Simultaneous Rayleigh-Quotient Minimization Methods for $Ax=\lambda Bx$

D. E. Longsine and S. F. McCormick*
Department of Mathematics
Colorado State University
Fort Collins, Colorado 80523

Submitted by Robert J. Plemmons

ABSTRACT

New simultaneous iteration techniques are developed for solving the generalized eigenproblem $Ax=\lambda Bx$, where A and B are real symmetric matrices and B is positive definite. The approach is to minimize the generalized Rayleigh quotient in some sense over several independent vectors simultaneously. In particular, each new vector iterate is formed from a linear combination of current iterates and correction vectors that are derived from either gradient or conjugate-gradient techniques. A Ritz projection or simultaneous iteration process is used to accelerate convergence. For one of the gradient versions, convergence and asymptotic rates of convergence are established. Also, some numerical experiments are reported that demonstrate the convergence behavior of these methods.

I. INTRODUCTION

This paper is concerned with the development of a gradient-type numerical method for solving problems of the form

$$Ax=\lambda Bx, \tag{1}$$

where A and B are real, symmetric matrices and B is positive definite. We are treating the case for which it is undesirable to factor either A or B, or a linear combination of A and B, into a product of matrices that lead to the direct solution of linear equations in these matrices. More precisely, we assume that A and B are large and sparse and that the direct solution of equations involving these matrices is either impossible or impractical. We also assume that the problem is to compute one or a few of the extreme eigenvalues and their eigenvectors. Problems of this type arise often in structural analysis and reactor problems, for example (cf. [1], [2]).

These assumptions prevent the *direct* use of many standard techniques such as the Lanczos [3] and power [4] methods. "Factorization-free"

*This work was supported by Air Force grant AFOSR76-3019 and National Science Foundation grant MCS78-03847.

52 Vanderbilt Ave., New York, NY 10017 0024-3795/80/060195+40$01.75

techniques that do apply include algorithms based on optimizing the Rayleigh quotient by such methods as coordinate relaxation [5], gradient [6], and conjugate gradient [7]. These methods apply to the computation of extreme eigenvalues only, although others can be computed with the aid of deflation, matrix transformations, or projection techniques. Of course, care must be taken to avoid destroying the sparsity of A and B. However, as we shall see, simultaneous iteration (cf. [8]) can be used in conjunction with gradient-type methods to compute several extreme eigenvalues and their eigenvectors. This approach, which is made somewhat like that for the power method, is the subject of this paper.

There are, of course, other factorization-free techniques for solving (1), including the outer-loop power, inverse iteration, or Lanczos methods used in combination with some inner-loop iteration scheme such as conjugate gradients. These combined techniques and the approaches of this paper are, in fact, closely related. For example, the single vector version of SIRQIT-G described below is essentially equivalent in form to a combined Rayleigh quotient iteration and gradient method using only one inner loop iteration per outer loop step. A promising combined algorithm is Lanczos/conjugate gradients which could be formulated in either the direct sense to compute the extreme eigenvalues or the inverse form for the interior one. We are presently comparing the technique with SIRQIT-CG, which is tested below.

The (conjugate) gradient methods for Rayleigh-quotient minimization are amenable to simultaneous acceleration as it was first proposed with the power method [8]. (The slightly more efficient approach developed by Rutishauser [9] does not apply, however, since we cannot in general assume that the iterates are derived from the image of a stationary linear operator. For a brief discussion of these two projection techniques, see [10].) The gradient method extended in this way is described in the next section, as is its relationship with a simultaneous coordinate overrelaxation scheme due to Schwarz [11]. In Sec. III, global convergence is established for $B=I$ and local convergence for general B. Rates are developed in Sec. IV. The conjugate-gradient version is described briefly in Sec. V, and numerical examples provided in Sec. VI.

II. SIMULTANEOUS GRADIENT METHODS

We introduce the following notation:

n: the dimension of A and B;

p: the number of vectors used simultaneously to compute p approximate solutions of (1);

Λ: $\operatorname{diag}(\lambda_1,\dots,\lambda_n)$, where $\lambda_1 \leqslant \cdots \leqslant \lambda_n$ are the eigenvalues of (1);
U: the matrix having the corresponding B-orthonormal eigenvectors $u_1,\dots,u_n$ as its columns, so that $U^TAU=\Lambda$ and $U^TBU=I$.

We partition Λ and U into

$$\Lambda=\begin{pmatrix}\Lambda_1 & 0\\ 0 & \Lambda_2\end{pmatrix} \quad \text{and} \quad U=\begin{pmatrix}U_1 & U_2\end{pmatrix},$$

respectively, where Λ_1 is the leading $p\times p$ minor of Λ, and U_1 is the $n\times p$ matrix consisting of the first p columns of U. Let (x, y) denote the Euclidean inner product of x and y, and let $(x, y)_B=(Bx, y)$ denote the inner product in the metric B.

The *s*imultaneous *R*ayleigh-*q*uotient *it*erative minimization methods (SIRQIT), which we now describe, represent a class of methods for approximating Λ_1 and U_1. (Of course, the largest eigenvalues can be computed by replacing A with $-A$, that is, by maximizing the Rayleigh quotient.) SIRQIT is based on certain properties of the Rayleigh quotient

$$R(x)=\frac{x^TAx}{x^TBx}, \qquad x\in\mathbb{R}^n, \quad x\neq 0 \tag{2}$$

and its gradient

$$\nabla R(x)=\frac{2}{x^TBx}\left[Ax-R(x)Bx\right].$$

For simplicity, since only the direction of $\nabla R(X)$ is important, we use

$$g(x)=Ax-R(x)Bx. \tag{3}$$

The extreme values of $R(x)$ are λ_1 and λ_n occurring at scalar multiples of u_1 and u_n, while $u_2,\dots,u_{n-1}$ are saddle points of $R(x)$ with values $\lambda_2,\dots,\lambda_{n-1}$. The critical points of $R(x)$ are precisely the eigenvectors of (1).

SIRQIT attempts to minimize $R(z_i)$ for each column, z_i, of an $n\times p$ matrix Z. This is obtained from the previous iterate Y, with columns $y_1,\dots,y_p$, and from $G(Y)$, with columns $g(y_1),\dots,g(y_p)$. A $p\times p$ matrix S is determined from a class of "permissible step matrices" so that the columns of $Y-G(Y)S$ in some sense represent the best set of solutions of (1) from $\operatorname{span}(Y, G(Y))$, the column span of $(Y, G(Y))$. (See Remark 2 below.) For example, the class of permissible step matrices might be restricted to the set of diagonal

matrices, with the effect that columns of $Y-G(Y)S$ are determined independently (and ignorantly) of each other. Most generally, the class might include all real $p\times p$ matrices, with the effect that the columns of $Y-G(Y)S$ approximately represent the best set of solutions of (1) in the column space of $(Y, G(Y))$. In SIRQIT we have chosen what amounts to a compromise between these two extremes. Specifically, we first compute $\hat{G}(y_i)=\hat{g}_i$ in turn by B-orthogonalizing $g(y_i)$ from $z_1,\dots, z_{i-1}$, and then form $Y-\hat{G}\hat{S}$ by restricting $\hat{S}$ to be a diagonal $p\times p$ matrix. Thus, iteration on the successive columns of Y can make intelligent use of information contained in the previous columns of Y and $G(Y)$. This has the net effect that the permissible step matrices applied to $G(Y)$ are just the upper triangular matrices. A general outline of SIRQIT is as follows:

(i) Let the tolerance $\varepsilon>0$ and initial $n\times p$ matrix $X^{(0)}=(x_1^{(0)},\dots, x_p^{(0)})$ be given such that $X^{(0)T}BX^{(0)}=I$. Let $i=0$.

(ii) Let $Y^{(i)}$ be the solution of the problem that results from orthogonally projecting (1) onto span$(X^{(i)})$, the span of the columns of $X^{(i)}$. (See Remark 1 below.) Note that $Y^{(i)T}AY^{(i)}=D^{(i)}$, a diagonal $p\times p$ matrix, and $Y^{(i)T}BY^{(i)}=I_p$.

(iii) Set $G(Y^{(i)})=AY^{(i)}-BY^{(i)}D^{(i)}$. Check for convergence by testing the condition

$$\left\|g\left(y_i^{(i)}\right)\right\|<\varepsilon,$$

where we have written $G(Y^{(i)})=(g(y_1^{(i)}),\dots, g(y_p^{(i)}))$.

