Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

SIAM J. NUMER. ANAL.
Vol. 5, No. 3, September 1968
Printed in U.S.A.

ITERATIVE SOLUTION OF IMPLICIT APPROXIMATIONS OF MULTIDIMENSIONAL PARTIAL DIFFERENTIAL EQUATIONS*

HERBERT L. STONE†

Summary. A new iterative method has been developed for solving the large sets of algebraic equations that arise in the approximate solution of multidimensional partial differential equations by implicit numerical techniques. This method has several advantages over those now in use. First, its rate of convergence does not depend strongly on the nature of the coefficient matrix of the equations to be solved. Second, it is not sensitive to the choice of iteration parameters, and as a result, suitable parameters can be estimated from the coefficient matrix. Finally, it reduces significantly the computational effort needed to solve a set of equations. For a typical set of 961 equations, it was found to reduce the number of calculations by a factor of three, when compared to the most competitive of the older methods. It is expected that this advantage will be even greater for larger sets of equations.

1. Introduction. Approximate solutions of multidimensional differential equations often are obtained by the application of implicit finite difference analogues. A difference equation is written for each grid point in the region of interest, and the resulting set of simultaneous equations must be solved for each time step.

Such sets of equations can be solved directly by elimination or by one of several iterative methods, such as relaxation, successive overrelaxation, or ADI (alternating direction iteration). The purpose of this paper is to describe a new iterative procedure that converges much faster than any of these methods.

The simplest method of solving these sets of equations is direct solution by elimination. Although this approach is the most efficient method available for small sets of equations, it is not for large sets. The procedure requires $2n^2$ arithmetic operations to solve n equations of the type being considered. When n becomes relatively large, it is more efficient to use an iterative procedure for solving the equations. In addition to computational efficiency, iterative procedures possess other advantages over the elimination method. In computer applications they require much less memory for for storage of intermediate data, and they are easier to program. Further, they are frequently applicable to nonlinear sets of equations, whereas solution by elimination is not. As a result of these several advantages, iterative procedures are generally preferred for solution of the moderate to large sets of equations encountered in this type of problem.

The most elementary iterative method available is the relaxation

* Received by the editors November 30, 1967.

† ESSO Production Research Company, Houston, Texas 77001.

method [12], also known as the Richardson [11] or point-Jacobi method [5]. This procedure is quite effective when applied manually to small problems where human judgment can continually be used to redirect the course of the calculations. It is much less effective in computer applications where it must be formalized into a cyclic, invariant procedure. For even the simplest problem, corresponding to flow of heat in a rectangular region of uniform conductivity (hereafter called the model problem), this method is quite slow and requires a computational effort proportional to the square of the number of equations in the set to be solved. This proportionality is exactly the same as that resulting from direct solution by elimination.

The Gauss-Seidel iterative method [7] is closely related to solution by successive overrelaxation [8], [15]. These two procedures are also known, respectively, as the Liebmann and extrapolated Liebmann methods [4], [6]. Each of these two methods involves the use of a single iteration parameter, and they differ only by the choice of this parameter. For the model problem, the Gauss-Seidel method can be shown to be simply a factor of two better than point-Jacobi iteration, regardless of the number of equations being solved. However, successive overrelaxation is significantly faster for this special problem and the amount of computing required is proportional to $n^{3/2}$, if the best choice of iterative parameter is made. Unfortunately, these good characteristics frequently do not apply to the more complex problems encountered in practice, where conductivity may vary from point to point in the region of interest.

ADI, the alternating direction iteration, was first proposed by Peaceman and Rachford [10], and analyzed by Douglas [2] in 1955. For the model problem, it is a great deal more efficient than either point-Jacobi iteration or overrelaxation. For this problem, the work requirement can be shown to be proportional to $n \ln n$. This is very near to the ultimate relationship, which would be simply work proportional to n, the number of equations to be solved. But again this efficiency extends only in part to more complex problems; in extreme cases, solutions obtained by ADI converge very slowly or not at all to the exact solution of the set of equations being solved. One reason for this slow convergence is that ADI requires the selection of a *set* of iteration parameters to be applied cyclically during the iteration. For the model problem there exists a theoretical basis for selection of a set that will give rapid convergence, but for the general case no practical basis exists. Thus, ADI frequently is applied with nonoptimal sets of iteration parameters.

A qualitative explanation of the higher convergence rates achieved by ADI is that this method is more implicit than the point-Jacobi or overrelaxation techniques. Stated another way, each step of ADI is more closely related to *direct* solution by elimination than are the steps of the other two

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

procedures. The new method described in this paper is even more strongly implicit than ADI, and its use results in convergence rates significantly greater than can be achieved by the ADI technique for all but the simplest problems. ADI converges somewhat faster when corresponding coefficients in each equation are equal and the region being considered is rectangular. But for nonrectangular regions, or when corresponding coefficients are widely different, ADI and many other well-known methods lose their effectiveness, whereas the new method does not. It is on these very difficult problems that the new method will be most useful. In some instances it has yielded solutions to problems for which ADI did not yield solutions.

The subject method is related to those of Oliphant [9] and Buleev [1] in that it involves the solution of a sparse matrix problem by Gaussian elimination. In all three approaches the matrices are such that their sparseness causes the computational work required for this solution to be small. However, each of the three methods employs different definitions of the sparse matrix, and only the method described in this paper employs several such matrices. It has been found that this difference is essential for rapid convergence in a wide class of problems.

Although computing experience with the subject method has been very encouraging, it has thus far defied rigorous analysis. During the time that this paper was being written, Dupont, Kendall and Rachford [3] have considered analysis of this and related methods. While little of consequence concerning the present method resulted, they were able to obtain significant results for a method which utilizes Buleev's definition of the sparse matrix, provided the problem to be solved is characterized by a symmetric matrix. In their method, a Chebyshev sequence of parameters is used, and analysis is used to relate the best set of parameters to the eigenvalues of the matrix. While the use of these parameters leads to a good rate of convergence, in practice this is helpful only if reliable estimates of the required eigenvalues can be obtained by the expenditure of a relatively small computational effort. In general, such is not the case. An additional point of interest is that their analysis, which permits the calculation of the parameters from the eigenvalues, is applicable only to symmetric matrices, whereas the subject method is applicable whether or not the matrix is symmetric.

The next section of this paper describes the nature of the sets of equations to be solved and then the new procedure is described. Following this description is a section evaluating the new method by comparing the convergence rates it achieves with those obtained by the older methods.

2. The problem. Heat conduction in a two-dimensional region is an example of a typical problem for which the subject method is applicable.

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

This heat conduction can be either transient or steady state. Since the set of simultaneous equations arising from the steady state problem is the more difficult to solve, it will be used as an illustrative problem. Equation (1) describes the temperature distribution for this case of steady state heat conduction:

$$\frac{\partial}{\partial x}\left[KX\frac{\partial T}{\partial x}\right] + \frac{\partial}{\partial y}\left[KY\frac{\partial T}{\partial y}\right] = -Q. \tag{1}$$

In this equation, x and y are the distance coordinates. The quantity Q is the strength of the local heat source, and KX and KY are the thermal conductivities in the x and y-directions, respectively. Each of the quantities Q, KX and KY is a known function of x and y. Under these conditions, (1) serves to define the temperature, T, as a function of x and y. Boundary conditions are that no heat is conducted across the boundaries of the region.

Equation (1) may be approximated by a suitable finite difference equation. In this illustrative problem a rectangular grid system will be used, but it should be understood that regions of any shape may be treated by use of such a grid simply by specifying zero values of the conductivity at the appropriate points. In fact, one of the significant features of the subject method is that the convergence rate is not seriously decreased by the consideration of irregularly shaped regions. The rectangular grid system is defined by the equations

$$x_j = j\Delta x, \quad 0 \leqq j \leqq J, \quad \Delta x = \frac{1}{J},$$

$$y_k = k\Delta y, \quad 0 \leqq k \leqq K, \quad \Delta y = \frac{1}{K}.$$

A point on this grid system will be indicated as (x_j, y_k), or more simply, as (j, k). With the convention that $T_{j,k} \equiv T(x_j, y_k)$, (1) is approximated by

$$\begin{aligned}&\frac{KX_{j+1/2,k}}{\Delta x^2}(T_{j+1,k} - T_{j,k}) - \frac{KX_{j-1/2,k}}{\Delta x^2}(T_{j,k} - T_{j-1,k})\\ &+ \frac{KY_{j,k+1/2}}{\Delta y^2}(T_{j,k+1} - T_{j,k}) - \frac{KY_{j,k-1/2}}{\Delta y^2}(T_{j,k} - T_{j,k-1})\\ &= -\frac{q_{j,k}}{\Delta x \Delta y}.\end{aligned} \tag{2}$$

Equation (2) relates the temperature at a point (j, k) to the temperatures at the four nearest grid points.

Equation (2) can be simplified in form to yield

$$B_{j,k}T_{j,k-1} + D_{j,k}T_{j-1,k} + E_{j,k}T_{j,k} + F_{j,k}T_{j+1,k} + H_{j,k}T_{j,k+1} = q_{j,k}, \tag{3}$$

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

where

$$B_{j,k} = -\frac{KY_{j,k-1/2}(\Delta x)}{\Delta y}$$

and other coefficients are similarly defined.

