MATHEMATICS OF COMPUTATION
VOLUME 41, NUMBER 164
OCTOBER 1983, PAGES 511-525

Block AOR Iteration for Nonsymmetric Matrices

By Theodore S. Papatheodorou

Abstract. We consider a class of matrices that are of interest to numerical applications and are large, sparse, but not symmetric or diagonally dominant. We give a criterion for the existence of (and we actually construct) the inverse matrix in terms of powers of a "small" matrix. We use this criterion to find that the spectral radius of the Jacobi iteration matrix, corresponding to a block tridiagonal partition, is in general $\geqslant 1$. We also derive conditions (that are satisfied in cases of interest to applications) for the Jacobi matrix to have spectral radius $= 1$. We introduce convergent "block AOR" iterative schemes such as extrapolated Jacobi and extrapolated Gauss-Seidel schemes with optimum (under) relaxation parameter $\omega = .5$. A numerical example pertaining to the solution of Poisson's equation is given, as a demonstration of some of our hypotheses and results. A comparison with SOR, applied to the 5-point finite difference method, is also included.

1. Introduction. We are concerned with the solution of a certain type of large linear systems that are encountered in some applications. One such instance, of importance to mathematical software, is the numerical solution of Poisson's equation on a square, with Dirichlet conditions, when the collocation method with Hermite bicubic elements is used. If the ordinary 5-point finite difference scheme is used for this problem, the resulting matrix is symmetric and diagonally dominant. Moreover, for the finite difference matrix, iterative methods have been developed and are well behaved and analyzed. However, there are instances in which one would prefer to use collocation instead of the standard 5-point difference scheme (Houstis et al. [4]) and at least one consideration (namely storage, cf. Rice [6]) makes it important to develop iterative methods for collocation matrices as well.

Unfortunately, the well-known iterative techniques, and their analysis, are not applicable to collocation matrices that are large and sparse but, in contrast to finite difference matrices, are not symmetric or diagonally dominant. For example, the point and the block tridiagonal Jacobi iteration matrices, for finite difference schemes, have eigenvalues that are real and less than one in modulus. On the other hand, these matrices are not even defined for collocation, due to the accumulation of zero entries on and around the main diagonal; cf. Figure 2, Section 5. Thus we introduce a modification of the standard collocation method (Section 5) that results in a well-defined block tridiagonal Jacobi matrix. Still, the only real eigenvalue of this matrix is zero, and all nonreal eigenvalues have modulus one. In this paper we develop a theory that predicts and explains such events, and, in spite of the lack of properties such as symmetry and diagonal dominance, we introduce and analyze

Received June 21, 1982; revised October 27, 1982.
1980 *Mathematics Subject Classification.* Primary 65F10.
Key words and phrases. AOR iteration.

0025-5718/83 $1.00 + $.25 per page

convergent iterative schemes for a class of matrices that includes those arising in the collocation method. We do that as follows:

In Section 2 we give two theorems that characterize the inverse matrix in terms of a (much smaller) "representative matrix R". In fact, a formula for constructing the inverse is also given, as a generalization of a result of this author [5].

In Section 3 we partition the matrix in block tridiagonal form, and, using the theorems of Section 2, we find that in general the spectral radius of the corresponding Jacobi matrix J is, disturbingly, $\geqslant 1$, since together with μ the inverse $1/\mu$ is also an eigenvalue. In the process we find a method for calculating the eigenvalues of J.

In Section 4, we introduce and investigate "block AOR (Accelerated Overrelaxation)" techniques based on the schemes of Hadjidimos [2], [3]. We give conditions that are satisfied in practice, for which all the nonreal eigenvalues of J are one in modulus and the only real eigenvalues are zero. We also derive a relation between the eigenvalues μ of J and the eigenvalues τ of the iteration matrix T of each scheme with proofs similar to those in Varga [7] and Young [8]. Then we show that convergent block AOR schemes do exist for our matrices, by exhibiting two such schemes, namely an extrapolated Jacobi and an extrapolated Gauss-Seidel scheme. For both, the best relaxation parameter is found to be $\omega = .5$.

In Section 5, we introduce a simple modification of the collocation method and demonstrate some facts and proofs of this paper, using a numerical application to Dirichlet's problem on the unit square. Finally, we give a comparison with SOR applied to the ordinary 5-point finite difference scheme.

The remainder of this introduction is used to describe the class of matrices G, examined in this paper. For reasons of economy we use some notational abbreviations:

Consider a 2×4 block

$$[A \mid B] = \left[\begin{array}{cc|cc} a_{11} & a_{12} & b_{11} & b_{12} \\ a_{21} & a_{22} & b_{21} & b_{22} \end{array}\right].$$

By repeating this block N times and shifting to the right by two columns at each repetition, form a $(2N) \times (2N+2)$ matrix

$$\begin{bmatrix} A & B & & & & \\ & A & B & & & \\ & & & \ddots & & \\ & & & & A & B \end{bmatrix}.$$

If we delete columns 1 and $2N-1$ of this matrix (elimination of boundary degrees of freedom) the resulting matrix is square $(2N) \times (2N)$:

$$\begin{bmatrix} a & B & & & & \\ & A & B & & & \\ & & & \ddots & & \\ & & & & A & B \\ & & & & & A & b \end{bmatrix}, \quad a = \begin{bmatrix} a_{12} \\ a_{22} \end{bmatrix}, \quad b = \begin{bmatrix} b_{12} \\ b_{22} \end{bmatrix}.$$

For such matrices G we use the abbreviated notation

$$G = [A \mid B]_{(2N)} \quad \text{or} \quad G = \left[\begin{array}{cc|cc} a_{11} & a_{12} & b_{11} & b_{12} \\ a_{21} & a_{22} & b_{21} & b_{22} \end{array}\right]_{(2N)}. \tag{1.1}$$

Matrices like this arise in one-dimensional collocation. For two-dimensional problems consider the previous definitions with each a_{ij}, b_{ij} being changed from a scalar to a matrix $(2N) \times (4N)$. Then A, B are of "block dimension" 2×2 and size $(4N) \times (4N)$. In order to remind ourselves that we are in the "block" or "two-dimensional" case we write, instead of (1.1),

$$G = [A \mid B]_{\otimes(2N)}. \tag{1.2}$$

Now G is $(4N)^2 \times (4N)^2$.

