CS44

RELAXATION METHODS FOR AN EIGENPROBLEM

BY

W. KAHAN

TECHNICAL REPORT NO. CS44
AUGUST 8, 1966

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

RELAXATION METHODS FOR AN EIGENPROBLEM

BY

W. KAHAN */

ABSTRACT

A theory is devloped to account for the convergence properties of certain relaxation iterations which have been widely used to solve the eigenproblem

$$(A - \lambda B)\,\underline{x} = 0\ , \qquad \underline{x} \neq 0\ ,$$

with large symmetric matrices A and B and positive definite B . These iterations always converge, and almost always converge to the right answer. Asymptotically, the theory is essentially that of the relaxation iteration applied to a semi-definite linear system discussed in the author's previous report (1966).

---

*/ Stanford University Computer Science Department and University of Toronto Departments of Mathematics and Computer Science.
Prepared under Contract Nonr-225(37) (NR-044-211) Office of Naval Research.

Relaxation Methods for an Eigenproblem

Given are two real symmetric $N \times N$ matrices A and B with positive definite B and very large N (> 500). Required are the minimum value λ_o of the Rayleigh Quotient

$$\Lambda(\underline{x}) \equiv \underline{x}^T A \underline{x} / \underline{x}^T B \underline{x} \quad \text{for} \quad \underline{x} \neq \underline{0} \ ,$$

and a vector $\underline{x}_o$ at which the minimum is achieved. In other words, a solution is required for the symmetric eigenproblem

$$(A - \lambda_o B) \underline{x}_o = \underline{0} \ .$$

The numerical solution of this last equation is complicated by the size of N ; the matrices A and B occupy so much storage that few of to-day's electronic computers could allow access to more than a few rows at a time. It is natural to consider a relaxation iteration that approximates $\underline{x}_o$ via a converging sequence $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n, \ldots$ in which $\underline{x}_{n+1}$ differs from $\underline{x}_n$ in Just one component, because such a process can make do with restricted access to the matrices A and B . But some questions arise. How best should a specified component of $\underline{x}_n$ be changed, and what will the consequences be? Does the iteration necessarily converge to the right answer?

Surprisingly, these convergence questions have not yet been discussed in print, although a variety of relaxation methods have been widely used for a long time. For example, see Shaw (1953) Ch. VIII and

Nesbet (1965). The object of this report is to shed some theoretical light upon the convergence questions. Since practical applications motivated this work, the hypotheses are only as weak as are likely to fit methods currently in use. Consequently, the conclusions are not as general as those of A. Ostrowski (1965), with whose independent work there is some overlap.

1.) Preliminaries

First, some convenient abbreviations. Since B is positive definite, it can be used to define an inner product

$$(\underline{x} , \underline{y}) \equiv \underline{x}^T B \underline{y}$$

and a norm

$$\|\underline{x}\| \equiv \sqrt{(\underline{x} , \underline{x})} = \sqrt{\underline{x}^T B \underline{x}}$$

with the usual properties. Hence

$$\Lambda(\underline{x}) \equiv \underline{x}^T A \underline{x} / \|\underline{x}\|^2 \quad .$$

It is also useful to have the residual vector function

$$\underline{r}(\underline{x}) \equiv (A - h(x) B) \underline{x}$$

because $\underline{r}(\underline{x})$ has the same direction as grad $\Lambda(\underline{x})$;

$$d\Lambda(\underline{x}) = 2 \underline{r}(x)^T d\underline{x} / \|\underline{x}\|^2 \quad .$$

This shows that $\Lambda(\underline{x})$ is minimized when

$$\begin{cases} \underline{r}(\underline{x}_o) = 0 & \text{for} \quad \underline{x}_o \neq 0 \ , \quad \text{and} \\ \\ \lambda_o = \Lambda(\underline{x}_o) & \text{is minimized} \ . \end{cases}$$

Incidentally, the relation $\underline{x}^T g(x) \equiv 0$ will be used without comment.

The next step is to replace the infinitesimal dx with a finite increment $\Delta\underline{x}$. Starting with some arbitrary $\underline{x} \neq 0$, and the corresponding

$$\lambda \equiv \Lambda(\underline{x}) \quad \text{and}$$

$$\underline{r} \equiv \underline{r}(\underline{x}) \quad ,$$

we consider the consequences of changing $\underline{x}$ to x + x . One consequence is that λ changes to

$$\lambda + \Delta\lambda \equiv \Lambda(\underline{x} + \Delta\underline{x}) \quad ;$$

$$\Delta\lambda = [2\,\Delta\underline{x}^T\underline{r} + (\Lambda(\Delta\underline{x}) - \lambda)\|\Delta\underline{x}\|^2]/\|\underline{x} + \Delta\underline{x}\|^2 \ .$$

We are incline&to prefer $(\underline{x} + \Delta x)$ to $\underline{x}$ 'whenever $\Delta\lambda < 0$.

Suppose now that $\Delta\underline{x}$ is restricted to the form

$$\Delta\underline{x} = \xi\underline{p}$$

where $\underline{p}$ is a given direction and ξ is a scalar which we hope to choose

in such a way that $\Delta\lambda < 0$. We shall abbreviate

$$\pi = \Lambda(\underline{p})$$

and assume, without loss of generality, that

$$\pi > \lambda \quad .$$

(Otherwise x and $\underline{p}$ should be exchanged.) It is convenient to assume further that

$$\pi > \lambda$$

for two reasons; first because much of what follows can be extended to the case $\pi = \lambda$ via a suitable limiting process with $\pi \to \lambda +$. Second, the vector p will range over coordinate vectors like

$$\underline{e}_i = (0 \ , \ 0 \ , \ldots , \ 0 \ , \ 1 \ , \ 0 \ , \ 0 \ , \ldots , \ 0)^T \ ,$$

with a 1 in the i^{th} position, so the value π will range over the quotients a_{ii}/b_{ii} of diagonal elements of A and B ; and an infinitesimal increase in a_{ii} can be of little practical importance. Now setting $\Delta\underline{x} = \xi\underline{p}$ produces

$$\Delta\lambda = (\pi - \lambda)\, \xi(\xi - 2\eta)\, \|\underline{p}\|^2 / \|\underline{x} + \xi\underline{p}\|^2$$

$$= (\pi - \lambda)(1 - 2\eta/\xi)\, \|\Delta\underline{x}\|^2 / \|\underline{x} + \Delta\underline{x}\|^2$$

where η is another abbreviation;

$$\eta = - \underline{p}^T \underline{r} / [(\pi - \lambda) \|\underline{p}\|^2] \quad .$$

The last few equations, together with the inequality $\pi - \lambda > 0$, imply that $\Delta\lambda < 0$ for any value of ξ strictly between 0 and 2η. Unless $\eta = 0$, there will be some latitude in the choice of ξ, and it is natural to look for the best choice.