Equations for the boundary points can also be put in the form of (3). For example, at the $y = 0$ boundary it is required that

$$\frac{\partial T}{\partial y} = 0.$$

This condition can be approximated by the well-known reflection technique which sets $T_{j,k-1} = T_{j,k+1}$ and $B_{j,k} = H_{j,k}$. Then $T_{j,k-1}$ in (3) can be eliminated in favor of $T_{j,k+1}$, yielding an equation similar to (3), but with one term missing. The result is that (3) written for this point has $B_{j,k} = 0$ and $H_{j,k}$ is double its normal value. Similar remarks apply at other boundaries.

In (3), known quantities are represented by $B_{j,k}$, $D_{j,k}$, $E_{j,k}$, $F_{j,k}$, $H_{j,k}$ and $q_{j,k}$. The T's represent unknown quantities. Since one such equation exists for each grid point (j, k), there is a total of

$$n = (J + 1)(K + 1)$$

equations in the n unknown temperatures. In the transient problem, one such set arises at each time step.

In the discussions to follow, the set of equations will be sequenced in the order of increasing values of j and k, with k being held constant at each of its values until j has taken on all values in its range. Thus the first $J + 1$ equations of the set correspond to $k = 0$, with j taking on values of 0, 1, 2, $\cdots$ and J. The second $J + 1$ equations are formed by setting $k = 1$ and letting j take on values from 0 to J. This continues until all $(J + 1)(K + 1)$ equations are formed.

For convenience in describing the development of the iterative procedure, the set of equations is written in matrix notation,

$$M\mathbf{T} = \mathbf{q}. \tag{4}$$

The expanded form of (4) is given in Fig. 1. The matrix M is a $(J + 1) \cdot (K + 1)$ square array composed of the coefficients of the temperatures in the sequenced equations. $\mathbf{T}$ is a vector composed of the sequenced temperatures, and $\mathbf{q}$ is a vector composed of the sequenced source strengths.

Each row of the matrix M consists of the coefficients of the unknown temperatures in one equation of the set. It is convenient to use the notation of (3) in M, so that each row of this matrix contains at most five nonzero elements; these are $B_{j,k}$, $D_{j,k}$, $E_{j,k}$, $F_{j,k}$ and $H_{j,k}$. It is important to note

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

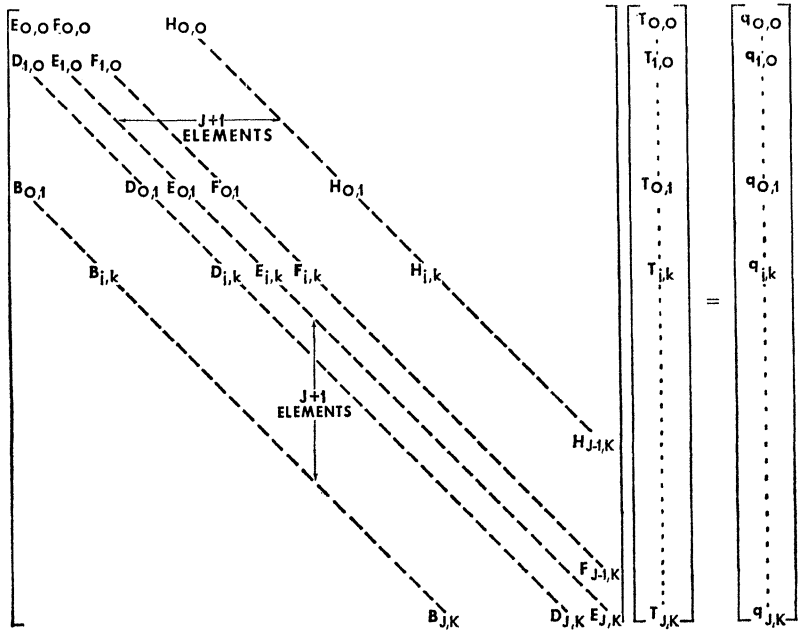

FIG. 1. *Expression of simultaneous equations in matrix notation*

that this notation does not conform to standard matrix notation; the subscript j, k refers to the grid system used in setting up the difference equations, rather than to location within the matrix.

Three of the five nonzero elements of each row of M are located on the principal diagonal and the two adjacent diagonals. The other two elements are located on diagonals situated $J + 1$ locations away on each side of the principal diagonal.

For the illustrative problem cited, the matrix M can be made symmetric. However, as noted in the Introduction, the technique being described in this paper may be applied to either symmetric or asymmetric matrix problems.

The problem being considered is how to solve a set of equations of the form of (4). The next section describes a way of achieving this solution.

3. Method of solution. Solution of large sets of equations like (4) by Gaussian elimination requires excessive computational effort. The iterative procedure which is the subject of this paper involves the solution by elimination of a set of equations formed by a modification of the original set, (4). The reasons why elimination is not satisfactory for the original set will be considered first, and this will motivate the way in which the original matrix may be modified to make elimination easy. It develops that there

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

are many possible modifications which will achieve this goal, and so one of these will be sought which will yield rapid convergence of the iterative procedure. The specification of the nature of this effective modification will be followed by a complete description of the new iterative method.

3.1. Direct solution by elimination. The first step of the solution of (4) by direct elimination is equivalent to factoring the matrix M into the product of a lower triangular matrix L' and an upper triangular matrix U', that is, $M = L'U'$. These triangular matrices, like M, will be square and of dimensions $(J + 1)(K + 1)$. In general, L' will have nonzero elements from the diagonals corresponding to the B-diagonal to the E-diagonal of Fig. 1. The matrix U' will have nonzero elements from the E- to the H-diagonal. Whereas each row of M has at most five nonzero elements, L' has $J + 1$ such elements, and U' has $K + 1$. In the elimination procedure, each of these elements of L' and U' must be computed and stored for later use. Therefore, for each point, approximately $J + K$ such elements must be computed. It is the generation of the large number of intermediate coefficients which makes elimination slow. However, it will be shown in the next section how the set (4) can be modified to make it amenable to direct solution.

3.2. Alteration of matrix M to achieve easier direct solution. The altered form of matrix M, which will be designated as $M + N$, is such that when $M + N$ is factored into a product LU, L and U have only three nonzero elements in each row, regardless of the size of J or K. Thus the work of solving this modified set is simply proportional to n, the number of equations in the set, and the constant of proportionality can be as low as six.

To show how such a modified matrix may be formed, the product matrix resulting from the multiplication LU will be considered. Fig. 2 depicts the definition of L, and Fig. 3 the definition of U. Fig. 2 indicates that L has nonzero elements in the diagonals corresponding to the B, D and E-diagonals of M, shown in Fig. 1. The matrix U has nonzero elements in those diagonals corresponding to E, F and H, with that corresponding to E, the principal diagonal, being everywhere equal to unity. To distinguish the elements of L and U from those of M, lower case symbols have been used in the first two, capitals in the latter. Note also that in all these matrices the subscripts refer to the grid-point system (j, k), and not to location within the matrix.

The product of premultiplying U by L is shown in Fig. 4. This resulting matrix, $M + N$, has seven nonzero diagonals, including five in locations corresponding to those of M, plus two which fall just inside the B- and H-diagonals. The elements of L and U cannot be selected in such a way that M is identical with $M + N$. This can be seen by noting that if the two

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

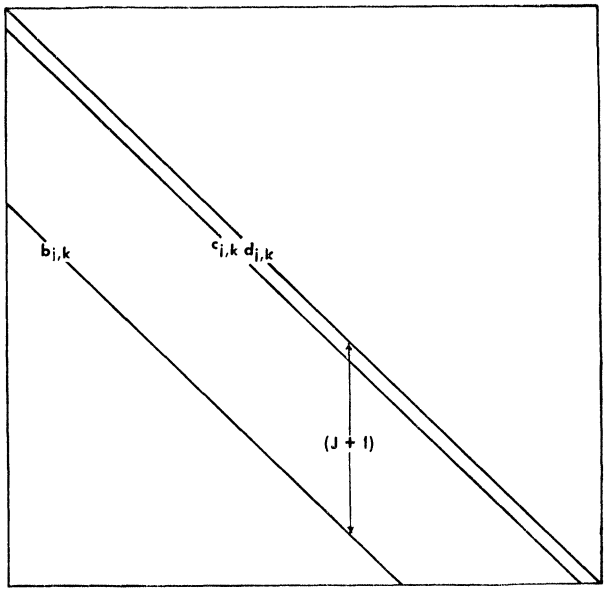

FIG. 2. *Lower triangular matrix L*

matrices are equal, corresponding elements of each must be equal. This would lead to the following set of relations for each grid point (j, k) (which corresponds to a row of each matrix):

$$b_{j,k} = B_{j,k}, \tag{5a}$$

$$b_{j,k}e_{j,k-1} = 0, \tag{5b}$$

$$c_{j,k} = D_{j,k}, \tag{5c}$$

$$d_{j,k} + b_{j,k}f_{j,k-1} + c_{j,k}e_{j-1,k} = E_{j,k}, \tag{5d}$$

$$d_{j,k}e_{j,k} = F_{j,k}, \tag{5e}$$

$$c_{j,k}f_{j-1,k} = 0, \tag{5f}$$

$$d_{j,k}f_{j,k} = H_{j,k}. \tag{5g}$$

In these seven equations, the capitalized coefficients are all known elements of M. A little reflection will show that all lower case coefficients with subscripts less than either j or k will have been found by applying the algorithm (5) to the grid point $(j - 1, k)$ or $(j, k - 1)$. This leaves but the five (j, k) subscripted coefficients undetermined, and so the seven relationships (5a)–(5g) cannot all be satisfied.