2. Theorems About the Inverse. For each matrix G of the form

$$G = [A \mid B]_{(2N)} \quad \text{or} \quad G = [A \mid B]_{\otimes(2N)}$$

we assume that B^{-1} exists, and we consider the "representative matrix"

$$R = -B^{-1}A. \tag{2.1}$$

Depending on the case, R is only 2×2 or "block 2×2". That is to say R is of the type

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix},$$

where S_{ij} are either scalars or $(2N) \times (2N)$ matrices. For the powers of such matrices we use the notation

$$S^k = \begin{bmatrix} S_{11}^{(k)} & S_{12}^{(k)} \\ S_{21}^{(k)} & S_{22}^{(k)} \end{bmatrix}, \tag{2.2}$$

with the warning that in general $S_{ij}^{(k)} \neq S_{ij}^k$. For simplicity $S_{ij}^{(1)} = S_{ij}$.

Consider first the scalar case, where A, B, R are 2×2, regardless of the size $(2N) \times (2N)$ of G. In a number of applications of interest (cf. [5])

$$R_{22} = R_{11}, \qquad \det(R) = R_{11}R_{22} - R_{21}R_{12} = 1. \tag{2.3}$$

A summary of some results in [5] is given by the following:

THEOREM 2.1. *If either*
(i) $R_{12}^{(N)} \neq 0$, *or*
(ii) $R_{11} \neq \cos(k\pi/N)$, $k = 1, \ldots, 2N$, *and* (2.3) *holds, then* G^{-1} *exists.*

A formula for constructing G^{-1} is also given in [5] and can be recovered as a special case of the following two-dimensional generalization: Define the block 2×2 matrices

$$G_{ij} = -R^{i-1} \begin{bmatrix} R_{12}R_{12}^{(N)^{-1}} & 0 \\ R_{22}R_{12}^{(N)^{-1}} & 0 \end{bmatrix} R^{N-j} + \begin{cases} R^{i-j} & \text{if } i \geq j, \\ 0 & \text{if } i < j, \end{cases}$$

$$i = 0, \ldots, N, j = 1, \ldots, N.$$

If $R_{12}^{(N)}$ is nonsingular, then G_{ij} is well defined. The matrix $\tilde{G} = \{G_{ij}B^{-1}\}$, $i = 0, \ldots, N$, $j = 1, \ldots, N$ is block $(2N+2) \times (2N)$, each block being of size $(2N) \times (2N)$. The matrix G' that results from $\tilde{G}$ if we delete row blocks 1 and $2N+1$ is

block $(2N) \times (2N)$, i.e. of size $(4N^2) \times (4N^2)$, the same as G. The following is a generalization of part (i) of Theorem (2.1):

THEOREM 2.2. *If $R_{12}^{(N)}$ is nonsingular, then G^{-1} exists and*

$$G^{-1} = G'.$$

Proof. By hypothesis G' is well defined. Verify that $G'G = I$. □

Generalizations in the direction of part (ii) of Theorem 2.1 are also possible but not needed here. We only give some properties leading to relations that are analogous to (2.3). For matrices encountered in our applications it happens that

$$[A \mid B] = \left[\begin{array}{cc|cc} A_1 & A_2 & A_3 & -A_4 \\ A_3 & A_4 & A_1 & -A_2 \end{array}\right], \tag{2.4}$$

each A_l being $(2N) \times (2N)$. Note that

$$B = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} A \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}, \tag{2.5}$$

which gives

$$R^{-1} = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} R \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} \tag{2.6}$$

and if R_{12}^{-1} exists, then

$$R^k = Q^{-1} S^k Q, \tag{2.7}$$

$$Q = \begin{bmatrix} I & 0 \\ 0 & R_{12} \end{bmatrix}, \qquad S = \begin{bmatrix} S_{11} & S_{21} \\ S_{12} & S_{22} \end{bmatrix} = \begin{bmatrix} R_{11} & I \\ R_{11}^2 - I & R_{11} \end{bmatrix}.$$

In analogy with (2.3) we note that

$$\det(R) = 1, \quad S_{11}S_{22} - S_{21}S_{12} = I, \quad S_{11} = S_{22}. \tag{2.8}$$

We close this section with a justification of our assumptions on the existence of inverses of smaller matrices involved, such as B and R_{12}. For the scalar case B is 2×2, R_{12} is an easily found scalar and these assumptions are easy to verify. For the two-dimensional case and for the applications of interest, not only (2.4) holds but also (cf. Section 5) each A_l is, in turn, a matrix of the same (scalar) type

$$A_l = [A \mid B]_{2N}.$$

Then Theorem 2.1 applies and the hypotheses of invertibility are again verified.

3. Eigenvalues of the Jacobi Matrix for the Block Tridiagonal Form. We consider the case (2.4), and we take $N = 2^l$, $l = 0, 1, 2, \ldots$. We use the following nonsingularity assumptions as observed in the applications and as justified at the end of the previous section: B^{-1} is nonsingular, hence R is well defined, $R_{12}^{(N)}$ is nonsingular so that G^{-1} exists (Theorem 2.2) and the matrix $R_{12}^{(1)} = R_{12}$ is nonsingular.

We partition G into block tridiagonal form

$$G = \begin{bmatrix} D_1 & -U_1 & & & & \\ -L_1 & D_1 & -U_1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -L_1 & D_1 & -U_1 & \\ & & & -L_1 & \overline{D}_1 & \end{bmatrix}, \tag{3.1}$$

where by (2.4)

$$D_1 = \begin{bmatrix} A_2 & A_3 \\ A_4 & A_1 \end{bmatrix}, \quad \bar{D}_1 = \begin{bmatrix} A_2 & -A_4 \\ A_4 & -A_2 \end{bmatrix}, \quad -L_1 = \begin{bmatrix} 0 & A_1 \\ 0 & A_3 \end{bmatrix}, \quad -U_1 = \begin{bmatrix} -A_4 & 0 \\ -A_2 & 0 \end{bmatrix}.$$

This partition corresponds to the splitting

$$G = D - L - U, \tag{3.2}$$

with the standard (and obvious) definitions of the block diagonal **D** and the lower (resp. upper) parts L (resp. U).

