ON PRECONDITIONING AND CONVERGENCE ACCELERATION
IN SPARSE MATRIX PROBLEMS

O. Axelsson
Propriété littéraire et scientifique réservée pour tous les pays du monde. Ce document ne peut 
btre reproduit ou traduit en tout ou en partie 
sans l’autorisation écrite du Directeur général 
du CERN, titulaire du droit d’auteur. Dans 
les cas appropriés, et s’il s’agit d’utiliser le 
document à des fins non commerciales, cette 
autorisation sera volontiers accordée.

Le CERN ne revendique pas la propriété des 
inventions brevetables et dessins ou modèles 
suscipients de dépôt qui pourraient être décrits 
dans le présent document; ceux-ci peuvent être 
librement utilisés par les instituts de recherche, 
les industriels et autres intéressés. Cependant, 
le CERN se réserve le droit de s’opposer à 
toute revendication qu’un usager pourrait faire 
de la propriété scientifique ou industrielle de 
toute invention et tout dessin ou modèle dé- 
crits dans le présent document.

© Copyright CERN, Genève, 1974

Literary and scientific copyrights reserved in 
all countries of the world. This report, or 
any part of it, may not be reprinted or trans-
lated without written permission of the 
copyright holder, the Director-General of 
CERN. However, permission will be freely 
granted for appropriate non-commercial use. 
If any patentable invention or registrable 
design is described in the report, CERN makes 
no claim to property rights in it but offers it 
for the free use of research institutions, manu-
facturers and others. CERN, however, may 
oppose any attempt by a user to claim any 
proprietary or patent rights in such inventions 
or designs as may be described in the present 
document.
ON PRECONDITIONING AND CONVERGENCE ACCELERATION
IN SPARSE MATRIX PROBLEMS

O. Axelsson
ABSTRACT

To simplify the storage of a matrix $A$ and to save computer storage, iterative methods are preferred in matrix problems. In this paper such methods are presented for the solution of a linear system of equations $Ax = f$. A combination of a simple preconditioning (of SSOR type) of a symmetric, positive definite matrix $A$ and a conjugate gradient convergence acceleration method, is shown to be faster than the fastest possible direct method for certain sparse matrices, arising in the discretization of partial differential equations. The preconditioning presented decreases the spectral condition number $\kappa(A)$ of $A$ to the order of magnitude of the square root $\sqrt{\kappa(A)}$, if a certain condition is satisfied.
CONTENTS

1. INTRODUCTION ................................................. 1
2. A CLASSICAL APPLICATION OF ITERATIVE METHODS .... 2
3. ONE-STEP CONVERGENCE ACCELERATION PROCESSES ... 4
4. PRECONDITIONING .............................................. 6
5. A DISCRETIZED SELF-ADJOINT ELLIPTIC PROBLEM .... 10
6. ON CONVERGENCE ACCELERATION ......................... 13
7. A NUMERICAL EXPERIMENT ................................. 19
REFERENCES ...................................................... 20
1. INTRODUCTION

In many applications, in particular in the discretization of partial differential equations, there arises the problem of the solution of a system of linear equations with a large sparse matrix $A = [a_{ij}]$. Often $A$ has banded form, i.e., $a_{ij} = 0$, $|i - j| > m$, where $m$ is a positive number, the bandwidth, and the number of non-zero elements in the band is small. An elimination method for the direct solution of the system of equations, will introduce "fill-in" of non-zero elements into a large portion in the band, where originally the elements were zero. (If a re-ordering of the equations is done in order to prevent too fast an increase of rounding errors, fill-in may also occur outside the band. Such re-orderings are however not necessary for positive definite matrices.) In order to keep the sparseness unchanged and thus to keep the demand for computer storage to a minimum, iterative methods must be used.

In fact, it is then not even necessary to store the matrix at all, instead a simple algorithm, like the averaging algorithm in the $(2n + 1)$ point difference approximation to an n-dimensional Poisson equation (cf. the difference equation in Section 2), can be applied. In a matrix arising from a more general finite-element approximation over a uniform mesh and a partial differential equation with constant coefficients, it is enough to store a small matrix, corresponding to only one mesh element [cf. Fried\(^1\)].

Thus, since the demand for storage is kept to a minimum, and since iterative algorithms are easily programmable, one can argue that they are advantageous, even if they are slower than some particular direct method. Actually there are even important applications where certain iterative algorithms are faster. Thus, for Hermitian, positive definite matrices, the iterative methods we are presenting are asymptotically faster than, or at least as fast as, the fastest elimination (LU-factorization) method if the mesh is fine enough [cf. Hoffman et al.\(^2\), George\(^3\) and Axelsson\(^4\)].

In this report we will present iterative methods in the form of convergence acceleration methods [semi-iterative methods, see Varga\(^5\)]. It will be shown that, for such methods, the number of necessary iterations to reach a certain accuracy, is directly proportional to the square root of the spectral condition number $\kappa(A)$ (the quotient between the largest and smallest eigenvalues) of the matrix $A$, if $A$ is Hermitian, positive definite.

It is also the purpose of this report to give a particular matrix $C$, consisting of sparse easily invertible factors, which under certain conditions have the effect of decreasing $\kappa(A)$ to $\kappa(C^{-1}A) \sim \kappa(A)^{\frac{1}{2}}$. Such a technique is an example of preconditioning of $A$, and the matrix $C$ is called a preconditioning matrix. The preconditioning we are referring to here is the simple SSOR method [see, for example, Habetler and Washspress\(^6\)] and Axelsson\(^7\)],

In a Banach space, where we do not necessarily have an inner product, in general the best convergence accelerator is the Chebyshev accelerator, and different ways of writing this method are mentioned below. This acceleration process minimizes the spectral radius (the largest absolute value of all eigenvalues) of the associated iteration matrix over an interval $[a,b]$, covering the spectrum of the preconditioned matrix $C^{-1}A$. In a Hilbert space, however, it is possible to give a simple method [a modification of the conjugate gradient method; see, for example, Hestenes and Stiefel\(^8\)], which minimizes the residual over-all possible convergence acceleration processes of a particular form.
2. A CLASSICAL APPLICATION OF ITERATIVE METHODS

Let us consider the simplest finite difference approximation of the Dirichlet problem

\[
\mathbf{Au} = \sum_{i=1}^{n} \frac{\partial}{\partial x_i} \left[ a_i(x) \frac{\partial u}{\partial x_i} \right] - g(x) = 0, \quad x \in \Omega,
\]

\[
u(x) = f(x), \quad x \in \partial \Omega,
\]

where \( \Omega \) is an open bounded region with boundary \( \partial \Omega \), \( a_i(x) > 0 \), \( x \in \bar{\Omega} \).

Let \( h \) be a mesh width parameter. Approximating the derivatives by central differences, we get a system of linear equations

\[
\mathbf{Au} = \mathbf{f}, \quad \mathbf{A} = \mathbf{D} + \mathbf{L} + \mathbf{U},
\]

where \( \mathbf{D} = \text{diag}(\mathbf{A}) \) and \( \mathbf{L}, \mathbf{U} \) are the lower and upper triangular parts of \( \mathbf{A} \), respectively; and are very sparse.

If \( a_i(x) \equiv 1 \), \( g \equiv 0 \) and \( n = 2 \), we get the discrete operator

\[
L_h \mathbf{u}_{jk} = u_{jk} - \frac{1}{4}(u_{j-1,k} + u_{j+1,k} + u_{j,k-1} + u_{j,k+1}) = 0,
\]

using an easily understandable notation.

