Abstract

We consider the use of fast direct methods as preconditioners for iterative methods for computing the numerical solution of non-self-adjoint elliptic boundary value problems. We derive bounds on convergence rates that are independent of discretization mesh size. For two-dimensional problems on rectangular domains, discretized on an nxn grid, these bounds lead to asymptotic operation counts of $O(n^2 \log n \log \epsilon^{-1})$ to achieve relative error ϵ and $O(n^2(\log n)^2)$ to reach truncation error.

Preconditioning by Fast Direct Methods for Non-Self-Adjoint Nonseparable Elliptic Equations

Howard C. Elman and Martin H. Schultz

Technical Report YALEU/DCS/TR-287

December 1983

This work was supported by the U. S. Office of Naval Research under grant N00014-82-K-0184.

Abstract

We consider the use of fast direct methods as preconditioners for iterative methods for computing the numerical solution of non-self-adjoint elliptic boundary value problems. We derive bounds on convergence rates that are independent of discretization mesh size. For two-dimensional problems on rectangular domains, discretized on an nxn grid, these bounds lead to asymptotic operation counts of $O(n^2 \log n \log \epsilon^{-1})$ to achieve relative error ϵ and $O(n^2(\log n)^2)$ to reach truncation error.

Preconditioning by Fast Direct Methods for Non-Self-Adjoint Nonseparable Elliptic Equations

Howard C. Elman and Martin H. Schultz
Technical Report YALEU/DCS/TR-287

December 1983

This work was supported by the U. S. Office of Naval Research under grant N00014-82-K-0184.

1. Introduction

The numerical solution of elliptic boundary value problems by finite difference methods requires the solution of systems of linear equations of the form

$$Au = f, \tag{1}$$

where A is a large sparse matrix. Fast direct methods are efficient techniques for solving (1) when the elliptic equation is posed on a rectangular domain and the differential operator is separable. In this paper, we show that nonseparable, non-self-adjoint elliptic problems can be solved efficiently using fast direct methods as preconditioners for iterative methods for nonsymmetric linear systems.

As a prototype, consider the two-dimensional Dirichlet problem

$$\begin{aligned} \boldsymbol{A}u &= f, \quad u \in \Omega, \\ u &= g, \quad u \in \partial\Omega, \end{aligned} \tag{2}$$

where Ω is a rectangular region in $\mathbf{R}^2$,

$$\boldsymbol{A}u \equiv -(au_x)_x - (bu_y)_y + cu_x + (cu)_x + du_y + (du)_y + eu, \tag{3}$$

and a(x,y), b(x,y), c(x,y), d(x,y), e(x,y), and f(x,y) are smooth functions defined on Ω, with a, b $>$ 0, e $\geq$ 0. Discretizing (2) by finite difference techniques on an nxn grid leads to a sparse system of linear equations of the form (1) where A is of order $N = n^2$. If c and d are zero, then $\boldsymbol{A}$ is self-adjoint; otherwise, $\boldsymbol{A}$ is non-self-adjoint and, in general, A is nonsymmetric. If

$$a = a(x),\ b = b(y),\ c = c(x),\ d = d(y),\ e = e_1(x) + e_2(y), \tag{4}$$

then $\boldsymbol{A}$ is separable. In the self-adjoint separable case, (1) can be solved by a variety of fast direct methods, such as the cyclic reduction and Fourier methods (surveyed in [7, 18]) and the generalized marching algorithm [2, 3]. In the non-self-adjoint case, the cyclic reduction method can still be used [17]. All of these methods require $O(n^2 \log n)$ arithmetic operations (i.e. floating point multiplications and divisions).[1]

If $\boldsymbol{A}$ is nonseparable, then fast direct methods can be used to solve (1) iteratively. For self-adjoint problems, Widlund [23] proposed the stationary method

$$u_{i+1} = u_i + \tau Q^{-1}(f - Au_i), \tag{5}$$

[1] In the self-adjoint case, this count can be reduced to $O(n^2 \log(\log n))$. See [18].

where Q is the discretization of a self-adjoint separable approximation of $\boldsymbol{A}$. Concus and Golub [5] and Bank [2] examined accelerating (5) by the Chebyshev and conjugate gradient methods respectively; equivalently, Q is a preconditioning matrix for these iterative methods. Convergence of all these methods is independent of the mesh size used in the discretization, so that a relative error of ϵ is achieved in $O(\log \epsilon^{-1})$ iterations. Since the dominant expense of each iteration is a fast direct solve, the asymptotic operation count is $O(n^2 \log n \log \epsilon^{-1})$.

Concus and Golub [4] and Widlund [22] extended this analysis to some particular non-self-adjoint problems using the generalized conjugate gradient method, which depends on the symmetric part of A as a preconditioner. Let

$$\boldsymbol{Mu} \equiv -(a\boldsymbol{u}_x)_x - (b\boldsymbol{u}_y)_y + e\boldsymbol{u}, \tag{6}$$

$$\boldsymbol{Ru} \equiv c\boldsymbol{u}_x + (c\boldsymbol{u})_x + d\boldsymbol{u}_y + (d\boldsymbol{u})_y,$$

so that $\boldsymbol{A} = \boldsymbol{M} + \boldsymbol{R}$. The symmetric part of A corresponds to the discretization of $\boldsymbol{M}$, the sum of the second and zero order terms of $\boldsymbol{A}$. The convergence of the generalized conjugate gradient method depends essentially on the spectral radius of the discrete analogue of the compact operator $\boldsymbol{M}^{-1}\boldsymbol{R}$, so that convergence is independent of mesh size. If $\boldsymbol{M}$ is separable, i.e. a, b, and e are as in (4), then fast direct methods can be used for the preconditioning, so that the asymptotic operation count is again $O(n^2 \log n \log \epsilon^{-1})$.