(iv) Set $Z^{(i)}=Y^{(i)}-G(Y^{(i)})S^{(i)}$, where $S^{(i)}$ is a $p\times p$ matrix chosen to improve the approximation to U_1 from span$(Y^{(i)}, G(Y^{(i)}))$. (See Remark 2 below.)

(v) Set $X^{(i+1)}=Z^{(i)}M^{(i)}$, where $M^{(i)}=M^{(i)}_{p\times p}$ is an upper triangular matrix constructed so that $X^{(i+1)T}BX^{(i+1)}=I_p$. (For example, $X^{(i+1)}$ might be constructed by a Gram-Schmidt process in the metric B.) Increment i by 1 and go to (ii).

REMARK 1. The orthogonal projection step in (ii) is an attempt to solve (1) projected onto span$(X^{(i)})$, which is equivalent to finding the classical eigenvalues and eigenvectors of the matrix $\mathcal{A}_1=X^{(i)T}AX^{(i)}$, where $X^{(i)T}BX^{(i)}=I$. This can be accomplished (cf. [1], [10], or [11]) by first finding a unitary matrix Q such that $Q^T\mathcal{A}_1Q=D=\operatorname{diag}(d_1,\dots, d_p)$, where $d_1\leqslant d_2\leqslant\cdots\leqslant d_p$, and then setting $Y^{(i)}=X^{(i)}Q$. Note that $Y^{(i)}BY^{(i)}=I_p$. The computation of Q can be accomplished by routines such as TRED2 and IMTQL2 in IMSL [12] or EISPACK [13].

REMARK 2. As we have indicated, the matrix S in step (iv) can be determined in a variety of ways. The following approach seems to be among the best in terms of numerical behavior. Specifically, a little reflection suggests many possible variations, depending on the process for B-orthogonalization, choice of step size, and Ritz step implementation. Many of these variations are not well founded, and others we tried did not perform as well in practice. (See, however, the discussion of SIRQIT-G2 below.) The following approach also facilitates the theoretical development of the convergence rates presented in Sec. IV. Dropping the subscripts and superscripts for convenience, S is chosen as a diagonal matrix where each diagonal entry, s_k, is found by minimizing $R(y_k - s\hat{g}_k)$ over s. Then the resulting z_k is made to be B-orthonormal to $z_1, \ldots, z_{k-1}$. Here, $\hat{g}_k$ is the result of B-orthogonalizing $g(y_k)$ from $z_1, \ldots, z_{k-1}$, namely,

$$\hat{g}_k = \begin{cases} g(y_1) & \text{for } k=1, \\ g(y_k) - \sum_{l=1}^{k-1} (g(y_l), z_l)_B z_l & \text{for } k>1, \end{cases}$$

which can be accomplished by the Gram-Schmidt method, for example. To derive a closed form expression for s, we set equal to zero the derivative of $R(y_k - s\hat{g}_k)$ with respect to s, which yields

$$as^2 + bs + c = 0, \tag{4}$$

where

$$a = (\hat{g}, \hat{g})_B (Ay, \hat{g}) - (\hat{g}, y)_B (A\hat{g}, \hat{g}), \tag{5}$$

$$b = (y, y)_B (A\hat{g}, \hat{g}) - (\hat{g}, \hat{g})_B (Ay, y), \tag{6}$$

$$c = (\hat{g}, y)_B (Ay, y) - (y, y)_B (Ay, \hat{g}). \tag{7}$$

These simplify be recalling that $y^T By = 1$ and $y^T Ay = d$. Solving (4), we have

$$s_{\pm} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}, \tag{8}$$

or equivalently,

$$s_{\pm} = \frac{-2c}{b \pm \sqrt{b^2 - 4ac}}. \tag{8'}$$

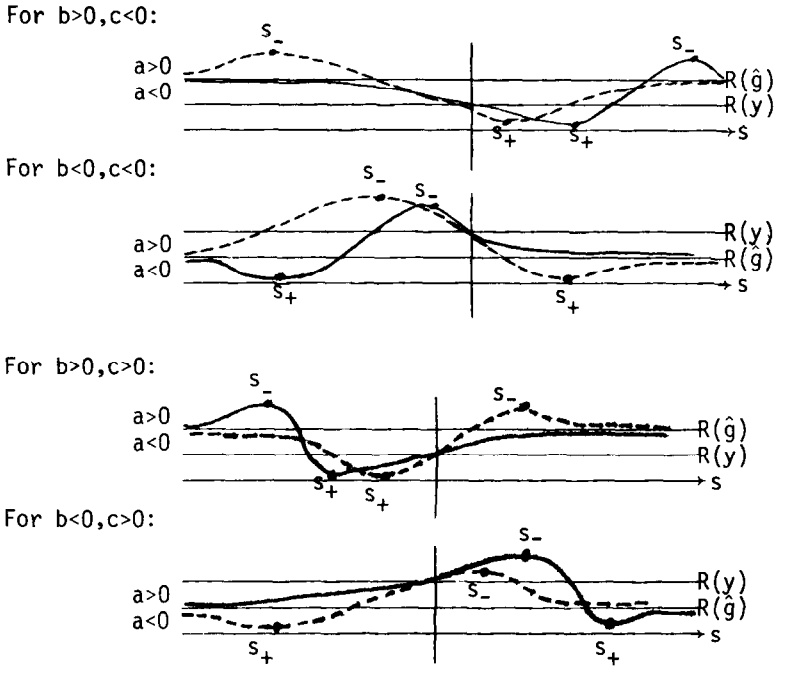

FIG. 1. The roles of s_+ and s_-.

Numerically, to avoid cancellation, it is best to use (8) when $b<0$ and (8′) when $b>0$, provided s_+ is used. To illustrate that s_+ (and not s_-) should be used to minimize $R(y-s\hat{g})$ over s, consider the expansion $R(y-s\hat{g})=R(y)-s\hat{g}^T g(y)+O(s^2)$. Noting that $c=-\hat{g}^T g(y)$ and $b=\hat{g}^T B\hat{g}(R(\hat{g})-R(y))$, a graphical display of $R(y-s\hat{g})$ as a function of s yields eight classes of functions as depicted in Fig. 1. This illustration, which can be made rigorous, shows clearly that the choice s_+ always yields the minimum. Note that when $B=I$, then $a>0$ and $c<0$, so that $\sqrt{b^2-4ac}>|b|$, which implies that $s_+>0$.

The complete SIRQIT process which uses (5) through (8) to compute S will henceforth be called SIRQIT-G.

An alternate and theoretically more powerful choice for S is achieved in effect by projecting (1) onto span$(Y, G(Y))$ and letting the columns of Z be the solutions of the projected problem that correspond to the p smallest projected eigenvalues. Specifically:

(a) B-orthonormalize $G(Y)$ from Y, calling the resulting matrix $\bar{G}$.
(b) Project (1) onto span$(Y, \bar{G})$ and solve in a manner similar to step (ii).

Let $\bar{Q}$ denote the eigenvectors of the projected $2p \times 2p$ matrix.

(c) Set X equal to the p columns of the matrix $(Y, \bar{G})\bar{Q}$ that correspond to the p smallest eigenvalues of the projected problem.

Note that $X^T B X = I_p$, so that step (v) in SIRQIT is not needed. For future reference, we shall refer to this version of SIRQIT that uses (a)–(c) to compute S as SIRQIT-G2.

There are a few minor numerical drawbacks of SIRQIT-G2. First, the dimension of the projected problem is twice that of SIRQIT-G, although the computation in (b) may be somewhat reduced by using the special properties of the matrix $\mathcal{Q}_2 = (Y, \bar{G})^T A (Y, \bar{G})$. Specifically, $\mathcal{Q}_2$ has the form

$$\mathcal{Q}_2 = \begin{bmatrix} D & \bar{G}^T G(Y) \\ G(Y)^T \bar{G} & \bar{G}^T A \bar{G} \end{bmatrix},$$

which is asymptotically equal to

$$\mathcal{Q}_2 = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \bar{G}^T A \bar{G} \end{pmatrix}.$$

Note that the columns of $\bar{G}$ are of unit length, while those of $G(Y)$ tend to zero as convergence occurs. Thus, the process of diagonalizing $\mathcal{Q}_2$ can be simplified somewhat as the iterates near the solution U_1. However, this is just the point at which SIRQIT-G may be used, since SIRQIT-G and SIRQIT-G2 exhibit similar asymptotic behavior. This was confirmed by the results of several numerical experiments. In practice, it may in fact be best to start with SIRQIT-G2 followed by the use of SIRQIT-G (with G replacing the use of $\hat{G}$; that is, with S as a diagonal matrix). However, the major purpose of this paper is to present a theoretical development of these methods in terms of convergence conditions and rates with numerical support for the results. Choice of these methods depends on the goals, application, and setting in which they are to be used. No concrete recommendations are made in this paper.