In order to solve \( \mathbf{Au} = \mathbf{f} \), many iterative procedures have been proposed in the literature during the past sixty years. They are all special cases of the following general one-step procedure

\[
\mathbf{Cu}^{(\ell+1)} = \mathbf{Cu}^{(\ell)} - \tau_{\ell \epsilon}(\mathbf{Au}^{(\ell)} - \mathbf{f}), \quad \ell = 0, 1, \ldots,
\]

or of a corresponding two-step version.

Here \( \mathbf{C} \) is a preconditioning matrix, the sequence of numbers \( \{\tau_1, \tau_2, \ldots\} \) are suitably chosen, and \( \mathbf{u}^{(0)} \) is any starting vector. For instance, with \( \mathbf{C} = \mathbf{I} \) (the unit matrix),

\[
\tau_{\ell+1} = \omega > 0, \quad \ell = 0, 1, \ldots,
\]

we get Richardson's method\(^{11}\), which, written in component form is

\[
u^{(\ell+1)} = u^{(\ell)} - \omega L_h u^{(\ell)}, \quad u^{(0)} \text{ arbitrary},
\]

and is also called a simultaneous relaxation method. (The special case \( \omega = 1 \) is called the Gauss-Jacobi method.) For any consistent norm, it is easily seen that in order to decrease the relative error vector by a factor \( \epsilon > 0 \), i.e. to satisfy the inequality

\[
||\mathbf{u}^{(p)} - \mathbf{u}|| = \epsilon ||\mathbf{u}^{(0)} - \mathbf{u}||,
\]

\( p = O(h^{-2} \log 1/\epsilon) \) iterations are required, since the smallest eigenvalue of the matrix \( \mathbf{A} \) above is \( O(h^2) \). (We will see later how the number of iterations depends on \( \omega \).) Since \( h \) is a fairly small number this is far from satisfactory.

In 1918, Liebmann\(^{12}\) considered a method, with \( \mathbf{C} = \mathbf{D} + \omega \mathbf{L} \), which was later called the successive over-relaxation method (SOR). If we order the equations in expression (1) in a
row-wise way (in other words the first component of the vector is $u_{j,k}^{(l)}$, the second is $u_{j,k}^{(l+1)}$, etc.), then his method may be written

$$u_{j,k}^{(l+1)} = u_{j,k}^{(l)} - \omega \left[ u_{j,k}^{(l)} - \frac{1}{2} \left( u_{j-1,k}^{(l+1)} + u_{j+1,k}^{(l+1)} + u_{j,k-1}^{(l+1)} + u_{j,k+1}^{(l+1)} \right) \right].$$

We observe that this use of "improved" (latest calculated) values as soon as they are computed, makes the efficiency of the SOR method a function of the ordering of the equations.

If $\omega = 1$, this method is also called the Gauss-Seidel method, and can then be shown to demand approximately half as many iterations as the Gauss-Jacobi method. There is a thorough theory of the SOR method, using Perron-Frobenius theory for non-negative square matrices [see Varga\textsuperscript{11} and Young\textsuperscript{12}]. For so-called consistently ordered equations there is an optimal value of $\omega$, $1 < \omega < 2$, $\omega_{opt} = 2 / (1 + \sqrt{1 - \rho(B)^2})$, $B = D^{-1} (L + U)$, where $\rho(B)$ is the spectral radius of $B$, which will minimize the spectral radius of the iteration matrix $(D + \omega L)^{-1} [(1 - \omega) D - \omega L]$. This is valid, independent of the ordering, as long as it is consistent.

For $\omega = \omega_{opt}$, the number of iterations is $O(h^{-1} \log 1/\epsilon)$. However, this number is sensitive to the choice of $\omega$, and furthermore, since the eigenvalues of the associated matrix of iteration are situated on the circle $|z| = \omega_{opt} - 1$ in the complex plane, no acceleration method as given below exists for the optimal SOR method, because, as we will see, these methods demand real eigenvalues.

However, Sheldon\textsuperscript{13} generalized a method of Aitken\textsuperscript{15}, passing alternatively backwards and forwards over the net with the SOR method. The corresponding preconditioning matrix $C$ will appear in Section 4. Since the corresponding matrix of iteration

$$M = (D + \omega U)^{-1} [(1 - \omega)D - L] (D + \omega L)^{-1} [(1 - \omega)D - U]$$

is similar to a symmetric matrix if $U = L^*$, this method is called the symmetric successive over-relaxation method (SSOR method). It is easily proved that if $A$ is positive definite and $0 < \omega < 2$, then the eigenvalues of $M$ are in the interval $[0,1]$, so the spectral radius is $< 1$ and the method is convergent. The efficiency of this method is dependent on the ordering, but only slightly on the parameter $\omega$. If the condition $\rho(L L^*) \leq 1$, where $L = D^{-1} L D^{-1}$, is satisfied, the number of necessary iterations is $O(h^{-1} \log 1/\epsilon)$. (We will return to this condition in a more general form later on.) This was proved by Habetler and Wachspress\textsuperscript{5}. Ehrlich\textsuperscript{13} generalized the SSOR method to a block SSOR method, using block-diagonal matrices $D$.

Since the eigenvalues of $M$ are real, it will follow that we can use an acceleration method based on the Chebyshev polynomials. The number of iterations is now only $O(h^{-1} \log 1/\epsilon)$. It is easily seen that the condition $\rho(L L^*) \leq 1$ is satisfied if the coefficients $a_i(x)$ are constant over $\Omega$. However, in this paper we will show that a less restrictive condition is sufficient for the successful applicability of the SSOR method and that this condition is also satisfied for variable, but smooth enough, coefficients $a_i(x)$.

The Richardson method was generalized by a number of authors, among them Young\textsuperscript{16}, using variable parameters $\tau_{k+1}$ [see Eq. (2)]. After a fixed and predetermined number $p$ of iterations was performed, these parameters were chosen to yield a Chebyshev acceleration,
equivalent apart from rounding errors to the multiplication by the Chebyshev polynomial $T_p$ (see next section). The straightforward application of this method, however, is numerically unstable, but a two-step version of the method is stable, and demands no predetermination of the number of iterations. A two-step version can be found in a paper of Stiefel \cite{17} and in Section 6 of this paper.

A new approach in accelerating iterative methods was given by Peaceman and Rachford \cite{18} in 1955: the alternating-direction-implicit (ADI) method. Asymptotically, for a model Poisson problem with constant coefficients over a rectangle, the number of iterations is then $O(\ln h^{-1})$. However, it was shown by Birkhoff and Varga \cite{19} that the theory only holds for the above model problem.

To end this section, it should be noted that the eigenvalues of the corresponding matrix of iteration in Eq. (2) are $1 - \tau_{k+1} \lambda_{C^{-1}A}$ and the optimal choice of a fixed parameter

$\tau_{k+1} = \omega$ is $\omega = 2/(\mu_0 + \mu_1)$ if $0 < \mu_1 \leq \lambda_{C^{-1}A} \leq \mu_0$, $\mu_1$, $\mu_0$ being the smallest and largest eigenvalues of $C^{-1}A$, respectively. In that case the spectral radius of the matrix of iteration is

$$\frac{1 - \mu_1/\mu_0}{1 + \mu_1/\mu_0}.$$ 

Thus we see that the best choice of the matrix $C$, which must be easily invertible, is that which minimizes the spectral condition number $\mu_0/\mu_1$ of $C^{-1}A$. This will be seen to be valid for all convergence acceleration methods.