In this paper we show that these asymptotic operation counts can be achieved even if the symmetric part of A does not come from a separable operator. Other iterative methods for nonsymmetric linear systems, such as the conjugate gradient method applied to the normal equations [13] and Orthomin(k) [8, 9], allow more general choices than the symmetric part for preconditioning matrices. Using an analysis of the finite difference discretization, we show that with symmetric positive-definite preconditioning matrices derived from other self-adjoint, separable operators $\boldsymbol{Q}$ that approximate $\boldsymbol{A}$, the asymptotic convergence rates of these iterative methods is independent of mesh size. Although this analysis holds only for symmetric preconditioning operators, we provide numerical examples that suggest that nonsymmetric preconditioners derived from non-self-adjoint separable approximations of $\boldsymbol{A}$ lead to the same asymptotic convergence properties. In Section 2 we review the basic properties and convergence bounds of the conjugate gradient method applied to the normal equations and Orthomin(k). In Section 3, we present the convergence analysis for general self-adjoint, separable preconditioning operators, and in Section 4, we demonstrate the performance of these techniques with both self-adjoint

and non-self-adjoint preconditioning operators.

2. Convergence Bounds for Iterative Methods

Given a nonsymmetric linear system of the form (1), let $A = M + R$, where $M = (A+A^T)/2$ is the symmetric part of A and $R = (A-A^T)/2$ is the skew-symmetric part of A. In this section, we present upper bounds on the convergence rates of iterative methods for nonsymmetric linear systems in terms of eigenvalues of M and R. We consider two iterative methods: the conjugate gradient method applied to the normal equations and, as representative of a recently developed collection of methods that avoid the use of the normal equations, Orthomin(k).

We first establish some conventions of notation. For any square matrix B, let $\kappa(B) \equiv \|B\|_2\|B^{-1}\|_2$ denote the condition number of B, let $\lambda_{min}(B)$ and $\lambda_{max}(B)$ denote the eigenvalues of B with smallest and largest modulus, respectively, and let $\rho(B)$ denote the spectral radius $|\lambda_{max}(B)|$. For symmetric positive-definite B, let $\|v\|_B$ denote the B-norm $(v,Bv)^{1/2}$.

A classical method for solving nonsymmetric linear systems of the form (1) is to apply the conjugate gradient method (CG) [13] to the normal equations

$$A^TAu = A^Tf.$$

That is, the nonsymmetric system is embedded into a system with a symmetric positive-definite coefficient matrix, to which the conjugate gradient method is applicable. We denote this method by CGN.[2] The convergence properties of CGN, derived from the standard error analysis of CG [6], are well understood. We summarize the results we need as follows. Let $\{u_i\}$ denote the sequence of iterates generated by CGN starting from an initial guess u_0, and let $\{r_i\}$ denote the residuals $\{f-Au_i\}$. The iterate u_i is the point in

$$u_0 + \text{span}\ \{A^Tr_0,(A^TA)A^Tr_0,\ldots,(A^TA)^{i-1}A^Tr_0\}$$

whose residual norm $\|r_i\|_2$ is minimum. As a result, the residuals satisfy [6]

$$\|r_i\|_2 \leq 2\left[\frac{1 - 1/\kappa(A)}{1 + 1/\kappa(A)}\right]^i \|r_0\|_2. \tag{7}$$

The following result relates this bound to the extreme eigenvalues of M and the spectral radius of R.

[2]A related technique is to solve $AA^Tu'=f$ by CG, with $u=A^Tu'$. This method has essentially the same convergence properties as CGN, see [9].

Theorem 1: If the symmetric part M of A is positive-definite, then

$$\lambda_{min}(A^TA) \geq \lambda_{min}(M)^2, \tag{8}$$

$$\lambda_{max}(A^TA) \leq [\lambda_{max}(M) + \rho(R)]^2. \tag{9}$$

Hence, the residuals $\{r_i\}$ generated by CGN satisfy

$$\|r_i\|_2 \leq 2\left[1 - \frac{2}{[\lambda_{max}(M)+\rho(R)]/\lambda_{min}(M) + 1}\right]^i \|r_0\|_2. \tag{10}$$

Proof: Let S denote the unique symmetric positive-definite square root of M, i.e. $S^2 = M$. Then

$$(v,A^TAv) = (Av,Av) = (S(S+S^{-1}R)v,S(S+S^{-1}R)v)$$
$$= ((S+S^{-1}R)v,M(S+S^{-1}R)v).$$

But for any real w,

$$(w,Mw) \geq \lambda_{min}(M)(w,w)$$

and

$$(w,Rw) = 0,$$

so that

$$(v,A^TAv) \geq \lambda_{min}(M)\,((S+S^{-1}R)v,(S+S^{-1}R)v)$$
$$= \lambda_{min}(M)\,[(Sv,Sv) + 2(v,Rv) + (S^{-1}Rv,S^{-1}Rv)]$$
$$= \lambda_{min}(M)\,[(v,Mv) + (S^{-1}Rv,S^{-1}Rv)]$$
$$\geq \lambda_{min}(M)^2\,(v,v)$$

Therefore,

$$\lambda_{min}(A^TA) = \min_{v\neq 0} \frac{(v,A^TAv)}{(v,v)} \geq \lambda_{min}(M)^2.$$

For the upper bound on $\lambda_{max}(A^TA)$,

$$\lambda_{max}(A^TA) = \|A\|_2^2 \leq [\|M\|_2 + \|R\|_2]^2 = [\lambda_{max}(M) + \rho(R)]^2,$$

where we have used the fact that $\|R\|_2 = \rho$ since R is skew-symmetric and hence normal [20].

Finally, note that the fraction in (7) satisfies

$$\frac{1 - 1/\kappa(A)}{1 + 1/\kappa(A)} = 1 - \frac{2}{\kappa(A)+1} .$$

Inequality (10) then follows from (7) - (9) and the fact that

$$\kappa(A) = \sqrt{\kappa(A^TA)} = \sqrt{\frac{\lambda_{max}(A^TA)}{\lambda_{min}(A^TA)}} \leq \frac{\lambda_{max}(M)+\rho(R)}{\lambda_{min}(M)} .$$

Q.E.D.