The Jacobi iteration matrix associated with this partition is

$$J = I - D^{-1}G = D^{-1}(L + U) \tag{3.3}$$

and is consistently ordered and weakly cyclic of index 2 (cf. [7]). Notice that μ is an eigenvalue of J if and only if the matrix

$$M_\mu = \mu D - L - U \tag{3.4}$$

is singular. But

$$M_\mu = \left[A_\mu \middle| B_\mu \right]_{\otimes(2N)}, \tag{3.5}$$

with

$$A_\mu = \begin{bmatrix} A_1 & \mu A_2 \\ A_3 & \mu A_4 \end{bmatrix}, \qquad B_\mu = \begin{bmatrix} \mu A_3 & -A_4 \\ \mu A_1 & -A_2 \end{bmatrix}. \tag{3.6}$$

By inspection of the block structure of J it is easy to see that $\mu = 0$ is an eigenvalue of J of multiplicity $4N$. Consider now $\mu \neq 0$ and the matrices $R = -B^{-1}A$ [representing G, with A, B defined in (2.4)] and $R_\mu = -B_\mu^{-1}A_\mu$ [representing M_μ, with A_μ, B_μ defined in (3.6)]. In analogy with (2.7)–(2.8) we find that the entries $R_{\mu,ij}^{(k)}$ of R_μ^k (recall notation of (2.2)) are related with the entries $R_{ij}^{(k)}$ of R^k through the formula

$$R_\mu^k = Q^{-1}S_\mu^k Q, \quad Q = \begin{bmatrix} I & 0 \\ 0 & R_{12} \end{bmatrix}, \quad S_\mu = \begin{bmatrix} \frac{1}{\mu}R_{11} & I \\ R_{11}^2 - I & \mu R_{11} \end{bmatrix}. \tag{3.7}$$

Thus, for the entries $R_{\mu,12}^{(N)}$, $S_{\mu,12}^{(N)}$ of the Nth power we obtain $R_{\mu,12}^{(N)} = S_{\mu,12}^{(N)}R_{12}$. By Theorem 2.2, for M_μ to be singular $R_{\mu,12}^{(N)}$ must be singular. Since R_{12} is invertible, $S_{\mu,12}^{(N)}$ must be singular. We now obtain a formula for $S_{\mu,12}^{(N)}$:

Define the polynomials

$$p_1(z) = z, \qquad p_{2^{k+1}}(z) = \left[p_{2^k}(z) \right]^2 - 2. \tag{3.8}$$

Use induction to show that

$$S_{\mu,12}^{(N)} = S_{\mu,12}^{(2^l)} = \left(S_{\mu,11}^{(2^{l-1})} + S_{\mu,22}^{(2^{l-1})} \right) \cdots \left(S_{\mu,11}^{(2)} + S_{\mu,22}^{(2)} \right)\left(S_{\mu,11} + S_{\mu,22} \right),$$

$$S_{\mu,11} + S_{\mu,22} = \left(\mu + \frac{1}{\mu} \right) R_{11}, \qquad S_{\mu,11}^{(2^{k+1})} + S_{\mu,22}^{(2^{k+1})} = \left(S_{\mu,11}^{(2^k)} + S_{\mu,22}^{(2^k)} \right)^2 - 2I,$$

and combine these relations with (3.8) to obtain

$$S_{\mu,12}^{(N)} = \prod_{k=0}^{l-1} p_{2^k}(\sigma R_{11}), \qquad (N = 2^l), \tag{3.9}$$

where $\sigma = \mu + 1/\mu$.

We seek all values of σ that make each of the factors in (3.9) singular. Thus, if r is an eigenvalue of R_{11}, then we seek the 2^k values of σ for which $p_{2^k}(\sigma r) = 0$, $k = 0,\ldots,l-1$. If we repeat for all $2N$ eigenvalues r of R_{11}, then we obtain $2N^2 - 2N$ values of σ. For each such σ we obtain two values of μ (μ and $1/\mu$). Hence we find all the remaining $4N^2 - 4N$ eigenvalues. Summarizing this discussion we have

THEOREM 3.1. *Let J be the Jacobi iteration matrix corresponding to the tridiagonal splitting* (3.1)–(3.3) *of G, with eigenvalues denoted by μ.*

(i) *$\mu = 0$ is an eigenvalue of multiplicity $4N$,*

(ii) *the remaining $4N^2 - 4N$ eigenvalues are found by first solving*

$$p_{2^k}(\sigma r) = 0, \tag{3.10}$$

(p_{2^k} defined in (3.8)), *for each eigenvalue r of R_{11} and each $k = 0,\ldots,l-1$, ($2^l = N$), and then solving*

$$\mu + 1/\mu = \sigma, \tag{3.11}$$

(iii) *together with each eigenvalue $\mu \neq 0$, $1/\mu$ is also an eigenvalue corresponding to the same σ, hence*

(iv) *the spectral radius of J is $\geqslant 1$.*

Thus, in view of part (iv), we are discouraged from using standard iterative techniques (although they may still converge), and we turn our attention to the more general AOR schemes of the following section.

4. Convergent Block AOR Schemes. AOR (Accelerated Overrelaxation) schemes are defined by Hadjidimos [2], [3]. Our block counterpart of these schemes for the case of G of (3.1) has the iteration matrix

$$T = (D - \rho L)^{-1}\big[(1-\omega)D + (\omega - \rho)L + \omega U\big],$$

and, of course, we are interested in schemes for which $|\tau| < 1$, for all eigenvalues τ of T. The pair (ω, ρ) consists of the "relaxation" and "acceleration" parameters, and well-known techniques are recovered for special combinations of ω, ρ. For instance, the pairs $(1,0)$, $(1,1)$ and (ω,ω) give the Jacobi, Gauss-Seidel and SOR methods. For more details see [3].

First, we relate the eigenvalues τ of T to the eigenvalues μ of the Jacobi matrix J by the following

THEOREM 4.1. *$\tau = 1 - \omega$ with multiplicity $4N$, and*

$$(\tau + \omega - 1)^2 = (\rho\tau + \omega - \rho)\omega\mu^2 \tag{4.1}$$

with multiplicity $2N^2 - 2N$.

