MATHEMATICS OF COMPUTATION, VOLUME 28, NUMBER 127, JULY 1974, PAGES 695-710

SOR-Methods for the Eigenvalue Problem with Large Sparse Matrices

By Axel Ruhe

Abstract. The eigenvalue problem $Ax = \lambda Bx$, where A and B are large and sparse symmetric matrices, is considered. An iterative algorithm for computing the smallest eigenvalue and its corresponding eigenvector, based on the successive overrelaxation splitting of the matrices, is developed, and its global convergence is proved. An expression for the optimal overrelaxation factor is found in the case where A and B are two-cyclic (property A). Further, it is shown that this SOR algorithm is the first order approximation to the coordinate relaxation algorithm, which implies that the same overrelaxation can be applied to this latter algorithm. Several numerical tests are reported. It is found that the SOR method is more effective than coordinate relaxation. If the separation of the eigenvalues is not too bad, the SOR algorithm has a fast rate of convergence, while, for problems with more severe clustering, the c-g or Lanczos algorithms should be preferred.

1. Introduction. In the present contribution, we set out to find solutions of the eigenvalue problem

$$(A - \lambda B)x = 0, \tag{1.1}$$

where A and B are $n \times n$ real symmetric matrices and B is positive definite. We will study direct iterative algorithms which are applicable in cases when A and B are large and sparse, so that neither inversions nor similarity transformations are convenient to apply. Usually, we are interested in one or a few of the smallest eigenvalues of (1.1) and their corresponding eigenvectors.

The eigenvalues are the stationary values of the *Rayleigh quotient*

$$\mu(x) := x^H Ax / x^H Bx, \tag{1.2}$$

and several of the algorithms proposed for the solution of (1.1) work by applying some optimization method to (1.2). Hestenes and Karush [12] studied the application of gradient methods (steepest descent); more sophisticated algorithms, such as conjugate gradients, were studied extensively subsequently [2], [8], [9], [18]. Applying the Ritz method to (1.2) yields the Lanczos algorithm [15], which has also been studied by several authors [17], [10], [18]. The conceptually simplest algorithm of all is *coordinate relaxation* where (1.2) is minimized by varying one coordinate of x at a time. Its origin is unknown but it has been described and studied by Faddeev and Faddeeva [6], Kahan [14], Schwarz [19], and Shavitt et al. [24].

Received June 21, 1973.

AMS (MOS) subject classifications (1970). Primary 65F15; Secondary 65H10, 65N20.

Key words and phrases. Eigenvalues, sparse matrices, overrelaxation.

The eigenvalue problem (1.1) is closely related to the linear systems problem, and most of the algorithms proposed for (1.1) have counterparts in the linear systems case. In that case, the theory of iterative methods has been developed much further, the results can be found in the recent monograph by Young [21]. It is interesting to note that the most successful algorithms for the solution of linear systems, such as the SOR method, also have counterparts for the eigenvalue problem, but that they have not been studied very much earlier in this context, with a few exceptions [3]. In the present contribution, we will develop the algorithms of SOR type for the eigenvalue problem up to the same level as they have been developed for the linear systems problem. We will show that the theory is closely related to that of the linear systems case, and the applicability and rates of convergence of the algorithms are also comparable. We will also show that the coordinate relaxation algorithm is closely related to the SOR method, to use the terminology of Ortega and Rheinboldt [16], who consider application of the SOR concept to nonlinear problems; the two methods are applications of nonlinear SOR and 1-step Newton SOR methods, respectively, to the eigenvalue problem. As far as the present author has seen, these generalized linear iterations have not been studied in detail, except for the case of a nonlinear operator being the sum of a linear and a diagonal one. Here, we extend the theory of such iterations to another important special case. It is interesting to note that methods of this kind are very often used *ad hoc* by nonspecialists, therefore it is valuable to analyze them theoretically in some frequently occurring cases.

We will start by formulating the algorithms and introducing some notations in Section 2. In Section 3, we prove under which conditions the algorithms are globally convergent. Though the conditions for convergence are more complicated than in the linear systems case, it is a relatively easy matter to make sure that they are fulfilled when applying the algorithms in practical cases. We continue by studying convergence rates and the choice of relaxation parameter ω in Section 4. It is possible to develop the theory fully when $A - \lambda B$ is 2-cyclic (property A), and we see that the rate of convergence is dependent on the separation of the smallest eigenvalue from the rest of the spectrum. An optimal choice of ω gives at least as great an improvement over simple iteration as does optimal Chebyshev semi-iteration, exactly as it does in the linear systems case [21]. Finally, in Section 5, we report several numerical tests and give a few examples with comparisons to other relevant algorithms. We compare the SOR methods to the Lanczos and c-g methods which were studied in an earlier report [18] and find that they give comparable results.

Our general analysis can easily be extended to cover application of other iterative algorithms for linear systems to the eigenvalue problem (1.1), provided that they are constructed by means of a convergent splitting [21]. Such methods, applicable to finite difference analogues to partial differential equations, are ADI methods (see [21]), symmetric factorizations [22], and direct Poisson solvers (see [23] and [25] for a similar application), as well as different semi-iterative methods [1], [21]. It is our purpose to report tests of some of these methods in a later report.

We also postpone the study of algorithms consisting of inverse iteration based on an algorithm for direct solution of sparse linear systems, since such methods, at least at the present state of the art, have to be specially tailored to the problem at hand [11].

2. Formulation of the Algorithms. If we count multiplicities, (1.1) has n real eigenvalues $\lambda_1, \cdots, \lambda_n$; we order them so that

$$\lambda_1 \leqq \lambda_2 \leqq \cdots \leqq \lambda_n.$$

Since B is positive definite, it defines a scalar product

$$(x, y)_B := x^H By$$

and, correspondingly, a norm

$$\| x \|_B := (x, x)_B^{1/2}.$$

We denote the B-orthonormalized eigenvectors by $u_1, \cdots, u_n$.

