BIT 12 (1972), 443–467

A GENERALIZED SSOR METHOD

O. AXELSSON*

Abstract.

To solve large sparse systems of linear equations with symmetric positive definite matrix $A = D+L+L^*$, $D = \mathrm{diag}(A)$, with iteration, the SSOR method with one relaxation parameter ω has been applied, yielding a spectral condition number approximately equal to the square root of that of A, if the condition $S(\tilde{L}\tilde{L}^*) \leqq \frac{1}{4}$, where $\tilde{L} = D^{-\frac{1}{2}}LD^{-\frac{1}{2}}$, is satisfied and if $0 < \omega < 2$ is chosen optimally. The matrix arising from the differenced Dirichlet problem satisfies in general the spectral radius condition given above, only if the coefficients of the differential equation are constant and if the mesh-width is uniform.

However, using one relaxation parameter for each mesh-point, the main result for the SSOR method, that the spectral condition number varies with a parameter $\zeta > 0$ like $O([\zeta^{-1}+\zeta/\lambda_1]h^{-1})$, $h \to 0$, where $\lambda_1 h^2$ is the smallest eigenvalue of $D^{-1}A$, carries over for variable smooth coefficients and even for certain kinds of discontinuities among the coefficients, if the mesh-width is adjusted properly in accordance with the discontinuity.

Since the resulting matrix of iteration has positive eigenvalues, a semi-iterative technique can be used. The necessary number of iterations is thus only $O(h^{-\frac{1}{2}})$.

Introduction.

The SSOR method for the iterative solution of large linear systems of equations was considered among others by Sheldon [1], Habetler and Wachspress [2] and Ehrlich [3]. It was shown in connection with the differenced Dirichlet problem that, if a certain condition is satisfied, a rate of convergence of $O(h^{-1})$ iterations, where h is a mesh-width parameter, was achieved and that it is possible to apply the Chebyshev acceleration technique to reach a rate of convergence of $O(h^{-\frac{1}{2}})$ iterations (see also Young [4]). In [2] a rather complicated formula for the calculation of an approximation to the optimal relaxation parameter was given. However, the speed of convergence is less influenced by the choice of the relaxation parameter than that for the SOR method (see, for example, Young [4]), where the speed of convergence $O(h^{-1})$ is only reached if an accurate enough value of the optimal relaxation parameter

Received July 11, 1972.

* On leave of absence from Chalmers University of Technology, Göteborg, Sweden.

is used. This parameter depends not only on the particular partial differential equation, but also on the particular boundary. Furthermore, it is not possible to use a semi-iteration technique to accelerate the rate of convergence of the optimal SOR method (see, for example, Young [4]).

The above results for the SSOR method can be derived in a simple way, using elementary inner product techniques (see Axelsson [5]). In the present paper we show that applying a modified SSOR method to the discretized multidimensional Dirichlet problem over a general region and using one relaxation parameter for each point in the mesh, the main result for the SSOR method carries over to a more general Dirichlet problem with Lipschitz continuous coefficients and a grid with not necessarily constant step-lengths. Lower and upper bounds of the eigenvalues of the iteration matrix are derived and applied in the Chebyshev semi-iterative method (see, for example, Young [4]) The speed of convergence is now much less dependent on the choice of the corresponding relaxation parameter. This is even more so in a variant of the conjugate gradient method (cf. Hestenes and Stiefel [6] and Lanczos [7]).

It is obvious that the method could be seen as a method of scaling by diagonal matrices (cf. Widlund [8]). It is also easily seen that the scaling derived in this paper, apart from minor details, is the same as that in Dupont [9]. The method in that paper is a variant of Stones method [10]. Among other methods in connection with the discretized Dirichlet problem, the method given in Gunn [11] should be mentioned. In this, a well-known iterative technique for dealing with non-linear problems is presented, each iteration step requiring the solution of a sparse linear system of equations. If the region is rectangular the ADI method can be applied, the necessary number of iteration thus being $O(\log h)$. However, this number is approximately proportional to the ratio

$$\max_{x,y} a(x,y)/\min_{x,y} a(x,y) ,$$

taken over the region where $a(x,y)$ is the variable coefficient in the Dirichlet problem. Since this number can be fairly large, even for smooth coefficients, the actual required number of iterations for a certain h can be very large. Using Chebyshev semi-explicit iteration however, gives a marked improvement. Still, the successful application of the method is sensitive to irregular regions, a well-known disadvantage with the ADI method. In connection with this, the scaling process presented in Widlund [12] might be useful. However, no numerical results are given, showing the practicability of that method. The principle, presented in

Gunn [11], for dealing with non-linear problems, is of course also applicable in connection with the present method (see also Dupont [9]).

Besides the technique presented in this paper for dealing with problems with discontinuous coefficients, the Woodbury formula for inverting modified matrices is also applicable. This formula is applied in connection with direct methods for the discretized Dirichlet problem in George [13] (see also Buzbee et al. [14]).

Model problem; self-adjoint Dirichlet problem.

Let the self-adjoint linear operator $\mathscr{L}$ be defined by

$$\mathscr{L}(u) = -\sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left(a_i(x)\frac{\partial}{\partial x_i}u\right) + q(x)u = f(x), \quad x \in \Omega \tag{2.1}$$

and

$$u = g, \quad x \in \partial\Omega\,,$$

where Ω is an open bounded set in R^n and $\partial\Omega$ is the boundary of Ω. Furthermore, let g and $a_i(x)$ be sufficiently smooth,

$$0 \leqq a_i(x) \leqq k_i, \quad \sum_{i=1}^{n} a_i(x) > 0\,,$$

and

$$q(x) \geqq 0, \quad x \in \Omega\,.$$

This Dirichlet problem is discretized in such a way that a system of linear equations is obtained, with a Hermitian coefficient matrix. This can, for instance, be done in the following way. We first introduce some notations.

Let $h_i^{(j)}$, $j = \ldots, -1, 0, 1, \ldots$, $i = 1, \ldots, n$ be small enough positive numbers, called step-lengths. Let the coordinate axes be divided so that $h_i^{(j)}$ is the distance between two consecutive points on the x_i axis. Let us construct a grid of lines through the points on the x_i axis and through all intersections of lines, parallel to the x_j axis, $j \neq i$, $i, j = 1, \ldots, n$. Let Ω_h be the set of points of the grid which are in Ω and let $\partial\Omega_h$ be the set of points of intersection between the grid-lines and the boundary $\partial\Omega$.

Let $\alpha = (j_1, \ldots, j_n)$ be a multi-index, i.e. let j_i be integers, and let e_i be the unit vectors in the x_i direction. By $\alpha \geqq 0$ we mean $j_i \geqq 0$, $i = 1, \ldots, n$ and by $\alpha \geqq \beta$ we mean $\alpha - \beta \geqq 0$. Now, order the points in the grid in accordance with the multi-index. For simplicity, we write h_{α,e_i} for the the step-length from point number α in the positive x_i direction.

However, we have to remember, that through the construction, the step-lengths h_{α,e_i} are constant along the grid-lines through α, perpendicular to the x_i axis.

We write

$$x_{\alpha+e_i} = x_\alpha + h_{\alpha,e_i}$$
$$x_{\alpha+\frac{1}{2}e_i} = x_\alpha + \tfrac{1}{2}h_{\alpha,e_i}$$
$$x_{\alpha-e_i} = x_\alpha - h_{\alpha-e_i,e_i}$$

and so forth.

We wish to find a grid function u_h, defined on $\Omega_h \cup \partial\Omega_h$, which satisfies a difference analogue of $\mathscr{L}(u)=f$ at each point of Ω_h and whose value is $g(x)$, $x \in \partial\Omega_h$. Let the difference analogue be

$$\mathscr{L}_h u_\alpha = b_\alpha u_\alpha - \sum_{i=1}^{n} (c_{\alpha,e_i} u_{\alpha+e_i} + c_{\alpha-e_i,e_i} u_{\alpha-e_i}) = f_\alpha, \quad x_\alpha \in \Omega_h . \tag{2.2}$$