Proof. The proof follows the same pattern as in [7]:

$$\mathrm{Det}\big[(\tau + \omega - 1)I - (\rho\tau + \omega - \rho)D^{-1}L - \omega D^{-1}U\big] = 0,$$

and, since $J = D^{-1}L + D^{-1}U$ is consistently ordered weakly cyclic of index 2,

$$\mathrm{Det}\left[(\tau + \omega - 1)I - \{(\rho\tau + \omega - \rho)\omega\}^{1/2} J\right] = 0.$$

By Romanovsky's Theorem

$$(\tau + \omega - 1)^{4N} \prod_{i=1}^{2N^2-2N} \left\{(\tau + \omega - 1)^2 - (\rho\tau + \omega - \rho)\omega\mu_i^2\right\} = 0,$$

where only one of the two eigenvalues $\pm\mu_i$ is counted. This completes the proof. □

In the special case of the SOR method, $\rho = \omega$, we recover the known formula (cf. [7])

$$(\tau + \omega - 1)^2 = \tau\omega^2\mu^2. \tag{4.2}$$

In our case however, we cannot proceed with assumptions that, e.g. μ is real. In fact, for the matrices we are interested in, the opposite is true, namely, the only real eigenvalue is $\mu = 0$. We proceed to give some realistic conditions (satisfied in our applications) that will result to some convergent block AOR schemes:

THEOREM 4.2. (i) *If all the eigenvalues r of R_{11} are real, then all the nonreal eigenvalues of the Jacobi matrix J lie on the circumference of the unit circle.* (ii) *If in addition $|r| \geq 1$, then the only real eigenvalue of J is $\mu = 0$.*

Proof. Recalling Theorem 3.1-(ii), fix r and consider solving (3.10). To do so, notice first that for each complex x there exists at least one y such that

$$q(y) := y + \frac{1}{y} = x. \tag{4.3}$$

Then calculate successively $p_1(x) = x = q(y)$, $p_2(x) = q(y^2)$, and inductively

$$p_{2^k}(x) = q\left(y^{2^k}\right) = y^{2^k} + 1/y^{2^k}.$$

Since $q(z) = 0$ if and only if $z = \pm i$, since the same x is produced in (4.3) by both y and $1/y$ and, finally, since $1/y = \bar{y}$ if y is such that $q(y^{2^k}) = 0$, the solutions of $p_{2^k}(\sigma r) = 0$ are found to be

$$\sigma r = 2\cos(\theta_{k,m}), \qquad \theta_{k,m} = \frac{(2m+1)\pi}{2^{k+1}}, \qquad m = 0, \ldots, 2^k - 1. \tag{4.4}$$

Hence, for each fixed r, we obtain from (3.11)

$$\mu + \frac{1}{\mu} = \sigma = \frac{2}{r}\cos(\theta_{k,m}). \tag{4.5}$$

If r is real, then $\mu + 1/\mu$ is real, which means that either μ is real or $|\mu| = 1$, proving part (i). In fact, setting $c = \cos(\theta_{k,m})$, we obtain from (4.5):

$$\mu = \frac{c}{r} \pm \sqrt{\left(\frac{c}{r}\right)^2 - 1}\,. \tag{4.6}$$

Thus, if $|c/r| < 1$, μ cannot be real. Since $c \neq \pm 1$ by definition of $\theta_{k,m}$ in (4.4) the condition $|r| \geq 1$ is sufficient for μ to be nonreal, proving part (ii). □

Fortunately, the hypotheses of Theorem 4.2 are observed to be true in some applications (see for instance, Section 5). With this in mind, we proceed to establish

the existence of converging AOR schemes by exhibiting two such schemes, namely:

Scheme EJ: Extrapolated Jacobi $(\omega, \rho) = (\omega, 0)$.

Scheme EGS: Extrapolated Gauss-Seidel $(\omega, \rho) = (\omega, 1)$.

For both schemes we have the initial standard restriction

$$0 < \omega < 2$$

imposed by the relationship $\tau = 1 - \omega$ of Theorem 4.1.

THEOREM 4.3. (i) $\mu = 1$ *and* $\mu = -1$ *are not eigenvalues of* J. (ii) *If all the eigenvalues* $\mu \neq 0$ *of* J *lie on the circumference of the unit circle* (*see Theorem* 4.2 *for sufficient conditions*), *then both schemes EJ and EGS converge for all* $0 < \omega < 1$ *with best* (*under*) *relaxation value* $\omega_b = \frac{1}{2}$.

Proof. $\mu = 1$ cannot be an eigenvalue of J because G is nonsingular. J is weakly cyclic of index 2, hence together with μ has the eigenvalue $-\mu$. Since $\mu = 1$ is not an eigenvalue, it follows that $\mu = -1$ is not an eigenvalue. To show part (ii) use the transformation

$$\tau = (1 - \omega) + \omega\beta \tag{4.7}$$

in (4.1) to find

$$\beta = \tfrac{1}{2}\left(\rho\mu^2 \pm \sqrt{\rho^2\mu^4 + 4(1 - \rho)\mu^2}\right). \tag{4.8}$$

For EJ, $\rho = 0$, hence $\beta = \pm\mu$. For EGS, $\rho = 1$, hence $\beta = 0$ or μ^2. In both cases $\beta \neq 1$ by part (i). Thus, the segment joining 1 and β in the complex plane is nontrivial, and β lies on the circumference of the unit circle for EJ and EGS. At the same time, by (4.7), τ lies on this segment (which is entirely inside the unit circle) for $0 < \omega < 1$ and therefore $|\tau| < 1$, with the smallest possible value of $|\tau|$ being obtained at the midpoint of the segment, i.e. with $\omega = \frac{1}{2}$. □

5. Application: Poisson's Equation on a Square. In this section we give an application of the preceding results to the problem

$$\nabla^2 u = f, \quad \text{on } \Omega := [0, 1] \times [0, 1], \tag{5.1}$$

$$u = g, \quad \text{on } \partial\Omega, \tag{5.2}$$

when the collocation method with Hermite bicubic elements is used. In order to do so we first give a brief description of this method and our modification of it that leads to matrices with the block structure of Sections 1 and 2.

