SIAM J. NUMER. ANAL.
Vol. 20, No. 1, February 1983

0036-1429/83/2001-0004 $01.25/0

FAST NUMERICAL SOLUTION OF THE BIHARMONIC DIRICHLET PROBLEM ON RECTANGLES*

PETTER BJØRSTAD†

Abstract. A new method for the numerical solution of the first biharmonic Dirichlet problem in a rectangular domain is presented. For an $N \times N$ mesh the complexity of this algorithm is on the order of N^2 arithmetic operations. Only one array of order N^2 and a workspace of size less than $10N$ are required. These results are therefore optimal and the algorithm is an order of magnitude more efficient than previously known methods with the possible exception of multi-grid. The method has an iterative part where a problem with different boundary conditions is used to precondition the original problem. It is shown that any initial error will be reduced by a factor ε after at most $k = \ln(2/\varepsilon)$ iterations using the conjugate gradient method. The conjugate gradient method is also shown to have a superlinear rate of convergence when applied to this formulation of the problem. The purpose of this paper is to provide a description and analysis of the new method.

1. Introduction. Consider the Dirichlet problem for the biharmonic operator in a rectangle R with boundary ∂R:

$$
\begin{aligned}
&\Delta^2 u(x, y) = f(x, y), \qquad && (x, y) \in R,\\
(1)\qquad &u(x, y) = g_1(x, y), && (x, y) \in \partial R,\\
&u_n(x, y) = g_2(x, y), && (x, y) \in \partial R.
\end{aligned}
$$

Here u_n denotes the normal derivative of u with respect to the exterior normal.

In linear elasticity, $u(x, y)$ can represent the Airy stress function or, as in the theory of thin plates, the vertical displacement due to an external force. In fluid mechanics, equation (1) describes the streamfunction of an incompressible two-dimensional creeping flow.

Let R be covered by a uniform mesh with mesh size h. For ease of exposition only, we take the number of interior gridpoints in both coordinate directions equal to an even number N. Thus there are N^2 interior gridpoints and unknowns in the discrete problem. Also, the mesh size h equals $1/(N+1)$ in both coordinate directions. The biharmonic operator is approximated using the 13-point stencil

$$
(2)\qquad \begin{aligned}
\Delta_{13}^2 u &\equiv \frac{1}{h^4}\begin{bmatrix} & & 1 & & \\ & 2 & -8 & 2 & \\ 1 & -8 & 20 & -8 & 1 \\ & 2 & -8 & 2 & \\ & & 1 & & \end{bmatrix} u \\
&= \Delta^2 u + \frac{h^2}{6}(D_1^6 + D_1^4 D_2^2 + D_1^2 D_2^4 + D_2^6)u + O(h^4)
\end{aligned}
$$

where $D_i \equiv \partial/\partial x_i$, $i = 1$ or 2. This stencil is only well defined for gridpoints P having all their (nearest) neighbors in the interior of R. For a gridpoint $P \in R$ with a neighbor

* Received by the editors July 6, 1981, and in revised form March 10, 1982. This research was supported by the Norwegian Research Council for Science and the Humanities, the U.S. Department of Energy under contracts DE-AC02-76ER03077, DE-AT03-76ER71030 and the Office of Naval Research under contract N00014-75-C-1132.

† Courant Institute of Mathematical Sciences, New York University, New York, New York 10012. Presently at Det Norske Veritas, N-1322 Høvik, Oslo, Norway.

$Q \in \partial R$ we use the normal derivative boundary condition at Q to formally get a local $O(h^3)$ accurate extrapolated value at the missing (exterior) point in the stencil (2). This results in a stencil of the form

$$\Delta_q^2 u(P) \equiv \frac{1}{h^4} \begin{bmatrix} & 1 & & \\ 2 & -8 & 2 & \\ -8 & 21 & -8 & 1 \\ 2 & -8 & 2 & \\ & 1 & & \end{bmatrix} u(P) + 2h^{-3} u_n(Q) = \Delta^2 u(P) + O(h^{-1}) \tag{3}$$

when applied to a point P next to the left boundary. It can be shown [4] that the discretization error $\|u - u_h\|$ is of order h^2. A discussion of alternative approximations and some of their properties is given in [3].

The discrete problem can be written as a linear system of algebraic equations

$$\frac{1}{h^4} A u_h = b \tag{4}$$

where the matrix A is defined by the stencils (2) and (3). The elements of the vector b can be computed from the data f, g_1 and g_2 evaluated at the appropriate meshpoints.

Many methods for the numerical solution of this linear system of algebraic equations have been proposed, see for example Bauer and Reiss [2], Buzbee and Dorr [6], Ehrlich [9], [10], [11], Golub [12], Greenspan and Schultz [14], Gupta [15], [16], Jacobs [18], McLaurin [21], Parter [23], Smith [27], [28], [29] and Vajteršic [31]. The operation count for these methods varies between $O(N^{5/2} \log N)$ and $O(N^4)$. They also require more storage than the method proposed in this paper. Furthermore, the methods with the most favorable complexity are all based on the coupled equation approach [21], [31] and the actual computational work is often comparable to, for example, the $O(N^3)$ capacitance matrix method presented in [6]. (See [11].)

In [5] Brandt proposes a multi-grid method for a class of boundary value problems. The solution of the biharmonic problem using this important method is mentioned. The method may have the same $O(N^2)$ complexity, but the analysis in [5] does not seem completely rigorous for this particular problem.