The "locally best" value for ξ is defined now as that value $\xi = \zeta$ for which $\Delta\lambda$ is a non-positive minimum. That value ζ always exists, even when $\pi = \lambda$, and satisfies the equation

$$\underline{p}^T \underline{r} / \|\underline{p}\|^2 + (\pi - \lambda)\zeta + [(\pi - \lambda)(\underline{p}, \underline{x}) - \underline{p}^T \underline{r}]\zeta^2 / \|\underline{x}\|^2 = 0 \, .$$

This equation is derived from the condition

$$\frac{d}{d\xi} \Lambda(\underline{x} + \xi \underline{p}) = 0 \quad \text{at} \quad \xi = \zeta \, ,$$

i.e. $$\underline{p}^T \underline{r}(\underline{x} + \zeta \underline{p}) = 0 \quad .$$

The last equation is similar to one satisfied by η in which $\Lambda(\underline{x} + \zeta \underline{p})$ is replaced by $\lambda = \Lambda(\underline{x})$:

$$\underline{p}^T (A - \lambda B)(x + \eta \underline{p}) = 0 \, .$$

The resemblance is also apparent in the formula

$$\zeta = 2\eta / [1 + \sqrt{1 + 4\eta(\eta \|\underline{p}\|^2 + (\underline{p}, \underline{x}))/\|\underline{x}\|^2}] \quad ,$$

which shows that ζ and η differ by a relatively small amount whenever

η is small or, more precisely, whenever $|\eta| << \|\underline{x}\|/\|\underline{p}\|$ . . This condition is satisfied when x is a sufficiently good approximation to $\underline{x}_o$, so there is some justification for simply choosing $\xi = \eta$ as is so often done in practice (e.g., by Nesbet (1965)). But when η is large then the choice $\xi = \zeta$ is much to be preferred. (Incidentally, $|\zeta| < \|\underline{x}\|/\|\underline{p}\|$.)

When $\xi = \zeta$ the change $\Delta\lambda$ satisfies

$$\Delta\lambda \leq -(\underline{p}^T\underline{r})^2/(3\, s\|\underline{p}\|^2\, \|\underline{x}\|^2) \quad \text{where}$$

$$s = \max(\Lambda(u) - \Lambda(v)) \quad \text{overall} \quad \underline{u} \neq 0 \text{ and } \underline{v} \neq 0 \ .$$

(i.e. s is the spread in the field of values of A .) A brief proof of the last inequality is given in Appendix I. That proof suggests that the choice of ξ is not very critical. In what follows, we shall assume only that ξ approximates ζ roughly, but well enough that

$$\Delta\lambda \leq -C(\underline{p}^T\underline{r})^2/(\|\underline{p}\|^2\|\underline{x}\|^2)$$

for some positive constant C which is independent of $\underline{p}$ and $\underline{x}$. The last inequality, weak though it is, suffices to establish convergence of the iterations described in the next section.

It is of considerable practical importance that the theory not restrict ξ to be either η or ζ , even though the latter value is the best value to use for any single step. Experience with relaxation processes suggests strongly that the best strategy for choosing values of ξ to be used in each of a long sequence of steps may well require that each value of ξ differ in some systematic way from the corresponding value of ζ

in each step. For example, a policy of overrelaxation, in which ξ/ζ is kept roughly constant and greater than 1 in each step, sometimes produces faster convergence than the policy of keeping $\xi/\zeta = 1$. Unfortunately, the theory of overrelaxation is not as well developed for the eigenproblem as it is for solving linear systems (cf. **Varga(1962)**).

2.) The Iteration

Let $\underline{e}_j$ **denote the** j^{th} coordinate vector,

$$\underline{e}_j = (0 , 0 ,\ldots, 0 ,1 , 0 ,\ldots, 0)^T \quad \text{for } 1 \leq j \leq N ,$$

with a 1 in the j^{th} position. Let

$$\underline{p}_1 , \underline{p}_2 ,\ldots, \underline{p}_n , \cdots$$

be a sequence constructed by choosing $\underline{p}_n = \underline{e}_j$ for some $j = J(n)$. Later, in section 3, more will be said about the way in which J(n) should behave. For the present, the notation for the sequence $\underline{p}_n$ is used merely to avoid a notation like

$$\underline{e}_{j_1} , \underline{e}_{j_2} ,\ldots, \underline{e}_{j_n} , \cdots$$

with subscripted subscripts.

Now we define the relaxation iteration to solve $Axx = \lambda_o B \underline{x}_o$ for $\underline{x}_o \neq 0$ and the smallest possible value of $\lambda_o = \Lambda(\underline{x}_o)$. Beginning with an arbitrary $\underline{x}_1$ with $\|\underline{x}_1\|^2 = \underline{x}_1^T B \underline{x}_1 > 0$ and $\lambda_1 = \Lambda(\underline{x}_1) = \underline{x}_1^T A \underline{x}_1 / \|\underline{x}_1\|^2$, we define for $n = 1 , 2 , 3 ,\ldots$

$$\underline{r}_n = \underline{r}(\underline{x}_n) = (A - \lambda_n B)\underline{x}_n \ ,$$

$$\pi_n = \Lambda(\underline{p}_n) \ ,$$

$$\eta_n = - \underline{p}_n^T \underline{r}_n / [(\pi_n - \lambda_n)\|\underline{p}_n\|^2] \ ,$$

$$\zeta_n = 2\eta_n / [1 + \sqrt{1 + 4\eta_n(\eta_n\|\underline{p}_n\|^2 + \underline{p}_n^T B\underline{x}_n)/\|\underline{x}_n\|^2}\,] \ ,$$

ξ_n is an acceptable approximation to ζ_n (see section 1) ,

$$\underline{x}_{n+1} = \underline{x}_n + \xi_n \underline{p}_n \ ,$$

$$\|\underline{x}_{n+1}\|^2 = \|\underline{x}_n\|^2 + 2\,\xi_n\,\underline{p}_n^T B\underline{x}_n + \xi_n^2\,\|\underline{p}_n\|^2 \ ,$$

$$\Delta\lambda_n = - (\pi_n - \lambda_n)\,\xi_n(2\eta_n - \xi_n)\,\|\underline{p}_n\|^2/\|\underline{x}_{n+1}\|^2 \ , \text{ and}$$

$$\lambda_{n+1} = \lambda_n + \Delta\lambda_n \ .$$

This computation is simpler than it seems because $\underline{p}_n$ is just one of the co-ordinate vectors $\underline{e}_j$. Therefore $\underline{p}_n^T A = \underline{a}_j^T$ is the j^{th} row of A , and $\underline{p}_n^T B = \underline{b}_j^T$ similarly. There is no need to compute $\underline{r}_n$ explicitly, merely $\underline{p}_n^T \underline{r}_n = \underline{a}_j^T \underline{x}_n - \lambda_n \underline{b}_{3-n}^T \underline{x}$. The number $\|\underline{p}_n\|^2 = b_{jj}$. is the j^{th} diagonal element of B , and $\pi_n = a_{jj}/b_{jj}$. Whenever $\eta_n << \|\underline{x}_n\|/\|\underline{p}_n\|$, ζ_n need not be computed but can be approximated by η_n ; this will happen for all sufficiently large n. Since the bulk of the work is concentrated in the computation of $\underline{p}_n^T A\,\underline{x}_n = \underline{a}_j^T \underline{x}_n$ and $\underline{p}_n^T B\,\underline{x}_n = \underline{b}_j^T \underline{x}_n$, the work required to go from $\underline{x}_n$ to $\underline{x}_{n+1}$ is roughly 2N multiplications and 2N additions, possibly much less if A and B are sparse. (If $\underline{p}_n$ were not a coordinate vector, the work

would be N times larger unless $\underline{p}_n^T A$ and $\underline{p}_n^T B$ were already known.)

By induction, $\lambda_n = \Lambda(\underline{x}_n)$ for all n . We can assume that $\pi_n > \lambda_1$ for the reasons given in section 1, and hence prove by induction that $\Delta\lambda_n \leq 0$. Therefore there must exist a limiting value

$$\lambda_\infty = \lim \lambda_n \quad \text{as} \quad n \to \infty$$

such that

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq \lambda_{n+1} \geq \cdots \geq \lambda_\infty \geq \lambda_o \quad .$$

Because of the way ξ_n was chosen, there is a constant $C > 0$ such that

$$(\underline{p}_n^T \underline{r}_n)^2 / [\|\underline{p}_n\|^2 \|\underline{x}_n\|^2] \leq - \Delta\lambda_n / C \to 0 \quad \text{as} \quad n \to \infty \quad .$$

Therefore $\eta_n \|\underline{p}_n\| / \|\underline{x}_n\| \to 0$ and hence $\xi_n \|\underline{p}_n\| / \|\underline{x}_n\|$ 40 as $n \to \infty$. Consequently

$$\|\underline{x}_{n+1} - \underline{x}_n\| / \|\underline{x}_n\| \to 0 \quad .$$

However, there is no reason yet to conclude that $\underline{x}_n$ or the normalized $\underline{x}_n^A = \underline{x}_n / \|\underline{x}_n\|$ approaches any limit. In the absence of further information about the sequence $\underline{p}_n$, the best that can be said is that the normalized sequence $\underline{x}_n^A$ has at least one point of condensation; and if there are more than one then these points of condensation form a continuum with no isolated points.