Standard Collocation. Consider a uniform grid with spacing $h := 1/N$, where N is the same as in the preceding sections. The coordinates of the nodes are (x_i, y_j), where $x_i = (i - 1)h$, $y_j = (j - 1)h$, $i, j = 1, \ldots, (N + 1)$. We define the generating cubic polynomials ϕ and ψ on $[0, 1]$ by

$$\phi(\sigma) := (1 - \sigma)^2(1 + 2\sigma), \qquad \psi(\sigma) := \sigma(1 - \sigma)^2, \qquad 0 \leqslant \sigma \leqslant 1. \tag{5.3}$$

Let t denote either x or y, introduce the "fictitious nodes" $t_0 := -h$ and $t_{N+2} := 1 + h$ and define the $2(N + 1)$ functions $\bar{B}_k$, $k = 1, \ldots, 2(N + 1)$ of $t \in \mathbf{R}$ as follows:

$$\bar{B}_{2m-1}(t) := \begin{cases} \phi(-\sigma_m(t)), & \text{if } t_{m-1} \leqslant t \leqslant t_m, \\ \phi(\sigma_m(t)), & \text{if } t_m \leqslant t \leqslant t_{m+1}, \\ 0, & \text{otherwise,} \end{cases}$$

$$B_{2m}(t) := \begin{cases} -h\psi(-\sigma_m(t)), & \text{if } t_{m-1} \leqslant t \leqslant t_m, \\ h\psi(\sigma_m(t)), & \text{if } t_m \leqslant t \leqslant t_{m+1}, \\ 0, & \text{otherwise,} \end{cases}$$

where $m = 1,\ldots,N+1$ and $\sigma_m(t) := (t - t_m)/h$.

It follows that $\bar{B}_k \in C^1(\mathbf{R})$, $k = 1,\ldots,2(N+1)$ and that

$$\begin{aligned} &\bar{B}_{2m-1}(t_j) = \delta_m^j, \quad \frac{d}{dt}\bar{B}_{2m-1}(t_j) = 0, \\ &\bar{B}_{2m}(t_j) = 0, \quad \frac{d}{dt}\bar{B}_{2m}(t_j) = \delta_m^j, \end{aligned} \qquad m, j = 1,\ldots,(N+1), \tag{5.4}$$

where δ_m^j denotes the "Kronecker delta".

We then seek an approximate solution of (5.1)–(5.2) in the form:

$$u_N(x, y) := \sum_{i,j=1}^{2(N+1)} \bar{\alpha}_{ij}\bar{B}_i(x)\bar{B}_j(y). \tag{5.5}$$

Note that by (5.4):

$$\begin{aligned} &\bar{\alpha}_{2i-1,2j-1} = u_N(x_i, y_i), \qquad \bar{\alpha}_{2i-1,2j} = D_y u_N(x_i, y_j), \\ &\bar{\alpha}_{2i,2j-1} = D_x u_N(x_i, y_j), \qquad \bar{\alpha}_{2i,2j} = D_{xy}^2 u_N(x_i, y_j). \end{aligned} \tag{5.6}$$

From (5.6) we see that four degrees of freedom (d.o.f.), or unknowns, $\bar{\alpha}_{k,m}$, $k = 2i-1, 2i$, $m = 2j-1, 2j$, are associated to each node (x_i, y_j) and that they represent values of u_N and its derivatives at this node. For the boundary nodes, some of the d.o.f. are eliminated beforehand by use of the boundary conditions. One way to do this is by interpolation of the function g, of (5.2). (A simple version of this, brief enough to be described here, is to use also the derivatives of g, if they exist: For example, for the boundary $x = 0$ we use $\bar{\alpha}_{1,2j-1} = g(0, y_j)$ and $\bar{\alpha}_{1,2j} = D_y g(0, y_j)$, $j = 1,\ldots,N+1$.) After the elimination of the boundary d.o.f. we are left with $n = 4N^2$ unknowns, and we need to construct the same number of equations, by use of the operator equation (5.1), in Ω ("interior collocation"). This is done by choosing 4 points in each of the N^2 elements $I_{ij} = [x_i, x_{i+1}] \times [y_j, y_{j+1}]$ and requiring that (5.1) is exactly satisfied by u_N of (5.5) at these points. These are the so-called "Gauss points" in each I_{ij}, $i = 1,\ldots,N$, $j = 1,\ldots,N$, i.e. their coordinates $(\xi_{i,k}, \eta_{j,m})$, $k, m = 1, 2$, are the roots of the Legendre polynomial of degree two, shifted over the corresponding subintervals. For example, $\xi_{i,k} = h(2i - 1 \pm \sqrt{3}/3)/2$, where the "$-$" is used for $k = 1$ and the "$+$" for $k = 2$. Note that we have a one-to-one correspondence between collocation points and equations. Thus, a numbering of the equations is produced when we number the collocation points. A numbering of the unknowns is produced when we number the nodes and count the unknowns associated with each node in the specific order that they appear in (5.6). Standard collocation uses the numbering demonstrated in Figure 1, for $N = 3$.

It should be clear that within each element I_{ij}, u_N is determined in (5.5) by use of only 16 d.o.f., the ones that are associated with the four nodes of I_{ij}. This is because the remaining basis functions $\bar{B}$ in (5.5) vanish inside I_{ij}. Therefore the large collocation matrix is banded, but otherwise, as we can observe in Figure 2, it has many zeros on and around the diagonal, it is not symmetric and, in summary, unfit for iteration. The situation worsens for larger N. Also, the entries of this collocation matrix depend on h.

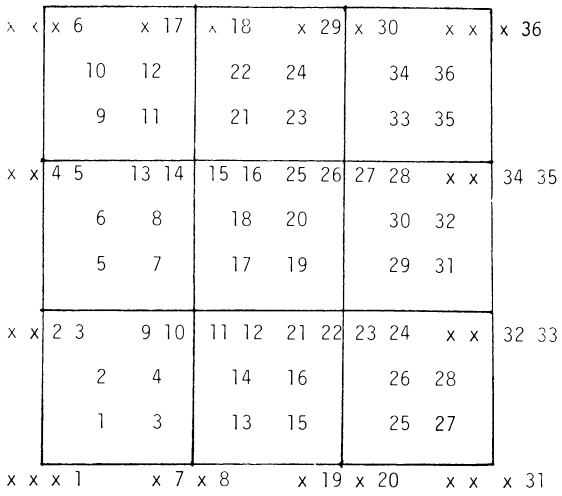

FIGURE 1 ($N = 3$)