A second minor drawback of SIRQIT-G2 is the lack of a guarantee that the columns of $G(Y)$ are linearly independent. In fact, quite independently of convergence, strong numerical dependence occurs quite frequently in practice. Thus, it is important either to reorthogonalize $(Y, \bar{G})$ in the B-inner-product sense or to discard those columns of $G(Y)$ that are computationally dependent. The latter can be accomplished by testing the norm of each successively computed column of $\bar{G}$.

Schwarz [11] uses a simultaneous coordinate-relaxation approach analogous to our gradient-type SIRQIT-G2. The method, in essence, chooses successive groups of columns of the identity matrix (instead of columns of G) and performs a Ritz projection on the space spanned by these directions together with the present approximations. The classical version of coordinate relaxation is attractive in that each single component of the present approximation is in turn altered by an easily computed correction. Unfortunately, because of the added complexity of the Ritz step, the simultaneous version appears to lose some of this attraction.

III. CONVERGENCE RESULTS

In this section, we assume that $\lambda_1 < \lambda_2$ and $B = I$. (See Remark 6 below for the general case $B \neq I$.) We begin the convergence proofs with a series of lemmas, the first of which shows that $X^{(i)}$ is of full rank whenever $X^{(0)}$ is, $i = 1, 2, \ldots$. This assures that step (ii) is feasible for each iteration of SIRQIT-G. In Lemma 2 we show that if $X^{(0)}$ is not completely deficient in u_1, that is, if $u_1^T X^{(0)} \neq 0$, then neither is $X^{(i)}$, $i = 1, 2, \ldots$. Lemmas 3 and 4 together imply that the Rayleigh quotient of each column of $X^{(0)}, X^{(1)}, \ldots,$ form monotically decreasing sequences bounded below in turn by $\lambda_1, \lambda_2, \ldots, \lambda_p$ and, in effect, that they cannot converge unless the corresponding $g(y_k^{(i)})$ tend to zero. Some of the consequences of Lemmas 3 and 4 are outlined immediately thereafter in order to provide a basis for Theorem 1, which is our main result on convergence.

REMARK 3. Let $Z_{(k-1)} = (z_1, \ldots, z_{k-1})$, and define the projection operator

$$P_k = I - Z_{(k-1)} Z_{(k-1)}^T.$$

Then $\hat{g} = P_k g(y_k)$, $P_k \hat{g}_k = \hat{g}_k$, $P_k z_k = z_k$, and $P_k y_k = y_k$. The last two equalities follow from the fact that the $z_1, \ldots, z_{k-1}$ are linear combinations of $g(y_1), \ldots, g(y_{k-1})$ and $y_1, \ldots, y_{k-1}$. Hence, if $g(y_k)$ is used in place of $\hat{g}_k$ in SIRQIT-G, then the effect of producing z_k can be absorbed into the step size s_k. More precisely, $\text{span}(Z) = \text{span}(\hat{Z})$, where $\hat{Z} = Y - G(Y)S$. Asymptotically, there is no difference between the use of $g(y_k)$ and $\hat{g}_k$, since the differences between them tend to zero. Hence, the effect of choosing s_k based on either $g(y_k)$ or $\hat{g}_k$ is asymptotically the same. This will be useful in the following proofs.

LEMMA 1. *If $X^{(0)}$ is of full rank, then so is $X^{(i)}$ for each $i=1,2,\ldots$.*

Proof. We proceed by induction. The case $i=0$ is true by assumption. Assume that $X^{(i)}$ is of full rank p so that $X^{(i)T}X^{(i)}=I$. Since $Y^{(i)}=X^{(i)}Q$, we then have that $Y^{(i)T}Y^{(i)}=I$. Then, since $Y^{(i)T}G(Y^{(i)})=0$, it follows that

$$\hat{Z}^{(i)T}\hat{Z}^{(i)}=I+S^{(i)T}G(Y^{(i)})^{T}G(Y^{(i)})S^{(i)},$$

which differs from I by a nonnegative definite matrix. Thus, $\hat{Z}^{(i)T}\hat{Z}^{(i)}$ and, hence, $Z^{(i)T}Z^{(i)}$ and $X^{(i+1)T}X^{(i+1)}$ are nonsingular. This implies that $X^{(i+1)}$ is of full rank p, and the lemma is proved. ■

LEMMA 2. *Assume that $X^{(0)}$ is not completely deficient in u_1, namely, that $u_1^TX^{(0)}\neq 0$. Then none of the $X^{(i)}$ are completely deficient in u_1, $i=1,2,\ldots$.*

Proof. Proceeding by induction, the case $i=0$ is true by assumption. Assume that $X^{(i)}$ is not completely deficient in u_1. Clearly, then, $Y^{(i)}$ is not completely deficient in u_1, so we may let k be the smallest index in $\{1,2,\ldots,p\}$ for which $u_1^Ty_k^{(i)}\neq 0$. But, then,

$$\begin{aligned}\left|u_1^T\left[y_k^{(i)}-g\left(y_k^{(i)}\right)s_k\right]\right| &= \left|u_1^Ty_k^{(i)}\left[1+\left(d_k^{(i)}-\lambda_1\right)s_k\right]\right| \\ &\geqslant \left|u_1^Ty_k^{(i)}\right| \\ &> 0.\end{aligned}$$

Hence, by Remark 3, we have $u_1^Tz_j^{(i)}\neq 0$ for some j possibly different from k. Clearly, then, $Z^{(i)}$ and hence $X^{(i+1)}$ are not completely deficient in u_1. The lemma is proved. ■

LEMMA 3. *For each column $z_k^{(i)}$ of $Z^{(i)}$ and each column $y_k^{(i)}$ of $Y^{(i)}$, we have*

$$\begin{aligned}R\left(z_k^{(i)}\right)-R\left(y_k^{(i)}\right) &= -\left\|\hat{g}_k^{(i)}\right\|^2 s_k^{(i)} \\ &= -\left\|P_k^{(i)}g\left(y_k^{(i)}\right)\right\|^2 s_k^{(i)}.\end{aligned}$$

Proof. Dropping the superscripts (i) for convenience, note that

$$R(z_k)=\frac{R(y_k)-2s_k(A\hat{g}_k, y_k)+s_k^2(A\hat{g}_k, \hat{g}_k)}{1+s_k^2(\hat{g}_k, \hat{g}_k)}$$

Hence,

$$R(z_k)-R(y_k)=\frac{-2s_k(A\hat{g}_k, y_k)+s_k^2\left[(A\hat{g}_k, \hat{g}_k)-(\hat{g}_k, \hat{g}_k)R(y_k)\right]}{1+s_k^2(\hat{g}_k, \hat{g}_k)}$$

$$=\frac{s_k^2\left[R(\hat{g}_k)-R(y_k)\right]-2s_k}{1+s_k^2(\hat{g}_k, \hat{g}_k)}(\hat{g}_k, \hat{g}_k), \tag{9}$$

since

$$(A\hat{g}_k, y_k)=(\hat{g}_k, g(y_k))=(P_k\hat{g}_k, g(y_k))=(\hat{g}_k, \hat{g}_k).$$

Differentiating the quantity $R(z_k)-R(y_k)$ with respect to the parameter s_k and setting it equal to zero implies that

$$2\left[1+s_k^2(\hat{g}_k, \hat{g}_k)\right]\left\{s_k\left[R(\hat{g}_k)-R(y_k)\right]-1\right\}-2\left\{s_k^2\left[R(\hat{g}_k)-R(y_k)\right]\right.$$
$$\left.-2s_k\right\}s_k(\hat{g}_k, \hat{g}_k)=0.$$

This simplifies to

$$s_k^2(\hat{g}_k, \hat{g}_k)+s_k\left[R(\hat{g}_k)-R(y_k)\right]-1=0. \tag{10}$$

Hence,

$$s_k^2\left[R(\hat{g}_k)-R(y_k)\right]=s_k-s_k^3(\hat{g}_k, \hat{g}_k).$$

From (9) we then have

$$R(z_k)-R(y_k)=-\frac{s_k+s_k^3(\hat{g}_k, \hat{g}_k)}{1+s_k^2(\hat{g}_k, \hat{g}_k)}(\hat{g}_k, \hat{g}_k)$$

$$=-(\hat{g}_k, \hat{g}_k)s_k,$$

and the lemma follows. ■

LEMMA 4. *For each* $k=1,2,\ldots,p$ *and each* $i=0,1,2\ldots$, *we have*

$$s_k^{(i)} \geqslant \frac{1}{\lambda_n-\lambda_1}.$$

Proof. Dropping the superscript (i), recall from Sec. II that each of the s_k's is positive when $B=I$. Hence, from Lemma 3 we can conclude that

$$\begin{aligned} R(z_k)-R(\hat{g}_k) &= R(z_k)-R(y_k)+R(y_k)-R(\hat{g}_k) \\ &= -\|\hat{g}_k\|^2 s_k + R(y_k)-R(\hat{g}_k). \end{aligned}$$