The simplest possible definition of $M + N$ which can be factored into

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

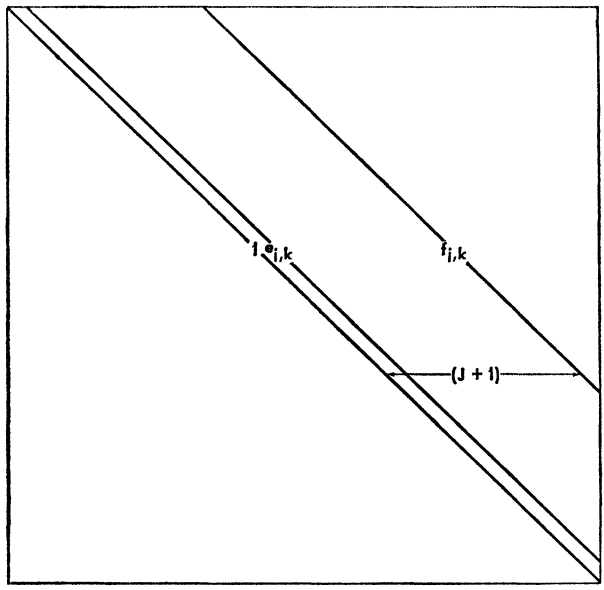

FIG. 3. *Upper triangular matrix U*

the LU form would result from ignoring (5b) and (5f) and using the remaining five relations to obtain the five (j, k) subscripted coefficients. This would result in the five nonzero diagonals in M being identical with the corresponding ones in $M + N$, but $M + N$ would contain two additional nonzero diagonals. This definition would achieve the objective that $M + N$ be easily solved by elimination, but it was found that such a matrix could not be used as the basis of a rapidly convergent iteration procedure. However, the above definition of the modified matrix is but one of a family of matrices which can be efficiently solved by elimination. This family is defined as follows:

Let

$$C_{j,k} = b_{j,k}e_{j,k-1} \quad \text{and} \quad G_{j,k} = c_{j,k}f_{j-1,k} . \tag{6}$$

Then the right-hand sides of each equation in (5), with the exceptions of (5b) and (5f), may be modified by adding linear combinations of $C_{j,k}$ and $G_{j,k}$. For example, (5a) could be modified to result in (7):

$$\begin{aligned} b_{j,k} &= B_{j,k} + \gamma_1 C_{j,k} + \gamma_2 G_{j,k} \\ &= B_{j,k} + \gamma_1 b_{j,k}e_{j,k-1} + \gamma_2 c_{j,k}f_{j-1,k} . \end{aligned} \tag{7}$$

Equation (7) is linear in the two unknowns $b_{j,k}$ and $c_{j,k}$, as will be the modified version of (5c). Thus these two equations may be solved simul-

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

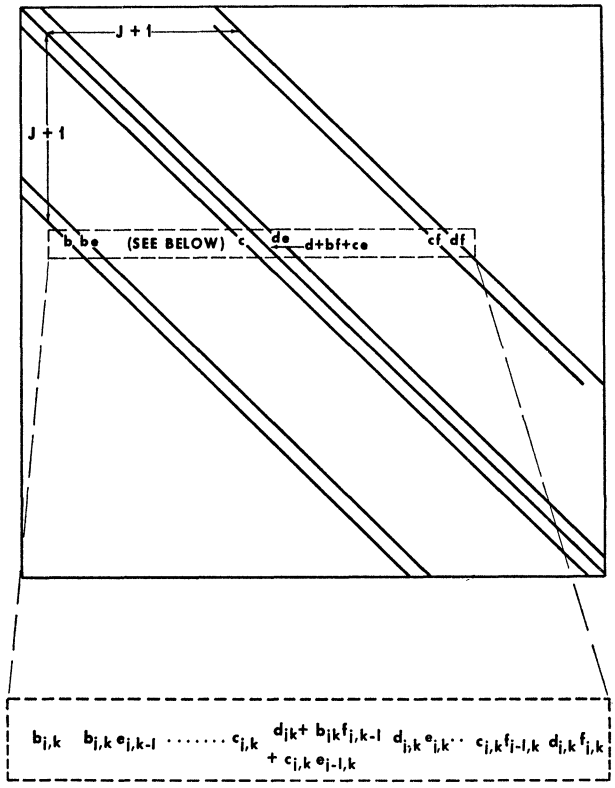

FIG. 4. *Product matrix* $M + N$

taneously to yield $b_{j,k}$ and $c_{j,k}$. Then the equations corresponding to (5d), (5e) and (5g) may be solved explicitly for $d_{j,k}$, $e_{j,k}$ and $f_{j,k}$. There are limits on the values of the constants γ_1 and γ_2 which can be employed, which depend upon the problem being solved. The family of possible definitions of $M + N$ results from assigning different values to the constants γ_1, γ_2 and their counterparts in the other four equations.

3.3. Alteration of matrix M to achieve rapid convergence of iteration. Attention is now focused upon selecting, from the family of easily factorable definitions of $M + N$, one which will result in rapid convergence if used as the basis of an iterative method.

Fig. 5 shows the grid-point system in the neighborhood of the point (j, k). Equation (3) for this point involves nonzero coefficients for the temperatures at (j, k) and its four nearest neighbors, as indicated by the solid circles on Fig. 5. The modified equation, corresponding to the matrix $M + N$, has nonzero coefficients not only for these temperatures, but also

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

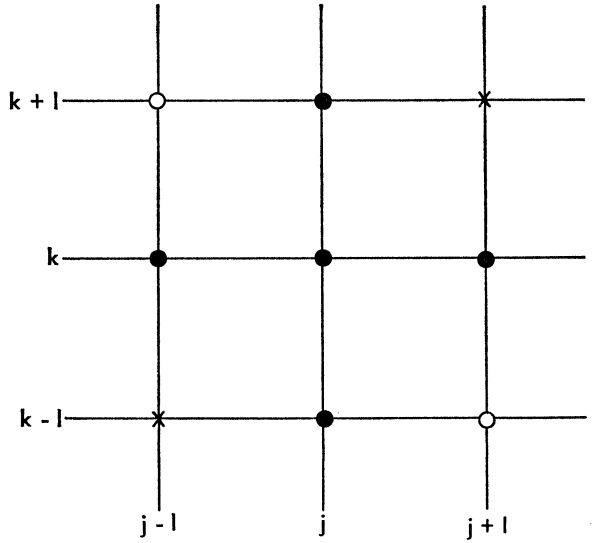

FIG. 5. *Temperatures involved in temperature equation of point* (j, k)

for temperatures at points $(j - 1, k + 1)$ and $(j + 1, k - 1)$. These points are indicated as open circles on Fig. 5. To minimize the influence of these new terms introduced in forming the modified equation, it is useful to balance them partially by subtracting approximately equal terms. Suitable terms involve only the five temperatures in the original equation and are obtained from a Taylor series expansion of the temperature in the vicinity of the point (j, k). Equation (8) is obtained by writing a Taylor series for point $T_{j-1,k+1}$ and subtracting from it similar expressions for $T_{j,k+1}$ and $T_{j-1,k}$, neglecting terms of the order of Δx^3, Δy^3, $(\Delta x \cdot \Delta y)$ and all higher order terms. Equation (9) is derived in a similar manner.

$$T_{j-1,k+1} = -T_{j,k} + T_{j,k+1} + T_{j-1,k}, \tag{8}$$

$$T_{j+1,k-1} = -T_{j,k} + T_{j+1,k} + T_{j,k-1}. \tag{9}$$

Since the $(\Delta x \cdot \Delta y)$ term which is neglected in arriving at (8) and (9) is of the same magnitude as the Δx^2 and Δy^2 terms, it would appear desirable to avoid neglecting this term. Such an approach will result in the temperature at additional grid points being involved in the analogues of (8) and (9). Approaches of this nature were tried, but without success.