The bound (7) implies that CGN is convergent for arbitrary nonsingular linear systems, and as we will show in Theorem 5, the corollary result (10) gives rise to strong asymptotic bounds under suitable conditions. However, the dependence of CGN on A^TA is a drawback. If A itself were symmetric positive-definite, then CG could be applied directly to (1). In this case, the bound on the relative error of the i'th iterate generated by CG is proportional to

$$\left[\frac{1 - 1/\sqrt{\kappa(A)}}{1 + 1/\sqrt{\kappa(A)}}\right]^i .$$

(see [6]), and the convergence of CG would be much faster than the convergence of CGN. Moreover, to avoid the explicit computation of A^TA, CGN requires two matrix-vector products per iteration, one by A and one by A^T. In contrast, CG requires only one matrix-vector product per iteration for symmetric positive-definite systems.

In recent years, many conjugate gradient-like iterative methods for nonsymmetric systems have been developed with the aim of avoiding these difficulties of CGN [1, 8, 9, 16, 21, 25]. In all of these methods, the i'th iterate u_i is chosen from the Krylov space $u_0 + K_i$, where

$$K_i \equiv \text{span}\{r_0, Ar_0, \ldots, A^{i-1}r_0\},$$

with the hope that the convergence depends on $\kappa(A)$ rather than $\kappa(A^TA)$. Each iteration requires one matrix-vector product of the form Av. In this paper, we will take the method known as Orthomin(k) to be representative of this class of iterative methods. Orthomin(k) chooses for u_i a point in $u_0 + K_i$ whose residual norm $\|r_i\|_2$ is minimum in a (k+1)-dimensional subspace of K_i (see [8]). Although its convergence properties are not as well understood as those of CG, the following result shows that when M is positive-definite, Orthomin(k) generates a sequence of approximate solutions that converge to $A^{-1}f$.

Theorem 2: The residuals $\{r_i\}$ generated by Orthomin(k) satisfy [8, 9]

$$\|r_i\|_2 \leq \left[1 - \frac{\lambda_{min}(M)}{\lambda_{max}(M) + \rho(R)^2/\lambda_{min}(M)}\right]^{i/2} \|r_0\|_2 . \qquad (11)$$

The work per step of the two methods considered in this section is [8, 9]

CGN: 5N operations plus two matrix-vector products, Av and $A^T v$;

Orthomin(k): (3k+4)N operations plus one matrix-vector product, Av.

3. Convergence Analysis for Separable Preconditioners

In this section, we show that if CGN or Orthomin(k) is combined with preconditioners based on certain separable operators to solve discretized elliptic problems, then convergence does not depend on the mesh size used in the discretization.

Let $\boldsymbol{A}$, $\boldsymbol{f}$ and $\boldsymbol{g}$ be defined as in (2), let Ω denote the unit square $0 \leq x,y \leq 1$, and let $\boldsymbol{A} = \boldsymbol{M} + \boldsymbol{R}$, as in (6). Discretizing (2) by centered differences on an nxn grid leads to a system of linear equations of the form (1). Let $h \equiv 1/(n+1)$. In terms of the contributions of the symmetric and skew-symmetric parts, the difference equations at a typical mesh point (x_i, y_j) have the form (after scaling by h^2)

$$[Au]_{ij} = [Mu]_{ij} + [Ru]_{ij} = h^2 f_{ij} ,$$

where

$$[Mu]_{ij} = [a_{i+1/2,j} + a_{i-1/2,j} + b_{i,j+1/2} + b_{i,j+1/2} + h^2 e_{ij}]u_{ij} \tag{12}$$

$$- a_{i+1/2,j}u_{i+1,j} - a_{i-1/2,j}u_{i-1,j} - b_{i,j+1/2}u_{i,j+1} - b_{i,j-1/2}u_{i,j-1},$$

$$[Ru]_{ij} = [(c_{i+1,j}+c_{ij})h/2]u_{i+1,j} - [(c_{ij}+c_{i-1,j})h/2]u_{i-1,j} + [(d_{i,j+1}+d_{ij})h/2]u_{i,j+1} \tag{13}$$

$$- [(d_{ij}+d_{i,j-1})h/2]u_{i,j-1}.$$

Hence, M corresponds to the discretization of $\boldsymbol{M}$, and R corresponds to the discretization of $\boldsymbol{R}$.

Let $\boldsymbol{Q}$ denote a self-adjoint elliptic operator on Ω which is spectrally equivalent to $\boldsymbol{M}$, i.e.

$$\alpha_1 \leq \frac{(\boldsymbol{v},\boldsymbol{Mv})}{(\boldsymbol{v},\boldsymbol{Qv})} \leq \alpha_2, \tag{14}$$

for all sufficiently smooth $\boldsymbol{v} \in L_2(\Omega)$ satisfying homogeneous Dirichlet boundary conditions, where

$$(\boldsymbol{v},\boldsymbol{w}) \equiv \int_{\Omega} \boldsymbol{v}\overline{\boldsymbol{w}} ,$$

and α_1 and α_2 are positive constants. For example, $\boldsymbol{Q}$ could be the negative Laplacian $-\left[\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right]$. The discretization of $\boldsymbol{Q}$ gives rise to a symmetric positive-definite matrix, Q, that is spectrally equivalent to M: α_1 and α_2 in (14) can be chosen so that

$$\alpha_1 \leq \frac{(v,Mv)}{(v,Qv)} \leq \alpha_2, \tag{15}$$

independent of the mesh size used in generating M and Q.

Since Q is symmetric positive-definite, it admits a factorization $Q = LL^T$ (here L is not necessarily lower triangular). Consider the *symmetrically* preconditioned linear system

$$\tilde{A}\tilde{u} = [L^{-1}AL^{-T}]\,[L^T u] = L^{-1}f = \tilde{f}, \tag{16}$$

in which the coefficient matrix $\tilde{A}$ has symmetric part $\tilde{M} = L^{-1}ML^{-T}$ and skew-symmetric part $\tilde{R} = L^{-1}RL^{-T}$. The extreme eigenvalues of $\tilde{M}$ are given by

$$\lambda_{min}(\tilde{M}) = \min_{v\neq 0} \frac{(v,L^{-1}ML^{-T}v)}{(v,v)}, \qquad \lambda_{max}(\tilde{M}) = \max_{v\neq 0} \frac{(v,L^{-1}ML^{-T}v)}{(v,v)}.$$

But for any $v\neq 0$, let $w = L^{-T}v$, so that

$$\frac{(v,L^{-1}ML^{-T}v)}{(v,v)} = \frac{(w,Mw)}{(w,Qw)}. \tag{17}$$

Hence, by (15),

$$\lambda_{min}(\tilde{M}) \geq \alpha_1, \qquad \lambda_{max}(\tilde{M}) \leq \alpha_2, \tag{18}$$

independent of mesh size.

