USE OF FAST DIRECT METHODS FOR THE EFFICIENT NUMERICAL SOLUTION OF NONSEPARABLE ELLIPTIC EQUATIONS

BY

PAUL CONCUS AND GENE H. GOLUB

STAN-CS-72-278

APRIL 1972

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

Also issued as Lawrence Berkeley Laboratory Report LBL 932

Use of fast direct methods for the efficient numerical solution of nonseparable elliptic equations

by

Paul Concus* and Gene H. Golub**

Abstract

We study an iterative technique for the numerical solution of strongly elliptic equations of divergence form in two dimensions with Dirichlet boundary conditions on a rectangle. The technique is based on the repeated solution by a fast direct method of a discrete Helmholtz equation on a uniform rectangular mesh. The problem is suitably scaled before iteration, and Chebyshev acceleration is applied to improve convergence. We show that convergence can be exceedingly rapid and independent of mesh size for smooth coefficients. Extensions to other boundary conditions, other equations, and irregular mesh spacings are discussed, and the performance of the technique is illustrated with numerical examples.

*Lawrence Berkeley Laboratory, University of California, Berkeley, California 94720.

**Department of Computer Science, Stanford University, Stanford, California 94305.

This research was supported by the Atomic Energy Commission, Project SU326 P30-17 .

Introduction. In recent years, fast direct methods have been developed for the numerical solution of the Poisson equation on a rectangle [1, 2]. By taking advantage of the special block structure of the approximating discrete equation on a uniform rectangular mesh, these methods obtain the solution with striking efficiency and accuracy. A comparison of fast direct methods with other methods can be found in [3], and the extension to more general separable elliptic equations in [4].

In this paper, we investigate a technique for using fast direct methods to solve iteratively more general formally self-adjoint strongly elliptic equations $\mathfrak{L}u = f$, which are not necessarily separable. We consider mainly Dirichlet conditions on the boundary of the rectangle, although the technique applies with slight modification to other boundary conditions for which fast methods are suitable. Our approach is to utilize a modified form of the iterative procedure

$$(1) \qquad -A u_{n+1} = -A u_n - \tau(\mathfrak{L}u_n - f), \qquad A \equiv \partial^2/\partial x^2 + \partial^2/\partial y^2$$

proposed for numerical computation in conjunction with alternating-direction methods by D'Yakonov[5] and discussed recently by Widlund [6]. This procedure, in addition to being of a form suitable for fast direct methods, has the desirable feature that for well-behaved problems its convergence rate is essentially independent of mesh size.

The iteration (1) as it stands, however, may be too slowly convergent to be of practical importance, even when optimal values of the parameter τ are used. The purpose of our paper is to discuss means for improving the iterative procedure so that it becomes a potent one for attacking a class of problems

arising frequently in applications. The means we employ are: (i) scaling the original problem $\mathfrak{L}u = f$ and iterating instead with the scaled problem $\mathfrak{M}w = q$; (ii) using, instead of (1), the shifted iteration

$$(2) \qquad (-\Delta+K)w_{n+1} = (-\Delta+K)w_n - \tau(\mathfrak{M}w_n - q) ,$$

where K is a suitably chosen constant; (iii) applying Chebyshev acceleration.

These means in themselves are not necessarily new; it is the effectiveness of their combination for solving this problem that we wish to investigate, We remark that algorithms for the fast direct solution of the discrete Poisson equation in a rectangle can handle iteration (2), which requires the repeated solution of a Helmholtz equation, with the same rapidity as they can (1).

In § 1 our 'basic iteration procedure for smooth coefficients is described and in § 2 its convergence studied. In §3 the generalization to non-smooth coefficients is discussed. In § 4 the results of numerical experiments are given to illustrate the behavior of the procedure. In the remaining sections, the question of scaling is covered, and generalizations to other equations and nonuniform mesh spacing are discussed.

Related iterative techniques for elliptic equations are studied in [7] in connection with alternating-direction methods and in [8, 9] in connection with Stone's sparse factorization method. This latter method is formally similar to ours; however, our technique has the desirable property of being based on a more natural splitting of the operator. In [10] a related approach to nonlinear ordinary differential equations is investigated.

1. The iterative procedure. In its simplest form, the iterative procedure considered in this paper solves numerically on a uniform rectangular mesh the problem

$$(3) \qquad \mathcal{L}u \equiv -\nabla\cdot [a(x,y)\nabla u] = f(x,y) \text{ on } \mathcal{R}$$

$$(4) \qquad u(x,y) = g(x,y) \text{ on } \partial\mathcal{R} ,$$

where $\mathcal{R}$ is the rectangle $0 < x < c$, $0 < y < d$ and $a(x,y)$ is strictly positive on $\mathcal{R}$ and its boundary $\partial\mathcal{R}$. We assume $a(x,y)$ and $g(x,y)$ to be sufficiently smooth so that the solution $u(x,y)$ is well behaved. The positivity of $a(x,y)$ implies that $\mathcal{L}$ is positive definite.

If $a(x,y)$ has bounded second derivatives on $\mathcal{R}\cup\partial\mathcal{R}$, which is the case of principal interest for the use of our procedure, the change of variable is performed

$$(5) \qquad w(x,y) = [a(x,y)]^{\frac{1}{2}}u(x,y) .$$

Then, after division by $a^{\frac{1}{2}}$, (3) becomes

$$(6) \qquad a^{-\frac{1}{2}}\mathcal{L}u = \mathcal{M}w \equiv -\Delta w + p(x,y)w = q(x,y) \text{ on } \mathcal{R} ,$$

where $p(x,y) = a^{-\frac{1}{2}}\Delta\,(a^{\frac{1}{2}})$ and $q(x,y) = a^{-\frac{1}{2}}f$. The effect of this scaling is to transform the operator $\mathcal{L}$ into one whose differential part is $-\Delta$. Note that the change of variable (5) does not alter the positive definiteness of $\mathcal{L}$, so that $\mathcal{M}$ is positive definite as well.

Substitution of (6) into (2) then yields as our iteration

$$(7) \qquad (-\Delta+K)w_{n+1} = (-\Delta+K)w_n - \tau(-\Delta+p)w_n + \tau q \text{ on } \mathcal{R} .$$

The boundary condition is

$$(8) \qquad w_{n+1} = H(x,y) \quad \text{on } \partial\mathcal{R} ,$$

where $H(x,y) = a^{\frac{1}{2}}g$.(1)

. In an attempt to make the operator $-\Delta+K$ on the left of (7) agree closely with $\mathcal{M}$, we choosethe constant K to approximate $p(x,y)$, The choice of central interest in our study is the minimax value,

$$(9) \qquad K = \tfrac{1}{2}(\beta+B) ,$$

where β is the minimum and B the maximum value of $p(x,y)$ on the closed *rectangle.* As will be shown in the next section, this choice leads to an estimate that the optimal value of the single parameter τ to give most rapid convergence in (7) is

$$(10) \qquad \tau = 1 .$$

For this value of τ , (7) becomes simply

$$(11) \qquad (-\Delta+K)w_{n+1} = (K-p)w_n + q \qquad \text{on } \mathcal{R} .$$

We have presented the iterative procedure in its underlying continuous form to bring emphasis to the point that the convergence properties should not be expected to depend significantly on the mesh size, at least for the case of twice differentiable $a(x,y)$. The discretized version of the iterative procedure (8, 9, 11) is discussed in subsequent sections. To obtain it, we place a uniform rectangular mesh on $\mathfrak{R}$ with spacing h in the x-direction and k in the y-direction and let W_{ij} correspond to $w(x,y)$ at the mesh points $x=ih$, $y=jk$. Corresponding to the operator $-A$ with Dirichlet boundary conditions we take the standard five-point approximation,

$$(12) \qquad -A_h W_{ij} \equiv h^{-2}(-W_{i-1,j}+2W_{ij}-W_{i+1,j}) + k^{-2}(-W_{i,j-1}+2W_{ij}-W_{i,j+1}) ,$$

$$i=1,2,\ldots,\frac{c}{h}-1 \; ; \; j=1,2,\ldots,\frac{d}{k}-1 .$$

Then the discrete form of iteration (8, 11) is given by

$$(13) \qquad (-\Delta_h+KI)W^{(n+1)} = (KI-P)W^{(n)} + Q ,$$

where P is a diagonal matrix with elements $P_{ij} = p(ih,jk)$, Q is a vector with elements $Q_{ij} = q(ih,jk)$, and I is the identity matrix. The solution of (13) is carried out in each iteration by using a fast direct method.