From (10) it follows that

$$0 = -\|\hat{g}_k\|^2 s_k^2 + s_k\left[R(y_k)-R(\hat{g}_k)\right]+1.$$

Thus,

$$R(z_k)-R(\hat{g}_k) = -\frac{1}{s_k}.$$

The lemma now follows from noting that

$$\lambda_1 \leqslant R(x) \leqslant \lambda_n$$

for any x in R^n. ∎

REMARK 4. From Lemma 3, it is clear that $\{R(x_k^{(i)})\}_{i=0}^{\infty}$ is a monotonically decreasing sequence for each $k=1,2,\ldots,p$. Since it is bounded below by λ_k, then it must converge. By Lemma 4, it follows that $\hat{g}_k^{(i)}$ tends to zero as i tends to ∞. Hence, examining each $k=1,2,\ldots,p$ in turn implies that each sequence $\{x_k^{(i)}\}_{i=0}^{\infty}$ is convergent and that each $\|g(y_k^{(i)})\|$ tends to zero. In particular, $x_k^{(i)}$ tends to u_{j_k} for some j_k in $\{1,2,\ldots,n\}$. The difficulty is now to show that $j_k=k$ for each k. We do this for $k=1$ in the next theorem, although the proof for general k has eluded us. (See Remark 6 below.)

REMARK 5. As $x_k^{(i)}$ approaches u_{j_k}, the scalar $\tilde{s}_k$ that *minimizes* $R(y_k-sg(y_k))$ over s becomes arbitrarily close to the scalar s_k that minimizes $R(y_k-s\hat{g}_k)$ over s. This is true because $g(y_k)$ is asymptotically orthogonal to

$Z_{(k-1)}$ for each k. To be more precise, dropping the superscript (i), note that

$$\frac{g(y_k)^T z_l}{\|g(y_k)\|} = \frac{g(y_k)^T y_l - s_l g(y_k)^T \hat{g}_l}{\|g(y_k)\|}$$

$$= -\frac{s_l[\, g(y_k)^T \hat{g}_l]}{\|\, g(y_k)\|}$$

for each $l=1,2,\ldots,k-1$. The right-hand side can be made arbitrarily small for large enough i, since $\hat{g}_l^{(i)}$ tends to zero. Roughly speaking, this implies that at convergence each column of X behaves almost as though no other vectors are used in the computation. We exploit this behavior in the proof of Theorem 1 by adapting the work of Faddeev and Faddeeva [14] to our setting. In particular, we note that convergence of $x_k^{(i)}$ to some eigenvector u_{i_k} implies roughly that the quantities $z_k^{(i)}$, $x_k^{(i)}$, and $y_k^{(i)}$ are asymptotically equal. More precisely, for any $\varepsilon>0$ there exists an $N<\infty$ such that, apart from a normalization factor near unity, we have

$$y_k^{(i)} = x_k^{(i)} - \sum_{\substack{l=1 \\ l\neq k}}^{p} \eta_{kl}^{(i)} x_l^{(i)}$$

and

$$x_k^{(i)} = z_k^{(i-1)} - \sum_{\substack{l=1 \\ l\neq k}}^{p} \theta_{kl}^{(i-1)} z_l^{(i-1)},$$

where (dropping the indices in the summation notation for convenience)

$$\eta_k^{(i)} = \sum |\eta_{kl}^{(i)}| < \varepsilon$$

and

$$\theta_k^{(i-1)} = \sum |\theta_{kl}^{(i-1)}| < \varepsilon$$

for all $i \geqslant N$. Combining these equations, with a possibly larger N we are then guaranteed that

$$y_k^{(i+1)} = z_k^{(i)} - \sum \kappa_{kl}^{(i)} z_l^{(i)}$$

where

$$\kappa_k = \sum |\kappa_{kl}^{(i)}| < \varepsilon$$

for all $i \geqslant N$.

THEOREM 1. *Assume that $X^{(0)}$ is not completely deficient in u_1. Then*

$$\lim_{i\to\infty} y_1^{(i)} = \lim_{i\to\infty} x_1^{(i)} = u_1.$$

Proof. To reach a contradiction, assume that $y_1^{(i)}$ does not converge to u_1. Thus, no column of $Y^{(i)}$ converges to u_1. In what follows, as in Remark 5, we ignore normalization factors of $X^{(i)}$, $Y^{(i)}$, and $Z^{(i)}$ since they are arbitrarily close to unity. For example, for each $k=1,2,\ldots,p$, we write

$$y_k^{(i)} = \varepsilon_k u_1 + u_{j_k} + \delta_k h_k$$

where $j_k \geqslant 2$, and h_k is of unit length and orthogonal to u_1 and u_{j_k}. For convenience, we drop the superscripts (i) in the coefficients for this proof. Then, by an argument analogous to that in [14], we are guaranteed of the existence of coefficients α_k and constants K, $\alpha_{\min}$, and $\alpha_{\max}$ such that

$$z_k = \varepsilon_k' u_1 + u_{j_k} + \delta_k' h_k',$$

where

$$\varepsilon_k' = (1+\alpha_k)\varepsilon_k,$$

$$0 < \alpha_{\min} \leqslant \alpha_k \leqslant \alpha_{\max},$$

and

$$|\delta_k'| \leqslant K|\delta_k|.$$

As in Remark 5, we write (assuming the worst case $h_l' = u_{j_k}$, $l \neq k$)

$$y_k^{(i+1)} = \varepsilon_k' u_1 + u_{j_k} + \delta_k' h_k' - \sum \kappa_{kl}(\varepsilon_l' u_1 + u_{j_l} + \delta_l' h_l')$$

$$= \Big(\varepsilon_k' - \sum \kappa_{kl}\varepsilon_l'\Big)u_1 + \Big(1 - \sum \kappa_{kl}\delta_l'\Big)u_{j_k} + \delta h,$$

where, as before and in what follows, we take the summation symbol Σ to include the limits $l=1$ to $l=p$, $l\neq k$. Let N be so large that

$$\kappa_k \leqslant \frac{\alpha_{\min}}{1+\alpha_{\max}+K}$$

and

$$\sum |\delta_k| \leqslant 1$$

for all $i \geqslant N$. Let r be the index that maximizes $|\varepsilon_k|$ over $k=1,2,\ldots,p$. Then

$$\begin{aligned}\left|\frac{\varepsilon_r' - \sum \kappa_{rl}\varepsilon_l'}{1-\sum \kappa_{rl}\delta_l'}\right| &= \left|\frac{(1+\alpha_r)\varepsilon_r - \sum \kappa_{rl}(1+\alpha_l)\varepsilon_l}{1+\sum \kappa_{rl}\delta_l'}\right| \\ &\geqslant \frac{(1+\alpha_{\min})|\varepsilon_r| - (1+\alpha_{\max})|\varepsilon_r|\kappa_r}{1+K\kappa_r\sum|\delta_l|} \\ &\geqslant \frac{(1+\alpha_{\min})-(1+\alpha_{\max})\kappa_r}{1+K\kappa_r}|\varepsilon_r| \\ &\geqslant |\varepsilon_r|.\end{aligned}$$

Thus, the ratio of the coefficient of u_1 in $y_r^{(i+1)}$ to the coefficient of u_{j_k} in $y_r^{(i+1)}$ has increased from the corresponding ratios for $y_r^{(i)}$. Of course, some $y_k^{(i+1)}$ with perhaps $k\neq r$ has the largest component in u_1 with respect to u_{j_k}, so the value of r might change with i. Nevertheless, a sequence $\{y_{r_i}^{(i)}\}_{i=N}^{\infty}$ can be constructed so that

$$\left|u_1^T y_{r_i}^{(i)}\right| \geqslant \left|u_1^T y_{r_N}^{(N)}\right|, \qquad i \geqslant N.$$

This contradicts the convergence assumption implying that one of the $y_k^{(i)}$'s must converge to u_1. Clearly, since the Ritz process orders the columns of $Y^{(i)}$ according to their Rayleigh quotients, the column that converges to u_1 must be $y_1^{(i)}$. Therefore, $x_1^{(i)}$ must also converge to u_1 and the theorem is proved. ■

REMARK 6. The proof of convergence of SIRQIT-G presented here is deficient in two respects. First, the proofs rest heavily on the assumption that $B=I$. Extension to more general B might be obtainable, although an alternative theoretical approach is probably necessary. The second deficiency is that the proof does not establish ordered convergence in the sense that the kth column of $x^{(i)}$ converges to u_k. This was done for $k=1$, and some compelling statements can be made for $k=2,3,\ldots,p$, but the complete proof appears to be beyond our grasp. Of course, in practice, after the first column of $x^{(i)}$ has been accepted, convergence of the second column to a slight perturbation of u_2 is assured. This follows because "deflated" iteration at this point is equivalent to unrestricted iteration on a slight perturbation of (1). This approach can thus be used in turn to establish a practical sort of ordered convergence.