3. ONE-STEP CONVERGENCE ACCELERATION PROCESSES

Let the eigenvalues of $A$ be $\{\lambda_{\ell}\}$, where $0 < a \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_b \leq b$. Let $u^0$ be an initial approximation to the solution $u$ of $Au = f$ and consider again the general one-step iterative method, but initially without preconditioning, i.e. with $C = I$,

$$u^{\ell+1} = u^\ell - \tau_{C}(Au^\ell - f), \quad \ell = 0, 1, \ldots.$$ 

Here $\{\tau_{C}\}$ is a parameter set, the choice of which gives a possible convergence accelerator process. In Section 6, two-step methods are given.

As noticed in the previous section, when $\tau_{k+1} = \omega$ is a fixed parameter the best choice of $\omega$ is to minimize the spectral radius $\rho$ of $I - \omega A$, i.e. $\omega = 2/(\lambda_0 + \lambda_1)$. Then

$$\rho(1 - \tau A) = \frac{1 - \lambda_1/\lambda_2}{1 + \lambda_1/\lambda_2}.$$ 

The number of iterations $p$ needed to decrease the relative error in the 2-norm (the Euclidean norm) to a number less than $\varepsilon$, is thus such that

$$\left(\frac{1 - \lambda_1/\lambda_2}{1 + \lambda_1/\lambda_2}\right)^p \leq \varepsilon.$$ 

This inequality is satisfied if

$$p \geq \frac{1}{2} \frac{\lambda_2}{\lambda_1} \ln \frac{1}{\varepsilon}. $$
If we use a variable parameter set \( \{ \tau_k \} \), we would like to minimize the spectral radius of the corresponding iteration matrix \( Q_p \), achieved after \( p \) iterations. Then we have

\[
e^p = Q_p(A)e^0,
\]

where \( e^p = u^p - u \) and

\[
Q_p(\lambda) = \prod_{k=1}^{p} (1 - \tau_k \lambda).
\]

\( Q_p(A) \) is the corresponding matrix polynomial. We observe that \( Q_p(0) = 1 \).

More generally, we would like to minimize the residual \( r^p = A^p - f \), or the error \( e^p = A^{-1}r^p \), in the norm

\[
\|u\| = (u, u)_v = (A^{-v}u, u)_v^{1/2}, \quad v \text{ an integer. (3)}
\]

Then we have

\[
\|r^0\| \leq \|Q_p(A)\| \|r^0\|,
\]

where, as is easily demonstrated,

\[
\|Q_p(A)\| = \max_{\lambda_i} |Q_p(\lambda_i)|.
\]

An upper bound for this quantity is

\[
\max_{\lambda_i \in S_{\lambda_0}} |Q_p(\lambda)|,
\]

which we thus want to minimize.

From elementary books on approximation theory, we know that the least maximum in an interval \( [\lambda_1, \lambda_0] \) over all polynomials \( Q_p \) of degree \( p \), normalized such that \( Q_p(0) = 1 \), is achieved by the Chebyshev polynomial. Thus we have

\[
\max_{\lambda_i \in S_{\lambda_0}} |Q_p(\lambda)| \geq c \max_{\lambda_i \in S_{\lambda_0}} |T_p(u)| = c,
\]

where

\[
T_p(z) = \begin{cases} 
(z + \sqrt{1 - z^2})^p & \text{if } |z| < 1 \\
(z + \sqrt{z^2 - 1})^p & \text{if } |z| \geq 1
\end{cases}
\]

is the Chebyshev polynomial over the interval \([-1, 1]\), and where we use the linear transformation

\[
u = -\frac{2}{\lambda_0 - \lambda_1} \left( \lambda - \frac{\lambda_0 + \lambda_1}{2} \right).
\]

From the normalization criterion \( Q_p(0) = 1 \), we have

\[
c = 1/T_p \left( \frac{\lambda_0 + \lambda_1}{\lambda_0 - \lambda_1} \right).
\]
Thus, the best choice (in the above sense) of \( \tau_\ell \) is that which gives

\[
\prod_{\ell=1}^p (1 - \tau_\ell\lambda) = c \ T_p \left[ \frac{2}{\lambda_0 - \lambda_1} \left( \lambda - \frac{\lambda_0 + \lambda_1}{2} \right) \right] = \theta_p(\lambda),
\]

i.e. comparing zeros,

\[
\frac{1}{\nu_\ell} = \frac{\lambda_0 - \lambda_1}{2} \cos \theta_\ell + \frac{\lambda_0 + \lambda_1}{2},
\]

where

\[
\theta_\ell = \frac{2\ell - 1}{2p} \pi, \quad \ell = 1, 2, \ldots, p.
\]

In practice, the smallest and largest eigenvalues are not known, so we need lower and upper bounds, \( a \) (\( a > 0 \)) and \( b \), respectively. Then we have to use the parameters

\[
\frac{1}{\nu_\ell} = \frac{b - a}{2} \cos \theta_\ell + \frac{b + a}{2}.
\]

We observe that in this process, which defines the classical Chebyshev acceleration method or Richardson method, we have to choose the value of \( p \) in advance. In Section 6 we give a two-step process which does not need this \textit{a priori} estimate of \( p \).

It is an easy matter to find that if

\[
p \geq \frac{1}{2} \sqrt{\frac{b}{a}} \ln \frac{2}{\varepsilon}, \quad \varepsilon > 0
\]

then

\[
\max_{\lambda \in [a,b]} |\theta_p(\lambda)| \leq \varepsilon,
\]

i.e. the relative 2-norm error is less than \( \varepsilon \) after a cycle of \( p \) iterations, as given above.

We have thus reached an \textit{effective spectral condition number} whose order of magnitude is the square root of that of \( A \). The above Chebyshev acceleration process can, however, be numerically unstable, unless we use some particular permutations of the parameters [see Lebedev and Finogenov\(^{(2)}\)], or we use a two-step version of the method (see Section 6).

4. PRECONDITIONING

For the choice of the matrix \( C \) in Eq. (2.2) we can use a fractional step technique [Marchuk\(^{(2)}\)]. Here we will use a two-step fractional step method, particular cases of which are the ADI and SSOR methods. Thus let

\[
A = B + H + V
\]

be a splitting of \( A \). Then if the matrices \( B + H \) and \( B + V \) are easily invertible from a practical point of view, like the product of triangular matrices (cf. Section 5), the following method is usable for the conditioning of \( A \) and the iterative solution of \( Au = f \):

\[
\begin{align*}
(B + H)u^{\ell+1} & = -Vu^\ell + f \\
(B + V)u^{\ell+1} & = -Hu^{\ell+1} + f, \quad \ell = 0, 1, \ldots.
\end{align*}
\]
From this we have

\[ u^{(x)} = T u^x + \ldots f, \]

where

\[ T = (B + V)^{-1} H (B + H)^{-1} V. \]

This can be viewed as an unaccelerated generalized AOI method. An accelerated version is the following, where we work with corrections \( \epsilon^x \),

\[
\begin{align*}
(B + H)\epsilon^{x+\frac{1}{2}} &= Au^x - f \\
(B + V)\epsilon^{x+\frac{1}{2}} &= B \epsilon^{x+\frac{1}{2}} \\
\epsilon^{x+1} &= u^x - \tau_{x+1} \epsilon^{x+\frac{1}{2}}, \quad x = 0, 1, \ldots
\end{align*}
\]

i.e.

\[ u^{x+1} = (I - \tau_{x+1} C^{-1} A) u^x + \tau_{x+1} C^{-1} f, \]

where

\[ C = (B + H) B^{-1} (B + V). \]

Thus

\[ u^x = \left[ \prod_{k=1}^{x} (I - \tau_k C^{-1} A) \right] (u^0 - A^{-1} f) + A^{-1} f. \]

\( \tau_k \equiv 1 \) gives the unaccelerated method, i.e.

\[ u^{(x)} = T u^x + C^{-1} f. \]

For the successful applicability of the Chebyshev convergence accelerator process, the matrix \( C^{-1} A \) must have eigenvalues > 0 and a spectral condition number much smaller than that of \( A \).