Finally, under the assumption that the eigenvalues of $(-\Delta_h+KI)^{-1}(KI-P)$ lie in the interval $[-\rho,\rho]$, Chebyshev acceleration is applied [14]:

$$(14) \qquad \tilde{W}^{(n+1)} = \omega_{n+1}(W^{(n+1)} - \tilde{W}^{(n-1)}) + \tilde{W}^{(n-1)} ,$$

where $\omega_0=1$, $\omega_1=2/(2-\rho^2)$, $\omega_{n+1}=(1-\rho^2\omega_n/4)^{-1}$ for $n=1,2,\ldots$, and $\tilde{W}^{(n+1)}$ is the improved value of $W^{(n+1)}$, where now $W^{(n+1)}$ satisfies

(13) with $W^{(n)}$ replaced by $\tilde{W}^{(n)}$ on the righthand side. This is equivalent to the use in (7) of a sequence $\{\tau_n\}$, rather than a single value of τ, in a manner that is numerically stable and does not require the total number of parameters in the sequence to be specified in advance. If in some cases memory limitations preclude the use of (14), then a fixed sequence $\{\tau_n\}$ could be used instead, ordered in the manner recommended in [11] for numerical stability.

2. Convergence properties. We return to iteration ('7, 8), in which the values of K and τ are not yet specified. Its convergence properties can be examined by standard methods in terms of the eigenvalues of the Laplace operator, which are known explicitly for the rectangle. We carry out here the analysis for the discrete form of the iteration; the continuous analysis proceeds in essentially the same manner (for example, as in [12]).

2.1 We give first, for comparison purposes, the behavior of the discrete form of iteration (1, 4) for the original problem (3, 4) without scaling or shifting. We place a uniform rectangular grid on the rectangle, as in the previous section, and obtain as the discrete form of (3,4)

$$(15)\qquad LU_{ij} \equiv h^{-2}(-a_{\bar{i}j}U_{i-1,j}+[a_{\bar{i}j}+a_{\bar{i}+1,j}]U_{ij} - a_{\bar{i}+1,j}U_{i+1,j}) +$$

$$+ k^{-2}(-a_{i\bar{j}}U_{i,j-1} + [a_{i\bar{j}} +a_{i,\bar{j}+1}]U_{ij} - a_{i\,\bar{j}+1}U_{i,j+1}) = F_{ij}\ ,$$

$$i=1,2,\ldots,\ \frac{c}{h}-1\ ;\ j=1,2\ ,\ldots,\ \frac{d}{k}-1\ ,$$

where $a_{i\cdot j}$ denotes $\frac{1}{2}(a_{\bar{i}\bar{j}} + a_{\bar{i},\bar{j}+1})$ and $a_{i\bar{j}}$ denotes $\frac{1}{2}(a_{\bar{i}\bar{j}} + a_{\bar{i}+1\,\bar{j}})$, $a_{\bar{i}\bar{j}}$ being the value of $a(x,y)$ at $x = (i-\frac{1}{2})h$, $y = (j-\frac{1}{2})k$. The vector element U_{ij} corresponds to $u(ih,jk)$ and F_{ij} is equal to $f(ih,jk)$. Then the discrete form of (1, 4) is

$$(16)\qquad -\Delta_h U^{(n+1)} = -\Delta_h U^{(n)} - \tau(LU^{(n)} - F) \quad \text{or, after premultiplying by } (-\Delta_h)^{-1}\ ,$$

$$U^{(n+1)} = (I-\tau[-\Delta_h]^{-1}L)U^{(n)} + \tau(-\Delta_h)^{-1}F\ ,$$

where, as is the case for $\mathcal{L}$, the positivity of $a(x,y)$ implies the positive

definiteness of L .

The spectral radius ρ of the iteration matrix $(I-\tau[-\Delta_h]^{-1}L)$ is expressed in terms of μ_m and μ_M , the minimum and maximum eigenvalues of the generalized eigenvalue problem

$$L\Phi = \mu(-\Delta_h)\Phi \ , \tag{17}$$

as

$$\rho(I-\tau[-\Delta_h]^{-1}L) = \text{Max}(|1-\tau\mu_m| \ , \ |1-\tau\mu_M|) \ . \tag{18}$$

Since L and $-\Delta_h$ are positive definite, $\mu_m > 0$. There follows the well-known result [22]

Lemma: Iteration (16) converges for any initial approximation $u^{(0)}$ if and only if $0 < \tau < 2/\mu_M$, and for a single parameter τ the optimal choice

$$\tau = \tau_0 \equiv 2/(\mu_m+\mu_M) \tag{19}$$

yields the smallest spectral radius

$$\rho = \rho_0 \equiv (\mu_M-\mu_m)/(\mu_M+\mu_m) \ . \tag{20}$$

It is straightforward. to show that the uniform estimate independent of h and k

$$0 < \alpha \leq \frac{\Phi^T L\Phi}{\Phi^T(-\Delta_h)\Phi} < A \tag{21}$$

holds for any vector Φ , where $\alpha = \min a(x,y)$ and $A = \text{Max } a(x,y)$ on the closed rectangle. There follows the corresponding estimate

(22) $$0 < \alpha \leq \mu_m \leq \mu_M \leq A \ ,$$

based on which we obtain, from (19) and (20), that for

$$\tau = 2/(\alpha + A)$$

there holds

(23) $$\rho \leq (A - \alpha)/(A + \alpha) \ .$$

The estimate (22), and hence (23), are, in fact, the sharpest possible uniform ones, as can be seen by taking- for Φ in the Rayleigh quotient (21) a vector that is zero except at the position corresponding to the maximum diagonal element of L, or, alternatively, is zero everywhere except at the minimum.

2.2 The discrete form of the shifted iteration (7, 8) for the scaled problem (6, 8) is

(24) $$(-\Delta_h+KI)W^{(n+1)} = (-\Delta_h+KI)W^{(n)} - \tau[(-\Delta_h+P)W^{(n)} - Q] \ .$$

We do not yet specify K to be the value (9), but require for now only that $K > -\lambda_m$, where λ_m is the smallest eigenvalue of $-\Delta_h$; hence $(-\Delta_h+KI)$ is positive definite. We assume also that $\mathcal{M}$ is sufficiently positive definite so that the discretization to $M \equiv -\Delta_h+P$ does not destroy the positive definiteness. Then, corresponding to (17) and (18), we have that if ν_m and ν_M are the minimum and maximum eigenvalues of

$$M\Phi = \nu(-\Delta_h+KI)\Phi \ ,$$

the spectral radius for iteration (24) is

$$(25) \qquad \rho(I-\tau[-\Delta_h+KI]^{-1}M) = \mathrm{Max}(|1-\tau\nu_m| \ , \ |1-\tau\nu_M|) \ ,$$

and the Lemma holds for iteration (24) with μ_m and μ_M replaced by ν_m and ν_M .