Example 1: $N = 3$, $B = I$,

$$A = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & -1 \\ 0 & -1 & 1 \end{pmatrix} , \qquad \underline{x}_1 = \begin{pmatrix} 1000 \\ 1000 \\ -1 \end{pmatrix} ,$$

and all $\underline{p}_n$ are chosen from the pair $\underline{e}_1$ and $\underline{e}_2$. Each of $\underline{e}_1$ and $\underline{e}_2$ is assumed to appear infinitely often in the sequence $\underline{p}_n$. Now it can be shown that $\lambda_n \searrow \lambda_\infty = 0$, and

$$\hat{\underline{x}}_n = \underline{x}_n / \|\underline{x}_n\| \to (\sqrt{1/2} , \sqrt{1/2} , 0)^T .$$

But $\underline{x}_n$ diverges towards $(+ \infty , + \infty , - 1)$ and

$$\lambda_\infty > \lambda_o = 1 - \sqrt{2} .$$

Example 2: Here $N = 6$ and $B = I$ and A is a diagonal sum of two matrices A from example 1 . Similarly,

$$\underline{x}_1 = (1000 , 1000 , -1 , 1000 , 1000 , -1)^T$$

and all $\underline{p}_n$ are chosen from the set $\{\underline{e}_1 , \underline{e}_2 , \underline{e}_4 , \underline{e}_5 \}$ . ... members of that set are assumed to occur infinitely often in the sequence $\underline{p}_n$. Now

$$\lambda_n \searrow \lambda_\infty = 0 > \lambda_o$$

as before, but neither $\underline{x}_n$ nor $\hat{\underline{x}}_n$ need converge. Instead, the points of condensation of the sequence $\hat{\underline{x}}_n$ constitute some arc of the curve traced by

$$(u\ ,\ u\ ,\ 0\ ,\ v\ ,\ v\ ,\ 0)^T$$

when $u > 0$, $v > 0$ and $u^2 + v^2 = 1/2$.

Example 3: N , A and B are as in example 1, but now $\underline{x}_1 = (1\ ,\ 0\ ,\ -1)^T$ and the sequence $\underline{p}_n$ is arbitrary. In this case all $\lambda_n = \lambda_1 = 1$, and all $\underline{x}_n = \underline{x}_1$. Here is a case of convergence to a wrong answer that is not obviously wrong. Fortunately, the limit $\underline{x}_1$ is unstable. $\underline{x}_1$ lies on the intersection of two planes which separate all space into four disjoint regions in which alternately $\Lambda(\underline{x}) > 1$ and $\Lambda(\underline{x}) < 1$. In any open spherical neighborhood of $\underline{x}_1$, no matter how small the sphere, there exist vectors $\underline{x}$ with $\Lambda(\underline{x}) < 1$, and if one of these is used to start the iteration then the iteration will converge to a new limit $\lambda_\infty < 1$.

Example 4: B is almost a unit matrix, in that all elements of B - 1 are very small; and A's diagonal terms a_{ii} differ by amounts that are very large compared with A's off-diagonal terms. i.e., each element of A - diag(A) is small compared with any difference $a_{ii} - a_{33}$ with $i \neq j$.

$\underline{x}_1 = \underline{e}_J$ for some J , and the sequence $\underline{p}_n$ consists of those coordinate vectors $\underline{e}_j$ with $j \neq J$ repeated infinitely often. Now $\underline{x}_n$ converges to an eigenvector whose eigenvalue is very near a_{JJ} . If a_{JJ} is the algebraically least diagonal element, then that eigenvalue is the desired λ_o .

The foregoing examples show how necessary it is to specify appropriate choices for $\underline{x}_1$ and the sequence $\underline{p}_n$ in order to secure convergence to the desired answer. In the following sections of this report, some assumptions are made regarding those choices. The assumptions are intended to be weak enough to be practical, yet strong enough to guarantee convergence to an answer which, if it is not the desired answer, can usually be checked and corrected.

3.) The Complete Iteration

This section discusses the consequences of choosing $\underline{p}_n$ from the set of N coordinate vectors $\underline{e}_j$ in such a way that each set of M consecutive vectors $\underline{p}_n$, $\underline{p}_{n+1}$, $\cdots$, $\underline{p}_{n+M-1}$ includes each coordinate vector $\underline{e}_j$ at least once. M is some fixed integer no smaller than N.

In the previous section it was shown that $\lambda_n \searrow \lambda_\infty \geq \lambda_0$ as $n \to \infty$, and that the sequence of normalized vectors $\hat{\underline{x}}_n = \underline{x}_n / \|\underline{x}_n\|$ possessed a continuum of points of condensation. Let $\hat{\underline{x}}_\infty$ be any one of these points of accumulation; it will be the limit of some subsequence of $\underline{x}_n$. Say

$$\underline{x}^n_{n_k} \to \underline{x}^n_\infty \quad \text{as} \quad n_k \to \infty \quad \text{for} \quad k = 1 , 2 , 3 , \cdots .$$

Evidently $\|\underline{x}^n_\infty\| = 1$. Furthermore, because

$$\|\underline{x}_{n+1} - \underline{x}_n\| / \|\underline{x}_n\| \to 0 \qquad \text{(see section 2)} ,$$

$$\|\underline{x}^n_{n+1} - \hat{\underline{x}}_n\| \to 0 \quad \text{too} ,$$

and therefore $\hat{\underline{x}}_{n_k+1} \to \hat{\underline{x}}_\infty$ as $k \to \infty$. Indeed,

$$\hat{\underline{x}}_{n_k+m} \to \hat{\underline{x}}_{\infty} \qquad \text{as} \qquad k \to \infty$$

for any fixed m , but we shall use this fact only for values of $m < M$.

Now let $\hat{\underline{r}}_n = \underline{r}_n / \|\underline{x}_n\|$ for all n ;

$$\overset{n}{\underline{r}}_n = \underline{r}(\hat{\underline{x}}_n) = (A - \lambda_n B)\hat{\underline{x}}_n \qquad \text{where}$$

$$\lambda_n = \Lambda(\hat{\underline{x}}_n) \quad .$$

Because of continuity,

$$\overset{n}{\underline{r}}_{n_k+m} \to \overset{n}{\underline{r}}_{\infty} \equiv r(\hat{\underline{x}}_{\infty}) \qquad \text{as} \qquad k \to \infty \quad .$$

The next objective is to deduce that $\overset{n}{\underline{r}}_{\infty} = 0$ by showing that $\underline{e}_j^T \hat{\underline{r}}_{\infty} = 0$ for $j = 1, 2, \ldots$ and N . The fact that $\underline{p}_n^T \hat{\underline{r}}_n \to 0$, proved in section 2, is exploited to prove $\underline{e}_j^T \hat{\underline{r}}_{\infty} = 0$ as follows.