Another well known method is sparse Gaussian elimination with a nested dissection ordering. The complexity of this direct method is $O(N^3)$ arithmetic operations and $O(N^2 \log N)$ storage locations. This and other sparse matrix methods for the given problem were studied and compared in [26]. The study indicates that both constants in the above estimates are quite large and that a regular band solver is very competitive even when the number of unknowns approaches one thousand.

On the basis of previously published algorithms, it was concluded in [24] that the solution of the first biharmonic problem was an order of magnitude more difficult than the solution of Poisson's equation, on parallel computers. Our results show that this is not the case.

We first describe a decomposition of the algebraic system. After a brief description of the algorithm, we shall in this paper concentrate on an analysis of its rate of convergence. The present paper is based on Chapter 3 of the author's Stanford University dissertation [3]. We plan to publish a paper describing efficient computer implementations of the method in the near future.

2. Decomposition of the linear system. We will show how to decompose the linear system (4) in a way that makes an efficient numerical solution possible. While we concentrate on the discrete case and provide a description that is quite close to the computer algorithm, we note that a similar analysis is also possible for the continuous problem. In that case the analysis is related to the solution of the separable problem where Δu, instead of u_n, is specified on two opposite parts of the boundary. We will see in § 4 that the discrete analysis provides precise estimates of the rate of convergence. The matrix A can be represented with the aid of two simple matrices. Thus, let the negative of a one-dimensional discrete Laplace operator be the symmetric, positive definite, tridiagonal $N \times N$ matrix

$$R = \text{tridiag}\,[-1, 2, -1]. \tag{5}$$

Let U be the $N \times 2$ matrix defined by

$$U = [e_1, e_N], \tag{6}$$

where e_i is the ith column of an $N \times N$ identity matrix I. The matrix A can be written

$$A = [(I \otimes R) + (R \otimes I)]^2 + 2(UU^T \otimes I) + 2(I \otimes UU^T). \tag{7}$$

The two last terms in (7) arise from the quadratic boundary extrapolation. The matrix

$$L = (I \otimes R) + (R \otimes I) \tag{8}$$

is the standard 5-point difference approximation of the negative Laplace operator in two dimensions. Let

$$B = L^2 + 2(UU^T \otimes I). \tag{9}$$

The matrix B represents a discrete approximation of the biharmonic problem with Δu specified rather than the normal derivative, on two opposite sides of the rectangle. For this problem separation of the variables is possible. We will show that this problem can be used in a special way, to precondition the original problem.

In the following let $P_{NM} \in R^{nm \times nm}$ be the permutation matrix such that if $D \in R^{m \times n}$, $E \in R^{k \times l}$ then

$$P_{KM}(D \otimes E)P_{LN}^T = (E \otimes D). \tag{10}$$

Notice that if the vector x with components x_{ij}, $i = 1, 2, \cdots, M$, $j = 1, 2, \cdots, N$ is defined on a grid with M rows and N columns, then $P_{MN}x$ is the permuted vector ordered along rows instead of columns. We also need the $N \times N$ orthogonal matrix

$$Q = \{q_{ij}\} = \sqrt{\frac{2}{N+1}}\left\{\sin \frac{ij\pi}{N+1}\right\}. \tag{11}$$

It is easy to show that the vectors q_i, $i = 1, 2, \cdots, N$ are the normalized eigenvectors of R and that

$$\begin{aligned} &QRQ = \Lambda, \\ &Q = Q^T = Q^{-1}, \\ &\Lambda = \text{diag}\,(\lambda_j), \\ &\lambda_j = 2\left(1 - \cos \frac{j\pi}{N+1}\right), \qquad j = 1, 2, \cdots, N. \end{aligned} \tag{12}$$

Q represents a real sine transform and $y = Qx$ can be computed in $O(N \log N)$ operations using the fast Fourier transform.

Using the Sherman–Morrison formula [8],

$$B^{-1}=L^{-2}(I-2(U\otimes I)\tilde{C}^{-1}(U^T\otimes I)L^{-2}), \tag{13}$$

where $\tilde{C}$ is the $2N\times 2N$ matrix

$$\begin{aligned}\tilde{C}&=I+2(U^T\otimes I)L^{-2}(U\otimes I)\\ &=I+2((QU)^T\otimes I)\tilde{S}^{-1}(QU\otimes I)\end{aligned} \tag{14}$$

with

$$\tilde{S}=[(I\otimes R^2)+2(\Lambda\otimes R)+(\Lambda^2\otimes I)]. \tag{15}$$

S is a block diagonal matrix, each block $\tilde{S}_k$, $k=1,\cdots,N$, being pentadiagonal.

THEOREM 1. *The solution of the linear system $\tilde{C}x=y$ can be reduced to the solution of the two linear systems*

$$\tilde{T}_1(x_1+x_2)=y_1+y_2,$$

$$\tilde{T}_2(x_1-x_2)=y_1-y_2,$$

where $x=\binom{x_1}{x_2}$, $y=\binom{y_1}{y_2}$ *and*

$$\tilde{T}_1=I+\frac{8}{N+1}\sum_{k=1,3,\cdots}^{N-1}\sin^2\frac{k\pi}{N+1}\tilde{S}_k^{-1},$$

$$\tilde{T}_2=I+\frac{8}{N+1}\sum_{k=2,4,\cdots}^{N}\sin^2\frac{k\pi}{N+1}\tilde{S}_k^{-1}.$$