Neither of these deficiencies is present in the *local* theory of convergence. In particular, a very simple proof of convergence can be given under the assumption that

$$R\left(y_k^{(0)}\right)<\lambda_{k+1}.$$

This alternate approach, which is straightforward, establishes *ordered* convergence for *general* $B\neq I$. However, this theory is restricted to local convergence and suffers somewhat from an esthetic point of view.

IV. RATES OF CONVERGENCE

Although we no longer require $B=I$, we shall assume throughout this section that ordered convergence of $X^{(i)}$ to U_1 is guaranteed for the initial guess $X^{(0)}$, namely, that

$$\lim_{i\to\infty} X^{(i)}=U_1.$$

[This is guaranteed, for example, when $B=I$ and when $R(y_k^{(0)})<\lambda_{k+1}$. (See Remark 6 in the previous section.)]

We present the asymptotic rates of convergence as asymptotic upper bounds for the ratios

$$\frac{\left|R\left(z_j^{(i)}\right)-\lambda_j\right|}{\left|R\left(y_j^{(i)}\right)-\lambda_j\right|},$$

$j=1,2,\ldots,p$. As we shall see, these rates depend upon the eigenvalues of the subproblems

$$C_{22}^{(j)}w=\mu w \tag{11}$$

and

$$\Gamma^{(j)}w=\rho\left(C_{22}^{(j)}\right)^2 w, \tag{12}$$

where

$$C_{22}^{(j)}=(\Lambda_2-\lambda_j I)^{1/2}U_2^T B^2 U_2(\Lambda_2-\lambda_j I)^{1/2},$$

$$\Gamma^{(j)}=(\Lambda_2-\lambda_j I)^{1/2}U_2^T B^2 W_j(D_j-\lambda_j I)W_j^T B^2 U_2(\Lambda_2-\lambda_j I)^{1/2},$$

W_j denotes the matrix with columns u_j, $u_{j+1},\ldots,u_p$, and D_j is a diagonal matrix with diagonal entries $\lambda_j, \lambda_{j+1},\ldots,\lambda_p$. The complexity of these expressions is the direct result of the fact that, in general, $U^TB^2U\neq I$. It therefore becomes necessary to examine the subspaces BU_1 and BU_2 to see how their lack of orthogonality effects rates of convergence. The problems (11) and (12) provide the basis for this determination.

When $B=I$, note that $C_{22}^{(j)}$ is just $\Lambda_2-\lambda_j I$. Thus, the problem (11) not only provides a measure of the distribution of the eigenvalues of (1), but also indicates in some sense how B differs from I on the subspace span(U_2). Further note that $\Gamma^{(j)}$ is zero when either $j=p$ or $B=I$. The problem (12) therefore provides a measure of how B distorts the orthogonality of the subspaces span$(B^{1/2}U_1)$ and span$(B^{1/2}U_2)$. The presence of the matrix $D_j-\lambda_j I$ in the definition of $\Gamma^{(j)}$ serves to dampen the effect of this distortion.

THEOREM 2. *Let* $\{\mu_{p+1}^{(j)}, \mu_{p+2}^{(j)},\ldots,\mu_n^{(j)}\}$ *and* $\{\rho_{p+1}^{(j)}, \rho_{p+2}^{(j)},\ldots,\rho_n^{(j)}\}$ *denote the eigenvalues of the problems* (11) *and* (12), *respectively, each listed in increasing algebraic order. Let*

$$r_1=\left[\frac{\mu_n^{(j)}-\mu_{p+1}^{(j)}}{\mu_n^{(j)}+\mu_{p+1}^{(j)}}\right]^2$$

and

$$r_2=\frac{\rho_n^{(j)}}{1+\rho_n^{(j)}}.$$

Assume that $\lambda_p < \lambda_{p+1}$. *Then*

$$\overline{\lim_{i\to\infty}} \left| \frac{R(z_j^{(i)}) - \lambda_j}{R(y_j^{(i)}) - \lambda_j} \right| \leqslant r_1 + r_2 - r_1 r_2. \tag{13}$$

When $B = I$, *these quantities reduce to*

$$r_1 = \left(\frac{\lambda_n - \lambda_{p+1}}{\lambda_n + \lambda_{p+1} - 2\lambda_j} \right)^2$$

and $r_2 = 0$.

Before we proceed with the proof of Theorem 2, we establish the following three lemmas. Lemma 5 shows that asymptotically SIRQIT produces correction vectors that lie essentially in the space spanned by the columns of BU_2. This is crucial to proving that rates of convergence depend upon p.

LEMMA 5. *For each* $j = 1, 2, \ldots, p$, *write*

$$\theta_j^{(i)} y_j^{(i)} = u_j + \delta_j^{(i)} v_j^{(i)} + \varepsilon_j^{(i)} h_j^{(i)},$$

where $v_j^{(i)} \in span(\{u_k\colon\ 1 \leqslant k \leqslant p,\ k \neq j\})$, $h_j^{(i)} \in span(U_2)$, $\theta_j^{(i)}$ *is a coefficient near unity, and*

$$\left\| v_j^{(i)} \right\|_B = \left\| h_j^{(i)} \right\|_B = 1.$$

Then there exists a constant $c > 0$ *such that*

$$\left| \delta_j^{(i)} \right| \leqslant \left| \varepsilon_j^{(i)} \varepsilon_k^{(i)} \right| c$$

for some $k = 1, 2, \ldots, p$, $k \neq j$.

Proof. For the moment we assume that the first p eigenvalues of (1) are distinct. Dropping the superscript (i) and fixing j, we then write

$$v_j = \sum_{k \neq j} \sigma_k u_k,$$

where

$$\sum_{k\neq j} \sigma_k^2 = 1.$$

For the moment, sums $\Sigma_{j\neq k}$ are to be taken over the integers $1,2,\ldots,p$. Now there must exist some index k such that $|\sigma_k|$ is no smaller than the arithmetic mean, that is,

$$|\sigma_k| \geq (p-1)^{-1/2}.$$

Consider

$$\theta_k y_k = u_k + \delta_k v_k + \epsilon_k h_k,$$

where

$$v_k = \sum_{l\neq k} \gamma_l u_l.$$

Then for $j\neq k$ we have

$$\begin{aligned} 0 &= (y_j, y_k)_B \\ &= \delta_k\gamma_j + \delta_j\sigma_k + \delta_j\delta_k(v_j, v_k)_B + \epsilon_j\epsilon_k(h_j, h_k)_B \end{aligned} \tag{14}$$

and

$$\begin{aligned} 0 &= (Ay_j, y_k) \\ &= \delta_k\gamma_j\lambda_j + \delta_j\sigma_k\lambda_k + \delta_j\delta_k(Av_j, v_k) + \epsilon_j\epsilon_k(Ah_j, h_k). \end{aligned} \tag{15}$$

Combining (14) and (15), we have

$$\delta_j\sigma_k(\lambda_k - \lambda_j) + \delta_k\delta_j K_1 - \epsilon_j\epsilon_k K_2 = 0, \tag{16}$$

where

$$K_1 = (Av_j, v_k) - \lambda_j(v_j, v_k)_B$$

and

$$K_2=\lambda_j(h_j, h_k)_B-(Ah_j, h_k).$$

Rewriting (16) yields

$$\delta_j\left[\sigma_k(\lambda_k-\lambda_j)+\delta_k K_1\right]=\varepsilon_j\varepsilon_k K_2.$$