Let j be fixed. Given k , it must be possible to solve

$$\underline{p}_{n_k+m} = \underline{e}_j$$

for $m = m_k$ between 0 and M-1 , because $\underline{e}_j$ appears at least once in the sequence

$$\underline{p}_{n_k}, \underline{p}_{n_k+1}, \ldots, \underline{p}_{n_k+M-1} \quad .$$

Therefore $\underline{e}_j^T \hat{\underline{r}}_{n_k+m_k} = \underline{p}_{n_k+m_k}^T \hat{\underline{r}}_{n_k+m_k} \to 0$ as $k \to \infty$, and hence

$\underline{e}_j^T \hat{\underline{r}}_\infty = 0$ as required.

Since $\hat{\underline{r}}_\infty = \underline{r}(\underline{x}_\infty) = 0$, $\hat{\underline{x}}_\infty$ must be an eigenvector and λ_∞ the corresponding eigenvalue of A with respect to B. This is so for every limit point $\hat{\underline{x}}_\infty$ in that continuum of limit points possessed by the sequence $\hat{\underline{x}}_n$. If λ_∞ is a simple eigenvalue, then $\hat{\underline{x}}_\infty$ must be one of the two normalized eigenvectors differing from each other only in sign, so in this case the normalized sequence $\hat{\underline{x}}_n$ must converge. But if λ_∞ is a multiple eigenvalue, the convergence of $\hat{\underline{x}}_n$ is a more difficult question which, together with the convergence of the unnormalized sequence $\underline{x}_n$, will be elaborated in the next section.

There is another question. Does $\lambda_\infty = \lambda_o$? It is remotely possible that $\lambda_\infty > \lambda_o$, but in this case λ_∞ cannot be a numerically stable limit. The next three paragraphs explain why.

Let C be the region containing all vectors v such that

$$\Lambda(\underline{v}) < \lambda_\infty = \Lambda(\underline{x}_\infty) .$$

C is easiest to describe with the aid of a coordinate system for

$$v \sim (v_1, v_2, \ldots, v_N)^T$$

in which B is represented by a unit matrix and A is represented by a diagonal matrix of its eigenvalues $\alpha_1, \alpha_2, \ldots, \alpha_N$. The eigenvectors of A with respect to B yield just such a coordinate system. Then $\Lambda(\underline{v}) < \lambda_\infty$ means

$$\sum_{\alpha_i > \lambda_\infty} (\alpha_i - \lambda_\infty) v_i^2 < \sum_{\alpha_i < \lambda_\infty} (\lambda_\infty - \alpha_i) v_i^2 ,$$

which describes the interior of a cone in that subspace complementary to the invariant subspace spanned by the coordinate vector(s) corresponding to the eigenvalue(s) $\alpha_i = \lambda_\infty$. The region C is the interior of the cylinder swept out by the cone as its vertex is translated throughout the invariant subspace.

Any open sphere $\mathcal{A}$ centered at $\hat{\underline{x}}_\infty$ intersects C in an open region $C \cap \mathcal{A}$ whose volume is a constant fraction f of the volume of $\mathcal{A}$ no matter how small $\mathcal{A}$ may be. And if attention is confined to the sphere $\mathcal{N}$ of normalized vectors $\hat{\underline{v}} = \underline{v}/\|\underline{v}\|$, then the area of $C \cap \mathcal{A} \cap \mathcal{N}$ is still the same fraction f of the area $\mathcal{A} \cap \mathcal{N}$. (I have used the words "sphere", "volume" and "area", as if the vectors v formed a three-dimensional space, with the intention that they apply to the corresponding N-dimensional generalizations.)

Instability, when $\lambda_\infty > \lambda_o$, stems from the situation of all limit points $\hat{\underline{x}}_\infty$ on the boundary of C . Since $\Lambda(\hat{\underline{x}}_n) \geq \lambda_\infty$, each member of the sequence $\hat{\underline{x}}_n$ must avoid the region C . But $\hat{\underline{x}}_n$ is at least as close to C as it is to any limit-point $\hat{\underline{x}}_\infty$, and $h(\hat{x})$ decreases faster when $\hat{x}$ is moved towards C than when $\hat{\underline{x}}$ is moved towards $\hat{\underline{x}}_\infty$, except possibly when $\underline{x}$ lies in the subspace complementary to that spanned by those eigenvectors corresponding to eigenvalues $\alpha_i < \lambda_\infty$. Therefore, it seems easy to imagine a force of attraction pulling each $\hat{\underline{x}}_n$ into C , and hard to imagine how the sequence can avoid succumbing to that attraction. The matter will be discussed further in the next section of this report.

Of course, the foregoing argument cannot be used to prove that $\lambda_\infty = \lambda_o$, because this is not necessarily so. The argument merely indicates how

unlikely it is that $\lambda_\infty > \lambda_0$. The risk is greater according as the second-least stationary value is closer to Λ's minimum value λ_0 , because when λ_∞ is very close to λ_0 then the region C is very narrow. This risk is not peculiar to the relaxation process. Given A , B , $\hat{\underline{x}}_\infty$ and λ_∞ such that $(A - \lambda_\infty B)\hat{\underline{x}}_\infty = 0$, and no other information, the only infallible algorithms known so far to decide whether $\lambda_\infty = \lambda_0$ or not all cost at least as much time as the triangular resolution

$$(A - \lambda_\infty B) = L D L^T ,$$

where L is a unit lower triangular matrix and D is diagonal. ($\lambda_\infty = \lambda_0$ if and only if no diagonal element of D is negative, however λ_∞ may have been obtained.) Fortunately for many applications, special information is frequently available to help avoid the risk. For example, one may know that $\underline{x}_0$ is the only eigenvector whose elements do not change sign (cf. boundary value problems with "pillow-shaped" eigenfunctions). Or one may know how to start the iteration with an $\underline{x}_1$ whose $\Lambda(\underline{x}_1)$ is less than Λ's second-least stationary value.

The last task of this sectionis to prove that

$$\hat{\underline{r}}_n = \underline{r}_n / \|\underline{x}_n\| \to 0 \qquad \text{as} \qquad n \to \infty .$$

Let $\mathcal{L}_\infty$ be the invariant subspace of eigenvectors $\underline{v}$ satisfying

$$(A - \lambda_\infty B)v = 0 ,$$

and let P_∞ be the B-orthogonal projector into that subspace; i.e.

$$(A - \lambda_\infty B)P_\infty \underline{x} = 0 \qquad \text{for every } \underline{x} \text{ , and}$$

$$(I - P_\infty)^T B P_\infty = 0 \ .$$

One way to construct P_∞ is to assemble all the normalized eigenvectors $\underline{v}_i^n$ satisfying

$$(A - \lambda_\infty B)\hat{\underline{v}}_i = 0 \quad \text{and} \quad (\hat{\underline{v}}_i , \hat{\underline{v}}_j) \equiv \hat{\underline{v}}_i^T B \hat{\underline{v}}_j = \delta_{ij}$$

into a sum

$$P_\infty = \sum_i \hat{\underline{v}}_i \hat{\underline{v}}_i^T B \ .$$

Note that $\|(I - P_\infty)\underline{x}\|$ represents the distance from $\underline{x}$ to $\mathcal{L}_\infty$.