Proof. Performing the matrix multiplications in (14) gives

$$\tilde{C}=I+2\begin{pmatrix}\sum_{k=1}^{N} q_{k1}^2\tilde{S}_k^{-1}, & \sum_{k=1}^{N} q_{k1}q_{kN}\tilde{S}_k^{-1}\\ \sum_{k=1}^{N} q_{k1}q_{kN}\tilde{S}_k^{-1}, & \sum_{k=1}^{N} q_{kN}^2\tilde{S}_k^{-1}\end{pmatrix}.$$

Partition the equations according to this block structure and use $q_{kN}=(-1)^{k+1}q_{k1}$. The final result is then obtained by adding and subtracting the block equations. □

Equation (13) shows that the solution of a linear system $Bx=y$ can be obtained if one can solve linear systems with coefficient matrices L and $\tilde{C}$. At most $O(N^2)$ operations are required to solve the linear system $\tilde{C}x=y$. This follows from Theorem 1 and the fact that QT_rQ is diagonal. An important observation is that several $O(N^2)$ methods for solving the discrete Poisson equation (matrix L) are known; see, for example, Bank and Rose [1] and Schröder, Trottenberg and Witsch [25]. Such a method must be used in order to obtain an $O(N^2)$ method for the present problem. Direct specialized algorithms for the solution of linear systems with coefficient matrix B can also be devised. Alternatively, one can proceed using a method based on Fourier transforms, since

$$\begin{aligned}B&=(I\otimes Q)P_{NN}^T[(\Lambda^2\otimes I)+2(\Lambda\otimes R)+(I\otimes R^2)+2(I\otimes UU^T)]P_{NN}(I\otimes Q)\\ &=(I\otimes Q)P_{NN}^TSP_{NN}(I\otimes Q),\end{aligned} \tag{16}$$

where

$$S=\tilde{S}+2(I\otimes UU^T). \tag{17}$$

S is block diagonal and the kth block of S is

$$S_k=\tilde{S}_k+2UU^T. \tag{18}$$

Using this decomposition, the solution takes $O(N^2 \log N)$ operations. Next, consider the formal inversion of the discrete biharmonic operator A

(19) $$A^{-1} = B^{-1}(I - 2(I \otimes U)C^{-1}(I \otimes U^T)B^{-1})$$

where

(20) $$C = I + 2(I \otimes U^T)B^{-1}(I \otimes U).$$

If an $O(N^2)$ algorithm for the solution of a linear system $Cx = y$ can be found, then the discrete biharmonic equation can be solved in $O(N^2)$ operations.

THEOREM 2. *The solution of the linear system $Cx = y$ can be reduced to the solution of the two linear systems,*

$$T_1(x_1 + x_2) = y_1 + y_2,$$
$$T_2(x_1 - x_2) = y_1 - y_2,$$

where

$$P_{2N}x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \qquad P_{2N}y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

and

$$T_1 = I + \frac{8}{N+1} \sum_{k=1,3,}^{N-1} \sin^2 \frac{k\pi}{N+1} S_k^{-1},$$

$$T_2 = I + \frac{8}{N+1} \sum_{k=2,4,}^{N} \sin^2 \frac{k\pi}{N+1} S_k^{-1}.$$

Proof. The proof is similar to that of Theorem 1 with $\tilde{S}$ replaced by S and a permutation P_{2N} of the variables. □

The next theorem provides the basis for the analysis of an iterative method for the solution of linear systems with coefficient matrix T_r, $r = 1, 2$.

THEOREM 3. *The following matrix has the block structure*

$$P_{2N}(Q\tilde{T}_rQ)^{-1/2}QT_rQ(Q\tilde{T}_rQ)^{-1/2}P_{2N}^T = I - \begin{bmatrix} (F^{r1})^T F^{r1} & 0 \\ 0 & (F^{r2})^T F^{r2} \end{bmatrix}$$

for $r = 1, 2$. Moreover, if we let $\psi_{ij} = (\lambda_i + \lambda_j)^2$, $i_r = 2(i-1) + r$, $j_r = 2(j-1) + r$ and

$$\alpha_k^r = 1 + \frac{8}{N+1} \sum_{j=r,r+2,\cdots}^{N} \sin^2 \frac{j\pi}{(N+1)} \Psi_{kj}^{-1},$$

then F^{rs} is the $N/2 \times N/2$ matrix with components

$$f_{ij}^{rs} = \frac{8}{N+1} \sin \frac{i_r\pi}{N+1} \sin \frac{j_s\pi}{N+1} \psi_{i_rj_s}^{-1} / \sqrt{\alpha_{i_r}^s \alpha_{j_s}^r}.$$

Proof. Write

$$QT_rQ = I + \frac{8}{N+1} \sum_{k=r,r+2,\cdots}^{N} \sin^2 \frac{k\pi}{N+1} QS_k^{-1}Q$$

$$= I + \frac{8}{N+1} \sum_{k=r,r+2,\cdots}^{N} \sin^2 \frac{k\pi}{N+1} \Psi_k^{-1}(I - QU(I_2 + 2(QU)^T\Psi_k^{-1}QU)^{-1}(QU)^T\Psi_k^{-1}),$$

where $\Psi_k = \text{diag}(\Psi_{kj})$, $j = 1, 2, \cdots, N$. It is clear that exactly the same kind of calculation that lead to Theorems 1 and 2 can be repeated, this time working with scalars. The calculation is straightforward but tedious. For more details we refer to [3]. □