To estimate ν_m and ν_M , *we* use the Rayleigh quotient for ν ,

$$(26) \qquad \frac{\Phi^T M\Phi}{\Phi^T(-\Delta_h+KI)\Phi} = \ , \ + \frac{\Phi^T(P-KI)\Phi}{\Phi^T(-\Delta_h+KI)\Phi} \ .$$

Thus

$$(27) \qquad 1 + \min\left(\frac{\beta-K}{\lambda_m+K} \ , \ \frac{\beta-K}{\lambda_M+K}\right) \leq \nu_m \leq \nu_M \leq 1 + \mathrm{Max}\left(\frac{B-K}{\lambda_m+K} \ , \ \frac{B-K}{\lambda_M+K}\right) ,$$

where λ_M is the largest eigenvalue of $-\Delta_h$.

The estimate for ρ obtained from (25) and (27) is least when a choice for K is made such that

$$(28) \qquad \beta-K \leq 0 \leq B-K \ ,$$

assuming $\beta > -\lambda_m$ holds. There results that for the corresponding optimal choice

$$(29) \qquad \tau = \frac{2(\lambda_m+K)}{2\lambda_m+B+\beta}$$

there holds

$$(30) \qquad P \leq \rho_u \equiv \frac{B}{2\lambda_m+B+\beta} \ .$$

To obtain a uniform upper bound on the spectral radius ρ , we note that the

smallest eigenvalue λ_m of $-\Delta_h$ is given by

$$\lambda_m = 4h^{-2}\sin^2\left(\frac{\pi h}{2c}\right) + 4k^{-2}\sin^2\left(\frac{\pi k}{2d}\right)$$

and that it satisfies

$$\lambda_m > \underline{\lambda}_m = \frac{\pi^2}{c^2}\left(1 - \frac{\pi^2 h_0^2}{24c^2}\right)^2 + \frac{\pi^2}{d^2}\left(1 - \frac{\pi^2 k_0^2}{24d^2}\right)^2 \tag{31}$$

for mesh spacings $0 < h \leq h_0$, $0 < k \leq k_0$. Substituting (31) into (30) yields the desired bound in terms of the upper bounds h_0 , k_0 on the mesh spacing.

We note also that λ_m is bounded above by

$$\lambda_m < \bar{\lambda}_m = \pi^2/c^2 + \pi^2/d^2$$

for all h , $k > 0$; the quantity $\bar{\lambda}_m$ is equal to the smallest eigenvalue of $-\Delta$ for Dirichlet boundary conditions on the rectangle. For most computational purposes it is not necessary to take into account the $0(h_0^2 + k_0^2)$ difference between $\underline{\lambda}_m$ and $\bar{\lambda}_m$, but instead to regard λ_m as being essentially equal to the simpler $\bar{\lambda}_m$.

The presence of the $^2\lambda_m$ term in the denominator of (30) can have the effect of there resulting a considerably smaller bound on ρ for the scaled and shifted problem than results from (23) for the original problem. Since (23) is essentially sharp such a smaller bound would imply a faster convergence rate. Thus we conclude that scaling and shifting are most effective when A/α is not especially close to one and p does not vary with excessive rapidity over the rectangle, in which case the resulting improvement in convergence rate could be substantial.

2.3 To illustrate the improvement in convergence rate for an ideal case, we consider the solution of $\nabla\cdot(e^{10(x+y)}\nu u) = f$. For this case, $\alpha = 1$ and $A = e^{20}$, so that for the unscaled, unshifted iteration (16), the estimate for the optimal spectral radius, from (23), is $\rho \lesssim 1 - 2e^{-20} \approx 1 - 0.4\times10^{-8}$. Thus it takes the order of 10^8 iterations to reduce the initial error by a factor $1/e$.

For the iteration (24), however, we have $p(x,y) = \Delta(e^{5(x+y)})/e^{5(x+y)} = 50$, so that $\beta = B = 50$; hence, if (9) holds (that is $K = 50$), the optimal spectral radius, from (30), is $\rho = 0$. Thus the problem is solved completely (to round-off accuracy) in only one iteration!

This example emphasizes the point that we solve directly a discrete Helmholtz equation (24) at each iteration.

2.4 We require in § 2.2 that β, the minimum of $p(x,y)$ on the rectangle, satisfy $\beta > -\lambda_m$. In the case for which $\beta \leq -\lambda_m$ (the positive definiteness of M does not preclude P dipping below $-\lambda_m$ over a portion of the rectangle) the estimate (27) no longer yields an upper bound on ρ that is less than one, hence it does not guarantee convergence. For the numerical examples of such cases given in § 4, the iteration (24) converges, but at a comparatively slower rate. We consider, then, as best candidates for our iterative procedure those cases for which $\beta > -\lambda_m$ (or uniformly $\beta > -\lambda_m$).

The choice of the particular value (9) for K out of the possible ones (28) yielding the best convergence rate estimate (30), corresponding to (29), is made for two reasons. One is that for the corresponding value $\tau = 1$, which is obtained from (29) for the shift (9), the resulting discrete Picard iteration (13) requires fewer computer operations than does the one

for general τ(24).[2] The other is that for this shift the actual convergence rate observed in our numerical experiments was somewhat more rapid than it was for shifts near the end points of the interval $[\beta,B]$. Centering the spectrum of P - KI at zero and taking $\tau = 1$ seemed to be a good strategy for reducing the spectral radius of $[I - \tau(-\Delta_h+KI)^{-1}(-\Delta_h+P)]$, at least for those problems for which $p(x,y)$ varied smoothly without rapid changes. If p varied more violently, the shift (9) was still effective, but in some cases an improvement could be realized by fixing τ at one and selecting another value for K in the interval $[\beta,B]$ that better approximated P ; see § 4. (For the discrete scaling of §3, one obtains the estimate (9) directly.)

Note that the primary effect of the shift K is the reduction of the norm of P - KI ; the effect on $(-A_h + KI)$ is usually slight and of little importance.

In some cases it may be more convenient and advantageous to use in (9) the sharper discrete bounds $\beta_h = \min P_{ij}$ and $B_h = \text{Max } P_{ij}$, instead of β and B . Specific shifts other than the min-max one (9), such as $K = (cd)^{-1}\iint_{\mathcal{R}} p(x,y)dxdy$ or its discrete equivalent, and shifts that change from one iteration to the next, are not considered here, but may be of practical interest.

For the choice (9) and for $-\underline{\lambda}_m < \beta \leq B < \infty$, we may summarize the behavior of the iteration procedure as follows:

Theorem: For mesh sizes $0 < h \leq h_0$, $0 < k \leq k_0$ the iteration (9, 13) converges with spectral radius $\rho \leq (B-\beta)/(2\underline{\lambda}_m + B + \beta)$.