An *iterative eigenvalue algorithm* produces a sequence of vectors $x_0, x_1, \cdots, x_s$, where

$$x_{s+1} := x_s - p_s, \qquad s = 0, 1, \cdots, \tag{2.1}$$

and, in case of convergence, the corresponding normalized vectors $\hat{x}_s := x_s / \| x_s \|_B$ converge towards some u_i; in most cases, we are interested in a sequence which converges towards u_1. The Rayleigh quotients

$$\mu_s := \mu(x_s) = x_s^H A x_s / x_s^H B x_s = \hat{x}_s^H A \hat{x}_s \tag{2.2}$$

at the same time converge to λ_1 and the normalized residual vectors

$$\hat{r}_s := r(\hat{x}_s) = (A - \mu_s B)\hat{x}_s \tag{2.3}$$

converge to zero.

We let the *rate of convergence* be measured by $\hat{r}_s$ by means of

$$R = \sup \left[\limsup_{s \to \infty} \| \hat{r}_s \|^{1/s} \right] \tag{2.4}$$

where the supremum is taken over all sequences converging to u_1. We see that $0 \leqq R < 1$ and that $-\ln(R)$ gives the rate sought.

Now we are ready to formulate our two applications of the SOR concept to the eigenvalue problem. A relaxation method is an iterative method where each iteration, or major step, is divided into n minor steps, in each of which one of the unknowns is modified in order to satisfy the corresponding equation. In case of overrelaxation, the correction is multiplied by an overrelaxation factor ω.

This can be described in matrix form, if we decompose the matrices occurring in a diagonal part D, a strictly lower triangular part E, and a strictly upper triangular part F

$$X = D(X) - E(X) - F(X). \tag{2.5}$$

The nonlinear equation $(A - \mu(x)B)x = 0$ will in each step be replaced by the linear equation $C_s x_{s+1} = 0$ with

$$C_s = A - T_s B, \tag{2.6}$$

where $T_s = \operatorname{diag}(t_{sp})$ consists of approximations to $\mu(x)$, t_{sp} being the approximation used in the sth major and pth minor step.

If we further denote the multiplied corrections by

$$W_s = \operatorname{diag}(\omega\alpha_{sp}) \tag{2.7}$$

and current parts by $D_s = D(C_s)$ etc., we get the following matrix equation:

$$D_s(x_{s+1} - x_s) = - W_s(D_s x_s - E_s x_{s+1} - F_s x_s).$$

Solving this for x_{s+1}, we obtain:

$$x_{s+1} = M(W_s, T_s)x_s = x_s - V_s^{-1}C_s x_s, \tag{2.8}$$

$$V_s = (W_s^{-1}D_s - E_s). \tag{2.9}$$

Different choices of the parameters T_s and W_s now give different algorithms. The optimal choice of parameters corresponds to the algorithm of *coordinate relaxation* [6], [14], [24] while there are several simpler choices that will be our main interest in this contribution.

Algorithm 1 (*Coordinate Relaxation*). *Modify t in each minor step*:

$$t_{sp} = \mu(x_{sp})$$

and choose α_{sp} optimally so that

$$\mu(x_{sp} + \omega^{-1}(x_{sp+1} - x_{sp})) = \min_{\xi} \mu(x_{sp} - \xi e_p). \tag{2.10}$$

In order to find the optimal α, we see that

$$\mu(x - \xi e_p) = (x^H A x - 2\xi e_p^H A x + \xi^2 e_p^H A e_p)/(\|x\|_B^2 - 2\xi(e_p, x)_B + \xi^2\|e_p\|_B^2)$$

which is minimized for ξ satisfying the equation [14]

$$a\xi^2 - \xi + \eta = 0, \tag{2.11}$$

$$\eta = e_p^H(A - \mu(x)B)x/(a_{pp} - \mu(x)b_{pp}), \tag{2.12}$$

$$a = [(e_p, x)_B - \|e_p\|_B^2\eta]/\|x\|_B^2. \tag{2.13}$$

The first-order approximation $\xi = \eta$ corresponds to $\alpha_{sp} = 1$, as can be seen from (2.7), and the optimal choice is $\alpha = 2/(1 + (1 - 4a\eta)^{1/2})$. We have to prescribe that $\mu(x) < \mu(e_p)$, otherwise we have to choose a different sign in solving (2.11) to get a minimum. For small η, we note that

$$\alpha = 1 + a\eta + O(\eta^2). \tag{2.14}$$

We use this observation to formulate our second algorithm.

Algorithm 2 (*SOR Method*). *Modify t in each major step*

$$t_{sp} = t_s = \mu(x_s), \qquad T_s = t_s I,$$

and choose $\alpha_{sp} = 1$ *so that* $W_s = \omega I$.

Now the matrix formulation (2.8) simplifies to

$$x_{s+1} = M(\omega, t_s)x_s = x_s - p_s = x_s - V_s^{-1}r_s, \tag{2.15}$$

where

$$V_s = (\omega^{-1}D_s - E_s), \tag{2.16}$$

$$\begin{aligned} M(\omega, t_s) &:= M(\omega I, t_s I) \\ &= (I - V_s^{-1}C_s) = (\omega^{-1}D_s - E_s)^{-1}((\omega^{-1} - 1)D_s + E_s^T). \end{aligned} \tag{2.17}$$

We note that it is possible to develop eigenvalue algorithms on the basis of many other splittings

$$A - t_s B = V_s - H_s, \tag{2.18}$$

where V_s is readily invertible and H_s has small norm—we just use the formulation (2.15). A closely related algorithm in which only A is split

$$V_s = \omega^{-1}D(A) - E(A) \tag{2.19}$$

has been described by Buffoni [3]. Now and then we will cite results for that algorithm and discuss its relation to Algorithm 2.