The coefficients of the operator $\mathscr{L}_h$ are obtained from $\mathscr{L}$ by replacing

$$\frac{\partial}{\partial x_i}\left(a_i(x)\frac{\partial}{\partial x_i}u\right) \quad \text{at} \quad x = x_\alpha,\ \alpha = (j_1, \ldots, j_n)$$

by

$$\frac{2}{h_{\alpha,e_i}+h_{\alpha-e_i,e_i}}\left[a_i(x_{\alpha+\frac{1}{2}e_i})\frac{u_{\alpha+e_i}-u_\alpha}{h_{\alpha,e_i}} - a_i(x_{\alpha-\frac{1}{2}e_i})\frac{u_\alpha-u_{\alpha-e_i}}{h_{\alpha-e_i,e_i}}\right] \tag{2.3}$$

and multiplying the equation at x_α by $\prod_{i=1}^{n}(h_{\alpha,e_i}+h_{\alpha-e_i,e_i})$. Thus we get

$$\begin{aligned}\mathscr{L}_h u_\alpha - f_\alpha = {} & 2\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}(h_{\alpha,e_j}+h_{\alpha-e_j,e_j})\,[a_i(x_{\alpha+\frac{1}{2}e_i})\,(u_\alpha-u_{\alpha+e_i})/h_{\alpha,e_i} \\ & + a_i(x_{\alpha-\frac{1}{2}e_i})\,(u_\alpha-u_{\alpha-e_i})/h_{\alpha-e_i,e_i}] \\ & + \prod_{j=1}^{n}(h_{\alpha,e_j}+h_{\alpha-e_j,e_j})\,[q(x_\alpha)u_\alpha - f(x_\alpha)] = 0, \quad x_\alpha \in \Omega_h\end{aligned}$$

and

$$u_\alpha = g(x_\alpha), \quad x_\alpha \in \partial\Omega_h .$$

We immediately see that this is of the form (2.1) and, in particular, that

$$b_\alpha \geqq \sum_{i=1}^{n}(c_{\alpha,e_i}+c_{\alpha-e_i,e_i}) , \tag{2.4}$$

where equality is valid if $q(x_\alpha)=0$. Furthermore, the coefficients b_α, c_{α,e_i}, and $c_{\alpha-e_i,e_i}$ are positive. The global discretization error is thus $O(h^2)$ (see Bramble, Hubbard and Thomée [15]). We write the system of linear equations as

$$Au = \tilde{f},$$

where u is a vector with the components u_α, $x_\alpha \in \Omega_h$ and $\tilde{f}$ comes from f and the boundary values g.

Let

$$\tilde{u}_\alpha = \begin{cases} u_\alpha, & \alpha \in \Omega_h \\ 0\ , & \alpha \in \partial\Omega_h \end{cases}$$

and let

$$\tilde{\mathscr{L}}_h u_\alpha = b_\alpha u_\alpha - \sum_{i=1}^{n} (c_{\alpha,e_i} \tilde{u}_{\alpha+e_i} + c_{\alpha-e_i,e_i} \tilde{u}_{\alpha-e_i}), \quad x_\alpha \in \Omega_h .$$

It follows that $\tilde{\mathscr{L}}_h u_\alpha$ gives the value of Au at the α component. It is now easily seen that the matrix A is Hermitian. If, namely, we order the equations in a generalized row-wise order, according to increasing indices α, the lower triangular part L of A satisfies

$$Lu_\alpha = (Lu)_\alpha = -\sum_{i=1}^{n} c_{\alpha-e_i,e_i} \tilde{u}_{\alpha-e_i},$$

where Lu_α is the value of Lu at the α component. Thus

$$L^*u_\alpha = (L^*u)_\alpha = -\sum_{i=1}^{n} c_{\alpha,e_i} \tilde{u}_{\alpha+e_i}$$

and

$$Au_\alpha = (Au)_\alpha = b_\alpha u_\alpha + (Lu)_\alpha + (L^*u)_\alpha .$$

With the restricted model problem we mean the model problem over a hyperrectangle with maximal side of 1 unit length, where the coefficients a_i and steplengths h_i are constant over the grid and where $q \equiv 0$. (It is easily seen that the assumption $q \equiv 0$ is no restriction, however.)

A scaling procedure.

In Axelsson [5] it is shown that the SSOR method can be seen as a particular case of the following fractional-step method (cf. Marchuk [16]),

$$\begin{aligned} (\tilde{D}+L)\xi^{(l-\frac{1}{2})} &= -(Aw^{(l)}-\tilde{f}) \\ (\tilde{D}+U)\xi^{(l)} &= \tilde{D}\xi^{(l-\frac{1}{2})} \\ w^{(l+1)} &= w^{(l)}+\beta_l \xi^{(l)} \end{aligned}$$

where $A = D+L+U$, $\tilde{D} = [d_\alpha] = [\omega_\alpha^{-1} b_\alpha]$, $D = [b_\alpha]$, L, U are the lower and upper triangular parts of A, respectively, and β_l is an iteration acceleration parameter. (In the following section we shall choose a two-step version of the acceleration formula.) In this section we shall describe

how to choose the matrix $\tilde{D}$, i.e. how to choose the set of real parameters $\{\omega_\alpha\}$, in order to get a spectral condition number $O(h^{-1})$ of the matrix of iteration $N^{-1}A$, where

$$N = (\tilde{D}+L)\tilde{D}^{-1}(\tilde{D}+U) = A+\tilde{D}-D+L\tilde{D}^{-1}U\,,$$

if the coefficients $a_i(x)$ are Lipschitz continuous in the x_i-direction and if we choose the step-lengths properly. We choose a natural ordering, i.e. row-wise or column-wise, of the model problem, since this enables us to deal with the general dimensional problem more easily.

Let

$$\tilde{c}_{\alpha,e_i} = \begin{cases} c_{\alpha,e_i}\,, & \alpha+e_i \in \Omega_h \\ 0\,, & \alpha+e_i \in \partial\Omega_h \end{cases} \tag{3.1}$$

$$\tilde{c}_{\alpha-e_i,e_i} = \begin{cases} c_{\alpha-e_i,e_i}\,, & \alpha-e_i \in \Omega_h \\ 0\,, & \alpha-e_i \in \partial\Omega_h \end{cases}$$

$$\tilde{\gamma}_\alpha = \sum_i \tilde{c}_{\alpha,e_i}/b_\alpha\,,$$

$$\gamma_\alpha = \left[\sum_i c_{\alpha,e_i} + \sum_i (c_{\alpha-e_i,e_i} - \tilde{c}_{\alpha-e_i,e_i})\right] \Big/ b_\alpha, \quad \alpha \in \Omega_h$$

and let

$$\omega_\alpha = 0, \quad \alpha \notin \Omega_h\,,$$

$$\tilde{\Omega}_h^{(I)} = \{\alpha \in \Omega_h;\ \alpha-e_i \notin \partial\Omega_h,\ i = 1, \ldots, n\}\,,$$

$$\tilde{\Omega}_h^{(II)} = \{\alpha \in \Omega_h;\ \alpha+e_i \notin \partial\Omega_h,\ i = 1, \ldots, n\}\,,$$

$$\Omega_h^{(I)} = \Omega_h/\{\alpha \in \Omega_h;\ \alpha-e_i \in \partial\Omega_h,\ i = 1, \ldots, n\}$$

$$\Omega_h^{(II)} = \Omega_h/\{\alpha \in \Omega_h;\ \alpha+e_i \in \partial\Omega_h,\ i = 1, \ldots, n\}$$

We observe, that according to (3.1) and (2.4), $0 < \tilde{\gamma}_\alpha \leqq \gamma_\alpha < 1$, $\alpha \in \Omega_h^{(I)} \cap \Omega_h^{(II)}$.

We have

$$(L^*u)_\alpha = -\sum_j \tilde{c}_{\alpha,e_j}\, u_{\alpha+e_j}\,,$$

$$(\tilde{D}^{-1}L^*u)_\alpha = -\sum_j d_\alpha^{-1}\, \tilde{c}_{\alpha,e_j}\, u_{\alpha+e_j}\,,$$

$$(L\tilde{D}^{-1}L^*u)_\alpha = \sum_i \tilde{c}_{\alpha-e_i,e_i}\,(\tilde{D}^{-1}L^*u)_{\alpha-e_i} = \sum_{i,j} d_{\alpha-e_i}^{-1}\, \tilde{c}_{\alpha-e_i,e_i}\, \tilde{c}_{\alpha-e_i,e_i}\, u_{\alpha-e_i+e_i}$$

and

$$(L\tilde{D}^{-1}L^*u,u) = \sum_{\alpha\in\Omega_h} \overline{(L\tilde{D}^{-1}L^*u)_\alpha}\, u_\alpha = \sum_\alpha \sum_{i,j} d_\alpha^{-1}\, \tilde{c}_{\alpha,e_i}\, \tilde{c}_{\alpha,e_j}\, \bar{u}_{\alpha+e_j}\, u_{\alpha+e_i}$$

$$= \sum_\alpha d_\alpha^{-1}\left[\sum_{i<j} \tilde{c}_{\alpha,e_i}\, \tilde{c}_{\alpha,e_j}\, 2\,\mathrm{Re}\,(\bar{u}_{\alpha+e_j}\, u_{\alpha+e_i}) + \sum_i \tilde{c}_{\alpha,e_i}^2\, |u_{\alpha+e_i}|^2\right].$$