If grid points are sufficiently close together and temperature variations between points are sufficiently smooth, (8) and (9) are good approximations. For the more general case, it proves useful to introduce additional flexibility into the method by multiplying the right-hand sides of (8) and (9)

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

by a variable iteration parameter, α, before considering them to be approximations of $T_{j-1,k+1}$ and $T_{j+1,k-1}$. If $0 < \alpha < 1$, then wherever the matrix modification introduces the temperature $T_{j-1,k+1}$ into the individual grid-point equation, this temperature is partially "canceled" by subtracting $\alpha(-T_{j,k} + T_{j,k+1} + T_{j-1,k})$. It would be possible to use an iteration parameter different from α in partially "canceling" $T_{j+1,k-1}$. However, as there is no obvious reason for doing this, the same parameter, α, will be employed for $T_{j+1,k-1}$. Using this approach, the left-hand, or unknown, side of (3) is modified to become (10):

$$\begin{aligned} B_{j,k}T_{j,k-1} + D_{j,k}T_{j-1,k} + E_{j,k}T_{j,k} + F_{j,k}T_{j+1,k} + H_{j,k}T_{j,k+1} \\ + C_{j,k}[T_{j+1,k-1} - \alpha(-T_{j,k} + T_{j+1,k} + T_{j,k-1})] \\ + G_{j,k}[T_{j-1,k+1} - \alpha(-T_{j,k} + T_{j,k+1} + T_{j-1,k})], \end{aligned} \tag{10}$$

where

$$C_{j,k} = b_{j,k}e_{j,k-1} \quad \text{and} \quad G_{j,k} = c_{j,k}f_{j-1,k}.$$

The first five terms in (10) are simply the left-hand side of the original temperature equation for point (j, k). The terms which multiply $C_{j,k}$ and $G_{j,k}$ are seen to be the partially canceled temperatures at $(j - 1, k + 1)$ and $(j + 1, k - 1)$ and so should be relatively small. Collecting the terms in (10) which involve $T_{j,k-1}$, there results $(B_{j,k} - \alpha C_{j,k})$. This is the element of $M + N$, Fig. 4, which corresponds to $b_{j,k}$, and hence

$$b_{j,k} = B_{j,k} - \alpha C_{j,k}.$$

Additional relations are obtained in a similar fashion and result in

$$\begin{aligned} c_{j,k} &= D_{j,k} - \alpha G_{j,k}, \\ d_{j,k} + b_{j,k}f_{j,k-1} + c_{j,k}e_{j-1,k} &= E_{j,k} + \alpha C_{j,k} + \alpha G_{j,k}, \\ d_{j,k}e_{j,k} &= F_{j,k} - \alpha C_{j,k}, \\ d_{j,k}f_{j,k} &= H_{j,k} - \alpha G_{j,k}. \end{aligned} \tag{11}$$

Equations (11) may be seen to be of the form of equations (7), which can be solved easily by elimination, and it also has been found to define a matrix $M + N$ which is effective if used in an iteration scheme.

3.4. The iterative method. With the matrix $M + N$ thus defined, the iterative method is derived by adding $N\mathbf{T}$ to both sides and $(M\mathbf{T} - M\mathbf{T})$ to the right-hand side of (4), to give:

$$(M + N)\mathbf{T} = (M + N)\mathbf{T} - (M\mathbf{T} - \mathbf{q}). \tag{12}$$

Since $M + N$ is easily factored, the left-hand side of (12) can be effi-

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

ciently solved for $\mathbf{T}$ if the right-hand side is known. Equation (12) therefore provides the basis for the general statement of the iteration procedure when written in the form[1]

$$(M + N)\mathbf{T}^{n+1} = (M + N)\mathbf{T}^n - (M\mathbf{T}^n - \mathbf{q}). \tag{13}$$

Values of $\mathbf{T}$ for the $(n + 1)$st iterative level can thus be calculated from a knowledge of $\mathbf{T}$ at the nth level.

It is beneficial from the standpoint of roundoff error [14] to change the form of (13) by defining the following vectors:

$$\boldsymbol{\delta}^{n+1} = \mathbf{T}^{n+1} - \mathbf{T}^n$$

and

$$\mathbf{R}^n = \mathbf{q} - M\mathbf{T}^n.$$

The quantity $\boldsymbol{\delta}^{n+1}$ is simply the change of the temperature in going from iteration level n to $n + 1$, and $\mathbf{R}^n$ is the residual at iteration level n.

The (j, k)th element of $\mathbf{R}^n$ is defined by (14):

$$\begin{aligned} R^n_{j,k} = q_{j,k} - (&B_{j,k}T^n_{j,k-1} + D_{j,k}T^n_{j-1,k} \\ &+ E_{j,k}T^n_{j,k} + F_{j,k}T^n_{j+1,k} + H_{j,k}T^n_{j,k+1}). \end{aligned} \tag{14}$$

Using the above definitions, (13) can be put in the following form:

$$(M + N)\boldsymbol{\delta}^{n+1} = \mathbf{R}^n. \tag{15}$$

Replacing $M + N$ by LU results in

$$LU\boldsymbol{\delta}^{n+1} = \mathbf{R}^n. \tag{16}$$

The elements of L and U are computed from the elements of M by the use of (11). A vector $\mathbf{V}$ is defined by (17):

$$L\mathbf{V} = \mathbf{R}^n. \tag{17}$$

Setting corresponding elements of the vectors represented by the left- and right-hand sides of this equation equal permits the elements of $\mathbf{V}$ to be computed from a knowledge of $\mathbf{R}^n$ and L. This equality, for grid point (j, k), is given as (18):

$$b_{j,k}V_{j,k-1} + c_{j,k}V_{j-1,k} + d_{j,k}V_{j,k} = R^n_{j,k}. \tag{18}$$

[1] A more general form of (13) would incorporate β as an additional iteration parameter, multiplying the term $(M\mathbf{T}^n - \mathbf{q})$ in (13). For problems in which the coefficients B, D, E, F and G vary radically from point to point, and for a variable iteration parameter α as described later, experience has shown that values of β other than unity do not in general improve the subject iterative method.

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Equations (16) and (17) are combined to eliminate $\mathbf{R}^n$, and the result is premultiplied by L^{-1}. The result is (19):

$$U\boldsymbol{\delta}^{n+1} = \mathbf{V}. \tag{19}$$

This equation is used to compute $\boldsymbol{\delta}^{n+1}$ from $\mathbf{V}$ and U. The (j, k)th element of this equality is:

$$\delta_{j,k}^{n+1} + e_{j,k}\delta_{j+1,k}^{n+1} + f_{j,k}\delta_{j,k+1}^{n+1} = V_{j,k}\,. \tag{20}$$

Equations (11), (14), (18) and (20) comprise the algorithm for application of the $(n + 1)$st step of the subject iterative method. This set of equations applies to all points in the grid system

$$0 \leqq j \leqq J \quad \text{and} \quad 0 \leqq k \leqq K.$$

It should be noted that the matrix M is such that

$$\begin{aligned} B_{0,k} &= 0 \quad \text{and} \quad H_{j,K} = 0 \quad \text{for} \quad j = 0, 1, \cdots, J, \\ D_{0,k} &= 0 \quad \text{and} \quad F_{J,k} = 0 \quad \text{for} \quad k = 0, 1, \cdots, K. \end{aligned} \tag{21}$$

As a result, special forms of the algorithm are not required for these boundary points.

3.5. Application of algorithm. The coefficients b, c, d, e, f, R^n and V are intermediate coefficients generated and used during the elimination process, and they are completely defined by (11), (14) and (18). Equation (18) is solved for $V_{j,k}$, and the resulting equation, along with (11) and (14), are applied successively to the points in the $k = 0$ line with j taking on the values 0, 1, 2, $\cdots$, J. Then k is incremented by 1, and j again takes on the above values. If this procedure is continued through $k = K$, then values of the intermediate coefficients for all grid points will be obtained. Equation (20) is solved for $\delta_{j,k}^{n+1}$, and the resulting equation is applied in the reverse order to each point to obtain the temperature change, $\boldsymbol{\delta}^{n+1}$.

Instead of using the above procedure for every step of the iteration, a slight variation is used for alternate steps. This variation consists simply of carrying out the above described calculation in the reverse order for k, while varying j in the same manner. Thus for even-numbered iterations, k takes on the values K, $K - 1$, $\cdots$, 0, while j varies from zero to J as in the odd-numbered steps. It is best to use the same value of the iteration parameter, α, for both an odd- and an even-numbered iteration. This reordering of the grid points has the effect of making the nonzero coefficients appear for the temperatures $T_{j-1,k-1}$ and $T_{j+1,k+1}$, as indicated by the crosses in Fig. 5, rather than for temperatures $T_{j-1,k+1}$ and $T_{j+1,k-1}$, as in the odd-numbered steps. Although this variation is not always essential to the convergence of

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

the new iterative procedure, it often does increase the rate of convergence dramatically.

Since rate of convergence is a critical measure of the value of an iterative method, the next section is devoted to analysis of the rate of convergence of the strongly implicit iterative method.

4. Analysis of rate of convergence for the model problem. The rate of convergence of the new procedure for a generalized version of the model problem can be determined by a Fourier analysis, with superposition of solutions, if certain idealizations are made. It must be assumed that the number of grid points involved is large, so that the influence of the (relatively few) boundary points is small when compared to that of the (many) interior points. Also, the analysis must be applied to a double step of the iteration, as described in the preceding section. In the first step, k takes on the values 0, 1, 2, $\cdots$, K, while in the second step it takes on values of $K, K-1, \cdots, 0$.