Observation (18) is the basis of the convergence results for the self-adjoint case [5]. To extend the analysis to the non-self-adjoint case, it is necessary to bound $\rho(\tilde{R})$. An intuitive approach is to note that $\tilde{R}$ is similar to $Q^{-1}R$, which is the discrete analogue of the continuous operator $\boldsymbol{T} \equiv \boldsymbol{Q}^{-1}\boldsymbol{R}$. For $\boldsymbol{Q}$ with sufficiently smooth coefficients, $\boldsymbol{T}$ is a compact operator whose eigenvalues are therefore bounded with 0 as their sole accumulation point [22, 24]; one would expect a similar bound on the eigenvalues of the discrete operator. We derive such a bound below. For simplicity, we present the analysis for the one-dimensional Dirichlet problem on the unit interval $\Omega \equiv [0,1]$, divided into n uniformly spaced interior mesh points, with h=1/(n+1). The analysis for the two-dimensional Dirichlet problem on a rectangular domain is identical, using the two-dimensional versions of the operators in (21) below.

The one-dimensional analogue of the first order operator $\boldsymbol{R}$ in (6) is

$$\boldsymbol{Ru} \equiv c\boldsymbol{u}_x + (c\boldsymbol{u})_x,$$

and the discretization analogous to that of (13) is

$$[Ru]_i = [(c_{i+1}+c_i)h/2]u_{i+1} - [(c_i+c_{i-1})h/2]u_{i-1} = \frac{h}{2}[c_i(u_{i+1}-u_{i-1}) + c_{i+1}u_{i+1} - c_{i-1}u_{i-1}], \tag{19}$$

$$1 \leq i \leq n.$$

Without loss of generality, we take Q to be the one-dimensional Laplacian, $Qu \equiv -u_{xx}$. (The argument can be modified easily for any operator spectrally equivalent to this choice of Q.) The discretization analogous to that of (12) is

$$[Qu]_i \equiv -u_{i+1} + 2u_i - u_{i-1}, \qquad 1 \le i \le n. \tag{20}$$

In (19) and (20), u_i, $[Ru]_i$ and $[Qu]_i$ are defined to be zero for i=0 and i=n+1.

Let the usual finite difference operators be given by

$$\begin{aligned} [D_+u]_i &\equiv \frac{u_{i+1}-u_i}{h}, \quad 0 \le i \le n, \qquad [D_+u]_{n+1} \equiv 0, \\ [D_-u]_i &\equiv \frac{u_i-u_{i-1}}{h}, \quad 1 \le i \le n+1, \qquad [D_-u]_0 \equiv 0. \end{aligned} \tag{21}$$

Note that $[D_+u]_0$ and $[D_-u]_{n+1}$ may be nonzero even if $u_0 = u_{n+1} = 0$.

To avoid repeatedly handling the 0'th and (n+1)'st indices as special cases, we adopt the following convention of notation. We identify C^n with the proper subspace of C^{n+2} consisting of complex (n+2)-vectors $v = (v_0,v_1,\ldots,v_n,v_{n+1})^T$ such that $v_0 = v_{n+1} = 0$. Thus, R and Q of (19) and (20) are defined on this representation of C^n without reference to indices 0 and n+1. The usual Euclidian inner product and norm on C^n are inherited by this space:

$$(v,w) \equiv \sum_{i=1}^{n} v_i\bar{w}_i, \qquad \|v\|_2 \equiv (v,v)^{1/2}. \tag{22}$$

We will also abuse notation slightly and define (v,w) and $\|v\|_2$ as in (22) for arbitrary vectors in C^{n+2}, noting that these functions *do not* define an inner product or norm on all of C^{n+2}. We refer to the true Euclidian inner product on C^{n+2} as the "extended" inner product

$$(u,v)_e \equiv \sum_{i=0}^{n+1} v_i\bar{w}_i.$$

The following result summarizes the properties of the finite difference operators that we need. Other results of this type can be found in [14].

Lemma 3:

(i) For v and w $\in C^{n+2}$, $(v,D_+w)_e = -(D_-v,w)_e + \frac{1}{h}[v_{n+1}\bar{w}_{n+1} - v_0\bar{w}_0]$;

For v and w $\in C^n$,

(ii) $(v,D_+w) = (v,D_+w)_e = -(D_-v,w)_e = -(D_-v,w)$;

(iii) $Qv = -h^2D_-D_+v = -h^2D_+D_-v$;

(iv) $(v,Qv) \ge h^2\|D_+v\|_2^2$, $\qquad (v,Qv) \ge h^2\|D_-v\|_2^2$.

Proof: For (i),

$$(v,D_+w)_e = \sum_{i=0}^{n} v_i \frac{\bar{w}_{i+1}-\bar{w}_i}{h} = \frac{1}{h}\left[\sum_{i=0}^{n} v_i\bar{w}_{i+1} - \sum_{i=0}^{n} v_i\bar{w}_i\right]$$

$$= \frac{1}{h}\left[\sum_{i=1}^{n+1} v_{i-1}\bar{w}_i - \sum_{i=1}^{n+1} v_i\bar{w}_i\right] + \frac{1}{h}\left[v_{n+1}\bar{w}_{n+1} - v_0\bar{w}_0\right]$$

$$= -(D_-v,w)_e + \frac{1}{h}\left[v_{n+1}\bar{w}_{n+1} - v_0\bar{w}_0\right]$$

Assertion (ii) follows immediately from (i) and the zero boundary conditions.

For the first equality of (iii),

$$[D_-D_+v]_i = \frac{[D_+v]_i - [D_+v]_{i-1}}{h} = \frac{1}{h}\left[\frac{v_{i+1}-v_i}{h} - \frac{v_i-v_{i-1}}{h}\right] = -\frac{1}{h^2}[Qv]_i .$$

The proof of the second equality is identical.