3. Global Convergence. Proofs of the global convergence of coordinate relaxation have been given [6], [14]; they also apply to overrelaxation with $\epsilon < \omega < 2 - \epsilon$, $0 < \epsilon < 1$. We therefore confine our attention to Algorithm 2. First, we will demonstrate under which conditions the Rayleigh quotients (2.2) $\mu_1, \mu_2, \cdots, \mu_s, \cdots$ form a decreasing sequence, since this gives a guarantee that μ_s converges to some value $\bar{\mu}$. Secondly, we study whether this $\bar{\mu}$ is an eigenvalue of (1.1) or, hopefully, the smallest eigenvalue.

First, we state an important identity that relates V_s defined by (2.16) to $C_s = A - t_s B$ and its main diagonal. We note that

$$V_s + V_s^T = C_s + (2 - \omega)\omega^{-1}D_s,$$

which implies that for any x

$$x^H V_s x = \tfrac{1}{2}[(2 - \omega)\omega^{-1}x^H D_s x + x^H C_s x]. \tag{3.1}$$

This identity is used in proving the following two lemmas which give the conditions for convergence of Algorithm 2.

LEMMA 3.1. *Let x_s and μ_s be computed by Algorithm* 2 (2.15). *Then*

$$\mu_{s+1} - \mu_s = -(2-\omega)\omega^{-1} p_s^H D_s p_s / \| x_{s+1} \|_B^2, \tag{3.2}$$

where D_s and p_s are defined by (2.7) *and* (2.1).

Proof. We calculate

$$\begin{aligned}(\mu_{s+1} - \mu_s) x_{s+1}^H B x_{s+1} &= x_{s+1}^H A x_{s+1} - \mu_s x_{s+1}^H B x_{s+1} \\ &= -2 p_s^H r_s + p_s^H (A - \mu_s B) p_s \\ &= -2 p_s^H V_s p_s + p_s^H (A - \mu_s B) p_s\end{aligned}$$

and apply (3.1) to get (3.2). Q.E.D.

Lemma 3.1 gives the condition for the sequence μ_s to decrease. Evidently, this takes place whenever the matrix $((2-\omega)/\omega) D(A - t_s B)$ is positive definite. Thus we have

LEMMA 3.2. *A sufficient condition for μ_s to form a nonstationary decreasing sequence is*

$$\text{(a)} \quad 0 < \omega < 2, \tag{3.3}$$

$$\text{(b)} \quad \mu_0 < \min_i a_{ii}/b_{ii} = \min_i \mu(e_i). \tag{3.4}$$

Note the relationship between condition (b) and the Rayleigh quotients of the coordinate vectors. By reversing (b) appropriately, we get an increasing sequence instead. When we use the algorithm given by (2.19), we have to add $-\mu_s \| p_s \|_B^2 / \| x_{s+1} \|_B^2$ to the right-hand side of (3.2) [3] and so we need that $(2-\omega) D(A) + \omega \mu_s B$ has to be positive definite. This gives us

LEMMA 3.3. *If V_s is given by* (2.19) *and $\lambda_1 > 0$, the μ_s form a decreasing sequence whenever*

$$\text{(a)} \quad 0 < \omega < 2 / \left(1 - \lambda_1 \cdot \lambda_{\min}(B) / \max_i (a_{ii})\right) \tag{3.5}$$

regardless of the starting vector.

We note that (3.5) gives a larger interval for ω than (3.3), and contains no assumptions on the starting vector. These lemmas can now be combined into:

THEOREM 1. *If* (3.3) *and* (3.4) *are satisfied, then Algorithm* 2 *converges in the following sense*

$$\mu_s \to \bar{\mu} = \lambda_i, \qquad \hat{r}_s \to 0.$$

Proof. Then Lemma 3.2 implies that $\mu_{s+1} \leqq \mu_s$, $\mu_s \geqq \lambda_1$; it is true that $\mu_s \to \bar{\mu}$ and $\mu_s - \mu_{s+1} \to 0$. But since (3.2) implies that

$$\begin{aligned}\mu_s - \mu_{s+1} &= (2-\omega)\omega^{-1} p_s^H D_s p_s / \| x_{s+1} \|_B^2 \\ &\geqq (2-\omega)\omega^{-1} \cdot \min_i (a_{ii} - \mu_0 b_{ii}) p_s^H p_s / \| x_{s+1} \|_B^2 \\ &\geqq \text{const} \| \hat{p}_s \|_B^2,\end{aligned}$$

it follows that $\|\hat{p}_s\|_B \to 0$. But by (2.15), $\hat{r}_s = V_s \hat{p}_s$ and, since $\| V_s \|$ is bounded, it follows that $\hat{r}_s \to 0$. Q.E.D.

Obviously, we have to prescribe that λ_i should be simple in order to be sure that $\hat{x}_s$ converge, otherwise the limit points of the sequence $\hat{x}_s$ may form a continuum in the subspace spanned by the eigenvectors corresponding to λ_i. Moreover, we cannot be sure of getting convergence towards λ_1 unless we choose such a starting vector that $\mu_0 < \lambda_2$.

4. Asymptotical Convergence Rates. First we show that, provided our general algorithm converges, it will approach a linear iteration in the limit, and then we study this linear iteration in order to see how we shall choose the parameter ω.