Since

$$2\,\mathrm{Re}\,(\tilde{u}_{\alpha+e_j}\,u_{\alpha+e_i}) = -|u_{\alpha+e_i}-u_{\alpha+e_j}|^2+|u_{\alpha+e_i}|^2+|u_{\alpha+e_j}|^2$$

we get

$$\begin{aligned}(L\tilde{D}^{-1}L^*u,u) &= \sum_{\alpha}\sum_{i,j} d_\alpha^{-1}\,\tilde{c}_{\alpha,e_i}\,\tilde{c}_{\alpha,e_j}\,|u_{\alpha+e_i}|^2 \\ &\quad -\sum_{\alpha}\sum_{i<j} d_\alpha^{-1}\,\tilde{c}_{\alpha,e_i}\,\tilde{c}_{\alpha,e_j}\,|u_{\alpha+e_i}-u_{\alpha+e_j}|^2 \\ &\leqq \sum_{\alpha}\sum_{i,j} d_{\alpha-e_i}^{-1}\,\tilde{c}_{\alpha-e_i,e_i}\,\tilde{c}_{\alpha-e_i,e_j}\,|u_\alpha|^2 \\ &= \sum_{\alpha}\sum_{i} \omega_{\alpha-e_i}\,\tilde{c}_{\alpha-e_i,e_i}\,\tilde{\gamma}_{\alpha-e_i}\,|u_\alpha|^2\,,\end{aligned}$$

if $d_\alpha>0$, i.e. if $\omega_\alpha>0$. Then

$$([\tilde{D}-D+L\tilde{D}^{-1}L^*]u,u) \leqq \sum_{\alpha\in\Omega_h}\left(\omega_\alpha^{-1}-1+\sum_i \omega_{\alpha-e_i}\,b_\alpha^{-1}\,\tilde{c}_{\alpha-e_i,e_i}\,\tilde{\gamma}_{\alpha-e_i}\right)b_\alpha\,|u_\alpha|^2\,.$$

Thus, if we choose

$$\omega_\alpha^{-1} = 1+\delta-\sum_i \omega_{\alpha-e_i}\,b_\alpha^{-1}\,\tilde{c}_{\alpha-e_i,e_i}\,\tilde{\gamma}_{\alpha-e_i},\ \delta\geqq 0\,,$$

then, if $\omega_\alpha>0$,

$$([\tilde{D}+L]\tilde{D}^{-1}[\tilde{D}+L^*]u,u) \leqq (Au,u)+\delta\sum_{\alpha\in\Omega_h} b_\alpha\,|u_\alpha|^2\,. \tag{3.2}$$

It follows from Lemma 3.1 that $\omega_\alpha>0$.

LEMMA 3.1. If

$$\omega^{-1} = 1+\delta-\sum_i \omega_{\alpha-e_i}\,b_\alpha^{-1}\,\tilde{c}_{\alpha-e_i,e_i}\,\tilde{\gamma}_{\alpha-e_i},\qquad \alpha\in\Omega_h\,,$$

where $\delta\geqq 0$, we have

$$(1+\delta)^{-1} \leqq \omega_\alpha < \gamma_\alpha^{-1},\qquad \alpha\in\Omega_h\,.$$

PROOF (induction). If $\alpha\in\Omega_h/\Omega_h^{(1)}$, then $\omega_\alpha=(1+\delta)^{-1}<\gamma_\alpha^{-1}$. Thus suppose $0<\omega_{\alpha-e_i}<\gamma_{\alpha-e_i}^{-1}$, $\alpha-e_i\in\Omega_h$, $i=1,\ldots,n$. Then, since $\gamma_\alpha\geqq\tilde{\gamma}_\alpha$,

$$\begin{aligned}\omega_\alpha^{-1} &= 1+\delta-\sum_i \omega_{\alpha-e_i}\,b_\alpha^{-1}\,\tilde{c}_{\alpha-e_i,e_i}\,\tilde{\gamma}_{\alpha-e_i} \\ &> 1-b_\alpha^{-1}\sum_i \tilde{c}_{\alpha-e_i,e_i} = b_\alpha^{-1}\left(b_\alpha-\sum_i \tilde{c}_{\alpha-e_i,e_i}\right).\end{aligned}$$

Thus from (2.4),

$$\omega_\alpha^{-1} > b_\alpha^{-1}\sum_i (c_{\alpha,e_i}+c_{\alpha-e_i,e_i}-\tilde{c}_{\alpha-e_i,e_i}) = \gamma_\alpha$$

i.e.

$$0 < \omega_\alpha < \gamma_\alpha^{-1} .$$

But from $\omega_{\alpha-e_i} \geqq 0$ it follows that $\omega_\alpha^{-1} \leqq 1+\delta$, i.e. $\omega_\alpha \geqq (1+\delta)^{-1}$.

To get a lower bound of $([\tilde{D}+L]\tilde{D}^{-1}[D+L^*]u,u)$ we first prove a Lemma (see Dupont [9]).

LEMMA 3.2. *Let* $c_i \geqq 0$, $i=1, \ldots, n$ *and let*

$$\sum_i^n c_i > 0 .$$

Let e *and* a_i, $i=1, \ldots, n$ *be complex. Then*

$$\sum_{1\leqq i<j\leqq n} c_i c_j |a_i - a_j|^2 \leqq \left(\sum_i c_i\right) \sum_i c_i |a_i - e|^2 .$$

PROOF. This Lemma is proved by first noting that the best discrete least square approximation by a constant is

$$e = \left(\sum c_i\right)^{-1} \sum c_i a_i .$$

An elementary calculation shows then that we have equality in the Lemma for this value of e.

Using this Lemma with $e = \beta_\alpha u_\alpha$, where

$$\beta_\alpha = \begin{cases} \omega_\alpha^{-1} \tilde{\gamma}_\alpha^{-1} \left(1 - \min\limits_\alpha (2-\omega_\alpha)\right), & \text{if } \omega_\alpha < 2, \text{ all } \alpha \\ 1 \qquad\qquad\qquad\qquad\quad , & \text{if } \omega_\alpha \geqq 2, \text{ any } \alpha , \end{cases}$$

we have for any $\{\omega_\alpha\}$ with $\omega_\alpha > 0$,

$$\begin{aligned} (3.3) \qquad (L\tilde{D}^{-1}L^*u,u) &= \sum_\alpha \sum_{i,j} d_{\alpha-e_i}^{-1} \tilde{c}_{\alpha-e_i,e_i} \tilde{c}_{\alpha-e_i,e_j} |u_\alpha|^2 \\ &\quad - \sum_\alpha \sum_{i<j} d_\alpha^{-1} \tilde{c}_{\alpha,e_i} \tilde{c}_{\alpha,e_j} |u_{\alpha+e_i} - u_{\alpha+e_j}|^2 \\ &\geqq \sum_\alpha \sum_{i,j} d_{\alpha-e_i}^{-1} \tilde{c}_{\alpha-e_i,e_i} \tilde{c}_{\alpha-e_i,e_j} |u_\alpha|^2 \\ &\quad - \sum_\alpha d_\alpha^{-1} \left(\sum_j \tilde{c}_{\alpha,e_j}\right) \sum_i \tilde{c}_{\alpha,e_i} |\beta_\alpha u_\alpha - u_{\alpha+e_i}|^2 \\ &= -\sum_\alpha \omega_\alpha \beta_\alpha^2 \tilde{\gamma}_\alpha^2 b_\alpha |u_\alpha|^2 \\ &\quad + 2 \sum_\alpha \sum_i \omega_\alpha \beta_\alpha \tilde{\gamma}_\alpha \tilde{c}_{\alpha,e_i} \operatorname{Re}(\bar{u}_\alpha u_{\alpha+e_i}) . \end{aligned}$$

Thus, since

$$(3.4)\qquad (Au,u) = \sum_{\alpha\in\Omega_h} [b_\alpha |u_\alpha|^2 - 2\sum \tilde{c}_{\alpha,e_i} \operatorname{Re}(\bar{u}_\alpha u_{\alpha+e_i})]$$

$$= \sum_{\alpha\in\Omega_h} \left[b_\alpha - \sum_i (\tilde{c}_{\alpha,e_i} + \tilde{c}_{\alpha-e_i,e_i}) \right] |u_\alpha|^2 + \sum_\alpha \sum_i \tilde{c}_{\alpha,e_i} |u_\alpha - u_{\alpha+e_i}|^2$$

we have

$$(3.5)\quad ([\tilde{D}+L]\tilde{D}^{-1}[\tilde{D}+L^*]u,u) = (Au,u) + ([\tilde{D}-D+L\tilde{D}^{-1}L^*]u,u)$$

$$\geqq \sum_{\alpha\in\Omega_h} \left[(\omega_\alpha^{-1} - \omega_\alpha \beta_\alpha^2 \tilde{\gamma}_\alpha^2) b_\alpha |u_\alpha|^2 - 2\sum_i (1-\omega_\alpha\beta_\alpha\tilde{\gamma}) \tilde{c}_{\alpha,e_i} \operatorname{Re}(\bar{u}_\alpha u_{\alpha+e_i}) \right]$$

$$\geqq \min(2-\omega_\alpha)\,(Au,u)\,,$$

if $\omega_\alpha < 2$, $\forall\alpha$, since then

$$\omega_\alpha^{-1} - \omega_\alpha\beta_\alpha^2\tilde{\gamma}_\alpha^2 = (\omega_\alpha^{-1} + \beta_\alpha\tilde{\gamma}_\alpha)(1-\omega_\alpha\beta_\alpha\tilde{\gamma}_\alpha)$$

$$= \omega_\alpha^{-1}\left[2 - \min_\alpha(2-\omega_\alpha)\right] \min_\alpha(2-\omega_\alpha) \geqq \min_\alpha(2-\omega_\alpha)\,.$$