For the first assertion of (iv),

$$\begin{aligned}
(v,Qv) &= (v,Qv)_e && \text{by the first equality of (ii)}\\
&= -h^2(v,D_-D_+v)_e && \text{by (iii)}\\
&= h^2(D_+v,D_+v)_e && \text{by the second equality of (ii)}\\
&\geq h^2\|D_+v\|_2^2 .
\end{aligned}$$

The proof of the second assertion of (iv) is identical.

Q.E.D.

Let $\tilde{R} = L^{-1}RL^{-T}$, where $Q = LL^T$ as above.

Theorem 4: There exists a constant $\beta \geq 0$, independent of the mesh size h, such that

$$\rho(Q^{-1}R) = \rho(\tilde{R}) \leq \beta. \tag{23}$$

Proof: Since $\tilde{R}$ is skew-symmetric and therefore normal, its spectral radius is given by the maximum value of the Rayleigh quotient:

$$\rho(\tilde{R}) = \max_{v\in C^n, v\neq 0} \frac{|(v,L^{-1}RL^{-T}v)|}{|(v,v)|} = \max_{w\in C^n, w\neq 0} \frac{|(w,Rw)|}{|(w,Qw)|} .$$

Writing Rw in terms of the finite difference operators of (21),

$$Rw = \frac{h^2}{2}[c\cdot(D_++D_-)w + (D_++D_-)(c\cdot w)] ,$$

where $c\cdot w$ denotes the vector with entries $\{c_iw_i\}_{i=0}^{n+1}$. Therefore

$$(w,Rw) = \frac{h^2}{2}\,[\,(w,c\cdot(D_+ + D_-)w) \;+\; (w,(D_+ + D_-)(c\cdot w))\,]$$

$$= \frac{h^2}{2}\,[\,(c\cdot w,(D_+ + D_-)w) \;-\; ((D_+ + D_-)w,(c\cdot w))\,],$$

by Lemma 3, (ii). Applying Cauchy-Schwarz and letting $C \equiv \max_{x\in\Omega} |c(x)|$,

$$|(w,Rw)| \le h^2\,\|c\cdot w\|_2\,\|(D_+ + D_-)w\|_2 \;\le\; C\,h^2\,\|w\|_2\,(\|D_+ w\|_2 + \|D_- w\|_2)$$

$$\le 2Ch\,\|w\|_2\,(w,Qw)^{1/2},$$

by Lemma 3, (iv).

Thus

$$\frac{|(w,Rw)|}{|(w,Qw)|} \le 2C\,\frac{h\|w\|_2}{(w,Qw)^{1/2}} \;\le \beta,$$

with β independent of h, since

$$\frac{(w,Qw)}{(w,w)} \ge \lambda_{\min}(Q) = O(h^2)\ [10].$$

Q.E.D.

C.f. [15] for an analysis of the convergence of the eigenvalues of $Q^{-1}R$ to those of $\boldsymbol{Q}^{-1}\boldsymbol{R}$.

We now return to the consideration of two-dimensional problems.

Note that applying either CGN or Orthomin(k) formally to (16) results in the residuals

$$\tilde{r}_i = \tilde{f} - \tilde{A}\tilde{u}_i = L^{-1}r_i,$$

so that the residual norm associated with (16) is $\|L^{-1}r_i\|_2 = \|r_i\|_{Q^{-1}}$. Hence, combining (15), (18) and (23) with the results of Theorems 1 and 2:

Theorem 5: If Q is a symmetric matrix that satisfies (15), (18) and (23), then the residuals generated by CGN with symmetric preconditioning by Q satisfy

$$\|r_i\|_{Q^{-1}} \le 2\left[\frac{\alpha_2-\alpha_1+\beta}{\alpha_2+\alpha_1+\beta}\right]^i \|r_0\|_{Q^{-1}}\,;$$

and the residuals generated by Orthomin(k) with symmetric preconditioning by Q satisfy

$$\|r_i\|_{Q^{-1}} \le \left[\,1 \;-\; \frac{\alpha_1^{\,2}}{\alpha_1\alpha_2+\beta^2}\,\right]^{i/2} \|r_0\|_{Q^{-1}}\,.$$

Hence, for discretized elliptic problems, the number of iterations needed to make $\|r_i\|_{Q^{-1}}/\|r_0\|_{Q^{-1}}$

$\leq \epsilon$ is independent of mesh size.

Proof: For the first inequality, by (18) and (23)

$$(\lambda_{max}(\widetilde{M})+\rho(\widetilde{R}))/\lambda_{min}(\widetilde{M}) \leq (\alpha_2+\beta)/\alpha_1.$$

Hence, applying Theorem 1,

$$\|r_i\|_{Q^{-1}} \leq 2\,[1 - \frac{2}{(\alpha_2+\beta)/\alpha_1 + 1}]^i \|r_0\|_{Q^{-1}} = 2\left[\frac{\alpha_2-\alpha_1+\beta}{\alpha_2+\alpha_1+\beta}\right]^i \|r_0\|_{Q^{-1}}.$$

The second inequality is proved in a similar manner using Theorem 2.

Q.E.D.

The extra work per step required for preconditioning comes in the matrix vector products $L^{-1}AL^{-T}v$ and $[L^{-1}AL^{-T}]^Tv$. Preconditioned CGN and Orthomin(k) can be implemented so that the only references to Q have the form of solving $Qw = v$, so that no factorization of Q is required (see [9]). The solve requires $O(n^2 \log n)$ operations, but the remaining work per step of both iterative methods is $O(N) = O(n^2)$. Hence, asymptotically, the preconditioning solves are the dominant cost per step.

Corollary 6: The number of arithmetic operations required by either CGN or Orthomin(k) with preconditioning by Q to solve (1) so that

$$\frac{\|r_i\|_{Q^{-1}}}{\|r_0\|_{Q^{-1}}} \leq \epsilon \tag{24}$$

is proportional to $n^2 \log n \log \epsilon^{-1}$.