By the choice of k,

$$\begin{aligned}|\sigma_k(\lambda_k-\lambda_j)+\delta_k K_1| &\geqslant|\sigma_k(\lambda_k-\lambda_j)|-|\delta_k K_1| \\ &\geqslant(p-1)^{-1/2}|\lambda_k-\lambda_j|-|\delta_k K_1| \\ &>0\end{aligned}$$

for $|\delta_k|$ sufficiently small. The lemma now follows for distinct eigenvalues of (1) by setting

$$c=\left|\frac{K_2}{\sigma_k(\lambda_k-\lambda_j)+\delta_k K_1}\right|.$$

For the case of multiple eigenvalues, let r represent the number of distinct eigenvalues in the set $\{\lambda_1, \lambda_2, \ldots, \lambda_p\}$. We assume that $1<r<p$. (The case $r=1$ follows trivially.) The proof just presented for distinct eigenvalues (i.e., $r=p$) now extends as follows. Let V_i represent the span of the eigenvectors corresponding to λ_i. Let m_i denote the dimension of V_i. We then have

$$\theta_j y_j=v_j+\delta_j\sum_{l\neq j}\sigma_l v_l+\varepsilon_j h_j,$$

where $v_l\in V_l$ and

$$\sum_{l\neq j}\sigma_l^2=1.$$

Here we take the sums over the range $l=1,2,\ldots, r$. Now for some k we may assume

$$|\sigma_k|\geqslant(r-1)^{-1/2}.$$

But there exists some index k' such that $\hat{v}_{k'}$ is in V_k,

$$\theta_{k'} y_{k'} = \hat{v}_{k'} + \delta_{k'} \sum_{l \neq k} \rho_l \hat{v}_l + \varepsilon_{k'} h_{k'},$$

and

$$|(\hat{v}_{k'}, v_k)_B| \geq m_k^{-1/2},$$

where each $\hat{v}_l \in V_l$. Taking inner products of y_j and $y_{k'}$ as in (14) and (15) introduces a multiple of δ_j contributed by the inner product of $v_{k'}$ and the sum $\Sigma_{l \neq j} \sigma_l v_l$. This multiple is thus bounded below by $|\lambda_k - \lambda_j|$ $[m_k(r-1)]^{-1/2} - O(\delta_{k'})$. A change in c defined above then allows the extension of this proof to multiple eigenvalues, and the lemma is proved. ■

The significance of Lemma 5 is that

$$\delta_j^{(i)} = O\left(\varepsilon_j^{(i)} \varepsilon_k^{(i)}\right),$$

so that y_j is asymptotically equal to $u_j + \varepsilon_j h_j$. Thus, y_j is nearly in the B-orthogonal complement of span$(\{u_k : 1 \leq k \leq p,\ k \neq j\})$. Making use of this together with the fact that z_j is arbitrarily close to y_j and hence u_j (as in Sec. III), we then have the following lemma.

LEMMA 6. *Define*

$$q_j^{(i)} = \begin{cases} 0, & j = 1, \\ -\sum_{k=1}^{j-1} \left((A - \lambda_j B) h_j^{(i)}, u_k\right)_B h_j^{(i)T} (A - \lambda_j B)^2 u_k, & j > 1. \end{cases}$$

With $\varepsilon_j^{(i)}$ given as in lemma 5, define the quantities

$$\eta_j^{(i)} = \left\| z_j^{(i)} - y_j^{(i)} \right\|$$

and

$$\delta_j^{(i)} = \left\| z_j^{(i)} - u_j^{(i)} \right\|.$$

Let the notation $O(\gamma)$ *denote terms of order* $\eta^{(i)}=\sum_{k=1}^{j-1}\eta_k^{(i)}$, $\delta^{(i)}=\sum_{k=1}^{j-1}\delta_k^{(i)}$, *and* $\varepsilon_j^{(i)}$ *collectively. That is,* $O(\gamma)$ *tends to zero as the maximum of these terms tends to zero. Then the step size for* SIRQIT-G *satisfies*

$$s_j^{(i)}=\frac{h_j^{(i)T}(A-\lambda_j B)^2 h_j^{(i)}}{h_j^{(i)T}(A-\lambda_j B)^3 h_j^{(i)}+q_j^{(i)}}+O(\gamma).$$

Hence,

$$\frac{R(z_j^{(i)})-\lambda_j}{R(y_j^{(i)})-\lambda_j}=1-\frac{\left[h_j^{(i)T}(A-\lambda_j B)^2 h_j^{(i)}\right]^2}{h_j^{(i)T}(A-\lambda_j B)h_j^{(i)}\left[h_j^{(i)T}(A-\lambda_j B)^3 h_j^{(i)}+q_j^{(i)}\right]}+O(\gamma).$$

Proof. Dropping the superscript (i) and writing $y_j=u_j+\varepsilon_j h_j$ (see the comment just before this lemma), it is easy to see that

$$R(y_j)=\lambda_j+\frac{\varepsilon_j^2 h_j^T(A-\lambda_j B)h_j}{1+\varepsilon_j^2}.$$

Thus,

$$\begin{aligned} g_j &\equiv g(y_j)\\ &=Ay_j-R(y_j)By_j\\ &=\varepsilon_j(A-\lambda_j B)h_j+O(\varepsilon_j^2). \end{aligned}$$

Note that

$$\hat{g}_j=\begin{cases} g_1, & j=1,\\ g_j-\sum_{k=1}^{j-1}(g_j,z_k)_B z_k, & j>1. \end{cases}$$

Clearly, $\|\hat{g}_j\|=O(\varepsilon_j)$ since $\|g_j\|=O(\varepsilon_j)$. Substitution into (5), (6), and (7) shows that $a_j=O(\varepsilon_j^3)$, $j\geqslant 1$. [Throughout this proof, a_j, b_j, and c_j will denote the quantities given in (5), (6), and (7), respectively, for the jth column of Y.]

Again, for $j=2,3,\ldots,p$, we have

$$\begin{aligned} c_j &= -(\hat{g}_j, g_j) \\ &= -\varepsilon_j^2 h_j^T (A-\lambda_j B)^2 h_j + \sum_{k=1}^{j-1} (g_j, z_k)_B (g_j, z_k) + O(\varepsilon_j^3) \\ &= -\varepsilon_j^2 h_j^T (A-\lambda_j B)^2 h_j + O(\varepsilon_j^2 \eta) + O(\varepsilon_j^3). \end{aligned}$$

Moreover,

$$c_1 = -\varepsilon_1^2 h_1^T (A-\lambda_1 B)^2 h_1 + O(\varepsilon_1^3).$$

Note that

$$\begin{aligned} b_j &= \hat{g}_j^T A \hat{g}_j - \hat{g}_j^T B \hat{g}_j R(y_j) \\ &= \hat{g}_j^T (A-\lambda_j B)\hat{g}_j + O(\varepsilon_j^4) \\ &= \varepsilon_j^2 h_j^T (A-\lambda_j B)^3 h_j - \sum_{k=1}^{j-1} (g_j, z_k)_B \\ &\quad \times \left[z_k^T (A-\lambda_j B)\hat{g}_j + g_j^T (A-\lambda_j B) z_k \right] + O(\varepsilon_j^3). \end{aligned}$$

That the first term in square brackets is negligible follows from the observation that

$$\begin{aligned} (g_j, z_k)_B z_k^T (A-\lambda_j B)\hat{g}_j &= (g_j, z_k)_B u_k^T (A-\lambda_j B)\hat{g}_j + O(\delta_k \varepsilon_j^2) \\ &= (g_j, z_k)_B (\lambda_k - \lambda_j) u_k^T B \hat{g}_j + O(\delta_k \varepsilon_j^2) \\ &= (g_j, z_k)_B (\lambda_k - \lambda_j) z_k^T B \hat{g}_j + O(\delta_k \varepsilon_j^2) \\ &= O(\delta_k \varepsilon_j^2). \end{aligned}$$

Note also that

$$-\sum_{k=1}^{j-1}(g_j, z_k)_B g_j^T(A-\lambda_j B)z_k = -\sum_{k=1}^{j-1}(g_j, u_k)_B g_j^T(A-\lambda_j B)u_k + O(\delta\varepsilon_j^2)$$

$$= -\varepsilon_j^2 \sum_{k=1}^{j-1}\left((A-\lambda_j B)h_j, u_k\right)_B h_j^T(A-\lambda_j B)^2 u_k$$

$$+ O(\delta\varepsilon_j^2) + O(\varepsilon_j^3)$$

$$= \varepsilon_j^2 q_j + O(\delta\varepsilon_j^2) + O(\varepsilon_j^3).$$

Hence,

$$b_j = \begin{cases} \varepsilon_1^2 h_1^T(A-\lambda_j B)^3 h_1 + O(\varepsilon_1^3), & j=1, \\ \varepsilon_j^2 h_j^T(A-\lambda_j B)^3 h_j + \varepsilon_j^2 q_j + O(\delta\varepsilon_j^2) + O(\varepsilon_j^3), & j>1. \end{cases}$$