Numbering of unknowns and equations for standard collocation

(x: d.o.f. *elminated beforehand*)

```
Interior collocation equations
                1         2         3
1234567890123456789012345678901234567890123456
1   dxx000xxxxxx.....................
2   xdx000xxxxxx.....................
3   xxd000xxxxxx.....................
4   xxx000xxxxxx.....................
5   .xxxd000xxxxxxxx.................
6   .xxxx000xxxxxxxx.................
7   .xxxx000xxxxxxxx.................
8   .xxxx000xxxxxxxx.................
9   ...xxx000000xxxxxx...............
10  ...xxx000000xxxxxx...............
11  ...xxx000000xxxxxx...............
12  ...xxx000000xxxxxx...............
13  ......xxxxxx000000xxxxxx.........
14  ......xxxxxx000000xxxxxx.........
15  ......xxxxxx000000xxxxxx.........
16  ......xxxxxx000000xxxxxx.........
17  ........xxxxxxxx0000xxxxxxxx.....
18  ........xxxxxxxx0000xxxxxxxx.....
19  ........xxxxxxxx0000xxxxxxxx.....
20  ........xxxxxxxx0000xxxxxxxx.....
21  ............xxxxxx000000xxxxxx....
22  ............xxxxxx000000xxxxxx....
23  ............xxxxxx000000xxxxxx....
24  ............xxxxxx000000xxxxxx....
25  ..................xxxxxx000000xxx...
26  ..................xxxxxx000000xxx...
27  ..................xxxxxx000000xxx...
28  ..................xxxxxx000000xxx...
29  ....................xxxxxxxx000xxxx.
30  ....................xxxxxxxx000xxxx.
31  ....................xxxxxxxx000xxxx.
32  ....................xxxxxxxx00dxxxx.
33  ......................xxxxxx000xxx
34  ......................xxxxxx000dxx
35  ......................xxxxxx000xdx
36  ......................xxxxxx000xxd
```

FIGURE 2 ($N = 3$)

Structure of the standard collocation matrix

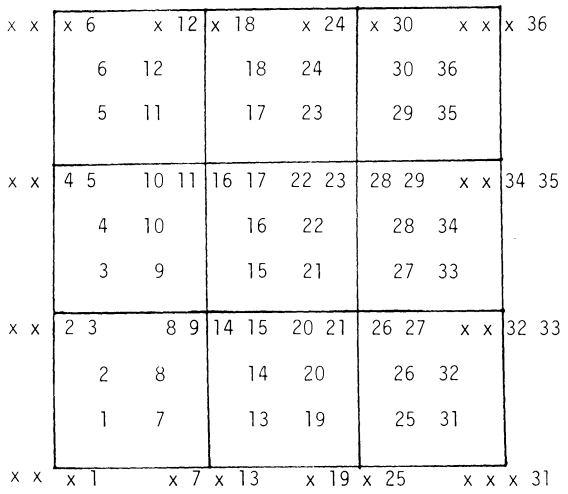

FIGURE 3 ($N = 3$)

Numbering of unknowns and equations for modified collocation

```
                  1           2           3
    123456 789012 345678 901234 567890 123456
 1  dxx... xxx... xxx... ...... ...... ......
 2  xdx... xxx... xxx... ...... ...... ......
 3  .xdxx. .xxxx. .xxxx. ...... ...... ......
 4  .xxdx. .xxxx. .xxxx. ...... ...... ......
 5  ...xdx ...xxx ...xxx ...... ...... ......
 6  ...xxd ...xxx ...xxx ...... ...... ......

 7  xxx... dxx... xxx... ...... ...... ......
 8  xxx... xdx... xxx... ...... ...... ......
 9  .xxxx. .xdxx. .xxxx. ...... ...... ......
10  .xxxx. .xxdx. .xxxx. ...... ...... ......
11  ...xxx ...xdx ...xxx ...... ...... ......
12  ...xxx ...xxd ...xxx ...... ...... ......

13  ...... xxx... dxx... xxx... xxx... ......
14  ...... xxx... xdx... xxx... xxx... ......
15  ...... .xxxx. .xdxx. .xxxx. .xxxx. ......
16  ...... .xxxx. .xxdx. .xxxx. .xxxx. ......
17  ...... ...xxx ...xdx ...xxx ...xxx ......
18  ...... ...xxx ...xxd ...xxx ...xxx ......

19  ...... xxx... xxx... dxx... xxx... ......
20  ...... xxx... xxx... xdx... xxx... ......
21  ...... .xxxx. .xxxx. .xdxx. .xxxx. ......
22  ...... .xxxx. .xxxx. .xxdx. .xxxx. ......
23  ...... ...xxx ...xxx ...xdx ...xxx ......
24  ...... ...xxx ...xxx ...xxd ...xxx ......

25  ...... ...... ...... xxx... dxx... xxx...
26  ...... ...... ...... xxx... xdx... xxx...
27  ...... ...... ...... .xxxx. .xdxx. .xxxx.
28  ...... ...... ...... .xxxx. .xxdx. .xxxx.
29  ...... ...... ...... ...xxx ...xdx ...xxx
30  ...... ...... ...... ...xxx ...xxd ...xxx

31  ...... ...... ...... xxx... xxx... dxx...
32  ...... ...... ...... xxx... xxx... xdx...
33  ...... ...... ...... .xxxx. .xxxx. .xdxx.
34  ...... ...... ...... .xxxx. .xxxx. .xxdx.
35  ...... ...... ...... ...xxx ...xxx ...xdx
36  ...... ...... ...... ...xxx ...xxx ...xxd
```

FIGURE 4 ($N = 3$)

Structure of the new collocation matrix

Modified Collocation. Our modification of the standard collocation method consists of the following simple changes: First, we use a different numbering of equations (collocation points) and unknowns as demonstrated in Figure 3, for $N = 3$. As a result the new matrix has the block structure of Sections 1 and 2, as can be seen from Figure 4.

Next, instead of the functions $\bar{B}$ we use

$$B_{2m-1} := \bar{B}_{2m-1}, \qquad B_{2m} := \frac{1}{h}\bar{B}_{2m},$$

and we compensate by replacing (5.5) and (5.6) accordingly:

$$u_N(x, y) = \sum_{i,j=1}^{N+1} \alpha_{ij} B_i(x) B_j(y), \tag{5.7}$$

$$\alpha_{2i-1,2j-1} = \bar{\alpha}_{2i-1,2j-1}, \qquad \alpha_{2i-1,2j} = h\bar{\alpha}_{2i-1,2j},$$

$$\alpha_{2i,2j-1} = h\bar{\alpha}_{2i,2j-1}, \qquad \alpha_{2i,2j} = h^2\bar{\alpha}_{2i,2j}.$$