Proof: By Theorem 5, the i'th residuals generated by both preconditioned methods satisfy

$$\frac{\|r_i\|_{Q^{-1}}}{\|r_0\|_{Q^{-1}}} \leq \eta\,\gamma^i,$$

where $\gamma<1$ is is a constant that is independent of mesh size, and $\eta=2$ for CGN, $\eta=1$ for Orthomin(k). Hence (ignoring the low order contribution of $\eta=2$), $i \geq \log(\epsilon^{-1})/\log(1/\gamma)$ iterations suffice to reduce the relative error of (24) to at most ϵ. Since the dominant cost of each iteration is the preconditioning solve, the total operation count is proportional to $n^2 \log n \log \epsilon^{-1}$.

Q.E.D.

The previous two results are expressed in terms of the Q^{-1}-norm of the residuals. It might be more practical to measure the relative Euclidean norm $\|r_i\|_2/\|r_0\|_2$. Alternatively, one might wish to reduce the error $s_i = u\text{-}u_i$ to within truncation error. Since the truncation error of the discretization is $O(n^{-2})$, it is sufficient to reduce the relative error $\|s_i\|_2/\|s_0\|_2$ by a factor of $O(n^{-2})$ to compute a solution u_i that is accurate to within truncation error.

Corollary 7: The number of arithmetic operations required by either CGN or Orthomin(k) with preconditioning by Q to solve (1) so that

$$\frac{\|r_i\|_2}{\|r_0\|_2} \leq \epsilon$$

is proportional to $(n^2 \log n)(\log n + \log \epsilon^{-1})$. The number of arithmetic operations required to compute a solution accurate to within truncation error is proportional to $n^2(\log n)^2$.

Proof: For the first assertion, note that

$$\|r_i\|_{Q^{-1}} \geq \frac{1}{\sqrt{\lambda_{max}(Q)}} \|r_i\|_2 , \qquad \|r_0\|_{Q^{-1}} \leq \frac{1}{\sqrt{\lambda_{min}(Q)}} \|r_0\|_2 ,$$

so that

$$\frac{\|r_i\|_2}{\|r_0\|_2} \leq \sqrt{\kappa(Q)} \frac{\|r_i\|_{Q^{-1}}}{\|r_0\|_{Q^{-1}}} . \tag{25}$$

Since $\kappa(Q) = O(n^2)$ [10], the left side of (25) is bounded by ϵ if

$$\frac{\|r_i\|_{Q^{-1}}}{\|r_0\|_{Q^{-1}}} \leq \epsilon_1 = O\left[\frac{\epsilon}{n}\right].$$

The result then follows from Corollary 6.

For the second assertion, note that the errors and residuals are related by $s_i = A^{-1}r_i$, $r_0 = As_0$. Hence,

$$\frac{\|s_i\|_2}{\|s_0\|_2} \leq \kappa(A) \frac{\|r_i\|_2}{\|r_0\|_2} \leq \kappa(A)\sqrt{\kappa(Q)} \frac{\|r_i\|_{Q^{-1}}}{\|r_0\|_{Q^{-1}}} .$$

To reduce the relative error by a factor of n^{-2}, it suffices to make

$$\frac{\|r_i\|_{Q^{-1}}}{\|r_0\|_{Q^{-1}}} \leq \epsilon_2 = O\left[\frac{n^{-2}}{\kappa(A)\sqrt{\kappa(Q)}}\right] . \tag{26}$$

But, by (8) and (9),

$$\kappa(A) \leq \kappa(M) + \frac{\rho(R)}{\lambda_{min}(M)}.$$

With M defined as in (12), $\lambda_{min}(M) = O(n^{-2})$ and $\lambda_{max}(M) = O(1)$, so that $\kappa(M) = O(n^2)$ [10]. Moreover, as defined in (13), all the nonzero entries of R lie in four off-diagonal bands, and their absolute values are bounded by Ch, where

$$C \equiv \max \left[\max_{(x,y)\in\Omega} |c(x,y)|, \max_{(x,y)\in\Omega} |d(x,y)| \right]$$

is independent of $h = 1/(n+1)$. By Gerschgorin's theorem [20], $\rho(R) = O(n^{-1})$. Therefore, $\rho(R)/\lambda_{min}(M) = O(n)$, so that $\kappa(A) = O(n^2)$. Substituting these results into (26),

$$\epsilon_2 = O(n^{-5}), \qquad \log \epsilon_2^{-1} = O(\log n).$$

The result again follows from Corollary 6.

Q.E.D.

The symmetric formulation of the preconditioned problem (16) requires that the preconditioning matrix Q be symmetric positive-definite. For non-self-adjoint problems, it seems preferable to allow Q to be nonsymmetric by including in it a discrete separable approximation to the first order terms in (2). Such preconditioning matrices can be used in either of the alternative preconditioned problems

$$AQ^{-1}\tilde{u} = f, \quad u = Q^{-1}\tilde{u}, \tag{27}$$

$$Q^{-1}Au = Q^{-1}f, \tag{28}$$

and the cyclic reduction method can be used for the preconditioning solves [17]. An additional advantage of (27) is that the residual norm minimized by the two iterative methods under consideration is

$$\|f - AQ^{-1}\tilde{u}_i\|_2 = \|f - Au_i\|_2 ,$$

i.e., it is independent of the preconditioning. However, we have been unable to extend the convergence analysis to these alternative problems (even for symmetric Q). In the analysis above, the application of Theorems 1 and 2 requires bounds on the extreme eigenvalues of the symmetric and skew-symmetric parts of $\widetilde{A}$ in (16). For (27), the symmetric part is given by $(AQ^{-1}+[AQ^{-1}]^T)/2$ and the skew-symmetric part by $(AQ^{-1}-[AQ^{-1}]^T)/2$. We have not been able to bound the eigenvalues of these matrices or those corresponding to (28) in a manner analogous to (18) and (23).

For an alternative approach to nonseparable M, see [11].

4. Numerical Experiments

In this section, we present numerical results that confirm the convergence analysis of Section 3 for (16) and suggest that similar behavior is exhibited for (27). All tests were run on a VAX11-780 in double precision (55 bit mantissa). The fast direct preconditioning solves were performed using the cyclic reduction method implemented in the routine BLKTRI in the FISHPACK subroutine package [19]. We consider two Dirichlet problems of the form (2) and one problem with mixed boundary conditions.