Observe that the two matrix problems associated with T_r $(r = 1, 2)$ have been split into four smaller problems. This reduces the required computer storage since we can process one problem at a time. The reduction into four subproblems is a consequence of the symmetry of the biharmonic operator on the rectangle R. Each subproblem corresponds to a subspace of the space of biharmonic eigenfunctions. Consider the square $R = \{(x, y): |x| < 1, |y| < 1\}$. The discrete biharmonic eigenfunctions with symmetry around the coordinate axis and symmetry or antisymmetry around the diagonals are generated by the matrix F^{11}, while the eigenfunctions with antisymmetry around the coordinate axis and symmetry or antisymmetry with respect to the diagonals are generated by F^{22}. The matrices F^{12} and F^{21} generate eigenfunctions which are antisymmetric under a rotation of π. This is a degenerate case and for each eigenvalue in this group there are two eigenfunctions of the same shape, one rotated $\pi/2$ relative to the other. A second important observation is that the elements f_{ij}^{rs} can be computed easily after some preprocessing of the quantities that appear in the above formula. This requires only $O(N^2)$ operations and $O(N)$ storage and provides an alternative to the implicit definition of T_r given in Theorem 2.

3. A preconditioned conjugate gradient method. A very attractive iterative scheme for the solution of a symmetric positive definite linear system $Ax = b$ is the conjugate gradient method. From an arbitrary initial vector x_0 the method generates a sequence of approximations $\{x_n\}$ to the solution x defined by

$$x_{n+1} = x_n + \alpha_n p_n, \qquad \alpha_n = (r_n, r_n)/(Ap_n, p_n),$$

$$p_{n+1} = r_{n+1} + \beta_n p_n, \qquad \beta_n = (r_{n+1}, r_{n+1})/(r_n, r_n),$$

where $r_n = b - Ax_n$ and $p_0 = r_0$. The method is due to Hestenes and Stiefel [17]. The iteration does not require knowledge of the matrix elements, since only matrix vector products are needed. It therefore follows from Theorem 2 that this iterative method can be used to solve the linear system $Cx = y$. It can be shown [3] that this method requires $O(N^{1/3})$ iterations resulting in an $O(N^{7/3})$ method for the biharmonic problem.

Suppose, instead of applying the conjugate gradient method directly to a matrix T, that we split T by writing

$$T = \tilde{T} - (\tilde{T} - T).$$

Assume that it is easy to solve linear systems with the matrix $\tilde{T}$. In this case the conjugate gradient method can be used with a preconditioning matrix $\tilde{T}$ corresponding to the above splitting of T. An analysis of this technique is given in Concus, Golub and O'Leary [7]. The process can be viewed equivalently as applying an ordinary conjugate gradient iteration to the transformed system $\tilde{T}^{-1/2} T \tilde{T}^{-1/2}$ with a change of variables. If $\tilde{T}^{-1}$ is an approximate inverse of T, then the convergence rate will be much improved. Two effects can contribute to this. First, the ratio between the largest and the smallest eigenvalue $\mu_{\max}/\mu_{\min}$ is often substantially reduced when we consider $\tilde{T}^{-1/2} T \tilde{T}^{-1/2}$ instead of T. Second, and often more important, the spectrum of $\tilde{T}^{-1/2} T$ $\tilde{T}^{-1/2}$ will usually have a more favorable distribution. Typically, there will be a cluster of eigenvalues close to one, and only a few outlying eigenvalues. We propose to solve the linear systems $T_r x = y$, using $\tilde{T}_r$ as a preconditioning matrix. We will show in the next section that both of the above mentioned effects are prominent in this case.

4. Convergence of the iterative method. We will in this section prove properties about the spectrum of $\tilde{T}_r^{-1}T_r$. This enables us to determine the rate of convergence of the iterative method proposed in the previous section. It follows from Theorem 3 that the spectrum of $\tilde{T}_r^{-1}T_r$ can be studied by considering the singular values of the four matrices F^{rs}, $r=1,2$, $s=1,2$. The following lemma enables us to express the quantity α_k^r, defined in Theorem 3, in closed form.

LEMMA 1. *For* $0<a<1$,

$$4a^2 \sum_{j=1,2,\cdots}^{N} \frac{\sin^2 \dfrac{j\pi}{N+1}}{\left(1+a^2-2a\cos\dfrac{j\pi}{N+1}\right)^2}$$

$$=2(N+1)\left\{\frac{a^2}{1-a^2}-\frac{(a^{N+1})^2}{1-(a^{N+1})^2}\left(\frac{2(N+1)}{1-(a^{N+1})^2}-\frac{1+a^2}{1-a^2}\right)\right\}$$

and

$$4a^2 \sum_{j=2,4,\cdots}^{N} \frac{\sin^2 \dfrac{j\pi}{N+1}}{\left(1+a^2-2a\cos\dfrac{j\pi}{N+1}\right)^2}$$

$$=(N+1)\left\{\frac{a^2}{1-a^2}-\frac{a^{N+1}}{1-a^{N+1}}\left(\frac{N+1}{1-a^{N+1}}-\frac{1+a^2}{1-a^2}\right)\right\}.$$

Proof. Let

$$f(x)=\frac{4a^2\sin^2 x}{(1+a^2-2a\cos x)^2}.$$

Poisson's summation formula gives the relation

$$\frac{1}{2}f(0)+\sum_{k=1}^{N} f\left(\frac{k\pi}{N+1}\right)+\frac{1}{2}f(\pi)=\frac{N+1}{\pi}\left[F_0+2\sum_{k=1}^{\infty} F_{2k(N+1)}\right], \tag{21}$$

where

$$F_k=\int_0^\pi f(x)\cos kx\,dx.$$

Integration by parts reduces this to well known integrals which can be found in [13]. When substituting the result back into (21), we are left with geometric series. The first result is obtained by performing the summations. The second result, where the sum extends over even integers only, follows in the same way by a change of variable. □

The sum over odd integers can now be found as the difference between the two expressions in Lemma 1. Together these results furnish closed form expressions for the individual matrix elements f_{ij}^{rs} defined in Theorem 3. The elements are positive and increase with N. The following important lemma gives the precise form of the limit matrix as the dimension N becomes large.