As can be seen, e.g. in [20], the eigenvalue approximations μ_s converge much faster than the residuals $\hat{r}_s$ or the vectors $\hat{x}_s$. Further, the parameters α_{sp} in Algorithm 1 soon approach their limiting value 1, as we see from (2.14). The limiting linear iteration will be (2.8)

$$x_{s+1} = M x_s, \qquad M = M(\omega, \lambda_1). \tag{4.1}$$

We now compare the computed x_{s+1} with what the linear iteration (4.1) would have produced. (V and C denote the limiting values of V_s and C_s (2.6), (2.9).)

$$\begin{aligned} x_{s+1} - M(\omega, \lambda_1) x_s &= (I - V_s^{-1} C_s) x_s - (I - V^{-1} C) x_s \\ &= (V^{-1} C - V_s^{-1} C_s) x_s \\ &= V^{-1}(C - C_s) x_s + (V^{-1} - V_s^{-1}) C_s x_s. \end{aligned} \tag{4.2}$$

Now we can bound

$$\| C - C_s \|_2 = \| \operatorname{diag}(\lambda_1 - \mu_{sp}) B \|_2 \leqq K_1 \cdot (\lambda_2 - \lambda_1)^{-1} \| r_s \|_B^2, \tag{4.3}$$

$$\begin{aligned} \| V^{-1} - V_s^{-1} \|_2 &= \| V^{-1} (V_s - V) V_s^{-1} \|_2 \\ &\leqq K_2 \| W_s - \omega I \|_2 + K_3 (\lambda_2 - \lambda_1)^{-1} \| r_s \|_B^2. \end{aligned} \tag{4.4}$$

The proof of (4.3) is a simple application of the results in [20] while, for (4.4), we additionally use (2.9) to get the identity:

$$\begin{aligned} V^{-1} - V_s^{-1} &= V^{-1} (V_s - V) V_s^{-1} \\ &= V^{-1} \{ -\omega^{-1} (W_s - \omega I) W_s^{-1} D_s + \omega^{-1} (D_s - D) - (E_s - E) \} V_s^{-1}. \end{aligned}$$

If we now prescribe that convergence has occurred so far that $\frac{1}{2} < \alpha_{sp} < 2$ in (2.7), constants K_1, K_2, and K_3 can be found that are independent of s. Since (2.12) implies $|\eta| \leqq K_4 \| r_s \|_B$, we see that

$$\| W_s - \omega I \|_2 \leqq K_5 \| r_s \|_B,$$

so we have bounded both terms of (4.2) by second order quantities in $\| r_s \|_B$. Consequently, we will get full information on the asymptotical behavior of the iterates by studying the limiting linear iteration (4.1).

We note that for any choice of V_s, u_1 is a fixed point of the iteration and, consequently, the matrix of the linear iteration (4.1) has u_1 as eigenvector to the eigenvalue 1, $M(\omega,\lambda_1)u_1 = u_1$ and so the iteration (4.1) converges to a vector in the subspace of u_1, provided that the remaining eigenvalues are less than 1 in magnitude, and the rate of convergence (2.4) is determined by

$$R = \max_{\lambda_i \neq 1} |\lambda_i(M(\omega,\lambda_1))|.$$

The fastest asymptotical rate of convergence will be obtained if ω is chosen so that R is minimized. The considerations we have to make are quite similar to those in the linear systems case: we have to determine ω so that all the eigenvalues of $M(\omega,\lambda_1)$, except the dominant one $\lambda_1 = 1$, will get the same absolute value. This is possible in the case that $A - \lambda_1 B$ is a 2-cyclic matrix (property A) [21].

THEOREM 2. *Suppose that* $\lim t_s = \lambda_1$ *and that* $C = A - \lambda_1 B$ *satisfies property* A. *Then the eigenvalues* η *of* $M(\omega,\lambda_1)$ *and* μ *of the Jacobi iteration matrix*

$$\mathscr{B} = I - D(C)^{-1/2}CD(C)^{-1/2} \tag{4.5}$$

ordered as

$$1 = \mu_1 > \mu_2 = 1 - \epsilon \geqq \mu_3 \geqq \cdots \geqq \mu_n = -1$$

satisfy the relation

$$(\eta - 1 + \omega)^2 = \omega^2\eta\mu^2. \tag{4.6}$$

If we choose $\omega = \omega_c$ *where*

$$\omega_c = 2/(1 + (1 - \mu_2^2)^{1/2}), \tag{4.7}$$

we get the best asymptotic rate of convergence

$$\begin{aligned} R = \omega_c - 1 &= [1 - (1 - \mu_2^2)^{1/2}]/[1 + (1 - \mu_2^2)^{1/2}] \\ &= 1 - 2\sqrt{2}\epsilon^{1/2} + O(\epsilon). \end{aligned} \tag{4.8}$$

Proof. In this case,

$$M(\omega,\lambda_1) = \left(\frac{1}{\omega} D(C) - E(C)\right)^{-1} \left(\frac{1}{\omega} D(C) + E(C)^T\right)$$

the ordinary SOR iteration matrix for C, and we can apply the theory from the linear systems case [21]. Since C is singular, the Jacobi iteration matrix $\mathscr{B}$ will have $\mu = \pm 1$ as a pair of eigenvalues and they will correspond to $\eta = 1$ and $\eta = (\omega - 1)^2$. For $\omega > \omega_c$ (4.7), all the other eigenvalues of B will correspond to complex eigenvalues η with $|\eta| = \omega - 1$, while if $\omega < \omega_c$, μ_2 will correspond to a real η with $|\eta| > \omega_c - 1$. In the case $\omega = \omega_c$, when all the eigenvalues of $M(\omega,\lambda_1)$ except two are situated at the circle $|\eta| = \omega_c - 1$, R is thus minimized and gets the value (4.8). Q.E.D.