As a result, after the number $1/(9h^2)$ is factored out, the entries of the new matrix G are independent of h. Now

$$G = \left[\begin{array}{cc|cc} A_1 & A_2 & A_3 & -A_4 \\ A_3 & A_4 & A_1 & -A_2 \end{array}\right]_{\mathcal{O}(2N)}, \tag{5.8}$$

where each A_i, $i = 1, 2, 3, 4$, is $(2N) \times (2N)$ and has the same structure

$$A_i = \left[\begin{array}{cc|cc} a_1 & a_2 & a_3 & -a_4 \\ a_3 & a_4 & a_1 & -a_2 \end{array}\right]_{(2N)}. \tag{5.9}$$

The values of the entries of each A_i are given in Table 1, where

$$t = 3 + \sqrt{3}, \quad r = 24 + 18\sqrt{3}, \quad s = 12 + 8\sqrt{3}, \quad q = 24, \quad v = 3 + 2\sqrt{3}, \tag{5.10}$$

and where $\bar{p}$ denotes the "conjugate" of $p = p_1 + p_2\sqrt{3}$, i.e. $\bar{p} = p_1 - p_2\sqrt{3}$.

TABLE 1

Entries of the $(2N) \times (2N)$ *blocks of* (5.8) *and* (5.9)

	a_1	a_2	a_3	$-a_4$
A_1	$-r$	$-s$	q	$-t$
A_2	$-s$	$-v$	$\bar{t}$	0
A_3	q	$\bar{t}$	$-\bar{r}$	$\bar{s}$
$-A_4$	$-t$	0	$\bar{s}$	$-\bar{r}$

We may now apply our theory to this G. One outcome is that of memory economization. If $N = 32$, for instance, and double precision on a 32-bit machine is used (single precision is of no use, in general, for the high order collocation with large N) the EGS scheme reduces the memory requirements from ≈ 7 million bytes of usual banded mode to only ≈ 33 thousand bytes. For more software details see [1].

For a numerical experiment we take $N = 4$. Easily extracted from the proof of Theorem 4.2 are the $4N = 16$ eigenvalues $\mu = 0$ and the $4N = 16$ eigenvalues $\mu = \pm i$ (8 each) of the Jacobi matrix J. The remaining $4N^2 - 32 = 32$ eigenvalues are shown in Table 2, from where we can verify that all have modulus one and that together with each eigenvalue μ we also have the eigenvalues $-\mu$, $1/\mu = \bar{\mu}$ and $-\bar{\mu}$.

TABLE 2
32 of the eigenvalues of J for N = 4

−0.533825 + 0.845595*i*	−0.037216 + 0.999307*i*
−0.533825 − 0.845595*i*	−0.037216 − 0.999307*i*
0.533825 + 0.845595*i*	−0.033636 + 0.999434*i*
0.533825 − 0.845595*i*	−0.033636 − 0.999434*i*
−0.282037 + 0.959404*i*	−0.035643 + 0.999365*i*
−0.282037 − 0.959404*i*	−0.035643 − 0.999365*i*
0.282037 + 0.959404*i*	0.054393 + 0.998520*i*
0.282037 − 0.959404*i*	0.054393 − 0.998520*i*
−0.135009 + 0.990844*i*	0.038500 + 0.999259*i*
−0.135009 − 0.990844*i*	0.038500 − 0.999259*i*
0.135009 + 0.990844*i*	0.033636 + 0.999434*i*
0.135009 − 0.990844*i*	0.033636 − 0.999434*i*
−0.054393 + 0.998520*i*	0.037216 + 0.999307*i*
−0.054393 − 0.998520*i*	0.037216 − 0.999307*i*
−0.038500 + 0.999259*i*	0.035643 + 0.999365*i*
−0.038500 − 0.999259*i*	0.035643 − 0.999365*i*

A geometric interpretation of the proof of Theorem 4.3 for the EJ scheme (replace μ by μ^2 for EGS) is given in Figure 5.

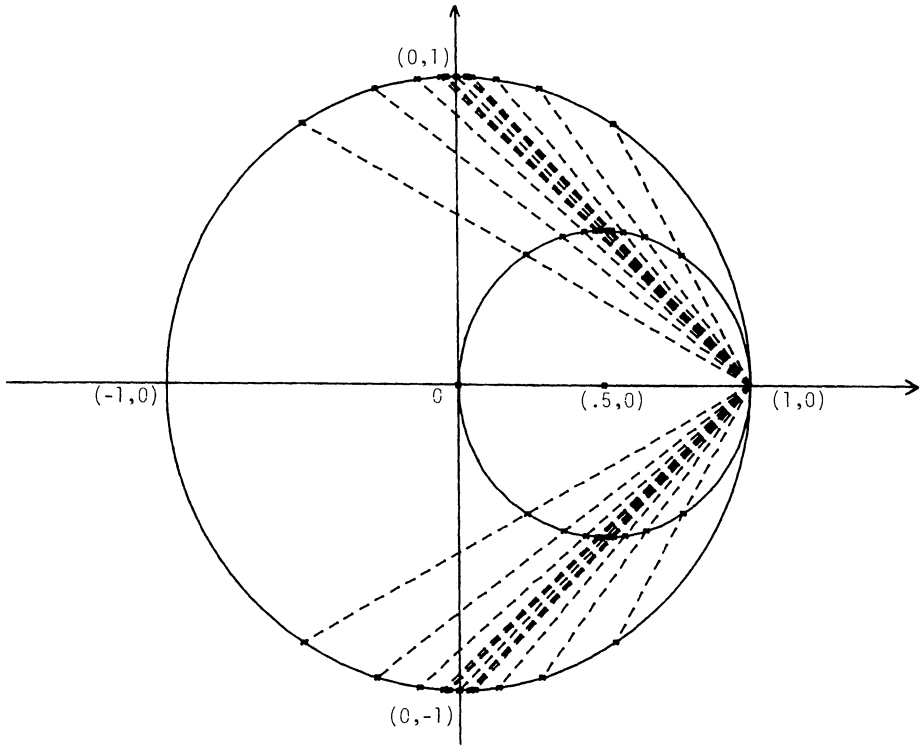

FIGURE 5
Eigenvalues of J and T_{EJ} for N = 4 and $\omega = \omega_b = .5$

On the unit circle we find the eigenvalues μ of J. All the eigenvalues τ of the iteration matrix T_{EJ} lie in the interior of the unit circle and on the smaller circle of radius 0.5 and center $0.5 + 0i$.

