_i, e_i} \neq 0,$$

$$(4.24) \quad \varphi_\alpha = c_1 h \text{ if for some } i \quad c_{\alpha - e_i, e_i} = 0,$$

where c_0 and c_1 are positive. Then we can choose $\{w_m\}$ such that $\{w_m\}$ given by (4.19) converges to the solution $w \in \mathcal{R}(A)$ of (4.2). Further, for $\frac{1}{2} \geq \epsilon > 0$ the w_m 's may be picked in such a fashion so that

$$(4.25) \quad \|w_k - w\|_A \leq \epsilon \|w_0 - w\|_A$$

where

$$(4.26) \quad k = O(h^{-\frac{1}{2}} \log \epsilon^{-1}).$$

(We remark that (4.22) is natural if each $a_i(x)$ is uniformly Lipschitz in x_i .)

Proof: First, observe that the form of φ_α and the hypothesis on ρ_α are sufficient to imply there exists $k_2 > 0$ such that

$$(4.27) \quad \theta \geq 1 + k_2 h.$$

To see this, note as in (3.35) that for each α

$$\tilde{\theta}_\alpha \geq 1 + \frac{1}{2} \{ \rho_\alpha + \varphi_\alpha (1 + \rho_\alpha)^2 [2\varphi_\alpha (1 + \rho_\alpha)^2 + (1 - \rho)^2]^{-\frac{1}{2}} \}.$$

Thus (4.26) holds no matter where the minimum of $\tilde{\theta}_\alpha$ is assumed.

Second, $\langle \tilde{D}u, u \rangle / \langle Au, u \rangle$ is bounded independently of h for $u \neq 0$, $u \in \mathcal{R}(A)$. In the arguments which show this it will be convenient to normalize the innerproduct as

$$\langle u, v \rangle = h^n \sum_{\alpha \in D_h} u_\alpha \bar{v}_\alpha; \text{ this does not effect the ratio of}$$

course. We shall write $\tilde{D} = \tilde{D}_1 + \tilde{D}_2$, where $\tilde{D}_1 w_\alpha = c_0 h^2 b_\alpha w_\alpha$ if $\tilde{D} w_\alpha = c_0 h^2 b_\alpha w_\alpha$ and $\tilde{D}_1 w_\alpha = 0$ otherwise, $\tilde{D}_2 w_\alpha = c_1 h b_\alpha w_\alpha$ if $\tilde{D} w_\alpha = c_1 h b_\alpha w_\alpha$ and $\tilde{D}_2 w_\alpha = 0$ otherwise. That $\langle \tilde{D}_1 u, u \rangle / \langle Au, u \rangle$ is bounded from above independently of h follows from the fact that the smallest nonzero eigenvalue of A is bounded below by a positive number independent of h .

In order to show that the smallest nonzero eigenvalue of A has this property we assume the contrary. Let $\{u_m\}$, $\{h_m\}$

be sequences such that u_m is a grid function on D_{h_m} ,

$\|u_m\|=1$, $\langle Au_m, u_m \rangle \rightarrow 0$, and $\langle u_m, \xi \rangle = 0$. Since the average value of u_m on D_i is bounded we can assume, by taking a subsequence, that the average value of u_m on D_i converges to r_i . Using $\langle Au_m, u_m \rangle \rightarrow 0$ and the fact that the smallest nonzero eigenvalue of A is bounded below by a positive number independent of h if D is a rectangular parallelepiped we see that $\|u_m - r_i\| \rightarrow 0$, where the norm is taken over D_i . From the connectedness of D we get that $r_i = r_j$ for all $i, j = 1, \dots, k$. This gives a contradiction since $\langle u_m, \xi \rangle = 0$ and $\|u_m\|=1$.