2.5 In applying Chebyshev acceleration (14) to iteration (13), one can either use the estimate (30) for the spectral radius or else obtain an estimate by observing the convergence rate when solving the problem first on a coarse grid. This latter procedure is often worth the small extra expenditure of computing effort, because the estimate (30) may be pessimistic and, since the iteration is essentially independent of mesh size, the observed value usually is more accurate. At any rate, the convergence of (14) is assured when $\rho < 1$.

If one uses a fixed sequence $\{\tau_n\}$ rather than (14), then it may be possible to speed convergence by utilizing the property that the largest eigenvalue of $-\Delta^{-1}$ is relatively isolated from its remaining eigenvalues, which cluster toward zero. For example, on the unit square with Dirichlet conditions the largest eigenvalue is $(2\pi^2)^{-1}$, whereas all the others lie between zero and $(5\pi^2)^{-1}$. The eigenvalues of $(-\Delta_h + KI)^{-1}(KI - P)$ exhibit a similar grouping for some problems, hence the special parameter selection method given in [13] for such cases could be employed.

A recent discussion of practical means for estimating Chebyshev acceleration parameters as an iteration proceeds is contained in [15].

We remark that obtaining the optimal Chebyshev acceleration parameters is not of central importance in our scheme. In many cases the scaling and shifting alone can yield a convergence rate that is so rapid that only a few iterations are required for convergence, thus leaving little room for any

substantial improvement to be made by further refinement of the Chebyshev parameters.

3. Non-smooth $a(x,y)$.

3.1. For the case in which $a(x,y)$ is only piecewise smooth, the situation generally is less favorable. The change of variable (5) cannot be carried out as described in § 1, since $\Delta(a^{\frac{1}{2}})$ does not exist everywhere on R (except in a generalized sense). It may still be possible, however, to improve on the convergence rate of (16) by performing the equivalent change of variable in discrete form.

A discrete scaling corresponding to (5) is the one transforming the diagonal elements of L into those of $-\Delta_h$,

$$(32) \qquad M_h = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} , \qquad W = D^{\frac{1}{2}} U ,$$

where D is the diagonal matrix with elements

$$(33) \qquad d_{ij} = \frac{1}{2(h^{-2}+k^{-2})} \left(\frac{a_{\bar{i}j}+a_{\bar{i}+1,j}}{h^2} + \frac{a_{i\bar{j}}+a_{i,\bar{j}+1}}{k^2} \right) .$$

The resulting scaled matrix operator M_h is then $M_h = -\Delta_h + R$, and the original discrete equation $LU = F$ becomes

$$(34) \qquad M_h W = D^{-\frac{1}{2}} F = Q_h .$$

The (symmetric) matrix R has zeros on its main diagonal and, in general, four non-zero diagonal bands. We have

$$(35) \qquad RW_{ij} = h^{-2} \left\{ \left[1 - \frac{2a_{\bar{i}j}(h^{-2}+k^{-2})}{d_{ij}^{\frac{1}{2}} d_{i-1,j}^{\frac{1}{2}}} \right] W_{i-1,j} + \left[1 - \frac{2a_{\bar{i}+1,j}(h^{-2}+k^{-2})}{d_{ij}^{\frac{1}{2}} d_{i+1,j}^{\frac{1}{2}}} \right] W_{i+1,j} \right\} +$$

$$+ k^{-2}\left\{\left[1 - \frac{2a_{ij}^{-}(h^{-2}+k^{-2})}{d_{ij}^{\frac{1}{2}} d_{i,j-1}^{\frac{1}{2}}}\right] w_{i,j-1}\right.$$

$$\left. + \left[1 - \frac{2a_{i,j+1}^{-}(h^{-2}+k^{-2})}{d_{ij}^{\frac{1}{2}} d_{i,j+1}^{\frac{1}{2}}}\right] W_{i,j+1}\right\} .$$

For the case in which $p(x,y) = a^{-\frac{1}{2}}\Delta(a^{\frac{1}{2}})$ exists at the point (i,j), RW_{ij} is a multipoint difference approximation of $p(x,y)w(x,y)$ there. That is, the matrix R is an alternative representation of $p(x,y)$, which is represented in § 1, 2 by the diagonal matrix P. Note that R requires approximately twice as much computer storage as does P.

For the case in which either a or ∇a has a simple discontinuity, the elements of R are not uniformly bounded for all mesh spacings in the way that the elements of P are bounded by β and B. For a given mesh spacing, however, R may be such that the convergence rate of the iteration corresponding to (24),

$$(36) \qquad (-\Delta_h + KI)W^{(n+1)} = (-\Delta_h + KI)W^{(n)} - \tau[(-\Delta_h + R)W^{(n)} - Q_h] ,$$

is satisfactorily rapid for suitable K and τ.

The best values for K and τ can be estimated in a manner similar to that of § 2.2. The' Rayleigh quotient analogous to (26) is

$$(37) \qquad \frac{\Phi^T M_h \Phi}{\Phi^T (-\Delta_h + KI)\Phi} = 1 + \frac{\Phi^T (R-KI)\Phi}{\Phi^T (-\Delta_h + KI)\Phi} .$$

Here, however, we do not as in § 2.2 choose K so that the spectrum of $(R - KI)$ is centered near zero (the spectrum of R is already centered at zero, since R has Property A and zero main diagonal), but we obtain the

proper estimate after first rewriting (37) as

$$\frac{\Phi^T M_h \Phi}{\Phi^T (-\Delta_h + KI)\Phi} = 1 - K\left[\frac{1}{2h^{-2}+2k^{-2}} \frac{\Phi^T (-\Delta_h)\Phi}{\Phi^T (-\Delta_h + KI)\Phi} \frac{\Phi^T (\frac{1}{K} R - \frac{1}{2h^{-2}+2k^{-2}} \Delta_h - I)\Phi}{\Phi^T (-\Delta_h + KI)\Phi}\right]$$

Neglecting the first term in the brackets on the right, which is $0(h^2 + k^2)$, and using the Gerschgorin estimate for the spectral radius of the matrix in the numerator of the second term we obtain the choice (9, 10), providing $\beta > -\lambda_m$, where now $\beta = \beta_R \equiv 4\{R\}_{min}$ and $B = B_R = 4\{R\}_{max}$; $\{R\}_{min}$ denotes the smallest element of R and $\{R\}_{max}$ the largest. These values for β and B are analogous to the corresponding ones for $p(x,y)$ in that they are minimum and maximum values of difference quotients approximating $a^{-\frac{1}{2}}\Delta(a^{\frac{1}{2}})$. Generally speaking, one expects these difference quotients to behave like $(h^{-1} + k^{-1})$ where ∇a has a simple discontinuity and like $(h^{-2} + k^{-2})$ where a itself is discontinuous.

Relevant numerical experiments are discussed in § 4.4. As might be expected, the behavior in this case is not totally as satisfactory as it is for the case of smooth $a(x,y)$ and the diagonal matrix P. The parameter estimates based on β_R and B_R are not as sharp, perhaps because the elements of R often vary sizably and are not always accurate approximations of $\frac{1}{4} a^{-\frac{1}{2}}\Delta(a^{\frac{1}{2}})$. An attempt to improve the value of τ, and even K, may be useful for such problems as information on the spectrum of $(-A_h + KI)^{-1}M_h$ is gained during the iteration.