If $\omega_\alpha \geqq 2$ for any α, then from (3.4) and (3.3) with $\beta_\alpha = 1$,

(3.6)

$$([\tilde{D}+L]\tilde{D}^{-1}[\tilde{D}+L^*]u,u) \geqq \sum_{\alpha\in\Omega_h} \Bigg\{ \left[b_\alpha - \sum_i (\tilde{c}_{\alpha,e_i} + \tilde{c}_{\alpha-e_i,e_i}) \right] |u_\alpha|^2$$

$$+ (\omega_\alpha^{-1} - 1 + \sum_i \omega_{\alpha-e_i} b_\alpha^{-1} \tilde{c}_{\alpha-e_i,e_i} \tilde{\gamma}_{\alpha-e_i})\, b_\alpha |u_\alpha|^2$$

$$+ \sum_i (1-\omega_\alpha\tilde{\gamma}_\alpha) \tilde{c}_{\alpha,e_i} |u_\alpha - u_{\alpha+e_i}|^2 \Bigg\}$$

$$> \min_\alpha (1-\omega_\alpha\tilde{\gamma}_\alpha)\,(Au,u)\,,$$

since, according to Lemma 3.1, $1 > 1-\omega_\alpha\tilde{\gamma}_\alpha \geqq 1-\omega_\alpha\gamma_\alpha > 0$, if

$$\omega_\alpha^{-1} - 1 + \sum_i \omega_{\alpha-e_i} b_\alpha^{-1} \tilde{c}_{\alpha-e_i,e_i} \tilde{\gamma}_{\alpha-e_i} = \delta \geqq 0\,.$$

Let

$$b_\alpha = \delta_\alpha b_\alpha + \sum_{i=1}^{n} (c_{\alpha,e_i} + c_{\alpha-e_i,e_i})\,,$$

where $\delta_\alpha \geqq 0$ comes from the term $q(x_\alpha)u_\alpha$ in (2.3), and let $\delta_1 = \min_{\alpha\in\Omega_h} \delta_\alpha$. Now choose

$$\delta = \max[0, Z^2/(1+\tau Z) - \delta_1]\,,$$

where

$$Z = \zeta h \geqq \frac{1}{\eta} - 1 ,$$

$\zeta > 0$, $\tau \geqq 1$ and where

$$\eta = \min_{\alpha \in \tilde{\Omega}_h^{(I)}} \sum_i c_{\alpha - e_i, e_i} / \sum_i \tilde{c}_{\alpha, e_i} \tag{3.7}$$

and (for example) the step-size parameter

$$h = \min_{\alpha \in \Omega_h} \sum_{i=1}^{n} h_{\alpha, e_i} / \sum_{\alpha \in \Omega_h} 1 > 0 .$$

We observe that if the h_{α, e_i} varies according to some function of h, which as the coefficients a_i are Lipschitz continuous in the x_i-direction, then $(1-\eta)/h$ is finite, $h \to 0$.

The following Lemma gives an upper bound of ω_α.

LEMMA 3.3. *Let* $\{\omega_\alpha\}$ *be defined by*

$$\omega_\alpha^{-1} = 1 + \delta - \sum_i \omega_{\alpha - e_i} \, b_\alpha^{-1} \, \tilde{c}_{\alpha - e_i, e_i} \, \tilde{\gamma}_{\alpha - e_i}$$

where we suppose that $Z > 0$ *satisfies*

$$\delta = \frac{Z^2}{(1 + \tau Z)} - \delta_1 \geqq 0$$

with $\tau = 1$, *and*

$$\min_{\alpha \in \Omega_h / \tilde{\Omega}_h^{(I)}} \left(\sum_i (c_{\alpha - e_i, e_i} - \tilde{c}_{\alpha - e_i, e_i}) \Big/ \sum_i \tilde{c}_{\alpha, e_i} \right) \geqq Z \geqq \eta^{-1} - 1 ,$$

where η *is defined in* (3.7). *Then if* $\alpha \in \Omega_h^{(II)}$, $\omega_\alpha^{-1} = (1 + Z_\alpha)\tilde{\gamma}_\alpha$, *where* $\tilde{\gamma}_\alpha$ *is defined in* (3.1) *and where* $Z_\alpha \geqq Z > 0$.

PROOF (induction). Suppose $Z_{\alpha - e_i} \geqq Z$. This is true if $\alpha - e_i \notin \tilde{\Omega}_h^{(I)}$, since then according to Lemma 3.1, with $\alpha' = \alpha - e_i$, $\omega_{\alpha'}^{-1} > \gamma_{\alpha'}$. Thus

$$(1 + Z_{\alpha'}) \sum_i \tilde{c}_{\alpha', e_i} / b_{\alpha'} = \omega_{\alpha'}^{-1} > \sum_i (c_{\alpha', e_i} + c_{\alpha' - e_i, e_i} - \tilde{c}_{\alpha' - e_i, e_i}) / b_{\alpha'} ,$$

i.e.

$$1 + Z_{\alpha'} > \sum_i (c_{\alpha', e_i} + c_{\alpha - e_i, e_i} - \tilde{c}_{\alpha' - e_i, e_i}) / \sum \tilde{c}_{\alpha', e_i} \geqq 1 + Z .$$

[We observe that the upper bound of Z is possible to satisfy for all non-pathological choices of the grid, since then this bound is $O(1)$, $h \to 0$.]

Now let $\alpha \in \hat{\Omega}_h^{(I)} \cap \Omega_h^{(II)}$. Then with

$$\omega_\alpha^{-1} = (1+Z_\alpha)\sum_i \tilde{c}_{\alpha,e_i}/b_\alpha$$

we have

$$\begin{aligned} 0 &= b_\alpha\omega_\alpha^{-1}+\sum_i \omega_{\alpha-e_i}\,\tilde{c}_{\alpha-e_i,e_i}\,\tilde{\gamma}_{\alpha-e_i}-(1+\delta)b_\alpha \\ &\leqq (1+Z_\alpha)\sum \tilde{c}_{\alpha,e_i}+\frac{1}{1+Z}\sum \tilde{c}_{\alpha-e_i,e_i} \\ &\quad -(\delta_\alpha+\delta)b_\alpha-\sum (c_{\alpha,e_i}+c_{\alpha-e_i,e_i})\,, \end{aligned}$$

and since $\tilde{c}_{\alpha,e_i}\leqq c_{\alpha,e_i}$, we have

$$Z_\alpha\sum_i \tilde{c}_{\alpha,e_i} \geqq \left[1-\frac{1}{1+Z}+\delta_1+\delta\right]\sum \tilde{c}_{\alpha-e_i,e_i}+(\delta_1+\delta)\sum \tilde{c}_{\alpha,e_i}\,,$$

i.e.

$$Z_\alpha \geqq Z\left(\eta+\frac{Z}{1+Z}\right) \geqq Z\,.$$

Running over the grid in the natural order shows that $Z_\alpha\geqq Z, \forall\alpha$.

For the restricted model problem we have

$$c_{\alpha,e_i} = c_{\alpha-e_i,e_i},\ \sum_i c_{\alpha,e_i}/b_\alpha = \tfrac{1}{2},\ \alpha\in\Omega_h \text{ and } \eta = 1\,.$$

Then we get a better bound of ω_α.

LEMMA 3.4. *Let*

$$c_{\alpha,e_i} = c_{\alpha-e_i,e_i},\ \sum c_{\alpha,e_i}/b_\alpha \geqq \tfrac{1}{2},\quad \alpha\in\Omega_h\,,$$

and let $\{\omega_\alpha\}$ *be defined as in Lemma* 3.3 *with* $\tau=\sqrt{2}$. *Let*

$$0 < \sqrt{2}\,Z \leqq \min_{\alpha\in\Omega_h/\Omega_h(I)} \sum_i \left(c_{\alpha-e_i,e_i}-c_{\alpha-e_i,e_i}\right)\Big/\sum_i \tilde{c}_{\alpha,e_i}$$

Then

$$\tilde{\omega}_\alpha^{-1} = \tfrac{1}{2}(1+Z_\alpha),\quad \textit{where } Z_\alpha \geqq \sqrt{2}\,Z > 0\,.$$

PROOF. This is proved in the same way as the previous Lemma.

Let

$$\Lambda_1 = \lambda_1 h^2 = \min_u\left((Au,u)/(Du,u)\right)\,. \tag{3.8}$$

We now state the main theorem.