To see that $\langle \tilde{D}_2 u, u \rangle / \langle Au, u \rangle$ is bounded independently of h we write $\tilde{D}_2$ as $\sum_{i=1}^n \tilde{D}_{2,i}$, where each $\tilde{D}_{2,i}$ operates on a part of the boundary of D which is perpendicular to the x_i direction. Define σ_i to be the set where $\tilde{D}_{2,i}$ is nonzero. It is clear that if $N_h = h^{-1} \min(\eta^i, \ell - \eta_{i,\ell})$ then for $\alpha \in \sigma_i$ and $0 \leq \ell < N_h$ $(\alpha + \ell e_i)h \in D_h$.

Taking $0 \leq \ell < N_h$ we see that

$$\begin{aligned} \left| \sum_{\alpha \in \sigma_i} w_\alpha^2 - \sum_{\alpha \in \sigma_i} w_{\alpha + \ell e_i}^2 \right| &= \left| \sum_{\sigma_i} \sum_{k=0}^{\ell-1} w_{\alpha + k e_i}^2 - w_{\alpha + (k+1) e_i}^2 \right| \\ &\leq \left[\sum_{\alpha \in \sigma_i} \sum_{k=0}^{\ell-1} |w_{\alpha + k e_i} - w_{\alpha + (k+1) e_i}|^2 \right]^{\frac{1}{2}} \\ &\quad \left[\sum_{\alpha \in \sigma_i} \sum_{k=0}^{\ell-1} |w_{\alpha + k e_i} + w_{\alpha + (k+1) e_i}|^2 \right]^{\frac{1}{2}}. \end{aligned}$$

Using this we see that

$$\begin{aligned}
\|w\|^2 &= h^n \sum_{D_h} w_\alpha^2 \\
&\geq h^n \sum_{\sigma_i} \sum_{\ell=0}^{N_h-1} w_{\alpha+\ell e_i}^2 \\
&\geq \{h^n N_h \sum_{\alpha \in \sigma_i} w_\alpha^2\} - \{2N_h [h^n \sum_{\alpha \in \sigma_i} \sum_{k=0}^{N_h-1} |w_{\alpha+ke_i} - w_{\alpha+(k+1)e_i}|^2]^{\frac{1}{2}} \cdot \\
&\quad [h^n \sum_{\alpha \in \sigma_i} \sum_{k=0}^{N_h} w_{\alpha+ke_i}^2]^{\frac{1}{2}}\} \\
&\geq C_2 \langle \tilde{D}_{2,i} w, w \rangle - C_3 \langle Aw, \tilde{w} \rangle^{\frac{1}{2}} \|w\|,
\end{aligned}$$

where C_2 and C_3 are positive numbers independent of h .

Thus

$$\begin{aligned}
\langle \tilde{D}_{2,i} w, w \rangle &\leq C_4 [\langle Aw, w \rangle^{\frac{1}{2}} \|w\| + \|w\|^2] \\
&\leq C_5 \langle Aw, w \rangle.
\end{aligned}$$

Hence we finally obtain that $\langle \tilde{D}u, u \rangle / \langle Au, u \rangle$ is bounded from above independently of h .

Thus we can apply Lemma 2, with e_1 independent of h and $e_2 = k_3 h^{-1}$, and obtain the stated result.

We may use the same trick which allowed us to drop the hypothesis on ρ_α in the Dirichlet case to drop it here.

Theorem 4: Let G and A be of the form (4.3) and satisfy (4.4). Suppose that both satisfy (4.21) and that A satisfies (4.22). Let $A+B=LL^*$ be obtained by (3.13) with φ_α given by (4.23) and (4.24). Then we can pick ω_i such that $\{w_m\}$, defined by

$$w_0 \in \mathcal{R}(A)$$

$$(4.28) \quad \text{PLL}^* w_{m+1} = \text{PLL}^* w_m - w_{m+1} (G w_m - \tilde{f})$$

$$w_{m+1} \in \mathcal{R}(A),$$

converges to $w \in \mathcal{R}(A)$; the solution of $Gw = \tilde{f}$. Further, for $\frac{1}{2} > \epsilon > 0$ the w_m 's may be picked such that

$$(4.29) \quad \|w_k - w\|_G \leq \epsilon \|w_0 - w\|_G,$$

where

$$(4.30) \quad k = O(h^{-\frac{1}{2}} \log \epsilon^{-1}).$$

Proof: The ratio $\langle Gu, u \rangle / \langle Au, u \rangle$ is easily bounded above and below by positive numbers which are independent of h and independent of u in $\mathcal{R}(A) = \mathcal{R}(G)$. Thus we can easily obtain bounds for $\langle Gu, u \rangle / \langle \text{PLL}^* u, u \rangle$ for u in $\mathcal{R}(A)$ and apply Lemma 2.