We remark that the method of this section could be used as well for smooth $a(x,y)$, as an alternative to the analytic calculation of $p(x,y) = a^{-\frac{1}{2}}\Delta(a^{\frac{1}{2}})$ and the subsequent numerical evaluation to obtain the

elements of P in (13). We do not recommend this alternative, however. If one wishes to avoid these calculations he should instead difference $a^{\frac{1}{2}}(x,y)$ directly, to obtain the approximate elements p^h_{ij} of P, $p^h_{ij} = \Delta_h a^{\frac{1}{2}}_{ij}/a^{\frac{1}{2}}_{ij}$, where $a^{\frac{1}{2}}_{ij} = [a(ih,jk)]^{\frac{1}{2}}$. The discretization error introduced by using p^h_{ij}, instead of $p(ih,jk)$, is of the same order as that already introduced by (12). The iteration (24) is generally preferable to (36) because it requires less storage and fewer computer operations per iteration and because, in our experience, the parameter estimates based on P are sharper than those based on R.

We remark also that discrete scalings other than (32, 33) might be used. For example, a closely related one is (32) with the choice, instead of (33), $d_{ij} = a(ih,jk)$. Alternatively, one could investigate the use of (24) with the elements of P equal to ph_{ij} in the case for which $a(x,y)$ is only piecewise smooth. This would be equivalent, for fixed mesh, to considering $a(x,y)$ to be a smooth, but locally rapidly changing, function. We hope to return to these matters in a future study. The question of scaling is discussed further in § 5.

3.2 When $a(x,y)$ is piecewise smooth with sizable discontinuities across sub-domain boundaries within the rectangle, it can be faster to solve the problem iteratively as a sequence of problems on each sub-domain than as a single problem over the entire rectangle. For example, consider the problem (3, 4) for which $a(x,y)$ is piecewise constant

$$a(x,y) = \begin{cases} a_0, & 0 < x \leq c/2 \\ a_1, & c/2 < x < 1 \end{cases}$$

with the matching condition that $a\partial u/\partial x$ is continuous at $x = c/2$.

We consider solving the problem numerically by the following scheme:

(i) In the sub-domain where $a = a_0$ solve by a fast direct method

$$\mathfrak{L}u_0^{(n)} = f ,$$

subject to (4) and $u_0^{(n)} = u^{(n)}$ on $x = c/2$.

(ii) In the sub-domain where $a = a_1$ solve similarly

$$\mathfrak{L}u^{(n+1)} = f ,$$

subject to (4) and $a_1 \partial u^{(n+1)}/\partial x = a_0 \partial u_0^{(n)}/\partial x$ on $x = c/2$.

Then one obtains that the error $\epsilon^{(n)}(y)$ in the value of u on $x = c/2$ at the n^{th} iteration satisfies $\epsilon^{(n)} = -\frac{a_0}{a_1}\epsilon^{(n-1)}$. Thus if a_0/a_1 is small, this scheme can be more rapidly convergent than the one given in § 3.1 for solving the problem on the entire rectangle at once. The scheme is equivalent to using on the original problem, instead of $D^{-\frac{1}{2}}$, a diagonal scaling that renders the discretized operator only weakly coupled between the sub-problems (i) and (ii) .

For the case in which a varies with x and y in each sub-domain, the iteration (i, ii) could be combined in some cases with that of § 1, 2 ; we hope to take up this matter in another paper. We give the results of a numerical experiment for piecewise constant $a(x,y)$ in § 4.4.

4. Numerical Examples. In this section we collect the results of numerical experiments for several cases of (3,4) to illustrate the contents of the previous sections.

4.1 The ideal case for the basic technique (13, 9) is one in which $p = a^{-\frac{1}{2}}\Delta(a^{\frac{1}{2}})$ is constant on the rectangle (e.g., $a = \cos^2(x+y)$, $a = J_0^2([x^2+y^2]^{\frac{1}{2}})$, etc.). Then, as is pointed out for one such example in § 2.3, only one iteration is required to solve the problem.

Since a numerical illustration of this property would have been trivial, we instead checked the correspondence of iterations (24) and (36) by solving several such cases using, instead of (24), the iteration (36), which is based on the discrete scaling (32). Using the value (9) for K (that is, $K = a^{-\frac{1}{2}}\Delta(a^{\frac{1}{2}}) = \text{const.}$) and $\tau = 1$, we found, as expected, that the spectral radius of the iteration matrix, as indicated by the observed convergence rate, was the order of magnitude of the discretization error and decreased with mesh spacing for a given a. When using, instead, the value $K = \frac{1}{2}(\beta_R+B_R)$ of § 3.1, we observed slightly slower rates of convergence, even though this value is derived from (36). The estimate (9) based on P was especially preferable in the cases for which the elements of R did not accurately approximate $\frac{1}{4}\, a^{-\frac{1}{2}}\Delta(a^{\frac{1}{2}})$ everywhere on the rectangle.

4.2 Other highly suitable cases for the basic technique are those not departing strongly from the ideal one. We include here the results for two such cases, one for non-negative and one for non-positive $p(x,y)$. We solved both numerically using (24) on the unit square $0 < x < 1$,

$0 < y < 1$ with uniform mesh spacing $h = k = 2^{-\ell}$, for the values $\ell = 4, 5,$ and 6. (The number of rows of interior mesh points should be $2^{\ell} - 1$, ℓ an integer, in at least one direction for fast direct methods to apply efficiently.) The righthand side $q(x,y)$ of (6), and similarly Q of (24), were taken to correspond to $w(x,y) = 2[(x-\frac{1}{2})^2 + (y-\frac{1}{2})^2]$, for which the solution of the discrete problem with boundary data $H(x,y) = 2[(x-\frac{1}{2})^2 + (y-\frac{1}{2})^2]$ agrees exactly with $w(x,y)$ at the mesh points. The elements of the initial approximation $W^{(0)}$ were taken to be either all zero or else pseudo-random numbers in $(0, 1)$, to permit the presence of different eigenvector blends in the initial error.

The first example is the one for which $a(x,y) = [1 + \frac{1}{2}(x^4+y^4)]^2$, hence $p(x,y) = 6(x^2+y^2)/a^{\frac{1}{2}}(x,y)$ and $\beta = 0$, $B = 6$. Thus the estimate (30) for the optimal spectral radius is $\rho \leq \rho_u = \frac{6}{2\lambda_m+6} \approx 0.132$ (using $2\pi^2$ for λ_m), and the shift (9) is $K = 3$. The results are summarized for five examples of parameter choices in Table 1-a.

The entries in Table 1 are the rounded values for a mesh with 64x64 interior points and for the initial approximation $W^{(0)} \equiv 0$ in the interior of R. For the 16x16 and 32x32 meshes the values differed only slightly, if at all, from those in the table, and for the random initial approximations the iterations behaved similarly. A value of K equal to 0 or to $\frac{1}{2}(\beta + B)$ was used, along with the corresponding value (29) for τ. When Chebyshev acceleration was included, either the estimate ρ_u from (30) or the experimentally observed estimate ρ_e was used to approximate the spectral radius ρ in (14) of $(I - \tau[-\Delta_h+KI]^{-1}[-\Delta_h+P])$.