LEMMA 2. *The matrix* F^{rs} *defined in Theorem* 3 *has elements*:

$$f_{ij}^{rs}=\frac{8}{\pi}\frac{(i_rj_s)^{3/2}}{(i_r^2+j_s^2)^2}\frac{a_i^{rs}a_j^{sr}}{b_i^{rs}b_j^{sr}}+O\left(\frac{1}{(N+1)^2}\right),\qquad r=1,2,\quad s=1,2,$$

where

$$i_r = 2(i-1)+r, \qquad j_r = 2(j-1)+r$$

and a_j^{rs} *and* b_j^{rs} *are exponentially close to one in j and given by*

$$a_j^{rs} = 1+(-1)^{s-1} e^{-j_r\pi},$$

$$b_j^{rs} = (1+2(-1)^{s-1} j_r\pi e^{-j_r\pi} - e^{-2j_r\pi})^{1/2}.$$

Proof. Derive Taylor expansions for each element f_{ij}^{rs} in the variable $1/(N+1)$ around zero. This is tedious to do by hand and the symbolic manipulation program MACSYMA [20] was used to derive the above expressions. □

As an illustration of Lemma 2, the 3×3 leading principal minors of the (infinite) limit matrix F_∞^{rs} are compared with the corresponding minors of F_{63}^{rs} for $N=63$ in Fig. 1. Notice that the approximation is quite good already for this value of N.

$$F_{63}^{11} = \begin{array}{|ccc} .545 & .122 & .038 \\ .122 & .209 & .125 \\ .038 & .125 & .123 \end{array} \qquad F_{\infty}^{11} = \begin{array}{|ccc} .546 & .122 & .039 \\ .122 & .212 & .128 \\ .039 & .128 & .127 \end{array}$$

$$F_{63}^{12} = \begin{array}{|ccc} .319 & .078 & .030 \\ .218 & .167 & .093 \\ .093 & .132 & .108 \end{array} \qquad F_{\infty}^{12} = \begin{array}{|ccc} .320 & .079 & .031 \\ .219 & .169 & .096 \\ .095 & .135 & .112 \end{array}$$

$$F_{63}^{22} = \begin{array}{|ccc} .323 & .144 & .065 \\ .144 & .156 & .107 \\ .065 & .107 & .101 \end{array} \qquad F_{\infty}^{22} = \begin{array}{|ccc} .325 & .146 & .067 \\ .146 & .159 & .111 \\ .067 & .111 & .106 \end{array}$$

FIG. 1. *Leading principal minors of* F_N^{rs} *for* $N=63$ *and* $N=\infty$.

The next lemma can be used to obtain bounds on the row and column sums of F_∞^{rs}.

LEMMA 3. *Let*

$$S_i^r = \sum_{j=1}^{\infty} \frac{(i_r j_r)^{3/2}}{(i_r^2+j_r^2)^2}, \qquad r=1 \text{ or } 2,$$

for some given $i \in (1,2,3,\cdots)$. *Then*

$$\frac{\pi\sqrt{2}}{16} - \frac{1}{i}\frac{25}{32}\left(\frac{3}{5}\right)^{3/4} \leqq S_i^r \leqq \frac{\pi\sqrt{2}}{16} + \frac{1}{i}\frac{25}{32}\left(\frac{3}{5}\right)^{3/4}$$

for all positive i and $r=1$ *or* 2.

Proof. Let

$$S_i = \sum_{j=1}^{\infty} \frac{i^{3/2} j^{3/2}}{(i^2+j^2)^2} = \frac{1}{i}\sum_{j=1}^{\infty}\frac{(j/i)^{3/2}}{(1+(j/i)^2)^2}.$$

Consider

$$f(x) = \frac{x^{3/2}}{(1+x^2)^2}$$

with

$$f_{\max} = f\left(\sqrt{\frac{3}{5}}\right) = \frac{25}{16}\left(\frac{3}{5}\right)^{3/4}, \qquad 0 \leqq x \leqq \infty,$$

and

$$\int_0^\infty f(x)\,dx = \frac{\pi\sqrt{2}}{8}.$$

Clearly

$$\lim_{i\to\infty} S_i = \int_0^\infty f(x)\,dx = \frac{\pi\sqrt{2}}{8}.$$

By considering the discrete sum for finite i and the fact that f is monotone on each side of its maximum, it follows that

$$\frac{\pi\sqrt{2}}{8} - \frac{1}{i} f_{\max} \leqq S_i \leqq \frac{\pi\sqrt{2}}{8} + \frac{1}{i} f_{\max}.$$

Doing the same analysis for the even sum S_{even},

$$S_{\text{even}} = \frac{1}{i} \sum_{j=2,4,\cdots}^{\infty} \frac{(j/i)^{3/2}}{(1+(j/i)^2)^2},$$

results in

$$\frac{\pi\sqrt{2}}{16} - \frac{1}{i} g_{\max} \leqq S_{\text{even}} \leqq \frac{\pi\sqrt{2}}{16} + \frac{1}{i} g_{\max}$$

where the appropriate function is

$$g(x) = \frac{2\sqrt{2}x^{3/2}}{(1+4x^2)^2}, \qquad \int_0^\infty g(x)\,dx = \frac{\pi\sqrt{2}}{16}$$

and

$$g_{\max} = g\left(\sqrt{\frac{3}{20}}\right) = f_{\max}.$$

Combining these two results proves the lemma. □

We can now give the following bound on the singular values of the matrices F^{rs}.