The eigenvalues of the Jacobi matrix $\mathscr{B}$ are closely related to those of the original problem (1.1). Consider the following special case

COROLLARY 1. *Suppose $B = I$ and $D(A) = dI$. Then*

$$\begin{aligned} \mathscr{B} &= I - (A - \lambda_1 I)/(d - \lambda_1), \\ \mu_k &= 1 - (\lambda_k - \lambda_1)/(d - \lambda_1) = 1 - 2(\lambda_k - \lambda_1)/(\lambda_n - \lambda_1). \end{aligned} \tag{4.9}$$

The last equality is a consequence of the fact that a matrix with property A and zero diagonal elements has pairs of eigenvalues $\lambda_k = -\lambda_{n-k}$. When B is not diagonal the situation is somewhat more complicated, but we believe that (4.9) will also give a decent approximation to μ_k in other cases where it is possible to apply the SOR theory.

When (4.9) holds, we get (4.8), $R = 1 - 4K^{-1/2} + O(K^{-1})$ where

$$K = (\lambda_n - \lambda_1)/(\lambda_2 - \lambda_1) \tag{4.10}$$

the condition number of $C = A - \lambda_1 B$. (Compare [18].)

If we use the V_s chosen in [3] (2.19) and make the same assumptions as in Corollary 1, we get the limiting iteration matrix

$$\begin{aligned} M &= (I - \omega E)^{-1}((1 - \omega)I + \omega E^T + \lambda_1 \omega d^{-1} I), \\ E &= d^{-1}E(A) \end{aligned}$$

and a simple analysis shows that the eigenvalue relation (4.6) is replaced by

$$(\eta - 1 + \omega(1 - \lambda_1 d^{-1}))^2 = \omega^2 \eta \mu^2$$

and we get

THEOREM 3. *Let V_s be chosen by* (2.19) *and A satisfy property* A *and $D(A) = dI$ and $B = I$. Then the best rate of convergence occurs for*

$$\omega = 2\beta/(1 + (1 - \beta^2 \mu_2^2)^{1/2}) \tag{4.11}$$

where

$$\begin{aligned} \beta &= d/(d - \lambda_1) = (\lambda_n + \lambda_1)/(\lambda_n - \lambda_1), \\ \mu_2 &= 1 - \lambda_2/d = 1 - 2\lambda_2/(\lambda_n + \lambda_1). \end{aligned}$$

The best rate is given by

$$R = \omega\beta^{-1} - 1 = 1 - 4K^{-1/2} + O(K^{-1})$$

where K is defined by (4.10).

We note that Theorem 3 is proved under stronger assumptions than Theorem 2. Also, here it is possible that the optimal $\omega > 2$ (cf. (3.5)) and it is no longer independent of how we shift the problem ($A' = A - \alpha B$ giving $\lambda_k' = \lambda_k - \alpha$). On the other hand, the best rate R is independent of such shifts and is the same as (4.8).

5. Numerical Examples. Several numerical tests of the algorithms described here have been performed. The programs have been run on a CD 3300 computer at the Umeå University Computing Center. It has 48 bits word length, 36 bits of which are the mantissa. The purpose of the tests has been to see how well the theory developed for special cases (property A etc.) holds in a more general situation, and to compare the SOR algorithms to other relevant algorithms, mainly the Lanczos and c-g algorithms. Tests of these algorithms have been reported earlier [18].

As a first very simple example, we consider the tridiagonal matrix

$$(5.1)\qquad A = T = \begin{bmatrix} 2 & -1 & & & 0 \\ -1 & 2 & \cdot & & \\ & & & \cdot & -1 \\ & & & \cdot & \\ 0 & & & & 2 \end{bmatrix}.$$

It has property A, and its eigenvalues are known to be

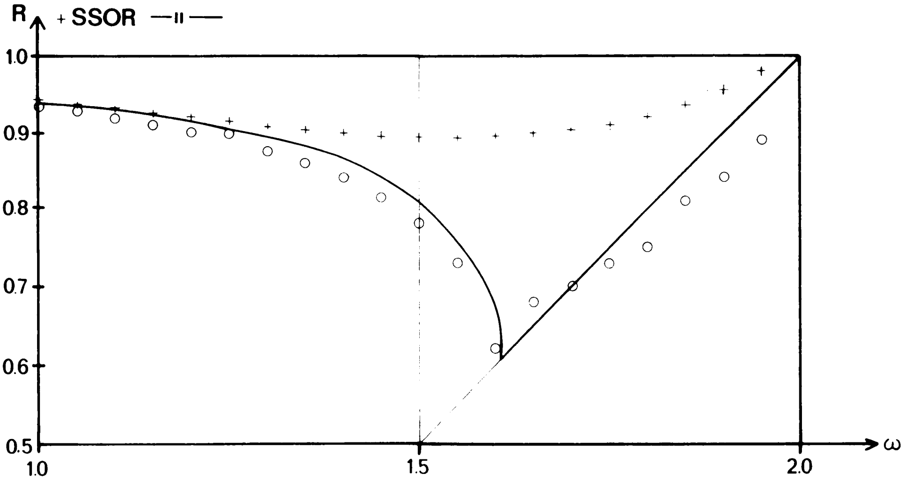
$$(5.2)\qquad \lambda_k = 4\sin^2(k\pi/2(n+1)), \qquad k = 1, 2, \cdots, n.$$

We performed Algorithm 2 for $n = 20$ and 100 with $\omega = 1(0.05)2$. The observed values of $R = \lim \|r_s\|_2^{1/s}$ are plotted in Fig. 1 and we see that they are well in agreement with the theoretical value suggested by the SOR theory. In the same figure, we also plotted the corresponding R obtained by the SSOR method, and we note that this latter method converged much more slowly. We also did several tests with (2.19) [3]; the results confirmed the the theory in Theorems 2 and 3 above. Theorem 2 gives $\omega_c = 1.59$ for $n = 20$ and $\omega_c = 1.90$ for $n = 100$.