We already know that every limit $\underline{x}_\infty$ of a convergent subsequence $\underline{x}_{n_k}^n$ lies in $\mathcal{L}_\infty$. This implies that $(I - P_\infty)\hat{\underline{x}}_n \to 0$ as $n \to \infty$; otherwise there would exist an infinite subsequence $\underline{x}_{n_k}^n$ for which $\|(I - P_\infty)\hat{\underline{x}}_{n_k}\| > \epsilon > 0$. This subsequence would contain a convergent sub-subsequence, say $\underline{x}_{n_k}^n$ itself, whose limit $\hat{\underline{x}}_\infty$ would have to satisfy

$$\|(I - P_\infty)\hat{\underline{x}}_\infty\| > \epsilon$$

too, contrary to what has already been proved. Finally,

$$\begin{aligned} \underline{r}_n^n &= (A - \lambda_n B)\hat{\underline{x}}_n \\ &= (A - \lambda_n B)(I - P_\infty)\hat{\underline{x}}_n + (A - \lambda_n B)P_\infty \hat{\underline{x}}_n \\ &= (A - \lambda_n B)(I - P_\infty)\hat{\underline{x}}_n + (\lambda_\infty - \lambda_n)B P_\infty \hat{\underline{x}}_n \\ &\to 0 \quad \text{as} \quad n \to \infty \ . \end{aligned}$$

4.) The Iteration Converges like a Geometric Series

In this section of the report, the previous section's conclusion, that $\lambda_n \to \lambda_\infty$ and $\underset{\sim}{r}_n \to 0$ as $n \to \infty$, is replaced by a stronger deduction: the sequence $\underset{\sim}{x}_n$ converges at least as quickly as a geometric series to an eigenvector $\underset{\sim}{x}_0$ corresponding to Λ's smallest stationary value λ_0, except in those rare and unstable cases when $\lambda_\infty = \lambda_0$. This deduction stems from the observation that

$$\lambda_n - \lambda_\infty = O(\hat{\underset{\sim}{r}}_n)^2 \quad \text{as} \quad \underset{\sim}{r}_n \to 0 .$$

(A more precise statement of the last equation is proved in Appendix II.) The natural thing to do is find out whether replacing λ_n by λ_∞ in the iteration formula for $\underset{\sim}{x}_n$ causes a significant change in the convergence properties of the iteration.

It is convenient to begin with an examination of η_n, ζ_n and ξ_n. We have

$$|\eta_n|/\|\underset{\sim}{x}_n\| = |\underset{\sim}{p}_n^T \hat{\underset{\sim}{r}}_n|/[(\pi_n - \lambda_n)\|\underset{\sim}{p}_n\|^2]$$

$$\leq \sqrt{\hat{\underset{\sim}{r}}_n^T B^{-1} \hat{\underset{\sim}{r}}_n}\ /[(\pi_n - \lambda_n)\|\underset{\sim}{p}_n\|]$$

$$\to 0 \quad \text{as} \quad n \to \infty .$$

Therefore, $\zeta_n/\eta_n \to 1$ except for those values of n when $\zeta_n = \eta_n = 0$. The value ξ_n must be so chosen that there exists a positive constant C such that

$$\Delta\lambda_n \leq - C(\underset{\sim}{p}_n^T \hat{\underset{\sim}{r}}_n)^2/\|\underset{\sim}{p}_n\|^2 \quad \text{for all } n$$

(cf. the second-last paragraph of section 1); this implies, by virtue of the last paragraph of Appendix I, that there exists some positive constant $d < 1$ such that

$$|\xi_n/\eta_n - 1| \leq d \quad \text{for all sufficiently large } n$$

except when $\xi_n = \eta_n = 0$. The last inequality plays an important role in what follows.

Now let us replace λ_n in the x-iteration by an independent variable μ_n. To diminish the possibility of confusion, we shall replace the letter x by y and re-define the iteration thus:

Given A, B, $\underline{p}_n$ and $\pi_n = \Lambda(\underline{p}_n)$ as before, replace $\underline{x}_n$ by a new sequence $\underline{y}_n$ defined with the aid of independent variables μ_n as follows. The residual $\underline{r}_n = (A - \lambda_n B)\underline{x}_n$ is replaced by

$$\underline{s}_n = (A - \mu_n B)\underline{y}_n .$$

The value $\eta_n = - \underline{p}_n^T \underline{r}_n / [(\pi_n - \lambda_n)\|\underline{p}_n\|^2]$ is

replaced by $\sigma_n = - \underline{p}_n^T \underline{s}_n / [(\pi_n - \mu_n)\|\underline{p}_n\|^2]$.

The value ξ_n is replaced by $\tau_n = \sigma_n \xi_n / \eta_n$, except that $\tau_n = \sigma_n$ when $\xi_n = \eta_n = 0$. Finally,

$$\underline{y}_{n+1} = \underline{y}_n - \tau_n \underline{p}_n .$$

Each vector $\underline{y}_n$ is an analytic function of μ_m for $m < n$ whose only singularities occur when $\mu_m = \pi_m$, but we shall restrict μ_m to the interval

$$\lambda_\infty \le \mu_m \le \lambda_m$$

from which π_m has already been excluded. Clearly, if $\underline{y}_n = \underline{x}_n$ and $\mu_m = \lambda_m$ for all $m > n$, then $\underline{y}_m = \underline{x}_m$ and $\underline{s}_m = \underline{r}_m$ for all $m > n$. On the other hand, if $\underline{y}_n = \underline{x}_n$ and $\mu_m = \lambda_\infty$ for all $m \ge n$, then the relations

$$\lambda_m - \lambda_\infty = O\,(\underline{r}_m/\|\underline{x}_m\|)^2 \qquad \text{and}$$

$$(\underline{x}_{m+1} - \underline{x}_m)/\|\underline{x}_m\| \to 0 \qquad \text{as } m \to \infty$$

imply that as $n \to \infty$

$$(\underline{y}_m - \underline{x}_m)/\|\underline{x}_m\| = O\,(\underline{r}_n/\|\underline{x}_n\|)^2 \qquad \text{and}$$

$$(\underline{s}_m - \underline{r}_m)/\|\underline{x}_m\| = O\,(\underline{r}_n/\|\underline{x}_n\|)^2$$

provided $m - n$ is bounded as $n \to \infty$. We shall be interested mainly in $m = n + M$. (M was introduced near the beginning of section 3.)

What is the special significance of setting $\mu_m = \lambda_\infty$? When this is done, the sequence $\underline{y}_m$ for $m > n$ coincides with what one would get if he used a relaxation iteration to solve the linear system

$$(A - \lambda_\infty B)\, y = 0$$

for a non-zero vector $\underline{y}$, starting the iteration from $\underline{y}_n = \underline{x}_n$. Something is known about the convergence of that iteration.

Consider first the stable case $\lambda_\infty = \lambda_o$, with $\mu_m = \lambda_o$ for all $m > n$. Now $(A - \lambda_o B)$ is a positive semidefinite matrix, for which

the theory of the relaxation iteration has already been developed in the author's previous report (1966). There it was shown that there exists a constant $K < 1$, dependent only upon $(A - \lambda_o B)$, M and the constant d, such that

$$|||\underline{s}_{n+M}||| < K|||\underline{s}_n||| \quad ,$$

where $|||=a|||$ represents a certain vector norm whose definition now requires a small digression.

Let $W = (\underline{w}_1, \underline{w}_2, \ldots, \underline{w}_N)$ be the orthogonal matrix $(W^TW = I)$ which diagonalizes $A - \lambda_o B$; say

$$A - \lambda_o B = W \operatorname{diag}(\theta_i) W^T = \sum_i \theta_i \underline{w}_i \underline{w}_i^T \quad .$$

Since $A - \lambda_o B$ is positive semi-definite, all $\theta_i \geq 0$. Then for any vector v define

$$|||\underline{v}|||^2 = \underline{v}^T \Big(\sum_{\theta_i} \theta_i^{-1} \underline{w}_i \underline{w}_i^T + \sum_{\theta_i = 0} \underline{w}_i \underline{w}_i^T \Big) \underline{v} \quad .$$

In the special case where s is a residual vector

$$s = (A - \lambda_o B)\underline{y} = \sum \theta_i \underline{w}_i \underline{w}_i^T \underline{y} \quad ,$$

$$|||\underline{s}|||^2 = \underline{y}^T (A - \lambda_o B)\underline{y} \quad ,$$

but this formula is of no use when $\underline{s}$ is a general vector.