For the generalized model problem, coefficients are constants for all points and are defined by the relations

$$B_{j,k} \equiv B = H_{j,k} = -KY\frac{\Delta x}{\Delta y},$$

$$D_{j,k} \equiv D = F_{j,k} = -KX\frac{\Delta y}{\Delta x},$$

$$E_{j,k} = -2(B + D).$$

For large j and k it is found that the intermediate coefficients b, c, d, e and f approach asymptotic values and that C approaches G. When $\alpha = 0$,

$$C_{j,k} \equiv C = \frac{BD}{-(B + D) + \sqrt{2BD}}, \tag{22a}$$

and when $\alpha = 1$,

$$C = \sqrt{BD}. \tag{22b}$$

Analysis requires the definition of an error in the nth level iterate as

$$\epsilon^n_{j,k} = T^n_{j,k} - T_{j,k}, \tag{23}$$

where $T_{j,k}$ is the true solution of (3). Equations (23) and (3) are used to eliminate temperatures and $q_{j,k}$ from the modified form of (3). A solution of the resulting error equation of the form

$$\epsilon^n_{j,k} = \xi^n A_{w,p}\, e^{i(w\pi j\Delta x + p\pi\Delta y)} \tag{24}$$

is assumed, where w and p each may be either positive or negative integers.

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Substitution of (24) into the error equation leads to the following equation defining ξ, the decay factor per step of iteration for the error component corresponding to w and p:

$$\xi = \frac{[0.5\,C(1-\alpha)J + C\alpha A]^2 - 4C^2R^2}{[0.5\,C(1-\alpha)J + C\alpha A + S]^2 - 4C^2R^2}, \tag{25}$$

where

$$J = \cos(w\pi\Delta x)\cos(p\pi\Delta y),$$

$$A = 2\sin^2\left(\frac{w\pi\Delta x}{2}\right)\sin^2\left(\frac{p\pi\Delta y}{2}\right),$$

$$R^2 = \left[\sin^2\left(\frac{w\pi\Delta x}{2}\right)\sin^2\left(\frac{p\pi\Delta x}{2}\right)\right]\left[1-\sin^2\left(\frac{w\pi\Delta x}{2}\right)\right]\left[1-\sin^2\left(\frac{p\pi\Delta y}{2}\right)\right],$$

$$S = -D\sin^2\left(\frac{w\pi\Delta x}{2}\right) - B\sin^2\left(\frac{p\pi\Delta y}{2}\right).$$

The quantities A and R^2 both arise from the incomplete canceling, or balancing, of the C and G terms induced in the modified matrix. If exact canceling were accomplished through exact factoring of M, then A and R would be zero and ξ would also be zero.

For the factorization employed by Buleev [1] and Dupont, Kendall and Rachford [3], the decay factor is similar to that given by (25) except that A is replaced by A', where

$$A' = A - \left[\sin^2\left(\frac{w\pi\Delta x}{2}\right) + \sin^2\left(\frac{p\pi\Delta y}{2}\right)\right].$$

In addition, the quantity $C(\alpha)$ for Buleev factoring is smaller than for the factoring described in this paper. Nonetheless, the decay factors given by (25) appear to have generally more desirable characteristics than those for Buleev factoring, especially for w and p corresponding to extremely different frequencies of error components. This behavior results from the fact that A' takes on negative values for these values of w and p, whereas A does not.

In addition to the above, (25) can be used to show the following:

(i) As either D or B (but not both) approaches zero, the decay factor approaches zero. This corresponds to KX or KY approaching zero, in which case the approximate factoring of the matrix M becomes exact and so the solution can be obtained without iteration.

(ii) If both B and D are large relative to $(\pi\Delta x/2)^2$ and $(\pi\Delta y/2)^2$, the repeated use of $\alpha = 1.0$ results in $|\xi| \geqq 1$. In fact, the re-

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

peated use of any α in the vicinity of 1.0 results in $|\xi| \geqq 1$ for some combinations of w and p.

(iii) For any combination of w and p, there exists an α lying in the range of zero to one which makes the corresponding decay factor small relative to unity. However, for other combinations of w and p, this choice of α may result in $|\xi|$ modestly greater than one. Thus the use of a cycle of parameters, each falling within the range of zero to one, is suggested. Values near unity tend to decay the fundamental, low frequency errors most rapidly, while values near zero decay the high frequency errors most rapidly. Values near one are more sensitive to the values of w and p; that is, if α is picked to minimize $|\xi|$ for a particular combination of w and p, a relatively small change in w and p will result in a much greater value of $|\xi|$. For this reason, it is to be expected that the values of α employed in a cycle will contain a relatively dense cluster near unity.

The exact value of the minimum parameter is not critical; zero is suitable. The maximum value used, however, is most important. If too great a value is used, divergence results; if too small a value, convergence may be slow. The best maximum value to be employed has been found to be dependent upon the particular problem being solved. Suitable maximum values can be computed for the case of constant KX, KY, Δx and Δy, by use of (26).

$$(1 - \alpha_{\max}) = \min \left[\frac{2\Delta x^2}{1 + \dfrac{KY\Delta x^2}{KX\Delta y^2}}, \frac{2\Delta y^2}{1 + \dfrac{KX\Delta y^2}{KY\Delta x^2}} \right], \tag{26}$$

where $\alpha_{\max}$ = maximum iteration parameter. This equation is similar to one which gives the minimum desirable parameters for the alternating direction technique.[2] For arbitrary variation of KX, KY, Δx and Δy, it has been found sufficient for a wide variety of problems to compute local values of $\alpha_{\max}$ for each grid point and to use an arithmetic average of these values for the entire set of equations. The local value of $\alpha_{\max}$ for the point (j, k) is computed as if the parameters in (26) are constants for the entire system, equal to the values existing in the neighborhood of this point.

It has been found desirable to use a minimum of four parameters, each used twice, per cycle. The individual parameters should be geometrically

[2] While (26) has been found adequate for computing the maximum value of α, it is unsatisfying to base the parameter calculation on alternating direction theory. Recently Weinstein [13] has been able to relate the maximum value of α to (25). Presumably, (26) was found adequate because the procedure is not overly sensitive to the maximum parameter.

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

spaced; (27) may be used to compute other members of a set from the maximum parameter:

$$1 - \alpha_m = (1 - \alpha_{\max})^{m/(M-1)}, \qquad m = 0, 1, \cdots, M - 1, \tag{27}$$

where M = number of parameters in a cycle. Equations (26) and (27) provide a method of generating a complete set of parameters for the most general form of (3). The order of application of the parameters within a cycle is not critical.

The Fourier convergence analysis presented in this section is at best only qualitatively significant. Computed results do agree with the analysis in a general way, and so the analysis is helpful in explaining why this strongly implicit iterative method is effective. Further evidence of its effectiveness is presented in the next section.

5. Evaluation of method. The new method has been evaluated by using it to solve a series of problems which span the range of difficulty normally encountered in application. The point-Jacobi, successive overrelation and alternating direction techniques were also applied to these same problems, so that the rate of convergence obtained by each of the four methods could be directly compared. These problems are described in the following section.

5.1. Description of problems. Most of the problems solved involved a square grid system with 31 points on each side. Three heat sources were located at grid points (3, 3), (3, 27), and (23, 4). The corresponding heat flux rates were 1.0, 0.5 and 0.6. Heat sinks were located at points (14, 15) and (27, 27), with rates of -1.83 and -0.27, respectively. No-flux boundary conditions were maintained at all exterior boundaries of the region. With one exception, as will be noted later, all initial temperatures were equal to zero.

Ordinarily heat fluxes were specified at sources or sinks, but in some cases either temperatures were specified or mixed boundary conditions were used. In these cases, the temperature at point (14, 15) was fixed at zero; and other temperatures, if fixed, were set at values that corresponded to the above-indicated flux rates, so that the true temperature distributions would be the same regardless of whether fluxes or temperatures were specified.

Some of the problems were solved for 11×11 and 21×21 grid-point systems. In these cases, the same sink or source strengths were used, and these were located as nearly as possible at the same relative location within the square as in the 31×31 systems.

Four different conductivity distributions were employed in the problems considered. Conductivities for the x-direction were specified independently from those in the y-direction, and in both cases corresponded to locations

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

half-way between grid points. The four distributions will be described in the order of ascending difficulty of solution.

The first distribution considered was the model problem, with KX and KY both equal to unity over the entire region. The second one was the closely related generalized model problem in which KX and KY are both constant but KX was 100-fold greater than KY.

The third case corresponded to more irregular geometry and will be described with the aid of Fig. 6. This figure depicts a 31 $\times$ 31 grid-point system. Heat sources are located at the points covered by open circles; heat sinks, at points covered by solid circles. In the region A, KX and KY were uniform and both equal to unity. In region B, KX was constant at a value of one, and KY, at a value of 100. In region C, these conductivities were reversed. In region D, both KX and KY were set equal to zero, representing an absolute barrier to heat flow. This problem might be characterized as one of heterogeneous permeability with homogeneous regions.