Fig.1 A=T n=20

Rate of convergence for different ω

○ SOR algorithm



We also tried more nontrivial examples with property A, such as the eigenvalues of the "standard" L shaped membrane (see e.g. [7, p. 334]). In

Table 1, we list theoretical and observed convergence quotients together with the total number of iterations for some values of ω.

TABLE 1. *L Shaped Membrane*

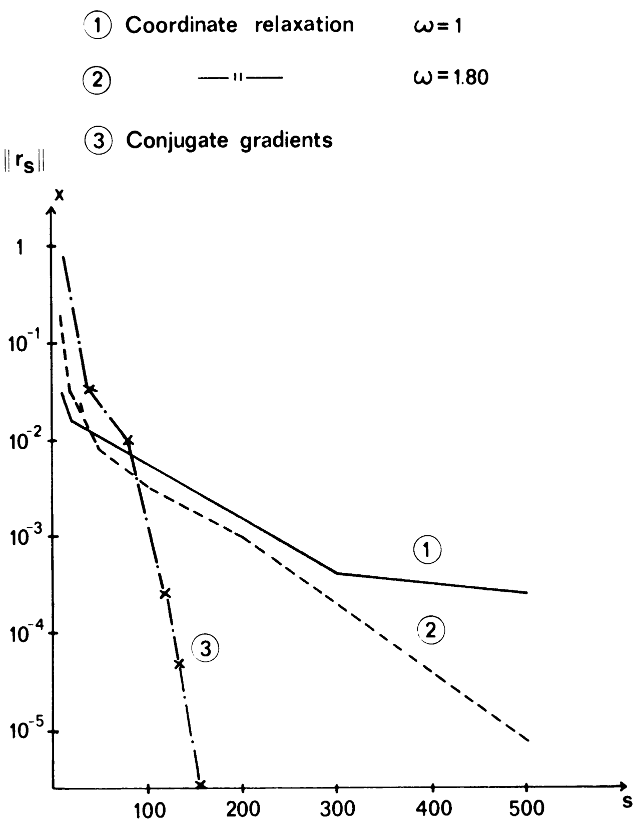
(a) $h = 1/8$ $n = 84$ $\omega_c = 1.65$
(b) $h = 1/16$ $n = 360$ $\omega_c = 1.812$

		Convergence quotient R		Iterations until
	ω	Theoretical	Observed	$\|r_s\| \leqq 10^{-5}$
(a)	1.00	0.956	0.922	111
	1.50	0.862	0.734	38
	1.55	0.836	0.696	33
	1.60	0.797	0.643	28
	1.65	0.703	0.665	29
	1.70	0.700	0.695	32
	1.80	0.800	0.770	45
	1.90	0.900	0.841	67
(b)	1.60	0.956	0.917	99
	1.70	0.934	0.870	72
	1.80	0.867	0.862	52
	1.81	0.836	0.842	54
	1.82	0.820	0.824	56
	1.83	0.830	0.824	58
	1.90	0.900	0.906	84

In order to show the behavior of the algorithms for matrices without property A, we list results for two such matrices, one with quite clustered eigenvalues and the other a Stieltjes matrix. We used the matrix $A = T^2$ (5.1), where we see that the eigenvalues are clustered in the lower end of the spectrum (5.2). Even if we content ourselves with choosing $n = 20$, we get $K = 2100$ (4.10) which gives $\omega_c = 1.92$ (4.9) and a theoretical quotient of $R = 0.92$. However, the best convergence we could get was obtained for $\omega = 1.80$ and was $R = 0.977$.

In Fig. 2, we list a diagram of the convergence for Algorithms 1 and 2. We also plotted results for the c-g algorithm [18], in order to get a comparison. Here, we note a definite advantage for the c-g algorithm and quite bad applicability of the SOR theory of Section 4 in this paper. The results agree quite well with those reported by Engeli et al. [5] for the linear equation problem. It might be of interest to study Fig. 3, where the eigenvalues of $M(\omega, \lambda_1)$ are plotted for some values of ω along with the circle $|z| = \omega - 1$. $M(\omega, \lambda_1)$ has the circle stretched in the positive real direction, $M(\omega, \lambda_n)$ in the negative direction.

Fig.2 $A=T^2$ smallest eigenvalue n=20

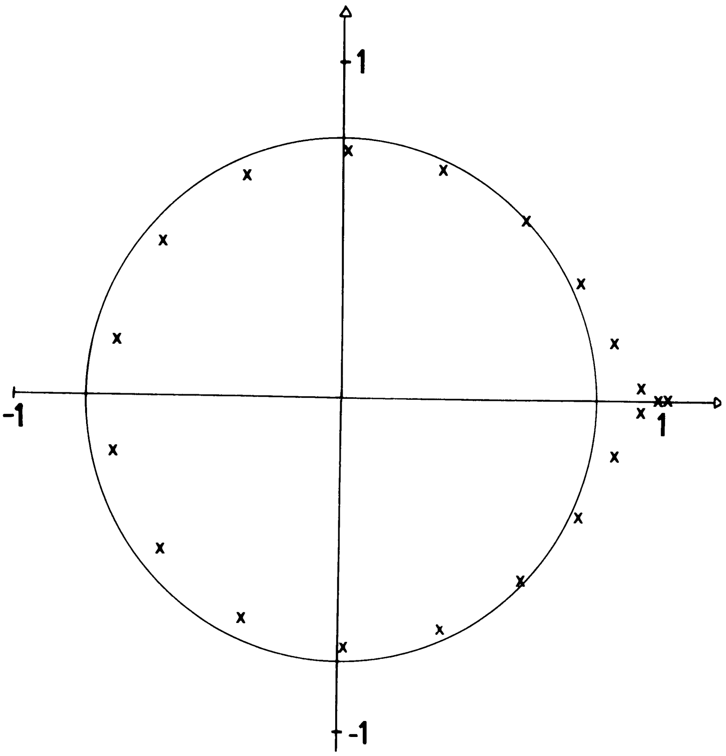


The two real eigenvalues of $M(\omega, \lambda_1)$ are not at 1 and $(\omega - 1)^2$, as Theorem 2 would have implied; on the other hand, they are both outside the circle, and their quotient, which is very close to unity, will determine the rate of convergence which is about 10 times slower than $\omega - 1$ would suggest.