Suppose now that the Hermitian transpose \( V^* = H \), i.e. \( C^* = C \), and that \( B \) is positive definite. Then \( C \) is positive definite [see formula (4) below] and \( C^{-1} A \) is similar to \( C^{-1} A C^{-1} \), which is Hermitian. Thus the eigenvalues of \( C^{-1} A \) are real. Furthermore, the eigenvalues of \( C^{-1} A \) are bounded below and above by lower and upper bounds, respectively, of the quotient \( (Ax, x)/(Cx, x) \) of the quadratic forms of \( A \) and \( C \). The upper bound, \( 1 \), is immediately found, since \( C = B + H + V + HB^{-1} V = A + V^* B^{-1} V \) and thus

\[ (Cx, x) \geq (Ax, x), \quad (4) \]

because

\[ (V^* B^{-1} V x, x) = (B^{-1} V x, V x) \geq 0. \]

Consider now the particular choice, the SSOR method, where \( A = D + L + U \), \( D \) is (block-) diagonal and is supposed to be positive definite, \( L = U^* \) is lower triangular and \( B = \left[ \left( 2/\omega - 1 \right) \right] D, 0 < \omega < 2 \). Then with

\[ H = \left( 1 - \frac{1}{\omega} \right) D + L, \quad V = \left( 1 - \frac{1}{\omega} \right) D + U \]
we have

\[ B + H = \frac{1}{\omega} D + L, \quad B + V = \frac{1}{\omega} D + U, \]

\[ C = (B + H)^{-1}(B + V) = \frac{1}{2 - \omega} \left( \frac{1}{\omega} D + L \right)^{-1} \left( \frac{1}{\omega} D + U \right). \]

Here the factor \(1/(2 - \omega)\) is a matter of convenience in the theoretical calculations only. We now examine how to choose the parameter \(\omega\) in some optimal way. Before proceeding, we remark that a more general choice is

\[ C = (B + L)B^{-1}(B + U), \]

where \(B\) is a (block-) diagonal matrix \([\text{for applications see, for example, Axelsson}^h]\).

After some simple algebraic calculations we get

\[ (2 - \omega)C = A + \frac{1}{4\omega} (2 - \omega)^2 D + \omega \left( LD^{-1} U - \frac{1}{4} D \right). \]  \hspace{1cm} (5)

Suppose now that there exist scalars \(\nu_0, K\) such that

\[ i) \quad \frac{\langle Ax, x \rangle}{\langle Dx, x \rangle} \geq \nu_0, \quad \forall x, \]

\[ ii) \quad (LD^{-1}Ux, x) - \frac{1}{4} \langle Dx, x \rangle \leq K \langle Ax, x \rangle, \quad \forall x, \]

or equivalently, since \(D\) is positive definite,

\[ (\bar{A}x, x) \geq \nu_0 (x, x), \quad \forall x \]

and

\[ (\bar{L}^* x, \bar{L}^* x) - \frac{1}{4} (x, x) \leq K (\bar{A}x, x), \quad \forall x, \]

where

\[ \bar{A} = D^{-\frac{1}{2}} AD^{-\frac{1}{2}} = I + \bar{L} + \bar{L}^*, \quad \bar{L} = D^{-\frac{1}{2}} LD^{-\frac{1}{2}}. \]

Then, from Eqs. (4) and (5),

\[ \frac{1}{\kappa(\omega)} = \frac{2 - \omega}{1 + \frac{1}{4\omega} (2 - \omega)^2 / \nu_0 + \omega K} \leq \frac{\langle Ax, x \rangle}{\langle Cx, x \rangle} \leq 1, \quad 0 < \omega < 2, \quad \forall x, \]

where \(\kappa(\omega)\) is thus an upper bound of the spectral condition number of \(C^{-1}A\). From an easy calculation we see that \(\kappa(\omega)\) is minimized if

\[ \omega = \bar{\omega} = \frac{2}{1 + \sqrt{2\nu_0 (1 + 2K)}} \]

and that

\[ \kappa(\bar{\omega}) = \sqrt{\frac{1 + 2K}{2\nu_0} + \frac{1}{2}}. \]
In practical problems the matrix \( A = A_h \) is associated with some parameter \( h \) (a step-length parameter in discretized partial differential equation problems). Then if

\[
\nu_0 = \lambda_0 h^\sigma, \quad \sigma > 0, \ h \to 0
\]

and

\[
K = \lambda_1 h^{-(\sigma-\varepsilon)},
\]

where \( \lambda_0 > 0, \lambda_1 > 0 \) are independent of \( h \), we have

\[
\kappa(\omega) = \sqrt{\frac{\lambda_1}{\lambda_0}} h^{-(\sigma-\varepsilon/2)} + O(h^{-\varepsilon/2}), \quad \text{if} \ \varepsilon < \sigma
\]

and

\[
\kappa(\omega) = \sqrt{\frac{1 + 2K}{2\lambda_0}} h^{-\sigma/2} + \frac{1}{2}, \quad \text{if} \ \varepsilon = \sigma.
\]

Thus, since the condition number of \( A \) in the important practical problems we consider, is \( O(1/\mu_0) \) [cf. Strang and Fix\(^2\)] and \( K \) is independent of \( h \) (\( \varepsilon = \varepsilon \)) we have achieved a spectral condition number whose order of magnitude is the square root of that of \( A \). In this case it is enough to choose

\[
\omega = \frac{2}{1 + \zeta h^{\sigma/2}}, \quad 0 < \zeta \ \text{independent of} \ h.
\]

Then

\[
\kappa(\omega) = \frac{1}{2} \left( \frac{1 + 2K}{\zeta} + \frac{\zeta}{2\lambda_0} \right) h^{-\sigma/2} + \frac{1}{2}.
\]

Thus it is not necessary to know a particularly accurate value of \( \tilde{\omega} \) (cf. Section 7). It ought to be mentioned (and is easily seen) that \( \tilde{\omega} = \omega_1 \) [see Young\(^2\), p. 464] if \( K = 0 \). If \( K \neq 0 \), however, then the parameter \( \omega_1 \) is not optimal with respect to minimizing the spectral condition number. This follows since in Young the spectral radius \( \rho \) of \( \Lambda \tilde{X} \) is used, which gives a more restrictive condition than (ii), unless \( (\Lambda \tilde{X}x,x)/(x,x) = \rho(\Lambda \tilde{X}) \) for some vector for which \( (\tilde{A}x,x)/(x,x) = \mu_0 \).