The next step is from $|||\underline{s}_{n+M}||| \leq K|||\underline{s}_n|||$ to

$$|||\underline{r}_{n+M}||| \leq K_1 |||\underline{r}_n||| \quad ,$$

where $K_1 < 1$ and K_1 is independent of n if n is sufficiently large. The step is valid because

$$\begin{aligned}
\|\underline{r}_{n+M}\| \, / \, \|\underline{x}_{n+M}\| &\le \|\underline{s}_{n+M}\| \, / \, \|\underline{x}_{n+M}\| + O(\underline{r}_n/\|\underline{x}_n\|)^2 \\
&\le K\|\underline{s}_n\| \, / \|\underline{x}_{n+M}\| + O(\underline{r}_n/\|\underline{x}_n\|)^2 \\
&\le K\|\underline{r}_n\| / \|\underline{x}_{n+M}\| + O(\underline{r}_n/\|\underline{x}_{n+M}\|)^2 .
\end{aligned}$$

(Recall that $\|\underline{x}_{n+M}\|/\|\underline{x}_n\| \to 1$ as $n \to \infty$.) It is necessary for n to be sufficiently large that the term $O(\underline{r}_n/\|\underline{x}_{n+M}\|)^2$ be small compared with $(1 - K)\|\underline{r}_n\| \, / \|\underline{x}_{n+M}\|$.

Thus do we see that $\underline{r}_n \to 0$ at least as quickly as the terms of some geometric progression with common ratio $K_1^{1/M} < 1$. And because

$$\begin{aligned}
\|\underline{x}_{n+1} - \underline{x}_n\| = |\xi_n| \cdot \|\underline{p}_n\| &\le 2|\eta_n| \, . \, \|\underline{p}_n\| \\
&\le 2\sqrt{\underline{r}_n^T \, B^{-1} \, \underline{r}_n} \, /(\pi_n - \lambda_n) \\
&\to 0 \quad \text{at least as quickly as} \quad K_1^{n/M} ,
\end{aligned}$$

the sequence $\underline{x}_n$ must converge to its limit $\underline{x}_o$ at least as quickly as some geometric series with common ratio $K_1^{1/M} < 1$.

This is just the result we want, especially since it shows that there is no need to compute the normalized sequence $\overset{n}{\underline{x}}_n = \underline{x}_n/\|\underline{x}_n\|$. Furthermore, the result is valid whether or not λ_o is a repeated eigenvalue with several linearly independent eigenvectors. But the method of proof conceals the closeness with which K_1 can approach 1 when λ_o

is merely the smallest of a cluster of nearly equal eigenvalues, in which case n may have to be exceedingly large before λ_n is closer to λ_o than to the next largest eigenvalue. Besides, when several eigenvalues are clustered near λ_o there is an enhanced risk that the limit of the sequence λ_n will be an eigenvalue $\lambda_\infty > \lambda_o$.

Consider now the unstable case $\lambda_\infty > \lambda_o$. The linear equation

$$(A - \lambda_\infty B)\underline{y} = 0$$

now has an indefinite matrix $(A - \lambda_\infty B)$, so the relaxation iteration is almost certain to produce a sequence of vectors $\underline{y}_m$ which diverge exponentially in such a way that

$$\underline{y}_m^*(A - \lambda_\infty B)\underline{y}_m \to -\infty \quad \text{exponentially as} \quad m \to \infty .$$

Some justification for claiming that divergence is almost certain can be found in section 4 of the author's previous report (1966). There it was shown that if the sequences $\underline{p}_m$ and τ_m/σ_m are chosen in advance of any knowledge about $\underline{y}_m$, then the sequence $\underline{y}_m$ must diverge exponentially unless the initial vector $\underline{y}_n$ is placed into a certain hyperplane $\mathcal{H}_n$ which depends upon $(A - \lambda_\infty B)$ and the sequences $\underline{p}_m$ and (τ_m/σ_m) in a practically undecipherable way. And since $(\underline{y}_m - \underline{x}_m)/\|\underline{x}_m\| = O(\underline{r}_n/\|\underline{x}_n\|)^2$, there is good reason to conclude that λ_m is most unlikely to converge to a limit $\lambda_\infty > \lambda_o$ unless the sequences $\underline{p}_m$ and ξ_m are correlated with $\underline{x}_m$ in some way designed to achieve what would otherwise be a rare event.

5.) Variations

At the cost of minor modification, the foregoing analysis can be extended to cope with two variations of the relaxation iteration which will be mentioned here.

One variation is 'block relaxation", in which each step $\underline{x}_{n+1} - \underline{x}_n = \xi_n \underline{p}_n$ is a suitable linear combination of some specified subset of the basis vectors $\underline{e}_j$. If the subset contains L vectors,, then the vector $\underline{x}_n + \zeta_n \underline{p}_n$ which minimizes $\Lambda(\underline{x}_n + \xi_n \underline{p}_n)$ is obtained by solving an $(L + 1) \times (L + 1)$ eigenproblem. Subject to this complication, the results in the foregoing sections can be applied to block relaxation-with no important changes. Regrettably, the techniques used in this report do not indicate when block relaxation is more efficient than the simpler iteration.

Another variation seems to have been motivated by the fear that $\underline{x}_n$ might diverge to infinity or converge to zero even though the normalized sequence $\underline{x}_n^n$ converges to the desired eigenvector. The discussion in the previous section of this report should put such a fear to rest, but in the absence of that discussion the natural thing to do is find some simple way to normalize the sequence $\underline{x}_n$. The simplest way is to fix some component, say the N^{th} , of all vectors $\underline{x}_n$ at some constant value, say 1 . This normalization is maintained by restricting each member of the sequence $\underline{p}_n$ to the subset

$$\{\underline{e}_1 , \underline{e}_2 , \ldots , \underline{e}_{N-1}\} ,$$

of which each element should appear at least once in each set of M

consecutive vectors

$$\underline{p}_n \, , \, \underline{p}_{n+1} \, , \, \underline{p}_{n+2} \, , \cdots , \, \underline{p}_{n+M-1} \, .$$

The analysis of sections 3 and 4 now requires some small modification to yield results which are outlined below.

Let A' and B' be obtained from A and B by deleting their respective last rows and columns, and let

$$\Lambda'(\underline{v}') \equiv \underline{v}'^{T} \underline{A}' \underline{v}' / \underline{v}'^{T} \underline{B}' \underline{v}' \text{ and}$$

$$\lambda_o' = \min \Lambda'(\underline{v}') \quad \text{over} \quad \underline{v}' \neq 0 \, .$$

There is some risk that the restricted relaxation iteration will converge to $\lambda_\infty > \lambda_o'$, but that risk is as negligible as before. The most likely event is that $\lambda_\infty = \lambda_o$ as before, and that $\underline{x}_n \to \underline{x}_o$ with $\underline{x}_o$ normalized by the condition $\underline{e}_N^T \underline{x}_o = 1$. There is also a non-negligible risk that the iteration may converge to $\lambda_\infty = \lambda_o' > \lambda_o$, in which case $\underline{x}_n$ will diverge to infinity although $\hat{\underline{x}}_n = \underline{x}_n / \|\underline{x}_n\|$ will converge to a vector $\hat{\underline{x}}_\infty$ such that $\underline{e}_N^T \hat{\underline{x}}_\infty = 0$. The first N - 1 components of $\hat{\underline{x}}_\infty$ will provide an eigenvector $\underline{v}' \neq 0$ satisfying

$$(A' - \lambda_\infty B')\underline{v}' = \underline{0} \, .$$

Example 1 of section 2 illustrates this possibility. The possibility that $\lambda_\infty = \lambda_o' > \lambda_o$ can easily be detected by performing a final relaxation $\underline{x}_{\infty+1} = \hat{\underline{x}}_\infty + \xi_\infty \underline{e}_N$, since $\lambda_o \leq \Lambda(\underline{x}_{\infty+1}) < \Lambda(\hat{\underline{x}}_\infty) = \lambda_\infty$ unless $\lambda_\infty = \lambda_o$. Therefore, the restricted and unrestricted relaxation

iterations can differ in only one important respect; one iterative method may converge faster than the other. The author's limited experience with both methods indicates that the unrest&ted iteration should normally be preferred despite the existence of rare examples (like that in section 3 of his earlier report (1966)) where the restricted iteration seems to be faster.