The fourth, and last, geometry was the same as the third one, except for region A. In this region KX and KY were varied in a random manner.

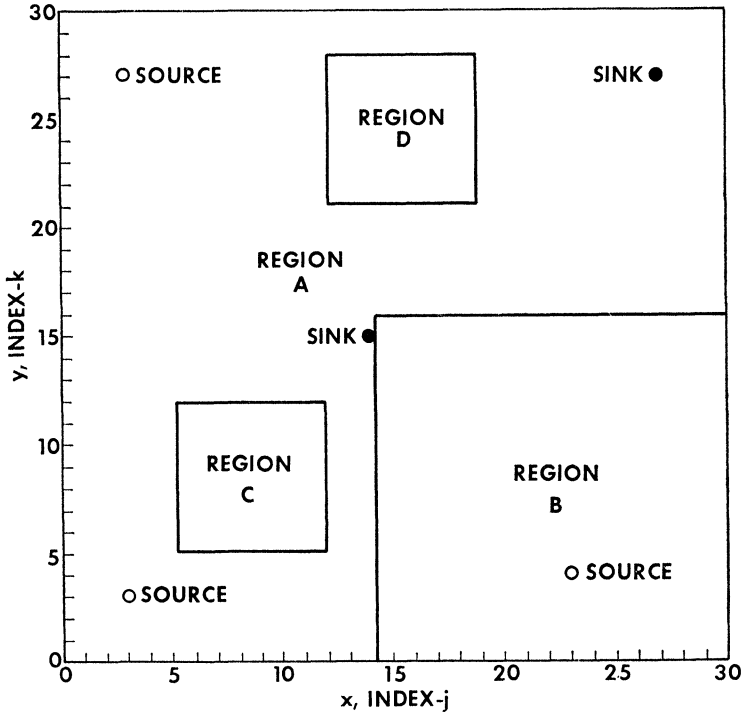

FIG. 6. *Areal geometry for heterogeneous test problem*

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

The values used were obtained by use of a standard subroutine which generates uniformly distributed random numbers falling within the range of zero to one. The ten per cent of the K-values thus generated which were less than one-tenth were replaced by zero, thus generating a large number of heat flow obstructions within this region.

5.2. Iteration parameters employed. For any of the four procedures used in this study, successful iteration is dependent upon the use of suitable values of the appropriate iteration parameters, or sets of parameters. It is therefore necessary to specify the values of parameters used with each method in the present study.

Point-Jacobi. The point-Jacobi method requires only a single parameter, ρ_J, as indicated by (28):

$$T_{j,k}^{n+1} = T_{j,k}^{n} - \frac{\rho_J}{E_{j,k}} (B_{j,k}T_{j,k-1}^{n} + D_{j,k}T_{j-1,k}^{n} + E_{j,k}T_{j,k}^{n} + F_{j,k}T_{j+1,k}^{n} + H_{j,k}T_{j,k+1}^{n} - q_{j,k}). \tag{28}$$

This parameter was taken to be unity.

Overrelaxation. Successive overrelaxation likewise requires only one parameter, ρ_S, which is defined by (29):

$$T_{j,k}^{n+1} = T_{j,k}^{n} - \frac{\rho_S}{E_{j,k}} [(B_{j,k}T_{j,k-1}^{n+1} + D_{j,k}T_{j-1,k}^{n+1}) + (E_{j,k}T_{j,k}^{n} + F_{j,k}T_{j+1,k}^{n} + H_{j,k}T_{j,k+1}^{n} - q_{j,k})]. \tag{29}$$

For each of the problems solved, the value of this parameter which resulted in the most rapid convergence was determined by testing a series of different parameters. Each parameter tested was greater than or equal to one, but less than two. For the 31×31 grid-point system, a value of 1.68 was found to result in most rapid convergence, and the data presented later were obtained using this value. This value is not the best value estimated by SOR theory to give the maximum asymptotic rate of convergence. This difference is attributed to the nature of the computational test in which roundoff error prevented the residual from being reduced by more than a factor of 10^6. Thus asymptotic rates of convergence are not being compared in these tests.

Alternating direction. The alternating direction method requires not one, but a set of iteration parameters. These parameters should be greater than zero but less than one.

Step 1.

$$T_{j,k}^{n+1/2} = T_{j,k}^{n} - \frac{1}{\rho_A E_{j,k}} (\Delta_x{}^2 T_{j,k}^{n+1/2} + \Delta_y{}^2 T_{j,k}^{n} - q_{j,k}).$$

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Step 2.

$$T_{j,k}^{n+1} = T_{j,k}^{n+1/2} - \frac{1}{\rho_A E_{j,k}} (\Delta_x{}^2 T_{j,k}^{n+1/2} + \Delta_y{}^2 T_{j,k}^{n+1} - q_{j,k}).$$

Here

$$\Delta_x{}^2 T_{j,k} = F_{j,k} T_{j+1,k} + EX_{j,k} T_{j,k} + D_{j,k} T_{j-1,k},$$

and $EX_{j,k}$ is that portion of $E_{j,k}$ due to x-direction contributions. The term $\Delta_y{}^2 T_{j,k}$ is similarly defined.

A set of six geometrically spaced parameters was used cyclically, with the largest ones being used first in each cycle. Each parameter was used in both Steps 1 and 2. The maximum parameter used was always one, but the best minimum parameter was determined by multiple solution of each problem and only the best results are reported herein. The use of the maximum ρ_A of unity and six parameters in a cycle is justified by the observed fact that the rate of convergence is not much affected by modest changes in these variables, provided that the minimum parameter is selected as just described.

Strongly implicit method. A set of nine geometrically spaced parameters was employed in the strongly implicit method of iteration, and each parameter was used twice in succession, as described earlier. This total of 18 parameters in a cycle causes no concern in difficult problems where 18 or more iterations are necessary to achieve suitable residuals. However, for easier problems that might yield a solution with less iterations, it is useful to break this set of 18 into three sets of 6 each, with each set of 6 more or less covering the entire range $0 \leqq \alpha \leqq 1$. Thus the following sequence of parameters was used, where the parameters are numbered in the order of increasing magnitude: 9, 9, 6, 6, 3, 3, 8, 8, 5, 5, 2, 2, 7, 7, 4, 4, 1, 1. The minimum parameter was zero, and the maximum parameter in each case was predicted by the technique described in an earlier section. No attempt was made to optimize the value of this maximum parameter for each problem.

5.3. Comparison of iterative methods. Each of the four iterative methods being studied was applied to each of four test problems that span the range of difficulty normally encountered. Calculations were made using a 31×31 grid-point system, with flow rates specified at all sources or sinks. The results, presented in Figs. 7 through 10, provide a basis for comparing the effectiveness of the various iterative methods.

Each figure (7–10) corresponds to one of the four test problems and presents results for all four iterative methods. The comparison is made on a work-required basis rather than number of iterations required. The ordinate in the figures is the absolute value of the maximum residual error in

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

(3), normalized by dividing by the sum of all heat source rates; the abscissa is the computational work required to achieve this value of the maximum error. The work unit is the amount of computing required for each step of iteration when using the strongly implicit procedure. This is approximately equal to the work required for three point-Jacobi or overrelaxation iterations, or for one application of both steps of the alternating direction method.

If residuals for every iteration were plotted in Figs. 7–10, the curves for both alternating direction and the strongly implicit procedures would oscillate and have a local maximum and minimum for each cycle of iteration. To facilitate visual comparison of the various rates of convergence, only the loci of local minima are shown for these two methods.

Fig. 7 presents data for the model problem in which $KX = KY = 1.0$ at all points. Although this ideal problem is rarely encountered in practice, it is of interest because in this case analyses may be made of the rela-

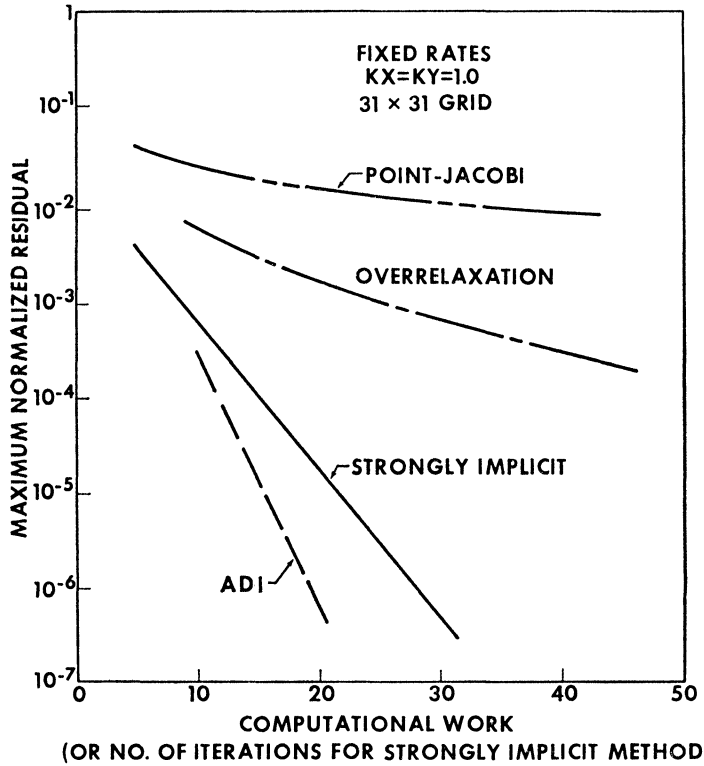

FIG. 7. *Comparison of computational work required for different iteration methods—model problem*

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

tive rates of convergence of the point-Jacobi, overrelaxation and ADI methods. The results of these analyses are, in general, borne out by the data of Fig. 7. These data show that both alternating direction and the strongly implicit method possess convergence rates a great deal faster than those of the other two methods. Alternating direction is slightly faster than the strongly implicit method, and overrelaxation is significantly more rapid than the point-Jacobi method.