For the first two problems, let the coefficients of (3) be given by

$$a(x,y) = e^{-xy}, \qquad b(x,y) = e^{xy}, \qquad c(x,y) = 0, \tag{29}$$

$$d(x,y) = \gamma(x+y), \quad e(x,y) = \frac{1}{1+x+y},$$

where γ is a scalar parameter. The operator A is nonseparable and, for $\gamma \neq 0$, nonsymmetric. The right hand side is determined by choosing the solution

$$u(x,y) = x\, e^{xy} \sin(\pi x) \sin(\pi y).$$

We pose the problem on the unit square $0 \leq x,y \leq 1$ with homogeneous Dirichlet boundary conditions, and we discretize using the five-point second order centered finite difference scheme on a uniform nxn grid, with $h=1/(n+1)$. We use the values $\gamma=5$, $\gamma=50$ and $h = 1/16, 1/32, 1/64, 1/128$, and for one test, $1/256$.

We consider both self-adjoint and non-self-adjoint separable approximations Q, which give rise to symmetric and nonsymmetric preconditioning matrices, respectively. For the self-adjoint approximation, the coefficients of (29) are approximated by

$$\tilde{a}(x,y) = a(x,.5), \qquad \tilde{b}(x,y) = b(.5,y), \qquad \tilde{c}(x,y) = 0,$$

$$\tilde{d}(x,y) = 0, \qquad \tilde{e}(x,y) = \tfrac{1}{2}e(x,.5)+\tfrac{1}{2}e(.5,y).$$

For the non-self-adjoint approximation, $\tilde{d}(y)=d(.5,y)$. We examine formulation (16) with symmetric Q (the only possible choice), and formulation (27) with both symmetric and nonsymmetric Q. The stopping criterion in all tests is

$$\frac{\|r_i\|}{\|r_0\|} \leq \epsilon = 10^{-6}.$$

where the norm used is

$\|r_i\|_{Q^{-1}}$ for (16), $\|r_i\|_2$ for (27).

The initial guess is $u_0=0$.

Tables 4-1 and 4-2 show the iteration counts for $\gamma=5$ and $\gamma=50$, respectively. In the tables, both the preconditioning formulation and its associated norm are listed.

To examine different boundary conditions, for the third problem we consider the differential equation [12]

$$-(u_{xx} + u_{yy}) + \beta u_x = 0 \tag{30}$$

on $\{(x,y) \mid x \geq 0, y \geq 0\}$, with boundary conditions

$$u(x,0) = 0, \quad u(0,y) = 1, \tag{31}$$
$$u(x,y) \text{ bounded as } |x| + |y| \to \infty.$$

The exact solution to (30) - (31) has a boundary layer at $y=0$ and is nearly identically one elsewhere. Following [12], for the discrete problem we restrict (30) to the unit square and add the boundary conditions

$$u(x,1) = 1, \quad u_x(1,y) = 0. \tag{32}$$

We discretize on a uniform nxn grid using centered differences for all terms except at the right boundary, where we use the first order approximation

$$0 = u_x(x_{n+1},y_j) \approx \frac{u_{n+1,j} - u_{nj}}{h}. \tag{33}$$

Incorporating (33) into the equations centered at $\{u_{nj}\}_{j=1}^{n}$ results in a linear system of order $N=n^2$. We consider the value $\beta=10$.

Because (30)-(32) is separable, the discrete problem can be solved directly by cyclic reduction. We therefore consider only symmetric preconditioning matrices based on the discrete Laplacian, with the discretization at the right boundary handled as in (33). The iteration counts for the same stopping criterion, mesh sizes and initial guess as above are shown in Table (4-3).

	CGN			Orthomin(1)		
	Sym. Q (16) $\|r_i\|_{Q^{-1}}$	Sym. Q (27) $\|r_i\|_2$	Nonsym. Q (27) $\|r_i\|_2$	Sym. Q (16) $\|r_i\|_{Q^{-1}}$	Sym. Q (27) $\|r_i\|_2$	Nonsym. Q (27) $\|r_i\|_2$
1/16	11	15	11	17	21	8
1/32	11	17	13	17	21	9
1/64	12	19	14	18	22	9
1/128	12	20	14	18	22	9

Table 4-1: Discretization of equation (29), γ=5. Iterations to reduce residual norm by factor of 10^{-6}.

	CGN			Orthomin(1)		
	Sym. Q (16) $\|r_i\|_{Q^{-1}}$	Sym. Q (27) $\|r_i\|_2$	Nonsym. Q (27) $\|r_i\|_2$	Sym. Q (16) $\|r_i\|_{Q^{-1}}$	Sym. Q (27) $\|r_i\|_2$	Nonsym. Q (27) $\|r_i\|_2$
1/16	38	69	34[3]	111	Fails	23[3]
1/32	43	101	22[3]	121	Fails	17[3]
1/64	44	137	17	124	Fails	14
1/128	45	166	18	126	Fails	14
1/256	--	188	--	--	--	--

Table 4-2: Discretization of equation (29), γ=50. Iterations to reduce residual norm by factor of 10^{-6}.

[3]For a given non-self-adjoint operator, cyclic reduction is applicable only for small enough h (see [17], p. 1142, or [19]). For γ=50, h = 1/16 and 1/32 are too large. These mesh sizes are handled in the experiments by reducing the contribution to Q of the first derivatives, taking $\tilde{d}(y)=\delta(y)d(.5,y)$ with $0<\delta(y)\leq 1$, so the approximation of R in Q is probably less accurate in these cases.

	CGN		Orthomin(1)	
	Sym. Q (16) $\|r_i\|_{Q^{-1}}$	Sym. Q (27) $\|r_i\|_2$	Sym. Q (16) $\|r_i\|_{Q^{-1}}$	Sym. Q (27) $\|r_i\|_2$
1/16	11	10	15	16
1/32	11	10	15	15
1/64	11	10	15	13
1/128	11	10	15	12

Table 4-3: Discretization of equation (30)-(32), β=10. Iterations to reduce residual norm by factor of 10^{-6}.