In what follows, the case $j=1$ can be accounted for by remembering that $q_1=0$. Substitution of the above expressions into (8) shows that

$$s_j = \frac{h_j^T(A-\lambda_j B)^2 h_j}{h_j^T(A-\lambda_j B)^3 h_j + q_j} + O(\gamma)$$

$$= -c_j/b_j + O(\gamma).$$

Hence,

$$R(z_j) = \frac{R(y_j) - 2s_j y_j^T A\hat{g}_j + s_j^2 \hat{g}_j^T A\hat{g}_j}{1 - 2s_j y_j^T B\hat{g}_j + s_j^2 \hat{g}_j^T B\hat{g}_j},$$

so that

$$R(z_i)-\lambda_i=\frac{\varepsilon_i^2 h_i^T(A-\lambda_i B)h_i-2s_i\hat{g}_i^T g_i+s_i^2\left[\hat{g}_i^T A\hat{g}_i-R(y_i)\hat{g}_i^T B\hat{g}_i\right]}{1+O(\varepsilon_i)}+O(\varepsilon_i^4)$$

$$=\frac{\varepsilon_i^2 h_i^T(A-\lambda_i B)h_i+2s_i c_i+s_i^2 b_i}{1+O(\varepsilon_i)}+O(\gamma)$$

$$=\frac{\varepsilon_i^2\left\{h_i^T(A-\lambda_i B)h_i-\left[h_i^T(A-\lambda_i B)^2 h_i\right]^2\left[h_i^T(A-\lambda_i B)^3 h_i+q_i\right]^{-1}\right\}}{1+O(\varepsilon_i)}+O(\gamma).$$

Therefore,

$$\frac{R(z_i)-\lambda_i}{R(y_i)-\lambda_i}=1-\frac{\left[h_i^T(A-\lambda_i B)^2 h_i\right]^2}{h_i^T(A-\lambda_i B)h_i\left[h_i^T(A-\lambda_i B)^3 h_i+q_i\right]}+O(\gamma)$$

and the lemma is proved. ■

The next lemma treats two special maximum problems needed for the proof of Theorem 2. They are applied in conjunction with the problems (11) and (12) to get the final result.

LEMMA 7. *Suppose E and F are $k\times k$ symmetric matrices. If E^2 is positive definite and F is nonnegative definite, then*

$$\max_{\substack{w\in R^k\\ w\neq 0}}\left[1-(w^TEw)^2(w^Tww^TE^2w)^{-1}\right]=\left(\frac{\mu_k-\mu_1}{\mu_k+\mu_1}\right)^2$$

and

$$\max_{\substack{w\in R^k\\ w\neq 0}} w^TFw\left[w^T(E^2+F)w\right]^{-1}=\frac{\rho_k}{1+\rho_k}.$$

Here, μ_k and μ_1 are the largest and smallest, respectively, of the eigenvalues of E, and ρ_k is the largest eigenvalue of the generalized eigenproblem

$$Fw=\rho E^2 w. \tag{17}$$

Proof. Let

$$f_1(w) = 1 - \frac{(w^T E w)^2}{w^T w w^T E^2 w}$$

and

$$f_2(w) = \frac{w^T F w}{w^T (E^2 + F) w}.$$

In order to use a Lagrange-multiplier argument, the constraint $w \neq 0$ is replaced by $w^T w - 1 = 0$, with gradient $2w$. Solutions of the maximum problems imply the existence of real numbers t_1, t_2 (not both zero) and t_3, t_4 (not both zero) such that

$$t_1 \nabla f_1(w) + t_2 = 0$$

and

$$t_3 \nabla f_2(w) + t_4 = 0.$$

But the homogeneity of f_1 and f_2 implies that their gradients at w are orthogonal to w. This implies that $t_2 = t_4 = 0$, so the maxima of f_1 and f_2 occur at the zeros of their gradients. We focus first on f_2, since the argument for it is simpler. Note that

$$\nabla f_2(w) = \frac{2w^T (E^2 + F) w \cdot Fw - 2w^T Fw (E^2 + F) w}{\left(w^T (E^2 + F) w\right)^2}$$

$$= \left(\alpha_2 F + \beta_2 E^2\right) w$$

for appropriate constants α_2 and β_2. Hence, the gradient of f_2 is zero if and only if w is an eigenvector of (17). But then

$$f_2(w) = \frac{\rho w^T E^2 w}{w^T E^2 w + \rho w^T E^2 w}$$

$$= \frac{\rho}{1 + \rho}.$$

Since

$$\rho = \frac{w^T F w}{w^T E^2 w} > 0,$$

then the maximum of f_2 is attained by setting $\rho = \rho_k$. This proves the lemma for f_2. The argument for f_1 is slightly more complicated. We first note that

$$\nabla f_1(w) = (\alpha_1 E^2 + \beta_1 E + \gamma_1 I) w$$

for appropriate constants α_1, β_1, and γ_1. Thus, the gradient of f_1 is zero if and only if w is a linear combination of at most two eigenvectors of E. Setting $w = \tau_l q_l + \tau_m q_m$, where q_l and q_m are eigenvectors of unit length of E associated with the eigenvalues μ_l and μ_m, respectively, and $\tau_l^2 + \tau_m^2 = 1$, we then have

$$f_1(w) = 1 - \frac{(\tau_l^2 \mu_l + \tau_m^2 \mu_m)^2}{(\tau_l^2 + \tau_m^2)(\tau_l^2 \mu_l^2 + \tau_m^2 \mu_m^2)}$$

$$= \frac{\tau_l^2 \tau_m^2 (\mu_l - \mu_m)^2}{\tau_l^2 \mu_l^2 + \tau_m^2 \mu_m^2}.$$

Assume that $\mu_m < \mu_l$, and set $\xi = \mu_m / \mu_l$ and $t = \tau_m^2$. Then

$$f_1(w) = \frac{(1-t)t(1-\xi)^2}{t(\xi^2 - 1) + 1},$$

where $0 < \xi < 1$ and $0 \leqslant t \leqslant 1$. The maximum of $f_1(w)$ over t occurs at $t = 1/(1+\varepsilon)$, so that

$$\max_{0 < t < 1} f_1(w) = \left(\frac{1-\xi}{1+\xi} \right)^2,$$

which is largest when ξ is as small as possible, namely, when $m = 1$ and $l = k$. This proves the assertion for f_1, and the lemma now follows. ■

Proof of Theorem 2. Define

$$f(h_j) = 1 - \frac{\left(h_j^T(A-\lambda_j B)^2 h_j\right)^2}{h_j^T(A-\lambda_j B)h_j\left[\, h_j^T(A-\lambda_j B)^3 h_j + q_j\right]},$$

where the q_j are defined as in Lemma 6. Dropping the subscript j for convenience, the proof now rests on the determination of the maximum value of f subject to the restriction that h lies in span(U_2). The difficulty for the case $B \neq I$ first appears here in the fact that span(U_2) is generally no longer an invariant subspace of $A-\lambda B$. However, the matrix $A-\lambda B$ is positive definite on span(U_2), a property central to the remainder of the proof.