The entries for the value of ρ_e in the table are the observed approximate limiting values of the ratio $\|W^{(n)} - W^{(n-1)}\|_\Delta / \|W^{(n-1)} - W^{(n-2)}\|_\Delta$,

Table 1

Results after 5 iterations

	a (x,y)	K	τ	Chebyshev Acceleration	ρ_e	Maximum Error
a)	$[1+\frac{1}{2}(x+y^4)^4]^2$	0	0.868	none	0.13	3.7(-5)
		3	0.868	using ρ_u	—	2.4(-6)
		0	1	none	0.039	3.9(-8)
		3	1	using ρ_u	—	1.1(-6)
		3	1	using ρ_e	—	4.3(-9)
b)	$[1+\sin\frac{1}{2}\pi(x+y)]^2$	0	16/15	none	0.066	1.2(-6)
		0	16/15	using ρ_u	—	7.7(-8)
		$-\pi^2/8$	1	none	0.061	2.3(-7)
		$-\pi^2/8$	1	using ρ_u	—	3.2(-8)
		0		using ρ_e	—	2.3(-8)
c)	$[2+\tanh 4(x+y-1)]^2$	0	0.829	none	0.31	3.4(-4)
			0.829	using ρ_u	—	5.9(-3)
		4.07	1	none	0.26	2.6(-4)
		4.07	1	using ρ_u	—	1.5(-3)
				using ρ_e	—	3.4(-5)

Table 2

Iteration details for Table 1-a with $K = 3$ and $\tau = 1$

Iteration n	Without Chebyshev Acceleration			Chebyshev Acceleration Using ρ_e		
	$\frac{\|W^{(n)}-W^{(n-1)}\|_\Delta}{\|W^{(n-1)}-W^{(n-2)}\|_\Delta}$	$\frac{\|W^{(n)}-W^{(n-1)}\|_\Delta}{\|W^{(n)}\|_\Delta}$	Max Error	$\frac{\|W^{(n)}-W^{(n-1)}\|_A}{\|W^{(n-1)}-W^{(n-2)}\|_\Delta}$	$\frac{\|W^{(n)}-W^{(n-1)}\|_A}{\|W^{(n)}\|_\Delta}$	Max Error
1	—	1	1.6(-2)	—	1	1.6(-2)
2	0.0057	5.7(-3)	6.4(-4)	0.0057	5.7(-3)	7.1(-4)
3	0.033	1.9(-4)	2.4(-5)	0.13	7.5(-4)	1.1(-5)
4	0.038	7.3(-6)	1.0(-6)	0.0047	3.5(-6)	2.7(-7)
5	0.039	2.8(-7)	3.9(-8)	0.080	2.8(-7)	4.3 (-9)
6	0.039	1.1(-8)	1.7(-9)	0.0063	W - 9)	1.2(-10)

where $\|W\|_\Delta = [W^T(-\Delta_h+KI)W]^{\frac{1}{2}}$. The maximum error, which is listed in the last column, is the maximum of the difference at the mesh points between $W^{(5)}$ and the solution $w(x,y)$. Note that the initial maximum error has the value of approximately one.

In Table 2 the iteration-by-iteration details are given for the third entry in Table 1a.

The second example is the one for which $a(x,y) = [1 + \sin\frac{1}{2}\pi(x+y)]^2$, for which $\beta = -\pi^2/4$, $B = 0$, and $\rho_u \approx 1/15$. The results analogous to Table 1a are given in Table 1b. In this case, the improvement obtained by using the shift $K = \frac{1}{2}(\beta + B)$, instead of $K = 0$, is not so great as it is for the first test problem,

The effect of scaling and shifting can be found by comparing the results for the two test problems with the estimate (23) for iteration (16). For both problems there holds $\alpha = 1$ and $A = 4$, so that the spectral radius estimate without scaling and shifting in each case is 0.6.

4.3 Cases that are less strikingly suitable for the basic technique are discussed in this sub-section. The example summarized in Table 1c is for the case $a(x,y) = [2 + \tanh 4(x+y-1)]^2$. The test problem is the same as the one for the examples of § 4.2, and the entries are analogous to the others in the table, except that in this case the task of calculating the actual extremal values of $p(x,y)$ on R was not carried out; instead, the discrete equivalents $\beta = \beta_h = \min P_{ij}$, $B = B_h = \max P_{ij}$ were used. For the 64x64 mesh, $\beta_h \approx -9.62$ and $B_h \approx 17.77$, for which $\rho_u \approx 0.575$. Note that in this example, $K = 0$ does not correspond to an end point of the interval $[\beta,B]$. As before, the results were insensitive to mesh size

and to which of the initial approximations was used.

An investigation of the non-sharpness of estimate (27) and non-optimality of (9) and (10), which are more important here than in a nearly ideal case, was carried out by fixing τ at the value one and observing the change in ρ_e as K was varied. A local minimum was found at approximately K = 3.0 , for which ρ_e is approximately 0.23.

For the case $a(x,y) = [2 + \tanh 10(x+y-1)]^2$, β_h and B_h become approximately -60 and 111, respectively. In this case $\beta < -\lambda_m$, hence the estimate (30) yields merely that $\rho \leq \rho_u > 1$. The iteration did converge, however, with the observed spectral radius $\rho_e \approx 0.63$ and a maximum error of 2.5×10^{-2} after five iterations for the usual test problem, with $K = \frac{1}{2}(\beta_h + B_h)$ and $\tau = 1$. With the inclusion of Chebyshev acceleration based on this value of ρ_e , the maximum error after five iterations was reduced to 6.3×10^{-3} . The value of ρ_e can be decreased in this case, with τ fixed at 1 , to a locally minimum value of approximately 0.54 at approximately K = 14.

The case $a(x,y) = \{1.5 + \sin[10(x+y)/\sqrt{2}]\}^2$, for which $\beta = -40 < -\lambda_m$ and B = 200 , was observed to converge also, but at a slower rate. Here, for the shift K = 80 and for $\tau = 1$, the value observed for ρ_e was 0.91. Even though this value is large, it is interesting to note that it is smaller, nevertheless, than the spectral radius estimate $(A - \alpha)/(A + \alpha) \approx 0.93$ for the iteration without scaling and shifting.

4.4 The cases included for non-smooth a(x,y) are $a(x,y) = (1 + 4|x - \frac{1}{2}|)^2$, for which there is a slope discontinuity at $x = \frac{1}{2}$, and $a(x,y) = \begin{cases} 1 , & x < \frac{1}{2} \\ 9 , & x > \frac{1}{2} \end{cases}$, for which there is a jump discontinuity at $x = \frac{1}{2}$. For both these cases,

the iteration without scaling and shifting (16) has the spectral radius estimate $(A - \alpha)/(A + \alpha) = 0.8$ (independent of h). The convergence properties of the scaled and shifted iteration are not essentially independent of h , however, as is the case for the examples of § 4.1 - 4.3.

The problems were solved numerically using the iterative procedure of § 3.1. The dependence on h for the first case is illustrated in Table 3. The relationship (29) between K and τ and the value (30) of ρ_u were computed using for β and B the observed quantities β_R and B_R , the rounded values of which are listed in the table. The value of ρ_u was essentially equal to the observed value ρ_e for K = 0 . Note that the maximum error after ten, not five, iterations is given in the table.

Although $B_R - \beta_R$ is large, the elements of R are essentially zero everywhere except at the mesh points on and adjacent to the line $x = \frac{1}{2}$, where they become large. This suggests that a value of K closer to zero than the minimax value $\frac{1}{2}(\beta_R+B_R)$ might result in more rapid convergence. Indeed, it was found that for h = 1/16 and $\tau = 1$ a local minimum for ρ_e occurred at approximately K = 13 (see the last row of Table 3).