5. A factorization procedure for solving a mildly nonlinear elliptic Dirichlet problem.

In this section we shall apply Lemma 3, Lemma 4, and the results of section 3 to obtain approximate solutions to the difference analogues of

$$\begin{aligned} \mathcal{L}(u) &= \sum_{i=1}^n - \frac{\partial}{\partial x_i} (a_i(x) \frac{\partial u}{\partial x_i}) + \sum_{i=1}^n b_i(x) \frac{\partial u}{\partial x_i} + f(x, u) = 0, \\ (5.1) \quad & x \in D, \\ & u = g, \quad \text{on } \partial D, \end{aligned}$$

where D is a bounded domain in $\mathbb{R}^n$.

We shall assume that u is known at each point of ∂D_h . At each point of D_h we shall replace $\mathcal{L}(u)=0$ by a difference operator defined below. We shall replace the divergence terms by one of the expressions (3.4), (3.5), (3.6). Let A denote the matrix which we get from the divergence terms. Let $A = \sum_{i=1}^n A_i$, where A_i corresponds to the divergence term in the x_i -direction. We shall approximate the first order terms $b_i \frac{\partial u}{\partial x_i}(x)$ by one of two methods. In the first method this term is replaced by

$$(5.2) \quad b_i(x) (2h)^{-1} [u(x+he_i) - u(x-he_i)].$$

Let $H_{1,i}$ denote the matrix which arises from (5.2). Let $H_1 = \sum_{i=1}^n H_{1,i}$. In the second method this term is replaced

by

$$(5.3) \quad b_i(x) h^{-1}[u(x+e_i h)-u(x)], \quad b_i(x) \leq 0$$

$$b_i(x) h^{-1}[u(x)-u(x-e_i h)], \quad b_i(x) \geq 0.$$

Let $H_{2,i}$ be the matrix resulting from this approximation.

Let $H_2 = \sum_{i=1}^n H_{2,i}$. The zero order term $f(x,u)$ is simply evaluated. Thus we have the following operator

$$(5.3) \quad \mathcal{L}_h(u) = Au + Hu + \hat{f}(u),$$

where H is H_1 or H_2 and $\hat{f}$ contains the terms which come from the boundary values as well as the $f(x,u)$ terms.

We want to find u_h such that

$$(5.4) \quad \mathcal{L}_h(u_h) = 0.$$

We shall use an iteration of the form of (2.14) to solve this problem. We let the A of (2.14) be the matrix which comes from the divergence terms. Let u_h be the solution to (5.4) and assume that f is continuously differentiable with respect to u for each x then

$$(5.5) \quad \mathcal{L}_h(w_h) = D(w_h)(w_h - u_h),$$

where

$$(5.6) \quad D(w_h)v = Av + Hv + f'(w_h)v.$$

The operator $f'(w_h)$ is a multiplication at each point x by $\frac{\partial}{\partial u}f(x, \xi)$, where ξ is between w_h and u_h at x .