There are two kinds of comparisions that we make: For the first, we use two different iterative schemes, EJ and EGS and compare them for the same matrix G and for each $N = 2, 4, 8, 16, \ldots$. By use of (4.7) with $\omega = .5$ we find that the spectral radii of T_{EJ} and T_{EGS} are

$$\text{(5.11)} \qquad \mathrm{sp}_N(T_{EJ}) = \sqrt{\frac{1 + \mathrm{Real}(\mu)_N}{2}}, \qquad \mathrm{sp}_N(T_{EGS}) = \mathrm{Real}(\mu)_N,$$

where $\mathrm{Real}(\mu)_N$ denotes the maximum real part of all the eigenvalues μ of J, for each N, and is always in $[0, 1)$. We conclude that $\mathrm{sp}_N(T_{EGS}) < \mathrm{sp}_N(T_{EJ})$, hence EGS is better than EJ. For $N = 4$, we find from Table 2 that $\mathrm{Real}(\mu)_N \simeq .533825$ and by (5.11)

$$\text{(5.12)} \qquad \mathrm{sp}_4(T_{EJ}) \simeq .875735, \qquad \mathrm{sp}_4(T_{EGS}) \simeq .533825.$$

In fact, we have numerical evidence to conjecture that for all N and for this G, the EGS scheme with $\omega = \omega_b = 0.5$ is the best of all block AOR schemes.

Our second comparison concerns two iterative schemes, used for two different methods of discretizing (5.1)–(5.2). In this case the relative performance of the two discretization methods is taken into account, in that we first prescribe an accuracy, within which the exact solution of the problem is to be approximated by the respective solutions of the two methods. It is assumed that a norm (such as the discrete $\|\cdot\|_\infty$) is specified. We then again compare the spectral radii of the corresponding iteration matrices, taking into account that matrix sizes are, in general, different for the two methods and for the same accuracy. For a numerical example consider (5.1)–(5.2) with

$$f(x, y) = 6xye^xe^y(xy + x + y - 3), \qquad g(x, y) = 0.$$

This example is taken from the "sample problem space" in Houstis et al. [4], where an extensive study of methods for solving more general elliptic problems may be found. The exact solution is $u(x, y) = 3e^xe^y(x - x^2)(y - y^2)$. We consider two methods, collocation and the ordinary 5-point finite difference method. Let N_C and N_F be the respective numbers of grid subdivision. The maximum errors ε_C and ε_F are given in Table 3.

TABLE 3

Maximum error for the finite difference and collocation methods

(*Houstis et al.* [4, *pp.* 343–344])

Finite Difference	N_F	5	7	10	13	15	17
	ε_F	.011000	.005190	.002780	.001630	.001220	.000965
Collocation	N_C	3	4	5	6	7	8
	ε_C	.000448	.000135	.000050	.000028	.000015	.000009

As expected theoretically, we observe that as N_F, N_C grow larger, the numerical results of Table 3 can be interpolated by the standard formulae

$$\varepsilon_F \approx b_F N_F^{-2}, \qquad \varepsilon_C \approx b_C N_C^{-4}, \tag{5.13}$$

where, for this problem, $b_F \approx .28$, $b_C \approx .036$. In order to proceed with the comparison we choose $N_C = 4$, we specify $\varepsilon_C = \varepsilon_F = .000135$ and we determine the corresponding N_F from (5.13) to be

$$N_F \approx \sqrt{b_F/\varepsilon_F} \approx 45.5.$$

Hence, let us use $N_F = 45$, with the observation that small variations in b_F, N_F do not significantly affect the outcome of the comparison.

The well-known SOR scheme (corresponding to the standard block tridiagonal partition) is better than EGS, for the finite difference matrix. Hence, for this matrix, we choose SOR and compare it with EGS for collocation. Let J_F, T_{SOR} be the matrices for the Jacobi and SOR schemes for finite differences. Recall (cf. [7]) that

$$c := \cos(\pi/N_F), \qquad \mathrm{sp}(J_F) = \frac{c}{2-c}, \tag{5.14}$$

$$\omega_b = \frac{2}{1+\sqrt{1-\mathrm{sp}(J_F)^2}}, \quad \mathrm{sp}(T_{\mathrm{SOR}}) = \omega_{b-1}.$$

Using $N_F = 45$, i.e. $c = .997564$, we calculate from (5.14) that

$$\mathrm{sp}(T_{\mathrm{SOR}}) = .820713,$$

which, as compared with the collocation results in (5.12), shows that the SOR scheme for the finite difference method is better than the EJ scheme but worse than the EGS scheme of collocation. If we measure the improvement by the ratio ([7])

$$\frac{-\ln(\mathrm{sp}(T_{\mathrm{EGS}}))}{-\ln(\mathrm{sp}(T_{\mathrm{SOR}}))} \simeq \frac{.627687}{.197582} \approx 3.18,$$

we find that SOR requires about three times more iterations than EGS in order to achieve a prespecified accuracy.

Department of Mathematics and Computer Science
Clarkson College of Technology
Potsdam, New York 13676

1. R. BALART, E. N. HOUSTIS & T. S. PAPATHEODOROU, *On the Iterative Solution of Collocation Method Equations*, Proc. IMACS World Congress, Montreal, Aug. 1982, Vol. 1, pp. 98–100.

2. A. HADJIDIMOS, "Accelerated overrelaxation method," *Math. Comp.*, v. 32 1978, pp. 149–157.

3. A. HADJIDIMOS, "A new method for the solution of linear systems arising from the discretization of P.D.E.'s," *Proc. Advances in Computer Methods for Partial Differential Equations* IV (Vichnevetsky and Stepleman, eds.,), IMACS, 1981, pp. 74–79.

4. E. N. HOUSTIS, R. E. LYNCH, T. S. PAPATHEODOROU & J. R. RICE, "Evaluation of numerical methods for elliptic partial differential equations," *J. Comput. Phys.*, v. 27, 1978, pp. 323–350.

5. T. S. PAPATHEODOROU, "Inverses for a class of banded matrices and applications to piecewise cubic approximation," *J. Comput. Appl. Math.*, v. 8, 1982, pp. 285–288.

6. J. R. RICE, *Matrix Computations and Mathematical Software*, McGraw-Hill Computer Science Series, New York, 1981.

7. R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, N.J., 1982.

8. D. M. YOUNG, *Iterative Solution of Large Linear Systems*, Academic Press, New York, 1981.