Finally, let us conjecture that the following alternative form of condition (ii) above might be useful, in particular in connection with some finite-element (generalized finite difference) approximations of elliptic problems (cf. Section 5). Thus let \( L = D_L + S \), where \( D_L \) is the Hermitian and \( S \) is the skew-Hermitian (i.e. \( S^* = -S \)) part of \( L \). Then

\[
U = L^* = D_L - S
\]

and

\[
A = D + L + U = 2D_L + D,
\]

i.e.

\[
D_L = \frac{1}{2} (A - D).
\]
Thus
\[
LD^{-1}U = \left[ \frac{1}{4}(A - D) + S \right] D^{-1} \left[ \frac{1}{4}(A - D) - S \right]
= \frac{1}{4} AD^{-1}A - \frac{1}{4} A + \frac{1}{4} D - SD^{-1}S - \frac{1}{4} SD^{-1}A + \frac{1}{4} AD^{-1}S .
\]

From the Cauchy-Schwarz inequality we have
\[
[(D^{-1}Ax, Sx)] = [(D^{-1}Ax, D^{-1}Sx)] \leq (D^{-1}Ax, Ax)^{\frac{1}{2}} (D^{-1}Sx, Sx)^{\frac{1}{2}} ,
\]
i.e.
\[
[LD^{-1}Ux, x] - \frac{1}{2} [dx, x] = \frac{1}{2} (D^{-1}Ax, Ax) + \frac{1}{2} (D^{-1}Ax, Sx) + \frac{1}{2} (Sx, D^{-1}Ax) + (D^{-1}Sx, Sx) - \frac{1}{2} (Ax, x)
\leq \left[ \frac{1}{2} (D^{-1}Ax, Ax) + (D^{-1}Sx, Sx) \right]^{\frac{3}{2}} - \frac{1}{2} (Ax, x) .
\]

Thus condition (ii) is satisfied if
\[
\frac{1}{2} (D^{-1}Ax, Ax) + (D^{-1}Sx, Sx) \leq k_0 (Ax, x) ,
\quad \forall x .
\]

That this condition is satisfied, at least for constant coefficient positive difference schemes (cf. Section 4) can be proved, using Lemma 3.2, p. 450 in Ref. 4.

5. **A DISCRETIZED SELF-ADJOINT ELLIPTIC PROBLEM**

Consider now
\[
E(u) = - \sum_{i=1}^{n} \frac{2}{3} x_i \left[ a_i(x) \frac{\partial^2}{\partial x_i^2} u \right] + q(x)u = f(x), \quad x \in \Omega ,
\]
\[
u = g, \quad x \in \partial \Omega ,
\]
where \( q(x) \leq 0 \). Suppose with no loss of generality that \( q \equiv 0 \) and suppose further that \( a_i(x) \geq k > 0 \) has a Lipschitz-continuous first derivative. Approximate (for notation see \( \text{Axelsson}^3 \))
\[
\frac{2}{3} a_i(x) \left[ \frac{\partial^2}{\partial x_i^2} u \right] \quad \text{by} \quad \frac{1}{h^2} \left[ a_i(x_h) \left( u_{x+e_i} - u_x \right) - a_i(x_h) \left( u_x - u_{x-e_i} \right) \right] .
\]

Then we get the following discretized operator
\[
E_h u = b_h u - \sum_{i=1}^{n} \left[ c_{a,e_i} u_{x+e_i} + c_{a-e_i,e_i} u_{x-e_i} \right] = f_h , \quad x_h \in \Omega_h , \quad b_h = \sum_{i=1}^{n} c_{a,e_i} + c_{a-e_i,e_i} .
\]

The crucial condition for the applicability of the SSOR preconditioning method is condition (ii) in Section 4. We will now sketch the proof that this condition is satisfied for the important problem given above. We will do this by giving an upper bound for the spectral radius of the matrix in the left-hand side of (ii), i.e.
\[
[LD^{-1}Ux, x] - \frac{1}{2} [dx, x] \leq \lambda_{LD^{-1}U-\frac{1}{2}D} = \rho_0 , \quad \forall x \{x, x\} = 1 .
\]
We will give an estimate of this spectral radius, using the classical Gerschgorin theorem [see, for example Varga]. We observe that all terms in $LD^{-1}L^*$ are non-negative, so it suffices to bound the row sums of the matrix $-LD + LD^{-1}L^*$. Thus

$$\rho \leq \max_{a} \left\{ -\frac{1}{4} b_a + \sum_{i=1}^{n} \left[ c_{a-e_i, e_i} b^{-1}_{a-e_i} \sum_{j=1}^{n} c_{a-e_j, e_j} \right] \right\}. $$

Let

$$b_a = \frac{\sum_{i=1}^{n} c_{a-e_i, e_i}}{\sum_{i=1}^{n} c_{a, e_i}}$$

and

$$\beta_a = \min_{i=1, \ldots, n} \frac{n}{b_a}.$$

Then

$$\rho \leq \max_{a} \left\{ \frac{1}{4} \left( \sum_{i=1}^{n} c_{a, e_i} \right) \left[ -1 + n_i \left( -1 + 4 \frac{1}{1 + \beta_a} \right) \right] \right\}.$$

Since $(A, x)(x, x) \geq O(h^2)$, $h \to 0$, it suffices to bound the term in braces from above by $Ch^2$ in order to show that $\rho \leq (A, x)(x, x)$. We rewrite this term as

$$\frac{1}{1 + n_i} \left[ -\left( 1 + n_i \right)^2 + 4n_i \right] + 4n_i \left[ \frac{1}{1 + \beta_a} - \frac{1}{1 + n_i} \right] = -\frac{(1 - n_i)^2}{1 + n_i} + \frac{4n_i}{(1 + \beta_a)(1 + n_i)} \left[ n_i - \beta_a \right].$$

For some $i = i_0$ we have

$$\beta_a = n_i - \beta_a.$$

Thus, since $a_i$ have a Lipschitz-continuous first derivative, we have

$$n_i = 1 - h \sum_{i=1}^{n} \frac{\partial}{\partial x_i} a_i(x + \epsilon e_i) + O(h^3)$$

and the desired bound of the row sums is achieved. The above analysis is mainly due to an old, but unpublished, result of Todd Dupont.

It is also possible to satisfy condition (ii) even if the coefficients $a_i(x)$ in the differential equation are only Lipschitz-continuous (in the $e_i$ direction), if a particular matrix $\tilde{D}$ (cf. remark in Section 4) is used [see Axelsson].

If we have discontinuous coefficients (in practice more often than not the case, since there are different material constants), then the following simple trick can be used.

Consider for simplicity the 5-point difference approximation of $Au = f$, when we have two orthogonal lines $\ell_1, \ell_2$, across which the coefficients are discontinuous (see the figure below). In a general case, it is advisable to use some simple finite element approximation with elements having one side along the lines of discontinuity.

We construct a uniform rectangular mesh in such a way that the corner $\ell_1 \cap \ell_2$ is not a mesh point, and so that the lines $\ell_1, \ell_2$ will fall half-way between two mesh lines (see the figure). For simplicity, we assume that the coefficients $a_1, \ell_2$ are constant in the two regions.
When constructing the 5-point approximation at the points numbered 1 and 2, we use that difference approximation to \( a(x) \partial u/\partial x \), which makes use of the point numbered 0 on the line of discontinuity \( \xi \), but with coefficients \( a_1 \) and \( a_2 \), respectively. To get an equation valid for this point (and to get a coupling between the two regions, divided by \( \xi \)), we use the law of continuous flow, i.e. the difference approximations of \( a(x) \partial u/\partial x \) on both sides of \( \xi \) are equal, where the \( x \) direction is orthogonal to \( \xi \). Then

\[
a_1 \frac{u_2 - u_0}{h/2} = a_2 \frac{u_2 - u_0}{h/2},
\]

i.e.

\[
u_2 = \frac{a_1 u_1 + a_2 u_2}{a_1 + a_2} \quad \text{and} \quad a_1 \frac{u_2 - u_1}{h/2} = 2 \frac{a_2 - a_1}{a_1 + a_2} \frac{u_2 - u_1}{h}.
\]

The part of the 5-point difference approximation matrix \( A \), corresponding to the coupling between \( u_1 \) and \( u_2 \), is then

\[
\begin{bmatrix}
-(3 + 2\delta) a_1 & 2a_1 \\
2(1 - \delta)a_2 & -(5 - 2\delta)a_2
\end{bmatrix}
\begin{bmatrix}
u_1 \\
u_2
\end{bmatrix},
\]

where \( \delta = a_2 / (a_1 + a_2) \), i.e. \( 0 < \delta < 1 \). We observe that \( 2(1 - \delta) a_2 = 2a_1 \), i.e. the matrix is symmetric.