Data for the generalized model problem, for which $KX/KY = 100$ throughout the grid system, are given in Fig. 8. In this case, the convergence rate for alternating direction shown in Fig. 8 might be improved by using different sets of parameters in each of its two steps. However, this modification is known to be helpful only for this special case of $KX/\Delta x^2$ and $KY/\Delta y^2$, each constant, but widely different. It has not been applied in the present study, since interest here is primarily focused upon more general problems. The data of Fig. 8 show that, for this slightly more general problem, only the strongly implicit method maintains the rate of convergence

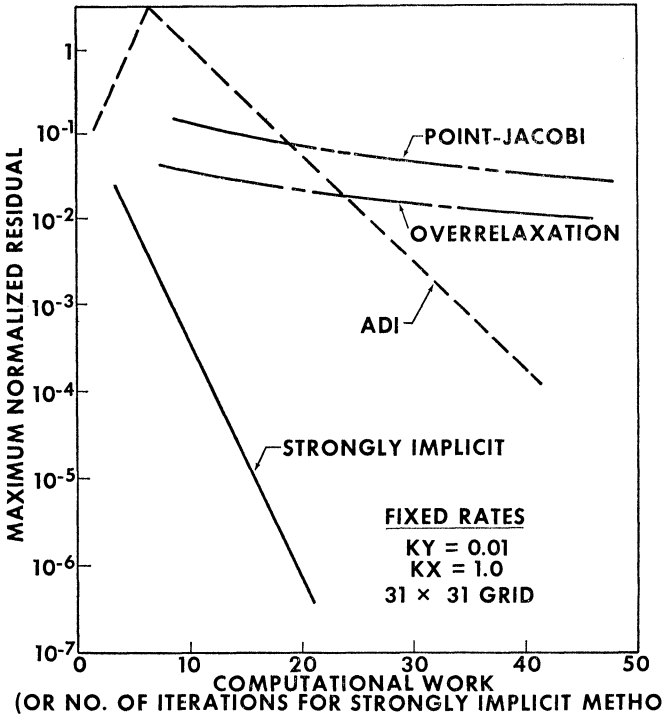

FIG. 8. *Comparison of computational work required for different iteration methods—generalized model problem*

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

demonstrated in the model problem. It is, in fact, more efficient here than in the preceding case. This behavior, which was predicted by the Fourier convergence analysis, also is subject to a more intuitive explanation. When $KX/\Delta x^2$ and $KY\Delta y^2$ are greatly different, the coefficients comprising the modification matrix N are much smaller than some of the coefficients of the matrix M. This means that the terms added to each equation to permit solution by elimination are relatively smaller, and so each step of iteration is more effective than for $KX/\Delta x^2 = KY/\Delta y^2$.

Data for the heterogeneous problem with homogeneous subregions are given in Fig. 9. The rate of convergence of the strongly implicit method is much more rapid than those of the other three, and this is also the case for the data of Fig. 10 which corresponds to the use of random conductivities in one subregion.

Of the three previous methods, alternating direction is seen to be the most effective. The data of Table 1 were derived from Figs. 7–10 to facilitate comparison of alternating direction and the strongly implicit method.

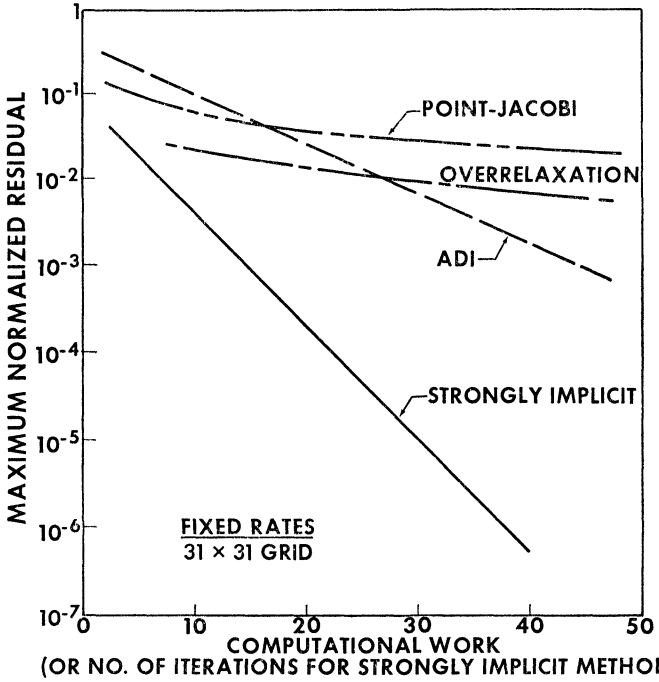

FIG. 9. *Comparison of computational work required for different iteration methods—heterogeneous model with homogeneous subregions*

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

This table shows the work required by both methods to achieve a maximum residual of 10^{-5} for each geometry tested. In three cases, straight line extrapolation of the curves of Figs. 7–10 was used to obtain the data of Table 1 for the alternating direction method.

It is apparent from the data of Table 1 that the effectiveness of the strongly implicit method is much less sensitive than ADI to the nature of the

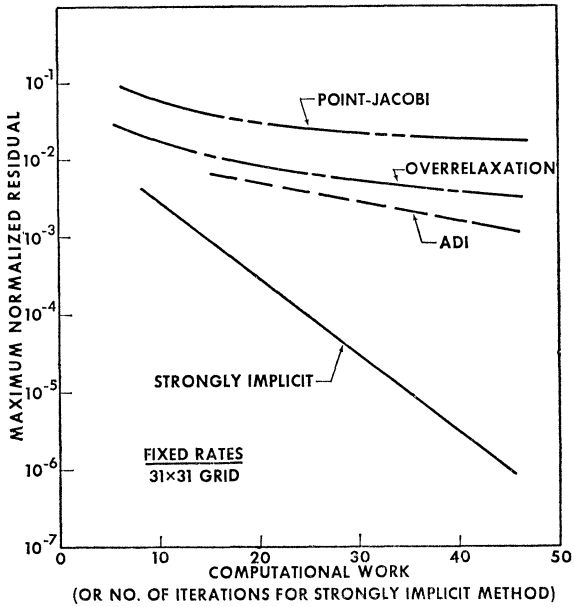

FIG. 10. *Comparison of computational work required for different iteration methods—heterogeneous model with random permeability variations in subregions*

TABLE 1

*Work requirements of iterative methods**

Problem Method	Model	Generalized Model	Heterogeneous	
			Subregions homogeneous	Random number subregions
Strongly Implicit	22	16	30	34
Alternating Direction	16	50†	80†	127†
Ratio: ADI/S.I.	0.73	‡	2.66	3.74

* To maximum residual of 10^{-5}—work equivalent to number of iterations for strongly explicit method.

† Based on straight line extrapolation.

‡ Omitted because best parameters not used for alternating direction.

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

problem being solved. For example, the ratio of work required to solve the random number problem to that required for the model problem is about 7.9 for ADI but is only 1.5 for the subject procedure.

The last line in Table 1 shows the ratio of work required when using ADI to that required when using the new method. No ratio is shown for the generalized model problem, since in this case the best parameters were not employed for ADI. For the model problem this ratio is 0.73, reflecting the fact that alternating direction is somewhat more efficient in this case. However, for the two heterogeneous cases, the ratio is 2.66 and 3.74, showing the new procedure to be about threefold more rapid than the most competitive method, ADI, for these more typical problems.