THEOREM 1. *Suppose that the conditions and definitions in Lemma* 3.3 *and Lemma* 3.4, *respectively, are valid. Let* μ_1, μ_2 *be the smallest and largest eigenvalues of the matrix* $\tilde{N}(\zeta)=N^{-1}A$, *respectively, where*

$$N = (\tilde{D}+L)\,\tilde{D}^{-1}\,(\tilde{D}+L^*)\,.$$

Then

$$\mu_1 \geqq [1+(Z^2/(1+\tau Z)-\delta_1)/\Lambda_1]^{-1}, \quad \mu_2 \leqq \frac{1+\tau Z}{\tau\tau_1 Z}\,,$$

where $\tau=\sqrt{2}$, $\tau_1=2$ *for the restricted model problem or otherwise* $\tau=\tau_1=1$ *and where* $Z=\zeta h$ *and* $\Lambda_1=\lambda_1 h^2$ *is defined in* (3.8).

PROOF. Since N and A are Hermitian and postive definite the lower and upper limits of $(Au,u)/(Nu,u)$ give the eigenvalues μ_1, μ_2 of $N^{-1}A$. From (3.2) we have

$$\begin{aligned}(Nu,u)/(Au,u) &\leqq 1+\left[\frac{Z^2}{1+\tau Z}-\delta_1\right]\max(Du,u)/(Au,u)\\ &= 1+[Z^2/(1+\tau Z)-\delta_1]/\Lambda_1\,.\end{aligned}$$

From (3.5) and (3.6) we have

$$(Nu,u)/(Au,u) \geqq \begin{cases}\min\limits_{\alpha}(2-\omega_\alpha), & \text{if } \omega_\alpha \leqq 2, \forall\alpha\\ \min\limits_{\alpha}(1-\tilde{\omega}_\alpha\gamma_\alpha), & \text{if } \omega_\alpha \geqq 2, \text{ any } \alpha\,.\end{cases}$$

From Lemma 3.1, we observe that $\omega_\alpha<2$ for the restricted model problem and from Lemma 3.4 we then have

$$2-\omega_\alpha = 2-2/(1+Z_\alpha) = \frac{2Z_\alpha}{1+Z_\alpha} \geqq \frac{2\tau Z}{1+\tau Z}$$

or else from Lemma 3.3

$$1-\omega_\alpha\tilde{\gamma}_\alpha = 1-\frac{1}{1+Z_\alpha} = \frac{Z_\alpha}{1+Z_\alpha} \geqq \frac{Z}{1+Z}\,.$$

This proves the theorem.

We thus see that the spectral radius $\varkappa(\tilde{N}(\zeta))$ of $\tilde{N}$ satisfies

$$\varkappa(N(\zeta)) \leqq \frac{1}{\tau\tau_1}\left[(Z^{-1}+\tau)\left(1-\frac{\delta_1}{\Lambda_1}\right)+Z/\Lambda_1\right] = O(h^{-1}),\ h\to 0$$

and this upper bound is minimized for $Z=Z_{\text{opt}}=\sqrt{\Lambda_1-\delta_1}$, since obviously $\delta_1<\Lambda_1$. We have

$$\varkappa(\tilde{N}(\zeta_{\text{opt}})) = \frac{1}{\tau\tau_1}\,[2\sqrt{\Lambda_1-\delta_1}/\Lambda_1+\tau(1-\delta_1/\Lambda_1)] \tag{3.9}$$

For the restricted model problem over a unit hyper-square, we have (see Axelsson [5]),

$$\lambda_1 = \frac{\pi^2}{2}[1+O(h^2)],$$

i.e. with $\zeta_{\text{opt}} = h^{-1}Z_{\text{opt}}$,

$$\zeta_{\text{opt}} = \frac{\pi}{\sqrt{2}}[1+O(h^2)]$$

and

$$\varkappa(\tilde{N}(\zeta_{\text{opt}})) = (\pi h)^{-1} + \tfrac{1}{2} + O(h), \quad h \to 0.$$

This is in accordance with the result (3.3) for the point-wise SSOR method in Axelsson [5].

It should be mentioned that in an actual computation over an irregular grid, we do not need to calculate the step-size parameter h. What is needed is only $Z = \zeta h$ and $\Lambda_1 = \lambda_1 h^2$. (In Section 4 we give a lower bound of Λ_1.) However, to judge the rate of convergence, when refining the grid, it is convenient to have this parameter.

The condition $\eta = 1 + O(h)$, in Theorem 1, is evidently satisfied if the coefficients a_i are continuously differentiable and if

$$|h_{\alpha,e_i} - h_{\alpha-e_i,e_i}|/h$$

is $o(1)$, $\alpha \in \tilde{\Omega}_h^{(I)}$, $h \to 0$. However, even for certain kinds of discontinuity among the coefficients it is also possible to satisfy this condition by matching the grid against the coefficients. Thus, along a plane of discontinuity, with normal in the e_i direction, and where a_i is constant on each side of the plane, we choose h_{α,e_i} and $h_{\alpha-e_i,e_i}$ so that

$$a_i(x_{\alpha+\frac{1}{2}e_i})/h_{\alpha,e_i} - a_i(x_{\alpha-\frac{1}{2}e_i})/h_{\alpha-e_i,e_i} = O(h),$$

where α is a point on the plane of discontinuity. To see why this satisfies the condition $\eta = 1 + O(h)$, we approximate the law of continuous flow, i.e. $a_i(x_\alpha)(\partial u/\partial x_i)$ constant on both sides of the plane, by central differences.

Thus

$$a_i(x_{\alpha-\frac{1}{2}e_i})\frac{u_\alpha - u_{\alpha-e_i}}{h_{\alpha-e_i,e_i}} = a_i(x_{\alpha+\frac{1}{2}e_i})\frac{u_{\alpha+e_i} - u_\alpha}{h_{\alpha,e_i}}$$

or

$$\left(\frac{a_i(x_{\alpha-\frac{1}{2}e_i})}{h_{\alpha-e_i,e_i}} + \frac{a_i(x_{\alpha+\frac{1}{2}e_i})}{h_{\alpha,e_i}}\right)u_\alpha - \frac{a_i(x_{\alpha-\frac{1}{2}e_i})}{h_{\alpha-e_i,e_i}}u_{\alpha-e_i} - \frac{a_i(x_{\alpha+\frac{1}{2}e_i})}{h_{\alpha,e_i}}u_{\alpha+e_i} = 0.$$

This equation is of the type (2.2). This technique, however, is not practical if the quotient between the coefficients on both sides of the line

of discontinuity is too large. In that case, the following technique may be feasible (cf. Gunn [11]). Let the matrix N correspond to a matrix B obtained from the same Dirichlet problem, but with some properly chosen smooth coefficients. From (3.4) it follows that the resulting coefficient matrix is spectrally equivalent to the given matrix A, corresponding to the discontinuous coefficients, i.e.

$$\varrho_1(Bx,x) \leqq (Ax,x) \leqq \varrho_2(Bx,x)$$

where ϱ_1, ϱ_2 are the least and largest quotients, respectively, between the coefficients c_{α,e_i} of A and c_{α',e_i} of B. Thus the condition number of $N^{-1}A$ is $\leqq \varrho_2/\varrho_1$ times that of $N^{-1}B$, which is $O(h^{-1})$. This technique of dealing with spectrally equivalent operators can apparently also be applied in connection with higher order approximations, like the "nine-point" difference operator.

$$\Delta_h^{(9)}u(x) = f(x)+\frac{h^2}{12}\Delta_h^{(5)}f(x)\,,$$

where $\Delta u(x)=f(x)$, for the Laplacian operator Δ (of Bramble and Hubbard [17]).

It ought to be mentioned that the method of choosing different relaxation parameters can also be seen as a procedure using scaling by diagonal matrices (cf. Widlund [8]). Thus, let

$$L_{\tilde{D}} = \tilde{D}^{-\frac{1}{2}} L\tilde{D}^{-\frac{1}{2}}, \quad A_{\tilde{D}} = \tilde{D}^{-\frac{1}{2}} A\tilde{D}^{-\frac{1}{2}}\,.$$

Then

$$\tilde{D}^{\frac{1}{2}}\tilde{N}\tilde{D}^{-\frac{1}{2}} = \tilde{D}^{\frac{1}{2}}(\tilde{D}+L^*)^{-1}\tilde{D}(\tilde{D}+L)^{-1}A\tilde{D}^{-\frac{1}{2}} = (I+L_{\tilde{D}}^*)^{-1}\,(I+L_{\tilde{D}})^{-1}A_{\tilde{D}}\,,$$

so that the matrix $\tilde{N}$ is similar to this matrix, consisting of scaled factors of the original matrix.

Finally, we notice that although the derivation of the scaling diagonal matrices in this paper is different from that in Dupont [9], it is easily seen that the above analysis, apart from minor details, has led to the same scaling as that in Dupont's paper. Here, however, an analysis is given of the spectral condition number, which permits us to choose the parameter ζ properly, and which also shows that the spectral condition number does not vary much with ζ. Also we have generalized the method to be valid for a general region and an irregular grid. In particular, since the step-lengths do not need to be constant, we can use the given boundary values directly, without needing to approximate the values on a boundary, chosen in some sense close to the given boundary. This latter procedure, adjusting the boundary, is common in many papers.