We now sketch briefly a proof that condition (ii) in Section 3 is satisfied with \( K = \Lambda_1 h^{-1} \). Let \( \Omega_1 = \Omega_1(h) \) consist of those points in the mesh which are neighbours of \( \xi \), and let \( \Omega_2 = \Omega_2(h) \) consist of the remaining points. It is obvious that \( A \) restricted to \( \Omega_2 \) (i.e. with all components of the vectors \( = 0 \) on \( \Omega_1 \) satisfies condition (ii) with \( K = 0 \).

Because we have with \( (x, x) = 1 \), \( \bar{L} = D^{-1} L D^{-1} L \),

\[
(\bar{L} \bar{L}^* x, x) \leq c(\bar{L} \bar{L}^*) \leq \|L\|_\infty \|\bar{L}^*\|_\infty = \frac{1}{4},
\]

since the elements in \( \bar{L} \) are \( \frac{1}{4} \) (at most two elements) and 0, where \( \|F\|_\infty \) is the maximal row-sum of elements in \( |F| = [f_{i,j}] \) (with \( F = [f_{i,j}] \)), and where we have used the well-known matrix property that all eigenvalues of a matrix are less than or equal to any natural norm. \( \|\cdot\|_\infty \) is the norm, corresponding to the norm \( \|x\|_\infty = \max |x_i| \) of the vector \( x \). From this we conclude that

\[
(\bar{L} D^{-1} L^* x, x) = \frac{1}{4}(Dx, x) \leq 0, \quad \text{if} \quad x_i = 0, \quad i \in \Omega_1.
\]
On the other hand, the contribution from $\Omega_1$ to this scalar product is $\leq K_1 (Dx, x)_{\Omega_1}$, where $K_1$ is a constant, independent of $h$, and $(Dx, x)_{\Omega_1}$ mean the scalar product over points in $\Omega_1$ only. But

$$
(Dx, x)_{\Omega_1}/(Ax, x)_{\Omega_1} \leq (Dx, x)_{\Omega_1}/(Ax, x)_{\Omega_1} \leq K_2 h^{-1},
$$

since the number of points in $\Omega_1$ are $O(h^{-1})$.

Thus condition (ii) is satisfied with $K = K_1 K_2 h^{-1}$ and the spectral condition number of the preconditioned matrix is thus $O(h^{-3/2})$.

6. ON CONVERGENCE ACCELERATION

As mentioned above, it is possible to give a two-step acceleration process

$$
u^{k+1} = \alpha_k u^k + (1 - \alpha_k) u^{k-1} - \beta_k r^k, \quad \kappa = 0, 1, \ldots,
$$

(6)

where $r^k = Au^k - f$ is the residual, which apart from rounding errors gives the same result as the one-step Chebyshev process. As in Section 3, we first introduce the method without preconditioning. The two-step process is however always stable and each approximation is chosen from a recursion formula which minimizes the spectral radius of the corresponding error iteration matrix $Q_k(A)$. As pointed out in Section 3, this was not so in the one-step method, where we had to choose the number $p$ of iterations in advance, and only the $p$th iteration gave an iteration matrix $Q_p(A)$ (see Section 3) with minimal spectral radius.

We will show that the corresponding sequence of numbers $\alpha_k, \beta_k$ is

$$
\alpha_k = \frac{a + b}{2}, \quad \beta_k = \frac{a + b}{2} - \left( \frac{b - a}{4} \right) \beta_{k-1}, \quad \kappa = 1, 2, \ldots
$$

(7)

$$
a_0 = 1, \quad \beta_0 = 4/(a + b).
$$

From Section 3, we have

$$
e^k = Q_k(A)e^0,
$$

where the error $e^k = u^k - A^{-1}f$. From the recursion formula (6) we thus have, observing that the formula is valid for all vectors $e^0$,

$$
Q_{\kappa+1}(A) - \alpha_{\kappa} Q_{\kappa}(A) + \beta_{\kappa} A Q_{\kappa}(A) + [\alpha_{\kappa} - 1] Q_{\kappa-1}(A) = 0, \quad \kappa = 1, 2, \ldots.
$$

Comparing this with the recursion formula

$$
T_{\kappa+1}(x) - 2 x T_{\kappa}(x) + T_{\kappa-1}(x) = 0, \quad \kappa = 1, 2, \ldots
$$

$$
T_1(x) = 1, \quad T_2(x) = x
$$

valid for Chebyshev polynomials, we see that if

$$
\alpha_k = 1 + T_{\kappa-1}(b')/T_{\kappa-1}(b')
$$

$$
\beta_k = \frac{4}{b - a} T_{\kappa}(b')/T_{\kappa-1}(b'), \quad b' = (b + a)/(b - a),
$$

valid for Chebyshev polynomials, we see that if

$$
\alpha_k = 1 + T_{\kappa-1}(b')/T_{\kappa-1}(b')
$$

$$
\beta_k = \frac{4}{b - a} T_{\kappa}(b')/T_{\kappa-1}(b'), \quad b' = (b + a)/(b - a),
$$

(7)
then

\[ Q_x(A) = \frac{T_x(x)}{T_x(b')} \], \quad x = \frac{1}{b-a} \left[ (b + a) I - 2A \right],

which is the result we wanted (cf. Section 3). Trivial calculations now give the recursion formula in Eqs. (7).

Finally we remark that it is possible to achieve the same asymptotic speed of convergence with certain fixed parameters \( \alpha_x, \beta_x \) (actually the limit of the sequences in Eqs. (7) [see Young\(^{21}\)]).

We observe, however, that in the Chebyshev process there are many reasons why this method is not optimal in minimizing the number of necessary iterations to reach a certain relative error. One reason is that we minimize the spectral radius over an interval \([a,b]\) covering the spectrum of the operator, instead of minimizing the spectral radius of the iteration operator over the discrete spectrum of the given operator. In practice, it is difficult to find good enough lower and upper bounds \( a, b \) of the eigenvalues. From what was said before, it follows that the resulting number of iterations is sensitive to the choice of these bounds, in particular to the choice of the lower bounds.

Another reason is that, since we work in a subspace spanned by vectors \( r^0, Ar^0, \ldots, A^Pr^0 \), we ought to minimize over the eigenvalues of the operator \( A \), restricted to this subspace. A discussion of these and related questions can be found in Kaniel\(^{21}\) and at the end of this section. Here we would like to point out that we can avoid the problems of giving upper and lower bounds of the spectrum of \( A \), or of the restriction of \( A \) to the subspace mentioned, by using a modification of a conjugate gradient method.

This latter method gives in fact the smallest residual in the norm \( \| r^0 \| \), defined in Eq. (5) over all possible acceleration processes of the form (6). From Eq. (6) we are able to construct a sequence of vectors

\[ u^{k+1} = \sum_{j=0}^{k} \alpha_j^{(k)} u_j - \beta_k r^k, \]

where \( r^k = Au^k - f \). According to consistency (i.e. \( u^1 = A^{-1} f \) is a solution), \( \sum \alpha_j^{(k)} = 1 \).