5.4. Influence of boundary conditions. The data presented in Fig. 11 are for the heterogeneous-random number conductivity distribution and were obtained to evaluate the effect of boundary conditions on the rate of convergence of the subject iterative method. The three curves correspond (i) to five temperatures fixed, (ii) to all source and sink strengths specified and (iii) to mixed conditions in which one temperature and four source and

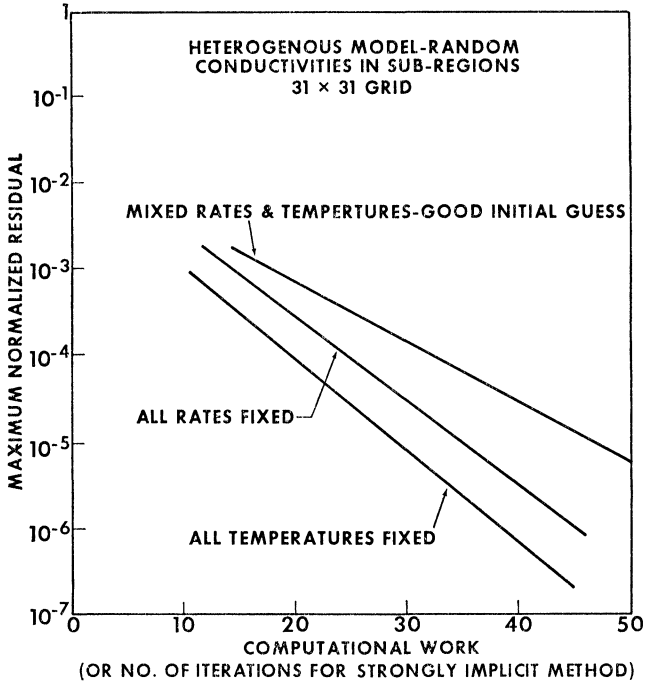

FIG. 11. *Effects of boundary conditions on computational work using strongly implicit iteration procedure*

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

sink strengths were fixed. In the latter case, a special initial condition was used wherein the initial temperature level was equal to the average temperature level of the true solution, while the initial maximum residual had essentially the same magnitude as in the other two cases.

These three cases required 30, 34 and 47 iterations to reduce the maximum residual to 10^{-5}. Thus, these data indicate that the boundary conditions are of only modest importance, provided the initial temperature guesses are as described above. This is usually the case when solving transient problems.

5.5. Influence of grid size. The data of Fig. 12 are presented to show the effect of grid size upon the rate of convergence of the strongly implicit method. The ordinate on this figure is the number of iterations required to achieve a maximum residual of 10^{-5}, while the abscissa is the number of grid points on each side of a square grid network. Calculations were for the square grid systems, with 11, 21 or 31 grid points on a side. Iteration parameters were predicted, as described earlier, and it should be noted that the data of Fig. 12 are influenced not only by the nature of the iterative process, but also by the appropriateness of the parameters thus predicted.

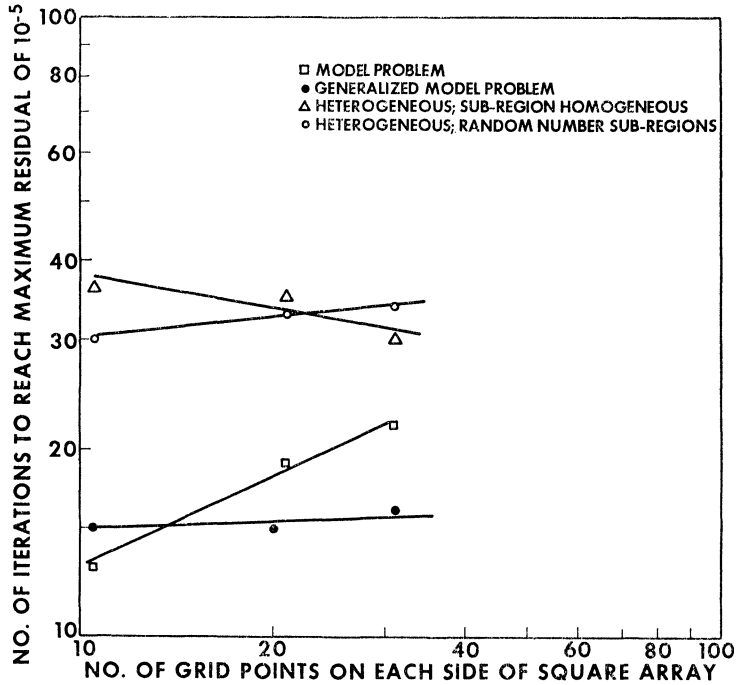

FIG. 12. *Effect of number of grid points upon rate of convergence of strongly implicit iterative procedure*

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

It can be seen that only for the model problem does the iterative work requirement depend appreciably upon the grid size. For the other cases, there appears to be little such dependence.

5.6. Effectiveness of parameter prediction. The strongly implicit method has been applied to a wide variety of problems, as described in the previous sections, and in all cases the parameters used were obtained by the procedure described earlier. Without exception, these parameters gave good convergence rates.

The effectiveness of the procedure in this respect was also demonstrated on other problems not described herein. These include:

(i) Transient temperature distribution calculations.
(ii) Rectangular regions, as opposed to the square regions reported previously.
(iii) Small regions of high conductivity embedded in large regions of low or zero conductivity. Several different shapes of nonzero regions were employed.
(iv) Different locations, numbers and strengths of sinks and sources.
(v) Conductivities specified only at grid points, instead of at points half-way between. The latter were computed from the former by arithmetic averaging.
(vi) Ratios of $KX \cdot \Delta y^2 / KY \cdot \Delta x^2 = 2500$.

6. Other applications. In the present paper, the strongly implicit iterative method has been used to obtain solutions of difference approximations of a single two-dimensional partial differential equation. In this case, each difference equation contains five unknowns. Similar equations which contain nine unknowns can be treated by a generalization of this procedure, as can difference approximations of three-dimensional problems. In addition, it is possible to treat situations in which there exist more than one equation and unknown per grid point. During the preparation of this paper, the case of two unknowns per grid point in two dimensions has been treated successfully. Currently, this case of two unknowns is being extended to treat three-dimensional problems.

7. Conclusions. The new method for solving the large sets of linear algebraic equations that arise in the approximate solution of partial differential equations has the following advantages:

(i) The method is only moderately sensitive to the nature of the boundary conditions to be applied.
(ii) The work required for each equation is not sensitive to the number of equations to be solved. Previous methods require a rapid increase in this work as the number of equations increases.
(iii) Iteration parameters suitable for application may be reliably esti-

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

mated. Similar parameters must be determined by trial and error for older techniques.

(iv) The method requires less computation than previous methods. For a typical problem involving 961 equations, the best competitive method requires approximately threefold more computation.

Acknowledgment. The author wishes to express his gratitude to the Esso Production Research Company for permission to publish this paper, the contents of which are essentially identical to a June 1, 1966, Company report.

REFERENCES

[1] N. I. Buleev, *A numerical method for the solution of two-dimensional and three-dimensional equations of diffusion*, Mat. Sb., 51 (1960), no. 2, pp. 227–238.

[2] J. Douglas, *On the numerical integration of* $\partial^2 u/\partial x^2 + \partial^2 u/\partial y^2 = \partial u/\partial t$ *by implicit methods*, J. Soc. Indust. Appl. Math., 3 (1955), pp. 42–65.

[3] Todd Dupont, Richard P. Kendall and H. H. Rachford, Jr., *An approximate factorization procedure for solving self-adjoint elliptic difference equations*, this Journal, 5 (1968), pp. 559–573.

[4] S. P. Frankel, *Convergence rates of iterative treatments of partial differential equations*, Math. Tables Aids Comput., 4 (1950), pp. 65–75.

[5] C. G. J. Jacobi, *Über eine neue Auflösungsart der bei der Methode der kleinsten Quadrate vorkommenden linearen Gleichungen*, Astronom. Nachr., 22 (1845), pp. 297–306.

[6] H. Liebmann, *Die angenäherte Ermittelung harmonischer Funktionen und konformer Abbildungen*, Bayer. Akad. Wiss. Math.-Phys. Kl. S.-B., 47 (1918), pp. 385–416.

[7] P. A. Nekrasov, *Die Bestimmung der Unbekannten nach der Methode der kleinsten Quadrate bei einer sehr grossen Anzahl der Unbekannten*, Mat. Sb., 12 (1885), pp. 189–204. In Russian.

[8] ———, *Zum Problem der Auflösung von linearen Gleichungssysteme mit einer grossen Anzahl von Unbekannten durch sukzessive Approximationen*, Ber. Petersburger Akad. Wiss., 69 (1892), no. 5, pp. 1–18.

[9] T. A. Oliphant, *An extrapolation process for solving linear systems*, Quart. Appl. Math., 20 (1962), pp. 257–267.

[10] D. W. Peaceman and H. H. Rachford, Jr., *The numerical solution of parabolic and elliptic differential equations*, J. Soc. Indust. Appl. Math., 3 (1955), pp. 28–41.

[11] L. F. Richardson, *The approximate arithmetical solution by finite differences of physical problems involving differential equations, with application to the stress in a masonry dam*, Philos. Trans. Roy. Soc. London. Ser. A, 210 (1910), pp. 307–357.

[12] R. V. Southwell, *Relaxation Methods in Theoretical Physics*, Clarendon Press, Oxford, 1946.

[13] H. G. Weinstein, personal communication.

[14] J. H. Wilkenson, *Rounding Errors in Algebraic Processes*, Prentice-Hall, Englewood Cliffs, New Jersey, 1963.

[15] David M. Young, *Iterative methods for solving partial difference equations of elliptic type*, Doctoral thesis, Mathematics Department, Harvard University, Cambridge, 1950.

Downloaded 12/27/12 to 150.135.135.70. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php