The problem parameter.

In order to estimate the value of the optimal parameter Z_{opt}, we need an estimation of the problem parameter Λ_1 in (3.9). At first we note that

$$\Lambda_1 = \min_u((Au,u)/(Du,u)) = 1-\max(D^{-1}[L+L^*]u,u) = 1-\varrho_1 ,$$

where

$$\varrho_1 = S(D^{-1}(L+L^*))$$

is the spectral radius of the Jacobian iteration matrix $D^{-1}(L+L^*)$. With the optimal relaxation parameter

$$\omega_{\text{opt}} = 2/[1+(1-\varrho_1^2)^{\frac{1}{2}}]$$

of the SOR method (see Young [4]), we have the convergence factor of the SOR method

$$\omega_{\text{opt}}-1 = [1-(1-\varrho_1^2)^{\frac{1}{2}}]/[1+(1-\varrho_1^2)^{\frac{1}{2}}] \sim 1-2\sqrt{2\Lambda_1} ,$$

and the necessary number of iterations to reach a relative accuracy of $1>\varepsilon>0$, is

$$\sim \ln \varepsilon^{-1}/[2\sqrt{2\Lambda_1}] .$$

If we use the iteration formula

$$\omega^{(l+1)} = \omega^{(l)}-\beta N^{-1}(A\omega^{(l)}-\tilde{f}) ,$$

with fixed iteration parameter

$$\beta = 2/(\mu_1+\mu_2) ,$$

we get a convergence factor $\sim 1-2\mu_1/\mu_2$, for which from (3.9) we have the upper bound $\sim 1-\tau\tau_1\sqrt{\Lambda_1}$, if $\delta_1=0$. In the constant coefficient case, where $\tau=2$, $\tau_1=\sqrt{2}$, we thus have a convergence factor which is less than that of the optimal SOR method.

Furthermore, as well known, the efficiency of the SOR method is very sensitive to the parameter ω, while the upper bound of the spectral condition number of $N^{-1}A$ varies like $O(Z^{-1}+Z/\Lambda_1)$, and is thus rather insensitive of the choice of $Z>0$. As will be apparent from the following sections, this is even more so, when the iterations are accelerated, in particular, when a variant of the conjugate gradient method is used. As is well known, it is not possible to accelerate the SOR iterations.

From the above, it follows that we need only a rough bound of Λ_1. A lower bound of Λ_1 gives an upper bound of the optimal spectral condition number, as follows immediately from (3.9). The following simple analysis yields such a bound.

Since D and A are Hermitian and positive definite, we have

$$\Lambda_1 = \min((Au,u)/(Du,u)) = \min_u((D^{-\frac{1}{2}}AD^{\frac{1}{2}}u,u)/(u,u))$$

$$\geqq \min_u \sum_{\alpha\in\Omega_h}\sum_i b_\alpha^{-\frac{1}{2}} b_{\alpha+e_i}^{-\frac{1}{2}} \tilde{c}_{\alpha,e_i} |\tilde{u}_\alpha - \tilde{u}_{\alpha+e_i}|^2 \Big/ \sum_\alpha |\tilde{u}_\alpha|^2$$

$$\geqq \min_u \sum_{\alpha\in\tilde{\Omega}_h^{(II)}} \sum_i c'_i |\tilde{u}_\alpha - \tilde{u}_{\alpha+e_i}|^2 \Big/ \sum_\alpha |\tilde{u}_\alpha|^2 ,$$

where

$$c'_{\alpha,e_i} = c'_i = \sum_{\alpha\in\tilde{\Omega}_h^{(II)}} (b_\alpha^{-\frac{1}{2}}\, b_{\alpha+e_i}^{-\frac{1}{2}}\, c_{\alpha,e_i})$$

and

$$\tilde{u}_\alpha = \begin{cases} u_\alpha & \alpha \in \Omega_h \\ 0\,, & \alpha \in \partial\Omega_h\,. \end{cases}$$

Let A′ be the matrix, which is obtained as A but from a region being an extension to a rectangular region R_h, covering $\Omega_h \cup \partial\Omega_h$ (see Fig. 1) and with coefficients c'_i and

$$b'_\alpha = \sum (c'_{\alpha,e_i} + c'_{\alpha-e_i,e_i})\,.$$

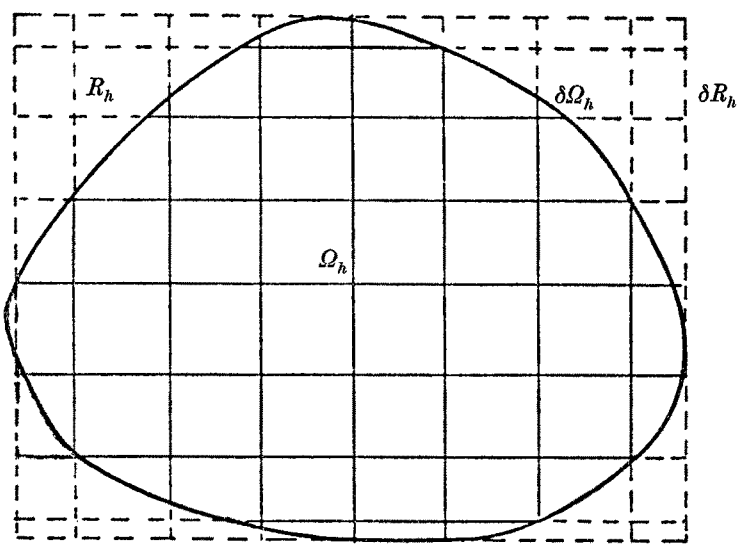

Fig. 1.

Then

$$\min((A'u,u)/\sum b'_\alpha |\tilde{\tilde{u}}_\alpha|^2) = \min_{\alpha\in R_h} \sum_i c'_i |\tilde{\tilde{u}}_\alpha - \tilde{\tilde{u}}_{\alpha+e_i}|^2 / \sum |\tilde{\tilde{u}}_\alpha|^2\,,$$

where

$$\tilde{\tilde{u}}_\alpha = \begin{cases} u_\alpha, & \alpha \in R_h \\ 0\,, & \alpha \in \partial R_h\,. \end{cases}$$

It is well-known that the eigenvalues of this matrix are

$$\sum_i c'_i 2(1-\cos p_i\pi/(n_i+1)) = \sum c'_i 4\sin^2\frac{p_i\pi}{2(n_i+1)}, \quad p_i = 1, 2, \ldots, n_i\,,$$

where R_h consists of n_i interior points in the e_i direction. Thus

$$\Lambda_1 \geqq \sum_i \min_{\alpha\in\tilde{\Omega}_h(II)} (b_\alpha^{-\frac{1}{2}} b_{\alpha+e_i}^{-\frac{1}{2}} c_{\alpha,e_i}) 4\sin^2\frac{\pi}{2(n_i+1)}\,. \tag{4.1}$$

Acceleration methods

In Young [18] a process of the form

$$w^{(l+1)} = w^{(l)} - \beta_l N^{-1} r^{(l)}$$

where $r^{(l)} = Aw^{(l)} - \tilde{f}$ is the residual, is tried to accelerate the iterations obtainable with a fixed parameter $\beta_l = \beta$. It is possible to choose the set $\{\beta_l\}$ so that

$$w^{(p)} = R_p w^{(0)} + C_p \tilde{f}\,,$$

where

$$I - R_p = C_p A$$

$$R_p = \frac{1}{T_p\left(\dfrac{b+a}{b-a}\right)} T_p\left(\frac{1}{b-a}\left[(b+a)I - 2N^{-1}A\right]\right),$$

and T_p is the Chebyshev polynomial of degree p with $0 < a \leqq \mu_1 < \mu_2 \leqq b$. This process, however, is numerically unstable, although Young [18] showed that the influence of the instability can be diminished by choosing the parameters β_l in a certain order. A two-step version,

$$w^{(l+1)} - \alpha_l w^{(l)} + (\alpha_l - 1) w^{(l-1)} = -\beta_l \xi^{(l)}\,,$$

where $\xi^{(l)} = N^{-1} r^{(l)}$, of that process, however, is stable (see Stiefel [19]). By elementary calculations (see Axelsson [20]) it is possible to show that if

$$\alpha_l = \frac{a+b}{2}\beta_l\,,$$

$$\beta_l^{-1} = \frac{a+b}{2} - \left(\frac{b-a}{4}\right)^2 \beta_{l-1}, \quad l = 1,2, \ldots$$

$$\alpha_0 = 1, \beta_0 = 4/(a+b),$$

apart from rounding errors, we get the same sequence of vectors as above.