The second case is, in a sense, an extreme version of the second one of § 4.3, for which a(x,y) changes rapidly from approximately 1 on one half of the region to approximately 9 on the other. Here $\beta_{R_1} < -\lambda_{min}$, so that again convergence cannot be guaranteed from the estimate (30). Very slow convergence was observed for this case, especially for the smaller mesh spacings. For the 64x64 mesh the observed values for β_R and B_R were approximately -5597 and 9057 , respectively, for which $\tau \approx 0.0113$ from (29) for K = 0 . There resulted $\rho_e \approx 0.988$ for these parameters. For the shift $K = \frac{1}{2}(\beta_R + B_R)$ and the corresponding value $\tau = 1$, ρ_e was not

Table 3

Results after 10 iterations for a = $(1+4|x-\frac{1}{2}|)^2$

h	K	τ	β_R, B_R	ρ_e	Max Error
1/16	0	0.270	0,107	0.73	3.3(-2)
1/32		0.145	0,234	0.86	1.8(-1)
1/64		0.075	0,489	0.93	4.3(-1)
1/16	53.5	1	as above	0.60	2.2(-3)
1/32	117			0.78	2.7 (-2)
1/64	244			0.88	8.6(-2)
1/16	13	1	as above	0.24	1.2(-7)

a great deal less.

When this same problem was solved numerically by the iterative scheme of § 3.2, a much more satisfactory situation resulted. The spectral radius was observed to be essentially $1/9$, independent of h, in agreement with the discussion there.

4.5 All the above experiments were carried out using a subroutine, written by Buzbee, which is based on Buneman's algorithm for odd-even reduction [4]. This subroutine solves the Helmholtz equation on a rectangle, and it includes the boundary-condition options required for our examples, The numerical solution of a problem on a 64x64 mesh requires approximately 0.06 seconds on the CDC 7600 computer.

Qualitative comparison of the computational requirements of our technique with those of other methods can be made using the operation-count table given in [3]. For example, odd-even reduction requires $(9/2)N^2 \log_2 N$ operations to carry out the direct solution of a problem on an $N \times N$ mesh. Setting up the righthand side of (13) requires another $2N^2$ operations per iteration, and, if Chebyshev acceleration (14) is used, another $3N^2$ operations are needed. Thus, according to the table, the operations required for one iteration of (13) are equivalent to those required for about 4 or $4\frac{1}{2}$ SOR iterations or $1\frac{1}{2}$ AD1 iterations if $N = 64$. The reduction of the initial error by a factor $N^{-2} \approx 2.5 \times 10^{-4}$ in the numerical solution of the Poisson equation is listed as requiring about 85 SOR or 7 ADI iterations for this N, when optimal parameters are used; the solution of (3) or (6) will generally require more. Further such comparisons can be made using the table.

The memory requirements of (13, 14) exceed those of SOR by about $3N^2$

locations if both P-KI and $\bar{W}^{(n-1)}$ are stored. This value can be reduced to N^2, however, in exchange for recomputing P-KI at each iteration and using a form of Chebyshev acceleration that requires, instead of $\bar{W}^{(n-1)}$, a sequence of parameters $\{\tau_n\}$.

We conclude from our numerical experiments that for well-suited cases, such as those in § 4.2, our basic technique is an extremely efficient one and compares very favorably with standard iterative and elimination methods. Its advantages are especially striking for problems with a large number of mesh points, since the number of iterations required is independent of h. For less well-suited and poorly suited problems, such as those in § 4.3 - 4.4, the scheme may be very satisfactory in some cases, but further study would be useful to clarify the best means for estimating the parameters and for utilizing the technique of § 3.2 in discontinuous cases.

5. Scaling. We presented the scaling (5, 6), or equivalently (32, 33), tacitly implying its suitability because the resulting operator resembles the one on which the iteration is based. We now make some remarks on the question of whether or not this scaling is in some sense the best one possible.

Since the optimal spectral radius for iteration (24) increases with the condition number ν_M/ν_m, the best sclaing is one that yields the minimum condition number for a given problem. In the discrete case, the choice (33), among all positive diagonal scalings (32), minimizes the condition number of M_h, which has Property A, but not necessarily that of $(-Ah + KI)^{-1}M_h$. The optimal diagonal scaling for these more general matrices is not known; a related discussion pertaining to scaling of alternating direction methods can be found in [16].

We have carried out calculations on some one-dimensional problems corresponding to (3, 4) to determine numerically the scaling necessary to minimize the condition number. We considered the standard three-point discretization on a uniform mesh $LU = F$ equivalent to (15) for the problem

$$\mathfrak{L}u \equiv -\frac{d}{dx}[a(x)\frac{du}{dx}] = f(x) , \qquad u(0) = u(1) = 0 .$$

The diagonal matrix D was calculated that minimized the condition number of the matrix $(-\Delta_h + KI)^{-1} D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$, where $-\Delta_h$ is the one-dimensional equivalent of (12). The minimization was carried out for several values of h using an algorithm of Osborne [17] and the minimization program of Fletcher [18]. The actual value of K that was used was, in general, not important, since the diagonal elements $2h^{-2}$ of $-\Delta_h$ were, by comparison, usually much larger. For the cases in which $\frac{d^2}{dx^2}(a^{\frac{1}{2}})/a^{\frac{1}{2}}$ was

constant, we found that the best D was essentially the same as the one we have used here, that is, proportional to the main diagonal of L. For the one-dimensional equivalents of the examples of § 4.2, in which $\frac{d^2}{dx^2}(a^{\frac{1}{2}})/a^{\frac{1}{2}}$ varied only moderately over the interval, we found that the best D departed from this value by only a comparatively slight amount. If $a(x,y)$ had a sizable discontinuity, however, then the best D was not as close to being proportional to the main diagonal of L; rather, it tended to smooth out the discontinuity (see also § 3.2). We concluded that for the problems with relatively smooth a, the ones for which our iterative technique has greatest potential, the scaling (5, 6) is adequate.

We remark that Gunn also observed notably improved convergence rates in certain cases when a variant of the scaling (5, 6) was used to solve (3, 4) by an iterative technique [19].

6. Extension to other problems. Our iteration procedure may be applied, as well, to problems other than (3, 4). One immediate extension is to the case in which the term $b(x,y)u$ is added to the left of (3), where $b(x,y)$ is such that $\mathcal{L} + b(x,y)$ remains positive definite. The transformation (5) still results in an equation of the form (6), to which the iterative procedure (7) applies directly.

Another extension is to the case in which on some of the edges of the rectangle there are specified periodic boundary conditions or boundary conditions of the form $\partial u/\partial n + au = b$ for which fast direct methods can be used. Then the value of λ_m changes, but otherwise the basic procedure is not altered so long as the boundary conditions remain suitable for these methods after the scaling (5,6) is performed.

We remark that the numerical solution of separable equations of the form

$$c(x)\frac{\partial}{\partial x}\left(a(x)\frac{\partial u}{\partial x}\right) + d(y)\frac{\partial}{\partial y}\left(b(y)\frac{\partial u}{\partial y}\right) + Ku = f(x,y) , \tag{38}$$

with suitable boundary conditions, can be carried out by fast methods with only the additional work of solving a tridiagonal eigenproblem, the dimension of which is the number of mesh points in a row. Thus it is not necessary to attempt to solve such problems iteratively, using the scaling and shifting procedures described here. Included in (38) is the radially symmetric Poisson equation on an annular region $0 < r_o \le r \le r_1$,

$$-\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial r}{\partial r}\right) - \frac{\partial^2 u}{\partial z^2} = f ,$$

which, after being multiplied by r, is also of the form (3).