In terms of residuals, we have

\[ r^{k+1} = \sum_{j=0}^{k} \alpha_j^{(k)} r_j - \beta_k Ar^k \]

or

\[ r^{k+1} = r^0 - \sum_{j=0}^{k} \delta_j^{(k)} A^j r^0. \] (8)

This is the general form of all acceleration processes of the kind given in Eqs. (6), using vectors from the subspace spanned by \( (r^0, Ar^0, \ldots, A^k r^0) \). Thus we see that the best linear least square approximation of \( r \), minimizes \( (r^{k+1}, r^{k+1}) \) over all sets \( (\delta_j^{(k)}) \). The corresponding normal equations contain the Gramian matrix \( (A^j r^0, A^k r^0) \), the solution of which thus gives us an impractical method. However, it is possible to achieve this minimum using a modification of a conjugate gradient method or a method of orthogonal directions.
To get such an algorithm, we first consider the minimizing problem using orthogonal polynomials [cf. Stiefel177]. Let $A \psi_i = \lambda_i \psi_i$, $(\psi_i, \psi_j) = \delta_{ij}$, $i, j = 1, \ldots, n$, and let $r^0 = E\psi_i \psi_i$, where we used the generalized inner product

$$(u, v)_\nu = (A^{-1}u, v).$$

The corresponding norm is denoted by $\|u\| = (u, u)^{1/2}_\nu$ [see Eq. (3)]. We have

$$A^i r^0 = \sum_{i=1}^n a_i A^i \psi_i$$

and thus

$$r^p = r^0 - \sum_{i=1}^n \beta^i_0 A^i \psi_i = \sum_{i=1}^n a_i [1 - q_p(\lambda_i)] \psi_i,$$

where $q_p$ is a polynomial of degree $\leq p$, such that $q_p(0) = 0$. We want to minimize

$$(r^p, r^p)_\nu = \sum_{i=1}^n [1 - q_p(\lambda_i)]^2 a_i^2/\lambda_i^\nu,$$

over all such polynomials $q_p(\lambda)$. Thus, let $(\pi_k(\lambda))_{k=1}^p$, be a set of mutually orthogonal polynomials with respect to the inner product

$$(\phi, \psi)_\nu = \sum_{i=1}^n \phi(\lambda_i) \psi(\lambda_i) a_i^2/\lambda_i^\nu,$$

such that $\pi_k(0) = 0$, $\pi_k$ of degree $k$.

Then we have

$$q_p(\lambda) = \sum_{k=1}^p \gamma_k \pi_k(\lambda),$$

and

$$(r^p, r^p)_\nu = (1 - q_p, 1 - q_p)_\nu,$$

and, as is well known from least square approximation theory, this is minimized when the error in the approximation of 1 with such polynomials $q_p$ is orthogonal to the space spanned by $(\pi_k(\lambda))_{k=1}^p$.

Let $s_j(\lambda) = \lambda [1 - q_j(\lambda)]$. Then, another way of characterizing the residuals is

$$(r^p, A r^p)_\nu = (1 - q_p, s_j)_\nu = 0, \quad j = 0, 1, \ldots, p - 1,$$

since $s_j(\lambda)$ is in the space of polynomials $(\pi_k(\lambda))_{k=1}^p$, the degree of $s_j$ being $\leq p$ and $s_j(0) = 0$. It is well known that we have uniqueness. In fact, since $(1 - q_p, q_p)_\nu = 0$, we have

$$(q_p, 1)_\nu = (q_p, q_p)_\nu,$$

or

$$(\pi_k, 1)_\nu = \gamma_k (\pi_k, \pi_k)_\nu, \quad k = 1, \ldots, p,$$

and the best least square approximation is given by

$$q_p(\lambda) = \sum_{k=1}^p \frac{(\pi_k, 1)_\nu}{(\pi_k, \pi_k)_\nu} \pi_k(\lambda).$$
We will, however, not use this explicit expression.

Now, let \( r^j = A u^j - f \) be the residuals. According to the above, we are going to determine the approximations in such a way that

\[
(r^i, Ar^j)_{u} = (r^i, r^j)_{u-1} = 0, \quad i \leq j .
\]

Thus let

\[
u^i = u^i - \beta_q r^i ,
\]

i.e.

\[
r^i = r^i - \beta_q Ar^i .
\]

Then

\[
(r^i, Ar^i)_{u} = 0 \Rightarrow \beta_q = \frac{(r^0, Ar^0)_{u}}{(Ar^i, Ar^i)_{u}} .
\]

Further

\[
u^{+1} = \alpha_\xi u^i + (1 - \alpha_\xi) u^{+1} - \beta_\xi r^i ,
\]

i.e.

\[
r^{+1} = \alpha_\xi r^i + (1 - \alpha_\xi) r^{+1} - \beta_\xi Ar^i .
\]

Then

\[
(r^{+1}, Ar^{+1})_{u} = 0 \Rightarrow \alpha_\xi = \beta_\xi \mu_\xi
\]

and

\[
(r^{+1}, Ar^{+1})_{u} = 0 \Rightarrow 1 - \alpha_\xi = \beta_\xi (Ar^{+1}, Ar^{+1})_{u}/\delta_{+1} = -\frac{\delta_\xi}{\delta_{+1}} \delta_{+1}/\delta_{+1} ,
\]

where

\[
\delta_{+1} = (Ar^i, r^j)_{u} = (r^i, r^j)_{u-1}
\]

and

\[
\mu_\xi = (Ar^{+1}, Ar^{+1})_{u}/(Ar^i, r^i)_{u} .
\]

It is then easily seen that

\[
(r^{+1}, Ar^{+1})_{u} = 0, \quad j = 0, 1, \ldots, \ell .
\]

Thus we have the following simple recursion formula for the construction of the best least square approximation in the norm \((\cdots)_{u}^{1/2}\).

\[
\begin{align*}
r^0 &= Au^0 - f, \quad \beta_q = \mu_0^{-1} , \\
u^1 &= u^0 - \beta_q r^0 , \quad r^1 = r^0 - \beta_q Ar^0 \\
\beta_\xi^{-1} &= \mu_\xi - \beta_\xi^{-1} \delta_{+1}/\delta_{+1} , \quad \alpha_\xi = \beta_\xi \mu_\xi \\
u^{+1} &= (\alpha_\xi - 1) (u^\xi - u^{+1}) - \beta_\xi r^i \\
r^{+1} &= (\alpha_\xi - 1) (r^\xi - r^{+1}) - \beta_\xi Ar^i , \quad \ell = 1, 2, \ldots .
\end{align*}
\]
(In order to reduce the build-up of rounding errors, it is advisable to work with differences, as shown in this algorithm.)

The algorithm given above requires the storage of four vectors \( u^k \), \( \Delta u^k = u^k - u^{k-1} \), \( r^k \), \( \Delta r^k = r^k - r^{k-1} \). As soon as a new component in \( Ar^k \) is calculated, the corresponding inner products, where \( Ar^k \) is one of the factors, are updated. Further, the corresponding component of \( \Delta r^k = r^{k+1} - r^k \) is calculated and stored in the corresponding position of \( \Delta r^{k-1} \).

If, however, a subroutine is used for the product of \( A \) by \( r^k \), we have also to store the vector \( Ar^k \). But the use of subroutines will increase the readability of the program.

The number of such products is only one per iteration in Eq. (10). However, it may be advisable to use the definition of the residual \( r^{k+1} = Au^{k+1} - f \), instead of using the recursion formula for \( r^{k+1} \) in Eq. (10), say at every 5th iteration step. This updating of a true residual will diminish the effect of rounding errors. Runs on simple test problems have however shown that the original (preconditioned) algorithm works quite satisfactorily without this updating.