Since we have Lemma 3 it is clear that all we need

to do to get the convergence of the iteration (2.14)
(and its rate) is to establish bounds for

$$(5.7) \quad \operatorname{Re}\langle D(w_h)u, u \rangle / \langle Au, u \rangle$$

and

$$(5.8) \quad |\operatorname{Im}\langle D(w_h)u, u \rangle / \langle Au, u \rangle|.$$

We shall make the following assumptions:

$$(5.9) \quad (a) \quad 0 < \mu_0 \leq \eta_i \leq a_i(x) \leq \mu_i$$

$$(b) \quad |b_i(x)| \leq M_1$$

$$(c) \quad \left| \frac{\partial}{\partial x_i} b_i(x) \right| \leq \theta \eta \lambda_i,$$

where λ_i is the minimal eigenvalue of A_i , $\eta \in [0, 1)$ and,
if $H=H_1$, $\theta=2$, and, if $H=H_2$, $\theta=1$,

$$(5.12) \quad (d) \quad 0 \leq \frac{\partial}{\partial u} f(x, u) \leq M_2.$$

Lemma 8: If (a), (b), (c), and (d) hold then

$$(5.13) \quad 0 < \gamma_1 \leq \operatorname{Re}\langle D(w_h)v, v \rangle / \langle Av, v \rangle \leq \gamma_2$$

$$(5.14) \quad |\operatorname{Im}\langle D(w_h)v, v \rangle / \langle Av, v \rangle| \leq \gamma_3,$$

for all w_h and $v \neq 0$. The γ_i 's are defined as

$$(5.15) \quad \gamma_1 = 1 - \eta$$

$$(5.16) \quad \gamma_2 = 1 + \eta + M_2 \left[\sum_{i=1}^n \lambda_i \right]^{-1}, \quad \text{if } H=H_1,$$

$$(5.17) \quad \gamma_2 = 1 + M_2 \left[\sum_{i=1}^n \lambda_i \right]^{-1} + M_1 \left[\frac{1}{2} \mu_0 \sum_{i=1}^n \lambda_i \right]^{-\frac{1}{2}}, \quad \text{if } H=H_2,$$

$$(5.18) \quad \gamma_3 = M_1 \left[2 \mu_0 \sum_{i=1}^n \lambda_i \right]^{-\frac{1}{2}}, \quad \text{if } H=H_1,$$

$$(5.19) \quad \gamma_3 = M_1 \left[\frac{1}{2} \mu_0 \sum_{i=1}^n \lambda_i \right]^{-\frac{1}{2}}, \quad \text{if } H=H_2,$$

Proof: First we establish that if $H=H_1$, then

$$(5.20) \quad |\operatorname{Re} \langle H v, v \rangle| \leq \eta \langle A v, v \rangle.$$

Note that

$$\begin{aligned} (\operatorname{Re} H_{1,i}) v_\alpha &= \frac{1}{2} \frac{1}{2h} \{ [b_i(\alpha h) - b_i((\alpha + e_i)h)] v_{\alpha + e_i} \\ &\quad + [b_i((\alpha - e_i)h) - b_i(\alpha h)] v_{\alpha - e_i} \}. \end{aligned}$$

Thus by a theorem of Gerschgorin (see Varga [7], page 16)

$$\|\operatorname{Re} H_{1,i}\| \leq \frac{1}{2} \max_{x \in D} \left| \frac{\partial}{\partial x_i} b_i(x) \right|$$

Thus

$$\begin{aligned} |\operatorname{Re} \langle H_{1,i} v, v \rangle| &= |\langle \operatorname{Re} H_{1,i} v, v \rangle| \\ &\leq \frac{1}{2} \max \left| \frac{\partial}{\partial x_i} b_i(x) \right| \|v\|^2 \\ &\leq \frac{1}{2} \max \left| \frac{\partial}{\partial x_i} b_i(x) \right| \lambda_i^{-1} \langle A_i v, v \rangle \\ (5.21) \quad &\leq \eta \langle A_i v, v \rangle. \end{aligned}$$

Thus (5.20) holds.