We have

$$r^{(p)} = R_p^* r^{(0)}. \tag{5.1}$$

Let

$$\langle u, v\rangle = (N^{-1}u, v),$$

where

$$(u, v) = \sum_{\alpha \in \Omega_h} \tilde{u}_\alpha v_\alpha .$$

In the Chebyshev acceleration process we minimize the spectral radius of the iteration error matrix R_p. Since

$$\|N^{-\frac{1}{2}} R_p^* N^{\frac{1}{2}}\|_2 = S(N^{-\frac{1}{2}} R_p^* N^{\frac{1}{2}}) = S(R_p^*) = S(R_p),$$

this is equivalent to minimizing the two-norm

$$\|R_p^*\|'_2 = \max_u (\langle R_p^* u, R_p^* u\rangle / \langle u, u\rangle)^{\frac{1}{2}} = \|N^{-\frac{1}{2}} R_p^* N^{\frac{1}{2}}\|_2$$

of the residual error matrix R_p^* in (5.1).

In this minimizing process we only use the information about lower and upper bounds, a, b of the eigenvalues μ_1, μ_2 respectively, of $N^{-1}A$. A simple calculation shows that the above norm of R_p^* is decreased to a value $\leqq \varepsilon$, $\varepsilon > 0$, if $p \geqq \frac{1}{2}\sqrt{b/a}\ln 2/\varepsilon$. Experiments show that this is a very accurate value of the actually needed number of iterations. The bounds given in Theorem 1, particularly in connection with a discretized variable coefficient Dirichlet problem, shows however to be too pessimistic and thus give an unnecessarily too slow rate of convergence (c.f. Section 6). However, using a conjugate gradient method (see Lanczos [7], Hestenes and Stiefel [6]), these bounds are not used and a faster procedure is obtained in a way almost as simple as in the Chebyshev acceleration method. In the variant of the conjugate gradient method considered here, we use weighted residuals,

$$(N^{-1}r^{(p)}, r^{(p)})$$

as above, which has computational advantages. We will show how to choose the set of parameters $\{\alpha_l\}$, $\{\beta_l\}$ in order to get residuals, which are mutually orthogonal with respect to $\langle , \rangle$. From Hestenes [21], it follows that this process yields the least square error $\langle e^{(l)}, e^{(l)}\rangle$, where $e^{(l)} = w^{(l)} - A^{-1}\tilde{f}$

is the iteration error, on the simplex with vertices $r^{(j)}$, $j=0,1,\ldots,l$, constructed by a linear acceleration process of the above kind. It is also possible to construct such an acceleration process, which gives the least square residual over the subspace generated by $\{r^{(j)}\}$. However, such a process is more complicated and, furthermore, experiments showed the number of necessary iterations to be the same. Thus, let

$$w^{(1)} = w^{(0)} - \beta_0 \xi^{(0)}$$

and

$$w^{(l+1)} - \alpha_l w^{(l)} + (\alpha_l - 1) w^{(l-1)} = -\beta_l \xi^{(l)}$$

where $\xi^{(l)} = N^{-1} r^{(l)}$ or

$$r^{(1)} = r^{(0)} - \beta_0 A \xi^{(l)}$$

$$r^{(l+1)} - \alpha_l r^{(l)} + (\alpha_l - 1) r^{(l-1)} = -\beta_l A \xi^{(l)}, \quad l = 1,2, \ldots .$$

Since N and A are Hermitian, we get from

$$\langle r^{(1)}, r^{(0)} \rangle = 0, \quad \text{i.e. } (r^{(1)}, \zeta^{(0)}) = 0$$

that

$$\beta_0 = (r^{(0)}, \xi^{(0)})/(A\xi^{(0)}, \xi^{(0)}) ,$$

and from

$$\langle r^{(l+1)}, r^{(l-j)} \rangle = 0, \quad j = 0,1$$

that

$$\alpha_l/\beta_l = \gamma_l$$

and

$$(1-\alpha_l)/\beta_l = (A\xi^{(l)}, \xi^{(l-1)})/\delta_l = (\xi^{(l)}, A\xi^{(l-1)})/\delta_l = -\beta_{l-1}^{-1} (\xi^{(l)}, r^{(l)})/\delta_l ,$$

where

$$\gamma_l = (A\xi^{(l)}, \xi^{(l)})/\delta_l ,$$

and

$$\delta_l = \langle r^{(l)}, r^{(l)} \rangle = (r^{(l)}, \xi^{(l)}) = (N\xi^{(l)}, \xi^{(l)}) .$$

It is immediately seen that in this way $r^{(l+1)}$ is also orthogonal to $r^{(l-j)}$, $j=0,1,\ldots$. Thus the sequence

$$\begin{aligned} \beta_0^{-1} &= \gamma_0 \\ \beta_l^{-1} &= \gamma_l - \beta_{l-1}^{-1} \delta_l/\delta_{l-1} \\ \alpha_l &= \gamma_l \beta_l, \quad l = 1,2, \ldots \end{aligned}$$

gives the desired least squaring process.

Since, as mentioned, this variant of the conjugate gradient method decreases the weighted residual error $(N^{-1} r^{(p)}, r^{(p)})$ in practice as fast as

the fastest of all linear acceleration methods, it has a rate of convergence, which is at least as fast as that for the Chebyshev acceleration method. We observe that since

$$(r^{(p)}, r^{(p)}) \leqq \max(\lambda_N)\ (N^{-1} r^{(p)}, r^{(p)})$$

and

$$\max(\lambda_N) \leqq \max_u \frac{(Nu,u)}{(Au,u)} \max_u \frac{(Au,u)}{(u,u)} = O(1), \quad h \to 0\,,$$

the two-norm of the residual is bounded upwards by a constant, independent of h, times the above minimized weighted residual.

Finally, it should be mentioned that in Marchuk [16], where only one parameter is needed to decrease the residuals, that parameter is chosen so as to minimize $(r^{(l+1)}, r^{(l+1)})/(r^{(l)}, r^{(l)})$, i.e. $\alpha_l = 1$ and $\beta_l = (A\xi^{(l)}, r^{(l)})/(A\xi^{(l)}, A\xi^{(l)})$ (see Axelsson [20]). This is thus the steepest descent method. Numerical experiments (see Section 6) confirm the superiority of the conjugate gradient acceleration method.

Numerical experiments

To test the described methods some examples on plane regions were run. In all of them $q(x) \equiv 0$, which evidently is no restriction. The iterations were interrupted when the relative error was $\leqq \varepsilon = 10^{-6}$, i.e. when $\|r^{(l)}\|_2 \leqq \varepsilon \|r^{(0)}\|_2$. The following regions were dealt with.

EXAMPLE 1. A unit square, $a_i \equiv 1$, $h_i = h$ and continuous boundary conditions.

EXAMPLE 2. As example 1, but with

$$a_1(x,y) = 1/(1+2x^2+y^2), \quad a_2(x,y) = 1/(1+x^2+2y^2)\,.$$

EXAMPLE 3. A square as in example 1, but with a square (symmetric) hole of side $\frac{1}{2}$.

EXAMPLE 4. An ellipse, with semi-axes 0.5 and 0.3.

With the matrix N and the bounds a, b chosen as described in Section 3, i.e.

$$a = 1/(1+\delta/\Lambda_1),\ b^{-1} = \max\left\{\min_\alpha (2-\omega_\alpha),\ \min_\alpha (1-\omega_\alpha \tilde{\gamma}_\alpha)\right\},$$

Example 1 was run with no acceleration, i.e. $\alpha_l = 1$ and $\beta_l = \beta$, with Chebyshev acceleration and with the least squaring acceleration. When

$\alpha_l = 1$ and the parameter $\beta_l = \beta$ is constant, this parameter was chosen to minimize the spectral radius of the matrix of iteration, i.e. $\beta = 2/(a+b)$.

The results, given in Table 1, confirm that the number of iterations for a fixed h varies as a function of $\zeta = Z/h$ like $\zeta^{-1} + \zeta/\lambda_1$, except for very small ζ, in the case of no acceleration and as $(\zeta^{-1} + \zeta/\lambda_1)^{\frac{1}{2}}$ for the Chebyshev acceleration method. Here λ_1 is defined in (3.8). Furthermore, the number of iterations for ζ_{opt} as defined in (3.9), varies with h as $h^{-\frac{1}{2}}$ for the Chebyshev and least squaring acceleration methods, in conformity with the theory. The upper bound $b = (1 + \tau Z)/(\tau\tau_1 Z)$, given in Theorem 1 is accurate in the case of constant coefficients for not too small ζ ($\zeta \geqq 1$). Even then, however, it was possible to shrink the interval $[a, b]$ and reach the desired accuracy in somewhat fewer iterations. For instance, in example 1a, 16 iterations were enough (using $\zeta = 2.0$). For smaller values of ζ, it was possible to gain more. For instance, the bounds calculated for $\zeta = 1.0$ was usable also for $\zeta = 0$, yielding 21 iterations. For the least squaring acceleration procedure the number of iterations is less than those for the Chebyshev acceleration method and is almost independent of ζ, in a fairly wide range. In particular, even $\zeta = 0$ (i.e. $\delta = 0$) gives the best (or almost the best) result. To explain this, more refined bounds of μ_1, μ_2 than those given in Section 3, are needed. Since the number of arithmetic operations in the least squaring method is approximately twice that in the Chebyshev acceleration method, the experiments showed that the Chebyshev acceleration method was at least as economical as the least squaring method.