THEOREM 4. *Let $\{\sigma_i\}_{i=1}^N$ be the singular values of one of the matrices F^{rs} defined in Theorem 3. Then*

$$0 \leqq \sigma_i < 0.8,$$

independent of N.

Proof. An upper bound for the largest singular value σ_1 of the matrices F^{rs} will be derived. The following elementary inequaltiy will be used:

$$\sigma_1 \leqq \|(F^{rs})^T F^{rs}\|_\infty^{1/2} \leqq (\|F^{rs}\|_1 \|F^{rs}\|_\infty)^{1/2} = \left[\left(\max_j \sum_i f_{ij}^{rs}\right)\left(\max_i \sum_j f_{ij}^{rs}\right)\right]^{1/2},$$

since all matrix elements are positive. It can be verified by calculation that

$$\sum_{j=1}^{\infty} f_{1j}^{11} < .759,$$

and that this row sum is larger than any other bound that can be obtained for small i (say $i < 20$). Lemma 3 shows that this value certainly cannot be exceeded for any larger i. (The factors a_i^{rs} and b_i^{rs} in Lemma 2 are exponentially small in i and present no difficulties.) □

Computations show that the largest singular value $\sigma_{\max}$ always belongs to F^{11}. A block Lanczos code written by Underwood [30] was used to compute this value for N ranging from 1 to 2047. The results, together with the bound from the proof of Theorem 4, are shown in Fig. 2. It should be noted that the smallest row sum of the matrices F^{rs} can be used to obtain a lower bound on $\sigma_{\max}$ as well. Calculations indicate that $\sigma_{\max} > .7$ as N tends to infinity.

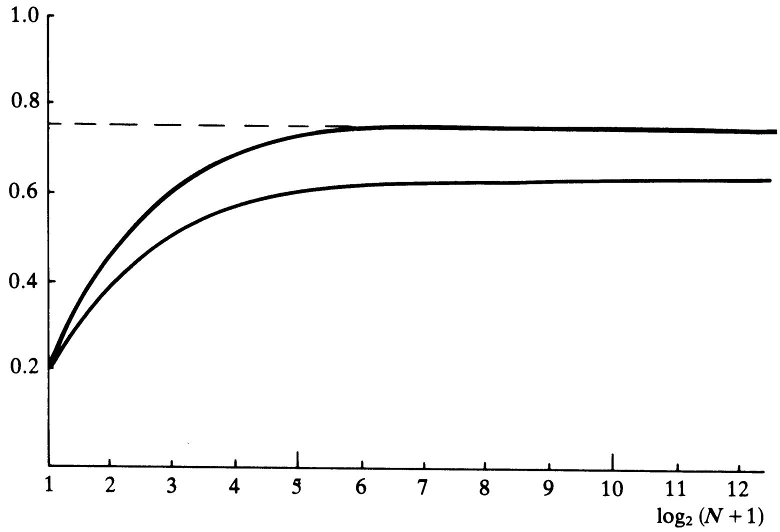

FIG. 2. *The largest singular value as a function of* $\log_2 (N+1)$ (*below*), *compared with the corresponding Gerschgorin bound* (*above*).

The next lemma shows that the singular values σ_i cluster at zero.

LEMMA 4. *The following bounds on the sum of the singular values* $\{\sigma_i\}$ *hold*:

$$\sum_{i=1}^{N/2} \sigma_i < \ln N \quad \text{if } \sigma_i \text{ belongs to } F^{11} \text{ or } F^{22},$$

$$\sum_{i=1}^{N/2} \sigma_i^2 < \ln N \quad \text{if } \sigma_i \text{ belongs to } F^{12} \text{ or } F^{21}.$$

Proof. Consider the matrix F^{11}. Since F^{11} is symmetric, it is sufficient to look at its trace.

$$\sum_{i=1}^{N} F_{ii}^{11} = \sum_{i=1}^{N} \frac{2}{\pi} \frac{1}{i_1} \left(\frac{a_i^{11}}{b_i^{11}} \right)^2 \leqq \frac{1}{\pi} \left(\gamma + \ln N + \delta + O\left(\frac{1}{N}\right)\right)$$

where γ is Euler's constant, $\gamma = .5772 \cdots$ and δ is the contribution from the small term a_i^{11}/b_i^{11}. Letting $N \to \infty$, this shows that the constant in front of the $\ln N$ term in the lemma (taken equal to 1 there) tends to $1/\pi$ as N becomes large. A similar argument gives the same result for F^{22}. It is an obvious conjecture that this result is true also for F^{12}, but since it is of little importance in this context a weaker statement is given. This can be proved by considering $\sum_{ij} (F_{ij}^{12})^2$ (the Frobenius norm of F^{12}). □

We now conclude this section with two theorems describing the rate of convergence of the conjugate gradient iteration proposed in § 3.