The last example is a sparse matrix with a more irregular pattern of filled-in elements; we have taken (the upper triangle of) the 54×54 matrix given by Curtis in [4] with the filled nondiagonal elements equal to -1 and the diagonal elements larger than the corresponding row sums. Here, the agreement with the SOR theory is much better, as can be seen from Fig. 4. Here, (4.9) gives $\omega_c = 1.61$ and the best convergence is observed for $\omega = 1.55$.

Our theory for Algorithm 2 does not guarantee convergence towards the smallest (or largest) eigenvalue, unless the starting vector is chosen quite carefully as indicated by Theorem 1. The tests performed indicate that, in practice, we get convergence either to the largest or the smallest eigenvalue, depending on $\mu(x_0)$. We tried the matrix $A = T^2$ with starting vectors consisting of ± 1 with different numbers of sign changes and got the smallest

Figure 3a: $A = T^2$, $n = 20$, lower end of spectrum
Eigenvalues of SOR-iteration matrix for $\omega = 1.80$



eigenvalue for $0-10$ and the largest for $11-19$ $(n=20)$ changes, respectively. The convergence was only dependent on the starting vector, not on the choice of ω. In some cases, convergence to eigenvalues other than those in the ends of the spectrum was observed, that took place for the Curtis matrix with starting vector $x_0 = (1, -1, 1, \cdots, -1)^T$, where the second largest eigenvalue was obtained.

When we come to compare coordinate relaxation, Algorithm 1, to the simpler Algorithm 2, the former gives a marginally faster convergence in the earlier iterations. However, it needs more work in each iteration, namely [19], $27n+z$ multiplications or divisions, where z is the total number of nonzero elements in A and B, while Algorithm 2 needs only $6n+z$, quite a substantial saving for the sparse matrices which we are considering $(z < 10n)$. The c-g algorithm is of a comparable speed, since it needs [18], $11n+z$ operations and we note that it most often converges in fewer iterations than the algorithms considered here. Thus, our tests indicate that SOR methods are the fastest only for very sparse matrices z around $5n$.

Figure.3b: A=T^2, n=20, upper end of spectrum.
Eigenvalues of SOR-iteration matrix for ω=1.50(0.05)1.65.

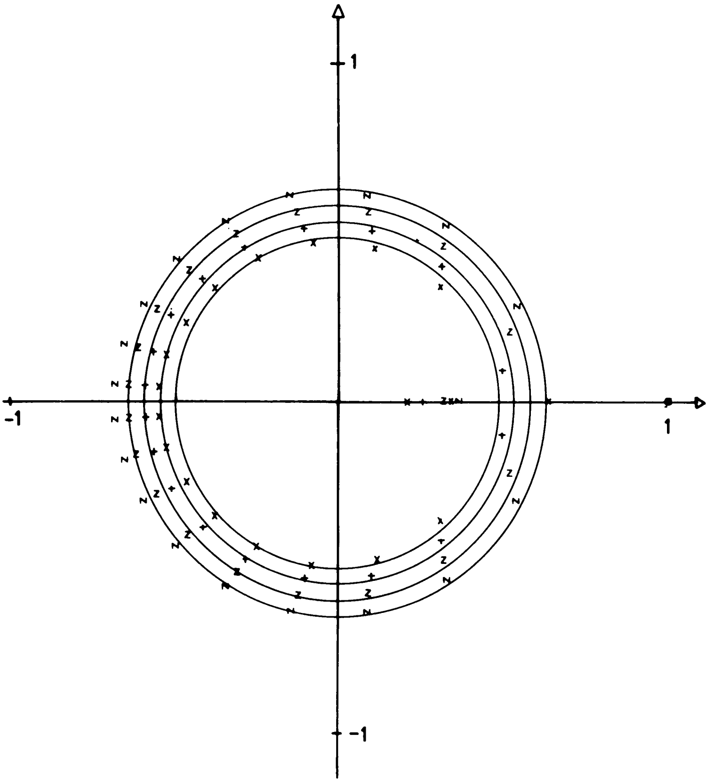
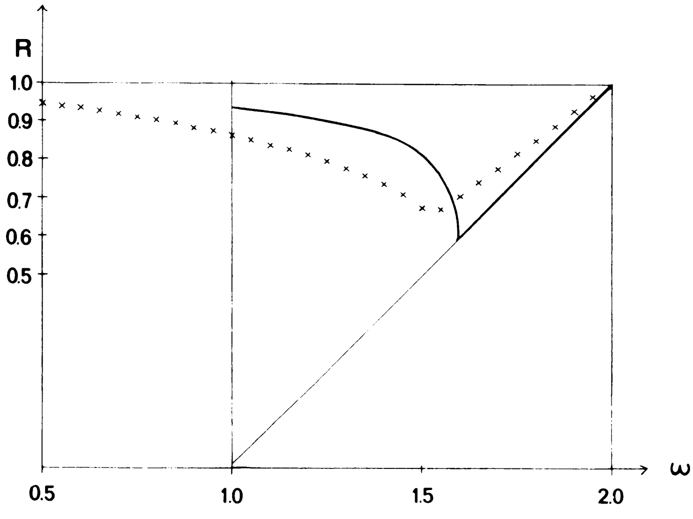


Fig.4 A=Curtis' matrix n=54
Smallest eigenvalue

Rate of convergence for different ω



Acknowledgements. The author gratefully acknowledges the help of several of his students in preparing this work. Per Byström and Sara Holm [13] did preliminary tests reported as term papers, and the computer tests reported in Section 5 were performed by Bo Kågström and Per Lindström. He also had valuable discussions with Professor H. R. Schwarz [19].