These experiments confirm the convergence analysis of (16) and, with one exception, suggest that convergence for (27) is independent of mesh size also, for both symmetric and nonsymmetric Q. The exception occurs with γ=50, for which the iteration count for CGN is still growing at h=1/256, and Orthomin(1) fails to converge for $h \leq 1/128$, indicating that (27) may require a very fine mesh if A is not well-approximated by a symmetric preconditioning matrix. (The convergence failures are due to the fact that the symmetric part $(AQ^{-1}+(AQ^{-1})^T)/2$ of the coefficient matrix of (27) is *indefinite* for the four mesh sizes considered.) The smaller iteration counts in Tables 4-1 and 4-2 for nonsymmetric preconditioning matrices suggest that nonsymmetric preconditioners offer some advantage. However, the operation counts for cyclic reduction on nonsymmetric matrices ($20n^2 \log n + O(n^2)$) are higher than those for fast direct methods for symmetric matrices (e.g., $5n^2 \log n + O(n^2)$ for cyclic reduction; see [2, 3, 7, 18]). For the two problems examined, the asymptotic counts for symmetrically applied symmetric preconditioners are actually lower.

Finally, we note that a comparison between these preconditioning techniques and others based on incomplete factorizations can be found in [9].

Acknowledgment. The authors wish to acknowledge the many helpful suggestions made by Joe Pasciak during the preparation of this paper.

References

[1] Owe Axelsson. Conjugate gradient type methods for unsymmetric and inconsistent systems of linear equations. *Linear Algebra and Its Applications* 29:1-16, 1980.

[2] Randolph E. Bank. Marching algorithms for elliptic boundary value problems. II: The variable coefficient case. *SIAM Journal on Numerical Analysis* 14:950-970, 1977.

[3] Randolph E. Bank and Donald J. Rose. Marching algorithms for elliptic boundary value problems. I: The constant coefficient case. *SIAM Journal on Numerical Analysis* 14:792-829, 1977.

[4] Paul Concus and Gene H. Golub. A generalized conjugate gradient method for nonsymmetric systems of linear equations. In R. Glowinski and J. L. Lions, Editors, *Lecture Notes in Economics and Mathematical Systems, Volume 134*, Springer-Verlag, Berlin, 1976, pp. 56-65.

[5] Paul Concus and Gene H. Golub. Use of fast direct methods for the efficient numerical solution of nonseparable elliptic equations. *SIAM Journal on Numerical Analysis* 10:1103-1120, 1973.

[6] James W. Daniel. *The Approximate Minimization of Functionals*. Prentice-Hall, New York, 1971.

[7] Fred W. Dorr. The direct solution of the discrete Poisson equation on a rectangle. *SIAM Review* 12:248-263, 1970.

[8] Stanley C. Eisenstat, Howard C. Elman and Martin H. Schultz. Variational iterative methods for nonsymmetric systems of linear equations. *SIAM Journal on Numerical Analysis* 20:345-357, 1983.

[9] Howard C. Elman. *Iterative Methods for Large, Sparse, Nonsymmetric Systems of Linear Equations*. Ph.D. Thesis, Department of Computer Science, Yale University, 1982. Also available as Technical Report 229.

[10] George E. Forsythe and Wolfgang R. Wasow. *Finite-Difference Methods for Partial Differential Equations*. John Wiley and Sons, New York, 1960.

[11] Gene H. Golub and Michael L. Overton. *Convergence of a Two-stage Richardson Iterative Procedure for Solving Systems of Linear Equations*. Technical Report 38, Computer Science Department, Stanford University, September 1981.

[12] G. W. Hedstrom and Albert Osterheld. The effect of cell Reynolds number on the computation of a boundary layer. *Journal of Computational Physics* 37:399-421, 1980.

[13] Magnus R. Hestenes and Eduard Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards* 49:409-435, 1952.

[14] Milton Lees. A priori estimates for the solutions of difference approximations to parabolic partial differential equations. *Duke Mathematical Journal* 27:297-312, 1960.

[15] Seymour V. Parter. On the eigenvalues of second order elliptic difference operators. *SIAM Journal on Numerical Analysis* 19:518-530, 1982.

[16] Y. Saad. Krylov subspace methods for solving large unsymmetric linear systems. *Mathematics of Computation* 37:105-126, 1981.

[17] Paul N. Swarztrauber. A direct method for the discrete solution of separable elliptic equations. *SIAM Journal on Numerical Analysis* 11:1136-1150, 1974.

[18] Paul N. Swarztrauber. The methods of cyclic reduction, Fourier analysis and the FACR algorithm for the discrete solution of Poisson's equation on a rectangle. *SIAM Review* 19:490-501, 1977.

[19] P. Swarztrauber and R. Sweet. *Efficient FORTRAN Subprograms for the Solution of Elliptic Partial Differential Equations*. Technical Report TN/IA-109, National Center for Atmospheric Research, 1975.

[20] Richard S. Varga. *Matrix Iterative Analysis*. Prentice-Hall, Englewood Cliffs, N.J., 1962.

[21] P. K. W. Vinsome. Orthomin, an iterative method for solving sparse sets of simultaneous linear equations. In *Proceedings of the Fourth Symposium on Reservoir Simulation*, Society of Petroleum Engineers of AIME, 1976, pp. 149-159.

[22] Olof Widlund. A Lanczos method for a class of non-symmetric systems of linear equations. *SIAM Journal on Numerical Analysis* 15:801-812, 1978.

[23] Olof B. Widlund. On the use of fast methods for separable finite difference equations for the solution of general elliptic problems. In D. J. Rose and R. A. Willoughby, Editors, *Sparse Matrices and Their Applications*, Plenum Press, New York, 1972, pp. 121-131.

[24] Kosaku Yosida. *Functional Analysis*. Academic Press, New York, 1965.

[25] David M. Young and Kang C. Jea. Generalized conjugate gradient acceleration of nonsymmetrizable iterative methods. *Linear Algebra and Its Applications* 34:159-194, 1980.