6.) Final Remarks

So far, the relaxation iteration for solving

$$(A - \lambda_o B)\underline{x}_o = 0$$

has been discussed without reference to rounding errors. Their most noticeable effect will appear in the sequences λ_n and $\|\underline{x}_n\|^2$ whose computed values are obtained indirectly during the iteration defined in section 2. Roundoff will prevent the computed values

$$\lambda_{n+1} = \lambda_n + \Delta\lambda_n \quad \text{(rounded)} \quad \text{and}$$

$$\|\underline{x}_{n+1}\|^2 = \|\underline{x}_n\|^2 + 2\,\xi_n\,\underline{p}_n^T B\underline{x}_n + \xi_n^2\,\|\underline{p}_n\|^2 \quad \text{(rounded)}$$

from precisely satisfying the equations

$$\lambda_{n+1} = \Lambda(\underline{x}_{n+1}) \quad \text{and}$$

$$\|\underline{x}_{n+1}\|^2 = \underline{x}_{n+1}^T B\underline{x}_{n+1}\,.$$

The remedy is simply to recompute the values λ_n and $\|\underline{x}_n\|^2$ directly

from their definitions once or twice during the course of the iteration. For example, one good time to recompute λ_n and $\|\underline{x}_n\|^2$ is when $(\lambda_n - \lambda_{n+1})$ has first remained no larger than about 10n units in the last place of λ_{n+1} for M consecutive values of n . Another good time to recompute λ_n and $\underline{r}_n = (A - \lambda_n B)\underline{x}_n$ is just before accepting $\underline{x}_n$ as an adequate approximation to $\underline{x}_o$; the smallness of $\underline{r}_n$ is a useful indication of the accuracy of $\underline{x}_n$ provided one has some information about the separation between λ_o and the next larger stationary value of A, and a bound for the size of B^{-1} . (See Appendix II, part (iii).)

Ultimately the iteration index n will become large enough that $(\lambda_n - \lambda_{n+1})$ is negligible although $\|\underline{x}_{n+1} - \underline{x}_n\|$ is not negligible yet. This occurs because

$$\lambda_n - \lambda_o = O\,(\underline{r}_n / \|\underline{x}_n\|)^2 \qquad \text{while}$$

$$\underline{x}_n - \underline{x}_o = O\,(\underline{r}_n)$$

as is shown in Appendix II. For all subsequent values of n it may be worthwhile to skip those parts of the calculation which up-date $\|\underline{x}_{n+1}\|^2$ and λ_{n+1} . The time saved is noticeable when A and B are such sparse matrices that the computation of

$$\underline{p}_n^T \underline{r}_n = (\underline{p}_n^T (A - \lambda_\infty B))\underline{x}_n$$

consumes fewer than a dozen arithmetic operations. λ_n can be held constant *for* several iterations, and recomputed from the definition $\lambda_n = \Lambda(\underline{x}_n)$ sufficiently infrequently that the time spent upon

recomputation is a small fraction of the time saved by not updating λ_n at each iteration. The conclusions in section 4 remain valid however λ_n may be defined provided that

$$|\lambda_n - \lambda_o| = O\left(\underline{r}_n / \|\underline{x}_n\|\right)^2 .$$

But the definition of λ_n given in section 2 (based upon the scheme used by Nesbet (1965)) is the nicest that the author has seen.

Appendix 1

Here is an outline of the proof that when $\xi = \zeta$,

$$\Delta\lambda \leq -(\underline{p}^T\underline{r})^2/(3s\ \|\underline{p}\|^2\ \|\underline{x}\|^2)$$

where $s = \max(\Lambda(u) - A(v))$ over $\underline{u} \neq 0$ and $\underline{v} \neq 0$.

For the sake of simplicity, and without loss of generality, it is assumed that

$$\|\underline{p}\| = \|\underline{x}\| = 1 \quad \text{and} \quad \pi \geq \lambda = 0 \ .$$

Now we abbreviate;

$$\alpha = \underline{p}^T\underline{r} \quad \text{and} \quad \beta = (\underline{p}\ ,\ \underline{x}) = \underline{p}^T B\underline{x} \ .$$

The numbers α , β and π can be bounded;

$$\beta^2 = (\underline{p}\ ,\ \underline{x})^2 \leq \|\underline{p}\|^2\ \|\underline{x}\|^2 = 1 \ .$$

$$\alpha^2 = (\underline{p}^T\underline{r})^2 \leq (\underline{p}^T B\underline{p})(\underline{r}^T B^{-1}\underline{r}) \text{ by the Schwartz inequality,}$$

$$= \underline{x}^T B^{1/2}\ (B^{-1/2} A\ B^{-1/2})\ B^{1/2}\underline{x} / \underline{x}^T B\underline{x} \leq s^2 \ .$$

$$\pi = \Lambda(\underline{p}) \leq s \ .$$

Now let us express $\Delta\lambda$ as a function of ξ :

$$\Delta\lambda = \pi\,\xi(\xi + 2\alpha/\pi)/(1 + 2\,\beta\xi + \xi^2) \ .$$

This is minimized when $\xi = \zeta$, at which point $d\Delta\lambda/d\xi = 0$. Therefore,

$$\Delta\lambda_{min} = (\pi\zeta + \alpha)/(\zeta + \beta) \quad ,$$

which may be combined with the former equation to yield

$$\Delta\lambda_{min} = \alpha/(\beta + 1/\zeta) \qquad \text{and}$$

$$(\pi\beta - \alpha)\zeta^2 + \pi\zeta + \alpha = 0 \quad .$$

Solving the last equation and substituting into the former yields

$$\zeta = -2\alpha/(\pi + \sqrt{\pi^2 - 4\beta\alpha\pi + 4\alpha^2}) \qquad \text{and}$$

$$\Delta\lambda_{min} = -2\alpha^2/(\pi - 2\beta\alpha + \sqrt{\pi^2 - 4\beta\alpha\pi + 4\alpha^2}) \quad .$$

An application of the bounds

$$\pi \leq s \quad , \quad -\alpha\beta < s \quad \text{and} \quad \alpha^2 < s^2$$

yields

$$\Delta\lambda_{min} \leq -\alpha^2/(3s) \quad ,$$

as was claimed.

Incidentally, if ξ/ζ is held constarit as $\zeta \to 0$, $\Delta\lambda/\Delta\lambda_{min} \to 1 - (\xi/\zeta - 1)^2$; in general, $\Delta\lambda$ approximates $\Delta\lambda_{min}$ with a relative error that is smaller than that with which ξ approximates ζ.