Second, if $H=H_2$ then

$$(5.22) \quad -\eta \langle Av, v \rangle \leq \operatorname{Re} \langle Hv, v \rangle.$$

Note that

$$(5.23) \quad \begin{aligned} \operatorname{Re} H_{2,i} v_\alpha &= \frac{1}{2h} \{ [-\max\{0, b_i(\alpha h)\} + \min\{0, b_i((\alpha - e_i)h)\}] v_{\alpha - e_i} \\ &\quad + 2 |b_i(\alpha h)| v_\alpha \\ &\quad + [\min\{0, b_i(\alpha h)\} - \max\{b_i((\alpha + e_i)h), 0\}] v_{\alpha + e_i} \}. \end{aligned}$$

Now counting cases we see that by Gerschgorin's theorem the smallest eigenvalue of $\operatorname{Re} H_{2,i}$ is not less than

$$-\max_{x \in D} \left| \frac{\partial}{\partial x_i} b_i(x) \right|.$$

Thus

$$(5.24) \quad \begin{aligned} \langle \operatorname{Re} H_{2,1} v, v \rangle &\geq -\max \left| \frac{\partial}{\partial x_i} b_i \right| \|v\|^2 \\ &\geq -\eta \langle A_i v, v \rangle. \end{aligned}$$

Thus (5.22) holds.

Next, note that $\sum_{i=1}^n \lambda_i$ is a lower bound on the smallest eigenvalue of A . Thus

$$(5.25) \quad \|v\|^2 \leq \langle Av, v \rangle \left[\sum_{i=1}^n \lambda_i \right]^{-1}.$$

Hence

$$(5.26) \quad 0 \leq \langle f'(w_n) v, v \rangle \leq M_2 [\sum \lambda_i]^{-1} \langle Av, v \rangle.$$

Combining (5.26) and (5.20) we see that (5.13) holds for $H=H_1$ with γ_1 and γ_2 defined by (5.15) and (5.16).

Next we see that

$$\begin{aligned}
|\operatorname{Im}\langle D(w_h)v, v \rangle| &= |\operatorname{Im}\langle Hv, v \rangle| \\
&\leq |\langle Hv, v \rangle| \\
(5.27) \qquad &\leq \|Hv\| \|v\|.
\end{aligned}$$

For $H=H_1$, using $|a-b|^2 \leq 2[|a-e|^2 + |e-b|^2]$ we get

$$\begin{aligned}
\|H_{1,i}v\|^2 &\leq \frac{M_1^2}{2\eta_i} \langle A_i v, v \rangle \\
&\leq \frac{M_1^2}{2\mu_0} \langle A_i v, v \rangle.
\end{aligned}$$

Thus for $H=H_1$

$$(5.28) \quad \|Hv\|^2 \leq \frac{nM_1^2}{2\mu_0} \langle Av, v \rangle$$

For $H=H_2$ we get directly

$$(5.29) \quad \|Hv\|^2 \leq \frac{2nM_1^2}{\mu_0} \langle Av, v \rangle.$$

In getting (5.28) and (5.29) we used the expression for $\langle Av, v \rangle$ given in (3.26). (The coefficient in (5.29) can be improved to $\max\{M_1^2/\mu_0, h \max_{i=1,n} \left| \frac{\partial}{\partial x_i} b_i(x) \right|\}$.)

We use (5.28), (5.29), and (5.25) to see that

$$|\operatorname{Im}\langle D(w_h)v, v \rangle| \leq \gamma_3 \langle Av, v \rangle,$$

for γ_3 as defined.

Further, from (5.29) and (5.25) we see that

$$(5.30) \quad |\operatorname{Re}\langle H_2 v, v \rangle| \leq M_1 \left[\frac{1}{2} \mu_0 \sum_{i=1}^n \lambda_i \right]^{-\frac{1}{2}} \langle A v, v \rangle.$$

Using (5.30), (5.26), and (5.24) we see that (5.13) holds with γ_1 and γ_3 defined by (5.15) and (5.17) for $H=H_2$. Thus the proof is complete.