The numbers \( \delta_{\nu} \), \( u^k \) are computable for \( \nu = 1, 0, -1, \ldots \). (For \( \nu = 2 \) the calculation of \( \delta_{\nu} \) demands \( A^{-1} f \), i.e. the solution of the system of equations we are trying to solve.) Only \( \nu = 0 \) and \( \nu = 1 \) give practical methods, however. For \( \nu = 0 \) we minimize the residuals in the Euclidian norm, and for \( \nu = 1 \), the residuals in the weighted norm
\[
([A^{-1} r^p, r^p])^{1/2} = ([e^p, r^p])^{1/2},
\]
where
\[
e^p = u^k - A^{-1} f
\]
is the error in \( u^k \).

Finally, let us consider the case where we have a preconditioning positive definite matrix \( C \). The inner product
\[
(u, v)' = ([CA]^{-1} u, v)
\]
will not increase the number of products of a matrix by a vector, as follows from the argument below. The algorithm (10) has now the form
\[
u^{k+1} - u^k = (a - 1) [u^k - u^{k-1}] - \beta C^{-1} r^k,
\]
\[
r^{k+1} - r^k = (a - 1) [r^k - r^{k-1}] - \beta C^{-1} r^k.
\]
Then
\[
([r^{k+1}, Ar^p]' = (C^{-1} r^{k+1}, r^p) = 0, \quad j \leq k,
\]
implies that [cf. Eq. (9)]
\[
u_x = (AC^{-1} r^x, C^{-1} r^y)/\delta_x,
\]
and
\[
\delta_j = (C^{-1} r^x, r^y).
\]
According to the previous analysis, this algorithm [which is used in Axelsson\(^4\)] will minimize
\[
([CA]^{-1} r^p, r^p) = ([C^{-1} r^p, e^p]) = ([A^{-1} C]^{-1} e^p, e^p).
Suppose now that we have a positive definite preconditioning matrix of the form $C = EE^*$ (cf. Section 4). Then we can also write our system of equations as

$$E^{-1}A(E^*)^{-1} E^* u = E^{-1} f,$$

or

$$\tilde{A} \tilde{x} = \tilde{f},$$

where

$$\tilde{A} = E^{-1}A(E^*)^{-1}, \quad \tilde{x} = E^* u, \quad \tilde{f} = E^{-1} f.$$

Then, since $\tilde{A}$ is Hermitian positive definite, the original procedure (10) can also be carried out with the inner product defined by Eq. (9) ($\nu = 0$ or 1) and with $\tilde{A}$ instead of $A$. However, the first given variant of preconditioning method seems simpler to use [cf. Evans'35].

At this point we ought to mention that we should of course never actually calculate the inverse of $C$ or $E$ explicitly, in order not to disturb the sparseness of the matrices. For instance, in order to calculate $E^{-1} f$, we merely solve the system of equations $E y = f$. This is easily solvable, since $E$ is a triangular sparse matrix (cf. Section 4), or some other suitable matrix.

The above variant of the conjugate gradient method appears to have the following advantage over the Chebyshev acceleration method:

i) it is the fastest of all methods for the minimization of $(r^p, r^p)_\nu$, and it takes into account the actual eigenvalue distribution of the initial errors,

ii) for $\nu = 0$ and 1, it demands approximately the same number of operations as the Chebyshev method for each iteration,

and, most important,

iii) it does not demand any estimation of lower and upper bounds of the spectrum of $A$.

The gain in using the conjugate gradient method is particularly noticeable if the operator $A$ has an eigenvalue distribution $0 < a < \lambda_i < \tilde{b}$, $i = 2, 3, \ldots$ and $b = \lambda_1 >> \tilde{b}$. Then, according to the analysis at the beginning of this section, if $r^0$ is constructed by the conjugate gradient method,

$$\|r^p\| \leq \max_{\lambda_1} \left[ 1 - q_\lambda(\lambda_1) \right] \|r^0\|$$

for any polynomial $q_\lambda$ of degree $\leq p$, such that $q_\lambda(0) = 0$. Let us choose $q_\lambda$ such that

$$1 - q_\lambda(\lambda) = \left( 1 - \frac{\lambda}{\lambda_1} \right) Q_{p-1}(\lambda),$$

where $Q_{p-1}(\lambda)$ is the best Chebyshev polynomial of degree $p - 1$ over the interval $[a, \tilde{b}]$ (cf. Section 3). Then

$$\max_{\lambda_1} \left[ 1 - q_\lambda(\lambda_1) \right] < 1/T_{p-1} \left( \frac{\tilde{b} + a}{\tilde{b} - a} \right),$$

since

$$\max_{\lambda_1} \left( 1 - \frac{\lambda}{\lambda_1} \right) < 1.$$
Thus, in this case, the conjugate gradient method converges with an effective spectral condition number $\sqrt{b/a}$, instead of $\sqrt{b'/a}$, as given by a Chebyshev acceleration method. The above is true for any number $< \sqrt{b'/a}$ of eigenvalues $> \bar{b}$.

From what has been said, it follows that the main use of the Chebyshev acceleration method is only theoretical, i.e. to get upper bounds for the necessary number of iterations needed by the conjugate gradient method.

7. A NUMERICAL EXPERIMENT

To test the preconditioning with one parameter $\omega = 2/(1 + \varepsilon h)$, $\varepsilon > 0$, a model problem on the unit square was run for the Laplacian equation (see the table below). The iterations were interrupted when the relative error was $\leq \varepsilon = 10^{-6}$, i.e. when $\|r^p\|_2 \leq \varepsilon \|r\|_2$. The following number $p$ of iterations were necessary in the preconditioned Chebyshev acceleration method [2-step - see Eqs. (6) and (7)] and in the preconditioned conjugate gradient method [Eqs. (12), (13) and (14)] and the recursion formula for $\alpha_k^*, \beta_k^*$ in Eq. (10), with the inner product defined in Eq. (11). In the table "Ch" and "gr" stand for values obtained in the first and second mentioned processes respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>Quantity measured</th>
<th>$\zeta$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.25</td>
</tr>
<tr>
<td>Ch</td>
<td>$p$</td>
<td>46</td>
</tr>
<tr>
<td>gr</td>
<td></td>
<td>21</td>
</tr>
<tr>
<td>Ch</td>
<td>$10^6|r^p|_2$</td>
<td>0.408</td>
</tr>
<tr>
<td>gr</td>
<td></td>
<td>0.274</td>
</tr>
<tr>
<td>Ch</td>
<td>$10^6|e^p|_2$</td>
<td>0.501</td>
</tr>
<tr>
<td>gr</td>
<td></td>
<td>0.158</td>
</tr>
<tr>
<td>Ch</td>
<td>$10^6|r^p|_{(CA)^{-1}}$</td>
<td>0.097</td>
</tr>
<tr>
<td>gr</td>
<td></td>
<td>0.042</td>
</tr>
</tbody>
</table>

Here $\|r^p\|_{(CA)^{-1}} = ([CA]^{-1}r^p, r^p)^{1/2}$. We notice the fairly flat curve of $p$ for the preconditioned conjugate gradient method. As was pointed out in Axelsson, this is even more apparent, when a scaling diagonal matrix $D$ is used instead of $D$ (cf. remark in Section 4). In that case also the difference between the preconditioned Chebyshev method and the conjugate gradient was even more noticeable than in the table given above.

Furthermore, we notice that the errors $\|e^p\|_2$ (in the example, the exact solution $u$ to $Au = f$ was known to be the exact solution of the Laplace equation), are even more in favour of the conjugate gradient method.

Acknowledgement

The readability of parts of this paper has increased thanks to the careful reading by Dr. Beno Schorr of CERN. I greatly appreciate his suggestions.
REFERENCES