Appendix II

For the sake of completeness, here is a short proof of the classical result that as $\underline{r}(\underline{x})/\|\underline{x}\| \to 0$ and $\Lambda(\underline{x}) \to \lambda_\infty$, a stationary value of $\Lambda(\underline{x})$,

$$\Lambda(\underline{x}) - \lambda_\infty = 0\,(\underline{r}(\underline{x})/\|\underline{x}\|)^2 \quad .$$

To be more precise, given real symmetric matrices A and B with B positive definite, and any vector $\underline{x}$, we define

$$\lambda = \Lambda(\underline{x}) \equiv \underline{x}^T A\underline{x}/\underline{x}^T B\underline{x} \quad ,$$

$$\underline{r} = \underline{r}(\underline{x}) \equiv (A - \lambda B)\underline{x} \quad ,$$

$$e = \sqrt{\underline{r}^T B^{-1}\underline{r}/\underline{x}^T B\underline{x}} \quad , \quad \text{and}$$

$$\lambda_\infty = \Lambda(\underline{x}_\infty) \text{ for some eigenvector } \underline{x}_\infty \neq 0$$
$$\text{such that } (A - \lambda_\infty B)\underline{x}_\infty = 0 \quad ;$$

and we prove that

(i) The functional A(x) has at least one stationary value λ_∞ between $\lambda - e$ and $\lambda + e$ inclusive.

(ii) If λ_∞ is the only stationary value in the aforementioned closed interval $[\lambda-e\ ,\ \lambda+e]$, and α_i are the other stationary values, then λ_∞ lies in the smaller interval

$$[\lambda - e^2/\min_{\alpha_i > \lambda_\infty}(\alpha_i - \lambda)\ ,\quad \lambda + e^2/\min_{\alpha_i < \lambda_\infty}(\lambda - \alpha_i)]\ ,$$

provided that, for example, when all $\alpha_i > \lambda_\infty$ then

$$e^2 / \min_{\alpha_i < \lambda_\infty} (\lambda - \alpha_i) = 0 \text{ by definition .}$$

(iii) If for $\delta > e$ we define $\mathcal{L}_\delta$ to be the subspace spanned by all the eigenvectors corresponding to eigenvalues (stationary values of A) in the interval $[\lambda-\delta , \lambda+\delta]$, and if θ is the angle between x and $\mathcal{L}_\delta$, then

$$\sin \theta \leq e / \min_{|\lambda - \alpha_i| > \delta} |\lambda - \alpha_i| \ .$$

(The angle θ is defined by

$$\cos^2 \theta = \max_{\substack{\underline{v} \in \mathcal{L}_\delta \\ \underline{v} \neq 0}} (\underline{x}^T B \underline{v})^2 / (\underline{x}^T B \underline{x} \ \underline{v}^T B \underline{v}) \ ,$$

and can be viewed as the smallest non-negative angle between $\underline{x}$ and a vector $\underline{v}$ in the hyperplane $\mathcal{L}_\delta$. When the interval $[\lambda-\delta , \lambda+\delta]$ contains only one eigenvalue, $\sin \theta$ is a measure of the error with which x approximates a corresponding eigenvector, even when the eigen- , value is repeated.)

The proof is essentially due to Kato (1949) with a few modifications. In particular, the results obtained here are valid even if λ_∞ is a "degenerate eigenvalue" whose eigenvectors span a subspace of dimension greater than 1 . Also, result (iii) is simpler than Kato's.

First, the positive definite symmetric matrix $B^{1/2}$ can be used to replace B by I ,

$$A \text{ by } A' = B^{-1/2} A B^{-1/2} ,$$

$$x \text{ by } x' = B^{1/2} x, \quad \text{and}$$

$$r \text{ by } r' = B^{-1/2} \underline{r} ,$$

with no other changes. Therefore, there is no loss in generality if B is assumed to be I at the outset.

Now suppose the interval $[\mu , \nu]$ contains no eigenvalue α of A. This means that $(\alpha - \mu)(\alpha - \nu) > 0$ if α is an eigenvalue of A, so $(A - \mu I)(A - \nu I)$ is positive definite. Therefore

$$\underline{x}^T (A - \mu I)(A - \nu I)x > 0 \qquad \text{or}$$

$$e^2 + (\lambda - \mu)(\lambda - \nu) > 0 .$$

Conversely, if $e^2 + (\lambda - \mu)(\lambda - \nu) = 0$, then $[\mu , \nu]$ contains at least one eigenvalue. The values $\mu = \lambda - e$ and $\nu = \lambda + e$ satisfy this equation and prove (i).

If the closed interval $[\lambda - e , \lambda + e]$ contains only one eigenvalue λ_∞, though λ_∞ may be a repeated eigenvalue, let α_i be the other eigenvalues and let

$$\mu \to \max_{\alpha_i < \lambda_\infty} \alpha_i + , \text{ or } -\infty \text{ if all } \alpha_i > \lambda_\infty .$$

As μ decreases to its limit, it becomes less than $\lambda - e$. Now set

$$\nu = \lambda + e^2/(\lambda - \mu) \searrow \lambda + e^2 / \min_{\alpha_i < \lambda_\infty} (\lambda - \alpha_i)$$

For the same reason as before, we conclude that $[\mu , \nu]$ contains at least one eigenvalue. But because

$$\max_{\alpha_i < \lambda_\infty} \alpha_i < \mu < \lambda - e < \lambda < \nu < \lambda + e$$

if μ is close enough to its limiting value, there can be only one eigenvalue in $[\mu , \nu]$, and that is ${}_\infty\lambda$. This proves part of (ii),, and a similar limiting process in reverse proves the rest of (ii).

Finally, write $\underline{x} = \underline{w} + \underline{u}$ where $w \in \mathcal{L}_\delta$ and $\underline{u} \in \mathcal{L}_\delta^\perp$. ($\mathcal{L}_\delta^\perp$ is the orthogonal complement of $\mathcal{L}_\delta$.) Since $\mathcal{L}_\delta$ and $\mathcal{L}_\delta^\perp$ are both invariant subspaces of A , in the sense that $A\mathcal{L}_\delta \subseteq \mathcal{L}_\delta$ and $A\mathcal{L}_\delta^\perp \subseteq \mathcal{L}_\delta^\perp$,

$$r = (A - \lambda I)\underline{x} = (A - \lambda I)\underline{w} + (A - \lambda I)\underline{u}$$

where $(A - \lambda I)w \in \mathcal{L}_\delta$ and $(A - \lambda I)u \in \mathcal{L}_\delta^\perp$. Therefore,

$$e^2 = [\underline{w}^T(A - \lambda I)^2\underline{w} + \underline{u}^T(A - \lambda I)^2\underline{u}]/\underline{x}^T\underline{x}$$

$$\geq [\underline{u}^T(A - \lambda I)^2\underline{u}/\underline{u}^T\underline{u}][\underline{u}^T\underline{u}/\underline{x}^T\underline{x}] .$$

Now, $\underline{u}^T\underline{u}/\underline{x}^T\underline{x} = \sin^2 \theta$; and

$$\underline{u}^T(A - \lambda I)^2\underline{u}/\underline{u}^T\underline{u} \geq \min_{|\alpha_i - \lambda| > \delta} (\alpha_i - \lambda)^2$$

because the restriction of A to $\mathcal{L}_\delta^\perp$ has no eigenvalues in the interval $[\lambda-\delta , \lambda+\delta]$. This proves (iii).

It is possible to show with examples that each of the bounds

implied by (i), (ii) and (iii) can be achieved, though not necessarily simultaneously. These results provide satisfactory error bounds except when the separation between adjacent eigenvalues is not much larger than the residual e , for which case see section 3 of Kato's paper (1949).

References

W. Kahan (1966) "Relaxation Methods for Semi-Definite Systems" Computer Science Department Technical Report CS45 August 9, 1966 - Stanford University, Stanford, California.

T. Kato (1949) "On the Upper and Lower Bounds of Eigenvalues" J'l Physical Soc. of Japan 4 pp. 334-y.

R. K. Nesbet (1965) "Algorithm for Diagonalization of Large Matrices" J'l of Chemical Physics 43, pp. 311-2.

A. Ostrowski (1965) "Contributions to the Method of Steepest Descent" (To appear).

F. S. Shaw (1953) "An Introduction to Relaxation Methods" Reprinted by Dover, New York.