Department of Information Processing
Institute of Mathematics and Statistics
University of Umeå
S-901 87 Umeå, Sweden

1. O. Axelsson, "A generalized SSOR method," *Nordisk Tidskr. Informationsbehandling (BIT)*, v. 13, 1972, pp. 443-467.

2. W. W. Bradbury & R. Fletcher, "New iterative methods for solution of the eigenproblem," *Numer. Math.*, v. 9, 1966, pp. 259-267. MR **36** # 2313.

3. G. Buffoni, "Evaluation of eigensolutions of discrete space diffusion equation," *Calcolo*, v. 4, 1967, pp. 169-177. MR **42** # 7077.

4. A. R. Curtis & J. K. Reid, "The solution of large sparse unsymmetric systems of linear equations," *Proc. IFIP Congress* (Ljubljana, 1971), C. V. Freeman (Ed.), North-Holland, Amsterdam, 1972.

5. M. Engeli, Th. Ginsburg, H. Rutishauser & E. Stiefel, "Refined iterative methods for computation of the solution and the eigenvalues of self-adjoint boundary value problems," *Mitt. Inst. Angew. Math. Zürich*, No. 8, 1959. MR **26** # 3218.

6. D. K. Faddeev & V. N. Faddeeva, *Computational Methods in Linear Algebra*, Fizmatgiz, Moscow, 1960; English transl., Freeman, San Francisco, Calif., 1963. MR **28** # 1742.

7. G. E. Forsythe & W. Wasow,*Finite-Difference Methods for Partial Differential Equations*, Appl. Math. Ser., Wiley, New York, 1960. MR **23** # B3156.

8. Isaac Fried, "Gradient methods for finite element eigenproblems," *AIAA J.*, v. 7, 1969, pp. 739-741.

9. M. Geradin, *Analyse Dynamique Duale des Structures par la Méthode des Eléments Finis*, Dissertation, Université de Liège, Belgium, 1972.

10. G. H. Golub, "Some uses of the Lanczos algorithm in numerical linear algebra," *Topics in Numerical Analysis*, J. H. Miller, Editor, Academic Press, London and New York, 1973, pp. 173-184.

11. F. G. Gustavson, "Some basic techniques for solving sparse systems of linear equations," *Sparse Matrices and Their Applications*, D. J. Rose & R. A. Willoughby, Editors, Plenum Press, New York, 1972, pp. 41-52.

12. M. R. Hestenes & W. Karush, "A method of gradients for the calculation of the characteristic roots and vectors of a real symmetric matrix," *J. Res. Nat. Bur. Standards*, v. 47, 1951, pp. 45-61. MR **13**, 283.

13. S. Holm, *Coordinate Overrelaxation Methods for the Eigenproblem*, Report UMINF-33.73, Department of Information Processing, Umeå, Sweden.

14. W. Kahan, *Relaxation Methods for an Eigenproblem*, Technical Report CS-44, Computer Science Department, Stanford University, Stanford, California, 1966.

15. C. Lanczos, "An iteration method for the solution of the eigenvalue problem of linear differential and integral operators," *J. Res. Nat. Bur. Standards*, v. 45, 1950, pp. 255-282. MR **13**, 163.

16. J. M. Ortega & W. C. Rheinboldt, "Local and global convergence of generalized linear iterations," *Studies in Numerical Analysis*. 2: *Numerical Solutions of Nonlinear Problems* (Sympos. SIAM, Philadelphia, Pa., 1968), Soc. Indust. Appl. Math., Philadelphia, Pa., 1970, pp. 122-143. MR **43** # 4242.

17. C. C. Paige, "Computational variants of the Lanczos method for the eigenproblem," *J. Inst. Math. Appl.*, v. 10, 1972, pp. 373-381.

18. A. Ruhe, *Iterative Eigenvalue Algorithms for Large Symmetric Matrices*, Report UMINF-31.72, Department of Information Processing, Umeå, Sweden.

19. H. R. Schwarz, "The eigenvalue problem $(A - \lambda B)x = 0$ for symmetric matrices of high order," *Comp. Meth. Appl. Mech. Engnrg.*, v. 3, 1974, pp. 11-28.

20. G. W. Stewart, "On the sensitivity of the eigenvalue problem $Ax = \lambda Bx$," *SIAM J. Numer. Anal.*, v. 9, 1972, pp. 669-686.

21. D. M. Young, *Iterative Solution of Large Linear Systems,* Academic Press, New York, 1971. MR **46** #4698.

22. A. Bracha-Barak & P. E. Saylor, "A symmetric factorization procedure for the solution of elliptic boundary value problems," *SIAM J. Numer. Anal.*, v. 10, 1973, pp. 190-206.

23. P. Concus & G. H. Golub, "Use of fast direct methods for the efficient numerical solution of nonseparable elliptic equations," *SIAM J. Numer. Anal.*, v. 10, 1973, pp. 1103-1120.

24. I. Shavitt, C. F. Bender, A. Pipano & R. P. Hosteny, "The iterative calculation of several of the lowest or highest eigenvalues and corresponding eigenvectors of very large symmetric matrices," *J. Computational Phys.*, v. 11, 1973, pp. 90-108.

25. O. B. Widlund, "On the use of fast methods for separable finite difference equations for the solution of general elliptic problems," *Sparse Matrices and Applications,* D. J. Rose & W. A. Willoughby, Editors, Plenum Press, New York, 1972, pp. 121-134.