THEOREM 5. *If the conjugate gradient algorithm is used to solve the linear system* $T_r x = y$ *with the splitting* $T_r = \tilde{T}_r - (\tilde{T}_r - T_r)$, *then the initial error will be reduced by a factor* ε *after at most*

$$k = \ln\left(\frac{2}{\varepsilon}\right)$$

iterations.

Proof. Let $\mu_1 \leqq \mu_2 \leqq \cdots \leqq \mu_N$ be the eigenvalues of $\tilde{T}_r^{-1} T_r$. It is well known from the theory of the conjugate method [19] that

$$\varepsilon \leqq \frac{1}{T_k\left(\dfrac{\mu_N + \mu_1}{\mu_N - \mu_1}\right)}$$

where T_k is the kth Chebyshev polynomial of the first kind. $T_k(x) = \cosh(k \cosh^{-1} x)$ for $x > 1$. Therefore

$$k \leqq \frac{\cosh^{-1}(1/\varepsilon)}{\cosh^{-1}\left(\dfrac{\mu_N + \mu_1}{\mu_N - \mu_1}\right)}.$$

Using $\cosh^{-1}(1/\varepsilon) < \ln(2/\varepsilon)$, $\mu_1 > 1 - .8^2 = .36$, and $\cosh^{-1}((1+.36)/(1-.36)) > 1$ gives the desired result. □

This theorem establishes convergence to any prescribed accuracy in a constant number of iterations independent of N. Since each iteration takes $O(N^2)$ arithmetic operations, the description of an $O(N^2 \log N)$ algorithm for the first biharmonic problem using (16) is complete. If the accuracy is required to increase with increasing N as N^{-p} for a fixed p, then $O(\log N)$ iterations are required and the overall asymptotic operation count remains unchanged. (In order to be consistent with a decreasing discretization error, p should be 2.)

However, under this assumption the use of an $O(N^2)$ Poisson solver will not make the overall algorithm any faster if the solution on the final grid is computed directly. In order to have an $O(N^2)$ method, it is necessary to compute the solution on a sequence of grids, reducing the error by a fixed amount on each grid. (The total work on all the coarser grids will only be $O(N^2)$.)

For practical computations ($N \leqq 2047$), the use of the computed spectral radius $\sigma_{\max} = .6343$ for $N = 2047$ (see Fig. 2) strengthens the above theorem to

$$k \leqq \frac{1}{2} \ln\left(\frac{2}{\varepsilon}\right).$$

As an illustration, with $\varepsilon = 10^{-10}$ this estimate gives $k \leqq 12$.

The above theorems show that the conjugate gradient iteration converges at a very fast linear rate. The next theorem complements this by showing that asymptotically the rate of convergence is in fact superlinear.

Recall that a sequence $\{e_k\}_{k=0}^{\infty}$ converges R-superlinearly to zero if and only if $\lim_{k\to\infty} \sup \|e_k\|^{1/k} = 0$. An excellent reference discussing the convergence of iterative processes is Ortega and Rheinboldt [22].

THEOREM 6. *The conjugate gradient method defined in Theorem 5 has an R-superlinear rate of convergence.*

Proof. Using the optimality property of the conjugate gradient iteration,

$$\|e_k\| \equiv (c_k)^k \|e_0\| \leqq \max_{\mu \in \{\mu_i\}_{i=1}^N} \prod_{j=1}^{k} \left|\frac{\mu_j - \mu}{\mu_j}\right| \|e_0\|$$

where $\|e_k\|$ is the error in the appropriate norm at iteration k. Let the set $\{\mu_i\}_{i=1}^N$ be ordered such that $\mu_i \leqq \mu_{i+1}$ for all i. Then

$$\begin{aligned}\|e_k\| &\leqq \max_{\mu \in \{\mu_i\}_{i=k+1}^N} \prod_{j=1}^k \left|\frac{\mu_j - \mu}{\mu_j}\right| \|e_0\| \\ &\leqq \max_{\sigma \in \{\sigma_i\}_{i=k+1}^N} \prod_{j=1}^k \frac{\sigma_j^2 - \sigma^2}{1-\sigma_j^2} \|e_0\| \\ &\leqq \prod_{j=1}^k \frac{\sigma_j^2}{1-\sigma_j^2} \|e_0\|.\end{aligned}$$

Using the arithmetic-geometric mean inequality, Lemma 4 and the fact that $\sigma_j < 1$ for all j gives

$$\|e_k\| \leqq \left(\frac{1}{k}\sum_{j=1}^k \frac{\sigma_j^2}{1-\sigma_j^2}\right)^k \|e_0\| \leqq \left(\frac{1}{k}\frac{\ln N}{1-\sigma_1^2}\right)^k \|e_0\|.$$

This inequality shows that the constant

$$c_k \leqq \frac{1}{k}\frac{\ln N}{1-\sigma_1^2}$$

tends to zero as k increases for fixed N.

However, since the concept of R-superlinear convergence is most meaningful in the case of an infinite number of iterations and the conjugate gradient method has finite termination on finite-dimensional problems, consider the limiting case as $N \to \infty$. Lemma 4 implies that $\lim_{k\to\infty} \sigma_k = 0$, and therefore

$$\lim_{k\to\infty} c_k = \lim_{k\to\infty} \left(\frac{1}{k}\sum_{j=1}^k \frac{\sigma_j^2}{1-\sigma_j^2}\right) = 0. \qquad \square$$

5. Acknowledgments. This work is part of the author's Stanford University dissertation. He is most grateful to Professor Gene Golub for stimulating discussions, support and encouragement. The author also acknowledges the use of MACSYMA, a symbolic manipulation system developed at MIT.