Theorem 5: Let $\mathcal{L}_h(u)$ be defined by (5.3). Let u_h be a solution of $\mathcal{L}_h(u_h)=0$. Suppose that (a),(b),(c), and (d) hold. Then for ρ and ω_i 's properly chosen the iteration (2.20) gives a sequence $\{w_k\}$ which converges to u_h , where $A+B = LL^*$ is obtained from the factorization (3.13) if the coefficients a_i are sufficiently smooth or $A+B$ is obtained by the procedure of Theorem 2 if the a_i 's are not sufficiently smooth. Further the number of arithmetic operations necessary to reduce the A-norm by a factor ϵ is

$$O(h^{-n-\frac{1}{2}} \log \epsilon^{-1}).$$

6. A computational factorization.

In this section we shall develop an algorithm for the approximate factorization of A into $\tilde{L}U$, where $\tilde{L}$ is lower triangular and U is upper triangular.

It is important to note that the factorization into $\tilde{L}U$ gives exactly the same operator as the factorization into LL^* so that the results in the previous sections remain valid.

Let $\tilde{L}$ be given by

$$(6.1) \quad \tilde{L}v_\alpha = f_\alpha v_\alpha + \sum_{i=1}^n c_{\alpha-e_i, e_i} v_{\alpha-e_i},$$

where the c_{α, e_i} 's are the coefficients of A as given in section 3. Let

$$(6.2) \quad Uv_\alpha = v_\alpha + \sum_{i=1}^n g_{\alpha, e_i} v_{\alpha+e_i},$$

If we let $B = \tilde{B} + \tilde{D}$ be given by (3.11) and (3.12) then we obtain the following set of relations for f and g_{α, e_i}

$$(6.3) \quad \begin{aligned} f_\alpha &= (1 + \varphi_\alpha) b_\alpha - \sum_{j=1}^n \{ c_{\alpha-e_j, e_j} [f_{\alpha-e_j}^{-1} \sum_{i=1}^n c_{\alpha-e_j, e_i}] \} \\ g_{\alpha e_i} &= f_\alpha^{-1} c_{\alpha, e_i}. \end{aligned}$$

Let F be the operator such that

$$(6.4) \quad Fv_\alpha = f_\alpha v_\alpha$$

and write

$$(6.5) \quad A = L_1 + L_1^* + D_1,$$

where L_1 is strictly lower triangular and D_1 is diagonal.

Then

$$(6.6) \quad \begin{aligned} A+B &= \tilde{L}U \\ &= (L_1+F)(I+F^{-1}L_1^*) \\ &= (L_1+F)F^{-1}(F+L_1^*). \end{aligned}$$

Thus in order to store the factorization $\tilde{L}U$ all we need to store is F . We can calculate F as follows

$$(6.7) \quad \begin{aligned} f_\alpha &= (1+\varphi_\alpha)b_\varphi - \sum_{j=1}^n c_{\alpha-e_j, e_j} z_{\alpha-e_j} \\ z_\alpha &= f_\alpha^{-1} \sum_{i=1}^n c_{\Sigma, e_i}. \end{aligned}$$

Note that in order to carry out the factorization we need only store a small fraction of the z_α 's at any time. If we carry out the factorization in the form (6.7) we do $n+1$ multiplications, $2n$ additions, and one division per point.

References

1. T. Dupont, R. P. Kendall, and H. H. Rachford, Jr., A factorization procedure for solving elliptic difference equations: part I, to appear.
2. E. G. D'jakanov, On an iterative method for the solution of finite difference equations, Dokl. Akad. Nauk. SSSR 138 (1961), 522-525.
3. G. E. Forsythe and W. R. Wasow, Finite-difference Methods for Partial Differential Equations, John Wiley and Sons, New York, 1960.
4. J. E. Gunn, the solution of elliptic difference equations by semi-explicit techniques, J. Soc. Indust. Appl. Math., Numer. Anal., Ser B, 2 (1964), 24-45.
5. E. L. Stiefel, Kernel polynomials in linear algebra and their numerical applications, Further Contributions to the Solution of Simultaneous Linear Equations and the Determination of Eigenvalues, Nat. Bur. Standards Appl. Math. Ser., 49, 1-22.
6. H. L. Stone, Iterative solution of implicit approximations of multidimensional partial differential equations, to appear.
7. R. S. Varga, Matrix Iterative Analysis, Prentice-Hall, Englewood Cliffs, New Jersey, 1962.