For comparison, the steepest descent method was also used. This procedure is slower than the least squaring method and the number of iterations varies with h as $O(h^{-1})$ for $\delta > 0$. However, for $\delta \leqq 0$ the experiments showed a minimum of 19 iterations in Example 1 for $\zeta \approx 0.5$ and this minimum seemed to vary with h as $O(h^{-\frac{1}{2}})$.

The number of iterations for Example 3, was much less than those for the restricted model problems in Example 1. It is wellknown, that it is difficult to get good estimates of ω_{opt} in connexion with the SOR method when dealing with regions having irregular shapes, holes etc. It is thus valuable to have procedures like the prescaled least squaring and the Chebyshev acceleration methods, which in practice needs no parameter-estimation or only a rough estimation, and which runs even faster when portions of a domain are cut off.

For the Chebyshev acceleration method, the bound given in (4.1) for Λ_1 was used. However, for irregular boundaries but with constant coefficients $a_i(x)$ and an internally uniform grid, this can lead to a too small estimate of Λ_1. In this case an approximation of Λ_1 of the form

$\Lambda_1 \approx \min_u((Au,u)/\max_{\alpha\in\Omega_h} b_\alpha\ (u,\ u))$ gives in practice better results. This approximation was used in Example 4.

Examples 1 and 2 were also run with a scaling using just one parameter $\omega = b_\alpha/d_\alpha$. As mentioned in the Introduction, according to the upper bounds derived in earlier papers, a spectral condition number $O(h^{-1})$ is in general possible to reach only for the constant coefficient case (Example 1). Experiments showed, however, that even in Example 2, with smooth variable coefficients, the same condition number was reached. The condition number varied, however, with the relaxation parameter with a rather sharp edged minimum, in contrast to the case for scaling with one parameter per mesh point as in Section 3.

Since the boundary in Example 4 is nonregular, the differentiation used in (2.3) has to be modified at the inner points with a neighbour on the boundary, in order to get a difference analogue of the form (2.2), and thus a Hermitian matrix A. If we use a convex combination of linear interpolations in the x_i-directions, such that

$$c_{\alpha,e_i} = \frac{h_{\alpha-e_i,e_i}}{h_{\alpha,e_i}} c_{\alpha-e_i,e_i}\,, \tag{5.1}$$

of $\alpha + e_i \in \partial\Omega_h$ (and correspondingly for other boundary points), however, we achieve the form (2.2). Furthermore, we still have the global discretization error $O(h^2)$, according to the general theorem in [15]. Since experiments showed minor differences in the number of iterations, using the differentiation (2.3) and using (5.1), it seems, however, as if (2.3) can be used instead of (5.1) even for irregular boundaries.

Acknowledgements.

This paper is one of those which I wrote during my stay at CERN, Geneva. I wish to express thanks to the DD-Division for its hospitality. The author also wishes to express thanks to Claes Örtendahl, Chalmers University of Technology, who prepared the experimental programs, and the Institute of Applied Mathematics, Stockholm, who granted Mr. Örtendahl's stay at CERN.

Number of iterations.

Table 1*a* (Example 1)
$h = 1/20, \quad \lambda_1 = \pi^2/2$

ζ	0	0.25	0.50	1.00	2.00	2.50	3.00	5.00	8.00
Acceleration: None	105	96	79	52	35	35	36	48	70
Chebyshev	28	26	25	21	18	18	18	21	25
Least squaring (conjugate gradient)	13	13	13	13	13	13	13	15	17
Steepest descent	24	25	28	26	24	28	31	34	40

Table 1*b* (Example 1)
$h = 1/40$

ζ	0	1.0	1.5	2.0	2.5	3.0	5.0	8.0
Acceleration: Chebyshev	42	29	27	26	26	26	29	35
Least squaring	18	18	18	17	17	18	20	23
Steepest descent	38	51	48	48	47	49	77	

Table 2*a* (Example 2)
$h = 1/20, \quad \lambda_1 = 4.01, \quad \eta = 0.974$

ζ	0	1.0	1.5	2.0	2.5	3.0	5.0
Acceleration: Chebyshev	32	22	20	20	20	20	24
Least squaring	14	14	13	13	13	13	15

Table 2*b* (Example 2)
$h = 1/40, \quad \lambda_1 = 3.98, \quad \eta = 0.986$

ζ	0	1.0	1.5	2.0	2.5	3.0	5.0
Acceleration: Chebyshev	50	31	28	28	28	29	34
Least squaring	19	18	18	18	18	18	21

Table 3*a* (Example 3)
$h = 1/20, \quad \lambda_1 = \pi^2/2$

ζ	0	0.5	1.0	1.5	2.0	2.5	3.0	4.0	8.0
Acceleration: Chebyshev	14	14	14	15	15	16	17	19	25
Least squaring	8	8	8	8	8	8	8	9	10

Table 3b (Example 3)
$h = 1/40$

ζ	0	0.5	1.0	1.5	2.0	2.5	3.0	4.0	8.0
Acceleration Chebyshev	22	22	22	23	24	24	25	28	36
Least squaring	12	12	12	12	12	12	12	12	13

Table 4a
$h = 1/20, \quad \lambda_1 = 9.3$

ζ	0	0.5	1.0	1.5	2.0	2.5	3.0	4.0	8.0
Acceleration: Chebyshev	23	21	18	17	16	15	15	16	19
Least squaring	12	12	12	11	11	11	11	11	13

Table 4b
$h = 1/40, \quad \lambda_1 = 9.3$

ζ	0	0.5	1.0	1.5	2.0	2.5	3.0	4.0	8.0
Acceleration: Chebyshev	33	29	25	24	23	21	22	21	27
Least squaring	16	16	15	15	14	14	14	14	16

REFERENCES

1. J. W. Sheldon, *On the numerical solution of elliptic equations*, MTAC 9 (1955), 101–111.
2. G. J. Habetler and E. L. Wachspress, *Symmetric successive overrelaxation in solving diffusion difference equations*, Math. Comp. 15 (1961), 356–362.
3. L. W. Ehrlich, *The block symmetric successive overrelaxation method*, J. Soc. Indust. Appl. Math. 12 (1964), 807–826.
4. D. M. Young, *Iterative solution of large linear systems*, Academic Press, New York and London, 1971.
5. O. Axelsson, *Generalized SSOR methods*, DD/72/8, CERN, Geneva, 1972.
6. M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Standards 49 (1952), 409–436.
7. C. Lanczos, *Solution of systems of linear equations by minimized iterations*, J. Res. Nat. Bur. Standards 49 (1952), 33–53.
8. O. B. Widlund, *On the effects of scaling of the Peaceman-Rachford method*, Math. Comp. 25 (1971), 33–41.
9. T. Dupont, *A factorization procedure for the solution of elliptic difference equations*, SIAM J. Numer. Anal. 5 (1968), 753–782.
10. H. L. Stone, *Iterative solution of implicit approximations of multi-dimensional partial differential equations*, SIAM J. Numer. Anal. 5 (1968), 530–558.
11. J. E. Gunn, *The solution of elliptic difference equations by semi-explicit iterative techniques*, J. SIAM Numer. Anal. Ser B 2 (1964), 24–45.

12. O. B. Widlund, *On the rate of convergence of an alternating direction implicit method in a noncommutative case*, Math. Comp 20 (1965), 500–515.
13. J. A. George, *The use of direct methods for the solution of the discrete Poisson equation on non-rectangular regions*, Computer Science Rep. 70–159, Stanford Univ., Stanford, Calif., 1970.
14. B. L. Buzbee, F. W. Dorr, J. A. George, G. H. Golub, *The direct solution of the discrete Poisson equation on irregular regions*, SIAM J. Numer. Anal. 8 (1971), 722–736.
15. J. H. Bramble, B. E. Hubbard and V. Thomée, *Convergence estimates for essentially positive type discrete Dirichlet problems*, Math. Comp. 23 (1969), 695–709.
16. G. I. Marchuk, *On the theory of the splitting up method*, SYNSPADE, Maryland, U.S.A. 1970.
17. J. H. Bramble and B. E. Hubbard, *On the formulation of finite difference analogues of the Dirichlet problem for Poisson's equation*, Numer. Math. 4 (1962), 313–327.
18. D. Young, *On Richardson's method for solving linear systems with positive definite matrices*, Journal Math. and Physics 32 (1954), 243–255.
19. E. L. Stiefel, *Kernel polynomials in linear algebra and their numerical applications*, Nat. Bur. Stand. Appl. Math. Ser. 49 (1958), 1–22.
20. O. Axelsson, *Lecture notes on iterative methods*, Report 72.04. Department of Computer Sciences, Chalmers University of Technology, Göteborg.
21. M. R. Hestenes, *The conjugate-gradient method for solving linear systems*, Proceedings of symposia in applied mathematics, vol VI, Mc Graw Hill, New York, 1956.

CERN
GENEVA
SWITZERLAND