REFERENCES

[1] R. E. BANK AND D. J. ROSE, *Marching algorithms for elliptic boundary value problems*, I: *The constant coefficient case*, this Journal, 14 (1977), pp. 782–829.

[2] L. BAUER AND E. L. REISS, *Block five diagonal matrices and the fast numerical solution of the biharmonic equation*, Math. Comp., 26 (1972), pp. 311–326.

[3] P. E. BJØRSTAD, *Numerical solution of the biharmonic equation*, Ph.D. Dissertation, Stanford Univ., Stanford, CA, 1980.

[4] J. H. BRAMBLE, *A second order finite difference analog of the first biharmonic boundary value problem*, Numer. Math., 9 (1966), pp. 236–249.

[5] A. BRANDT, *Multi-level adaptive solutions to boundary-value problems*, Math. Comp., 31 (1977), pp. 333–390.

[6] B. L. BUZBEE AND F. W. DORR, *The discrete solution of the biharmonic equation on rectangular regions and the Poisson equation on irregular regions*, this Journal, 11 (1974), pp. 753–763.

[7] P. CONCUS, G. H. GOLUB AND D. P. O'LEARY, *A generalized conjugate gradient method for the numerical solution of elliptic partial differential equations*, in Proc. Symposium on Sparse Matrix Computations, J. R. Bunch and D. J. Rose, eds., Academic Press, New York, 1976.

[8] G. DAHLQUIST AND Å. BJÖRCK, *Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1974.

[9] L. W. EHRLICH, *Solving the biharmonic equation as coupled finite difference equations*, this Journal, 8 (1971), pp. 278–287.

[10] ———, *Coupled harmonic equations, SOR and Chebyshev acceleration*, Math. Comp., 26 (1972), pp. 335–343.

[11] ———, *Solving the biharmonic equation in a square: A direct versus a semidirect method*, Comm. ACM, 16 (1973), pp. 711–714.

[12] G. H. GOLUB, *An algorithm for the discrete biharmonic equation*, unpublished (see the appendix of L. W. Ehrlich, 1973 [11]).

[13] I. S. GRADSHTEYN AND I. M. RYZHIK, *Table of Integrals, Series and Products*, Academic Press, New York, 1965.

[14] D. GREENSPAN AND D. SCHULTZ, *Fast finite difference solution of biharmonic problems*, Comm. ACM, 15 (1972), pp. 347–350.

[15] M. M. GUPTA, *Discretization error estimates for certain splitting procedures for solving first biharmonic boundary value problems*, this Journal, 12 (1975), pp. 364–377.

[16] M. M. GUPTA AND R. P. MANOHAR, *Direct solution of the biharmonic equation using noncoupled approach*, J. Comp. Phys., 33 (1979), pp. 236–248.

[17] M. R. HESTENES AND E. STIEFEL, *Methods of conjugate gradients for solving linear systems*, Nat. Bur. Standards J. Research, 49 (1952), pp. 409–436.

[18] D. A. JACOBS, *The strongly implicit procedure for biharmonic problems*, J. Comp. Phys., 13 (1973), pp. 303–315.

[19] D. G. LUENBERGER, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, 1973.

[20] The Mathlab group, *MACSYMA Reference manual*, version 9, Laboratory for Computer Science, Mass. Institute of Technology, Cambridge, MA. July 1977.

[21] J. W. MCLAURIN, *A general coupled equation approach for solving the biharmonic boundary value problem*, this Journal, 11 (1974), pp. 14–33.

[22] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[23] S. V. PARTER, *On two line iterative methods for the Laplace and biharmonic difference equations*, Numer. Math., 1 (1959), pp. 240–252.

[24] A. H. SAMEH, S. C. CHEN AND O. J. KUCK, *Parallel Poisson and biharmonic solvers*, Computing, 17 (1976), pp. 219–230.

[25] J. SCHRÖDER, U. TROTTENBERG AND K. WITSCH, *On fast Poisson solvers and applications*, in Proceedings of a Conference on Numerical Treatment of Differential Equations, Lecture Notes in Mathematics, 631, Springer-Verlag, Berlin, 1978, pp. 153–187.

[26] A. H. SHERMAN, *On the efficient solution of sparse systems of linear and nonlinear equations*, Ph.D. Thesis, Dept. Computer Science, Yale Univ., 1975.

[27] J. SMITH, *The coupled equation approach to the numerical solution of the biharmonic equation by finite differences*, I, this Journal, 5 (1968), pp. 323–339.

[28] ———, *The coupled equation approach to the numerical solution of the biharmonic equation by finite differences*, II, this Journal, 7 (1970), pp. 104–111.

[29] ———, *On the approximate solution of the first boundary value problem for $\nabla^4 u = f$*, this Journal, 10 (1973), pp. 967–982.

[30] R. R. UNDERWOOD, *An iterative block Lanczos method for the solution of large sparse symmetric eigenproblems*, Ph.D. Thesis, Stanford Univ., Stanford, CA. 1975.

[31] M. VAJTERŠIC, *A fast algorithm for solving the first biharmonic boundary value problem*, Computing, 23 (1979), pp. 171–178.