Also included in (38) is the case in which the iteration (11) is discretized on a rectangular mesh with nonuniform spacing. Let $h_{\bar{i}} = x_i - x_{i-1}$ and $k_{\bar{j}} = y_j - y_{j-1}$ be the mesh spacings; then the resulting five-point discretization of $-\Delta + K$ at an interior point is

$$(-\tilde{\Delta}_h+K)W_{ij} = \frac{k_{\bar{j}} + k_{\bar{j}+1}}{2}\left[-\frac{1}{h_{\bar{i}}}W_{i-1,j} + \left(\frac{1}{h_{\bar{i}}} + \frac{1}{h_{\bar{i}+1}}\right)W_{ij} - \frac{1}{h_{\bar{i}+1}}W_{i+1,j}\right]$$

$$+ \frac{h_{\bar{i}} + h_{\bar{i}+1}}{2}\left[-\frac{1}{k_{\bar{j}}}W_{i,j-1} + \left(\frac{1}{k_{\bar{j}}} + \frac{1}{k_{\bar{j}+1}}\right)W_{ij} - \frac{1}{k_{j+1}}W_{i,j+1}\right]$$

$$+ \frac{K}{4}(h_{\bar{i}} + h_{\bar{i}+1})(k_{\bar{j}} + k_{\bar{j}+1})W_{ij}.$$

After multiplying each equation through by $4(h_{\bar{i}} + h_{\bar{i}+1})^{-1}(k_{\bar{j}} + k_{\bar{j}+1})^{-1}$, or by performing the transformation that preserves symmetry [20],

$$D^{-\frac{1}{2}}(-\tilde{\Delta}_h + K)D^{-\frac{1}{2}}, \quad D = \operatorname{diag}\left\{\frac{(h_{\bar{i}} + h_{\bar{i}+1})(k_{\bar{j}} + k_{\bar{j}+1})}{4}\right\},$$

one obtains separable equations that can be treated by the direct methods suitable for (38). It also would be possible, of course, to apply the techniques of §3 to the original problem (3) discretized on the nonuniform mesh.

Finally, we remark that if the domain on which the equation is to be solved is not itself a rectangle, but is, instead, a union of rectangles, then our iterative technique might be combined efficiently with the fast methods suitable for such domains [21]. These might then, in turn, also be combined with iteration (i, ii) of § 3.2 for the case in which a(x,y) is piecewise smooth over such subdomains.

We plan to study these extensions in the future and to consider, as well, application of the iterative technique to nonlinear equations.

Acknowledgements. We wish to express our appreciation to B. L. Buzbee, F. W. Dorr, J. A. George, and R. W. Hockney for making available to us their collection of computer programs for solving the Poisson equation by fast direct methods and to M. R. Osborne for making available the minimization program described in § 5. We thank also Margaret Wright for programming the scaling experiments discussed in that section. Part of the work reported here was done in England during the first author's tenure as a Senior Visiting Fellow at Reading University under a Science Research Council grant and as a Visiting Research Fellow at the National Physical Laboratory in Teddington, during 1970-71. The work was supported in part by the U. S. Atomic Energy Commission.

Footnotes

(1) The boundary data for the operator $(-\Delta+K)$ need not necessarily be the same as for $\mathcal{M}$. Other boundary data, such as 0 , may be computationally more convenient for some problems.

(2) We note also, without comment on possible relevancy, that the underlying iteration operator (2) becomes completely continuous when $\tau = 1$.

References

[1] O. Buneman, "A compact non-iterative Poisson solver," Report 294, Stanford University Institute for Plasma Research, Stanford, California, 1969.

[2] R. W. Hockney, "The Potential Calculation and Some Applications," Methods in Computational Physics, vol. 9, B. Adler, S. Fernbach, and M. Rotenberg, eds., Academic Press, New York and London, 1969, pp. 136-211.

[3] F. W. Dorr, "The direct solution of the discrete Poisson equation on a rectangle," SIAM Rev., 12 (1970), pp. 248-263.

[4] B. L. Buzbee, G. H. Golub, and C. W. Nielson, "On direct methods for solving Poisson's equations," SIAM J. Numer. Anal., 7 (1970), pp. 627-656.

[5] E. G. D'Yakonov, "On an iterative method for the solution of finite difference equations," Dokl. Akad. Nauk SSSR, 138 (1961), pp. 522-525.

[6] O. B. Widlund, "On the use of fast methods for separable finite difference equations for the solution of general elliptic problems," Sparse Matrices and Applications, D. J. Rose and R. A. Willoughby, eds., Plenum Press, New York, 1972, pp. 121-134.

[7] J. E. Gunn, "The numerical solution of $\nabla \cdot a \nabla u = f$ by a semi-explicit alternating direction iterative method," Numer. Math., 6 (1964), pp. 181-184.

[8] H. L. Stone, "Iterative solution of implicit approximations of multi-dimensional partial differential equations," SIAM J. Numer. Anal., 5 (1968), pp. 530-558.

[9] T. Dupont, R. P. Kendall, and H. H. Rachford, Jr., '*An approximate factorization procedure for solving self-adjoint elliptic difference equations," SIAM J. Numer. Anal., 5 (1968), pp. 559-573.

[10] L. F. Shampine, "Boundary value problems for ordinary differential equations," SIAM J. Numer. Anal., 5 (1968), pp. 219-242.

[11] V. I. Lebedev and S. A. Finogenov, "On the order of choice of the iteration parameters in the Chebyshev cyclic iteration method," Zhur. Vych. Mat. i Mat. Fiz., 11 (1971), pp. 425-438. English translation in Report , Computer Science Department, Stanford University, Stanford, California, 1972.

[12] T. Dupont, "A factorization procedure for the solution of elliptic difference equations," SIAM J. Numer. Anal. 5 (1968), pp. 753-782.

[13] V. I. Lebedev, "Iterative methods for the solution of operator equations with their spectrum lying on several intervals," Zhur. Vych. Mat. i Mat. Fiz., 9 (6) (1969), pp. 1247-1252. English translation in Report , Computer Science Department, Stanford University, Stanford, California, 1972.

[14] R. S. Varga, Matrix Iterative Analysis, Prentice-Hall, Englewood Cliffs, New Jersey, 1962, p. 141, prob. 8.

[15] M. A. Diamond, 'An economical algorithm for the solution of finite difference equations," Report R-71-492, Department of Computer Science, University of Illinois, Urbana, Illinois, 1971.

[16] O. B. Widlund, "On the effects of scaling of the Peaceman-Rachford method," Math. Comp., 25 (1971), pp. 33-41.

[17] M. R. Osborne, private communication.

[18] R. Fletcher, "A survey of algorithms for unconstrained optimization," Report TP456, Theoretical Physics Division, A.E.R.E. Harwell, Didcot, Berks, England, 1971.

[19] J. E. Gunn, "The solution of elliptic difference equations by semi-explicit iterative techniques, SIAM J. Numer. Anal. 1 (1965), pp. 24-45.

[20] E. L. Wachspress, Iterative Solution of Elliptic Systems, Prentice-Hall, Englewood Cliffs, New Jersey, 1966, § 6.4.

[21] B. L. Buzbee, F. W. Dorr, J. A. George, and G. H. Golub, "The direct solution of the discrete Poisson equation on irregular regions," SIAM J. Numer. Anal., 8 (1971), pp. 722-736.

[22] E. L. Stiefel, "Über einige Methoden der Relaxationsrechnung," Z. angew. Math. Phys., 3 (1952), pp. 1-339