The constraint that h is in span(U_2) is accounted for by setting $h = Uv$, where $v = (0,0,\ldots,0, v_{p+1},\ldots, v_n)^T$. Setting $\bar{v} = (v_{p+1},\ldots, v_n)^T$, the maximum problem can then be written as

$$\max_{\bar{v}^T\bar{v}=1} f(Uv).$$

Let $\Theta = U^T B^2 U$. Then Θ is positive definite on R^n and $BU = U\Theta$. Hence it follows that

$$f(Uv) = 1 - \frac{\left(v^T U^T(A-\lambda B)^2 Uv\right)^2}{v^T U^T(A-\lambda B)Uv\left[\, v^T U^T(A-\lambda B)^3 Uv + q\right]}$$

$$= 1 - \frac{\left(v^T\bar{\Lambda}\Theta\bar{\Lambda}v\right)^2}{v^T\bar{\Lambda}v\left[\, v^T\bar{\Lambda}\Theta\bar{\Lambda}v + q\right]},$$

where $\bar{\Lambda} = \Lambda - \lambda I$. We shall make use of the partitions that correspond to $U = (U_1 \;\; U_2)$ given by

$$\Theta = \begin{pmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{21} & \Theta_{22} \end{pmatrix}$$

and

$$\bar{\Lambda} = \begin{bmatrix} \bar{\Lambda}_1 & 0 \\ 0 & \bar{\Lambda}_2 \end{bmatrix}.$$

The entries of $\bar{\Lambda}_2$ are strictly positive, and we define $\bar{\Lambda}_2^{1/2}$ in the same way. Let $\bar{\Lambda}_1^-$ and $\bar{\Lambda}_1^+$ denote the diagonal $p\times p$ matrices with diagonal entries given by

$$d_k^- = \begin{cases} -|\lambda_k-\lambda_j|^{1/2}, & k<j, \\ (\lambda_k-\lambda_j)^{1/2}, & k\geqslant j, \end{cases}$$

and

$$d_k^+ = \begin{cases} |\lambda_k-\lambda_j|^{1/2}, & k<j, \\ (\lambda_k-\lambda_j)^{1/2}, & k\geqslant j, \end{cases}$$

respectively. Recalling that $\lambda=\lambda_j$, then it is easily seen that $\bar{\Lambda}=\bar{\Lambda}^-\bar{\Lambda}^+$, where $\bar{\Lambda}^{\pm}$ is the $n\times n$ diagonal matrix whose first $p\times p$ block is $\bar{\Lambda}_1^{\pm}$ and whose remaining diagonal entries are those of $\bar{\Lambda}_2^{1/2}$. Note that (18) now becomes

$$f(Uv)=1-\frac{(v^T\bar{\Lambda}\Theta\bar{\Lambda}v)^2}{v^T\bar{\Lambda}v\left[v^T\bar{\Lambda}\Theta\bar{\Lambda}\Theta\bar{\Lambda}v+q\right]}$$

$$=1-\frac{(v^T\bar{\Lambda}^+\bar{\Lambda}^-\Theta\bar{\Lambda}^+\bar{\Lambda}^-v)^2}{v^T\bar{\Lambda}^+\bar{\Lambda}^-v\left[v^T\bar{\Lambda}^+\bar{\Lambda}^-\Theta\bar{\Lambda}^+\bar{\Lambda}^-\Theta\bar{\Lambda}^+\bar{\Lambda}^-v+q\right]}$$

$$=1-\frac{(K^T\bar{\Lambda}^-\Theta\bar{\Lambda}^+K)^2}{K^TK\left[K^T(\bar{\Lambda}^-\Theta\bar{\Lambda}^+)^2K+q\right]},$$

where $K=\bar{\Lambda}^+v=\bar{\Lambda}^-v$. To simplify the expression further, let $C=\bar{\Lambda}^-\Theta\bar{\Lambda}^+$, so that

$$f(Uv)=1-\frac{(K^TCK)^2}{K^TK\left[K^TC^2K+q\right]}.$$

To use the block representation $K=(0^T\ (\bar{\Lambda}^{1/2}\bar{v})^T)^T=(0^T\ w^T)^T$, we partition

C according to

$$C=\begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

$$=\begin{bmatrix} \bar{\Lambda}_1^{-}\Theta_{11}\bar{\Lambda}_1^{+} & \bar{\Lambda}_1^{-}\Theta_{21}\bar{\Lambda}_2^{1/2} \\ \bar{\Lambda}_2^{1/2}\Theta_{21}\bar{\Lambda}_1^{+} & \bar{\Lambda}_2^{1/2}\Theta_{22}\bar{\Lambda}_2^{1/2} \end{bmatrix}.$$

We may then write $(K^TCK)^2=(w^TC_{22}w)^2$ and $K^TC^2K=w^T(C_{22}^2+C_{21}C_{12})w$. Define

$$f_r(w)=1-\frac{\left(w^TC_{22}w\right)^2}{w^Tw\left[\,w^T(C_{22}+C_{21}C_{12})w+q\right]}.$$

Clearly,

$$\max_{\bar{v}^T\bar{v}=1} f(Uv)=\max_{w\neq 0} f_r(w).$$

To examine q in (18), define the n-vector ϕ by its entries

$$\phi_l=\begin{cases} 0, & l\geqslant j, \\ ((A-\lambda B)h, u_l)_B, & l<j. \end{cases}$$

Then

$$q=-h^T(A-\lambda B)^2U\phi.$$

Let $\bar{\phi}$ denote the $(j-1)$-vector whose entries are the first $j-1$ of ϕ. Let $U_{(j-1)}$ denote the matrix consisting of the first $j-1$ columns of U, and $\bar{\Lambda}_{(j-1)}$ be the leading $(j-1)\times(j-1)$ minor of $\bar{\Lambda}$. Then

$$q=-v^TU(A-\lambda B)^2U\phi$$

$$=-v^T\bar{\Lambda}U^TB^2U\bar{\Lambda}\phi$$

$$=-\tilde{v}^T\bar{\Lambda}_2U_2^TB^2U_{(j-1)}\bar{\Lambda}_{(j-1)}\bar{\phi}.$$

But

$$\bar{\phi} = U_{(j-1)}^T B(A - \lambda B)h$$

$$= U_{(j-1)}^T B^2 U_2 \bar{\Lambda}_2 \bar{v}.$$

Again let $w = \bar{\Lambda}^{1/2} v$, so that

$$q = -w^T \bar{\Lambda}_2^{1/2} U_2^T B^2 U_{(j-1)} \bar{\Lambda}_{(j-1)} U_{(j-1)}^T B^2 U_2 \bar{\Lambda}_2^{1/2} w.$$

Thus,

$$w^T C_{21} C_{12} w + q = w^T \Gamma_1^T \Gamma_1 w$$

$$= w^T \Gamma w,$$

where Γ_1 is the matrix that consists of the last $p-j+1$ rows of C_{12}. We therefore have

$$\max_{w \neq 0} f_r(w) = \max_{w \neq 0} \left[1 - \frac{\left(w^T C_{22} w\right)^2}{w^T w w^T \left(C_{22}^2 + \Gamma\right) w} \right]$$

$$= 1 - \min_{w \neq 0} \left[\frac{\left(w^T C_{22} w\right)^2}{w^T w w^T C_{22}^2 w} \cdot \frac{w^T C_{22}^2 w}{w^T \left(C_{22}^2 + \Gamma\right) w} \right]$$

$$\leqslant 1 - \left[1 - \max_{w \neq 0} \left[1 - \frac{\left(w^T C_{22} w\right)^2}{w^T w w^T C_{22}^2 w} \right] \right] \left[1 - \max_{w \neq 0} \frac{w^T \Gamma w}{w^T \left(C_{22}^2 + \Gamma\right) w} \right].$$

Lemma 7 now applies with $E = C_{22}$ and $F = \Gamma$. That the hypotheses are satisfied follows from the observations that C_{22} and hence C_{22}^2 are positive definite and Γ is nonnegative definite. This last assertion is true because $\Gamma = \Gamma_1^T \Gamma_1$. This proves Theorem 2. ■

REMARK 7. Since the rank of $\Gamma^{(j)}$ is no larger than $p-j$, the null space of $\Gamma^{(j)}$ is very large when $p \ll n$. Further, the matrix C_{22}^2 involves only the larger eigenvalues of (1). Thus, the contribution of $\Gamma^{(j)}$ to the value of the bounds on the rates will often be quite small. Of course, when B is badly conditioned, this contribution may be on the same order of magnitude as

that for C_{22}^2. This is clarified in the corollary below, which relates the ratios r_1 and r_2 to Λ and $K=K(B)$, the condition number of B.

COROLLARY 1. *Let*

$$r_1' = \left(\frac{K(\lambda_n - \lambda_j) - (\lambda_{p+1} - \lambda_j)}{K(\lambda_n - \lambda_j) + (\lambda_{p+1} - \lambda_j)} \right)^2$$

and

$$r_2' = \frac{(K-1)^2(\lambda_p - \lambda_j)}{(K-1)^2(\lambda_p - \lambda_j) + 4K(\lambda_{p+1} - \lambda_j)}.$$

Then $r_1 \leqslant r_1'$, $r_2 \leqslant r_2'$, *and* (13) *becomes*

$$\lim_{i \to \infty} \left| \frac{R(z_j^{(i)}) - \lambda_j}{R(y_j^{(i)}) - \lambda_j} \right| \leqslant r_1' + r_2' - r_1' r_2'.$$

Proof. We first examine the bound r_1'. Let $v = \bar{\Lambda}_2^{1/2} w$, and write

$$BU_2 v = U_{(j-1)} c_{(j-1)} + W_j c_j + U_2 c_2,$$

where c_j and c_2 are vectors of dimension $p-j+1$ and $n-p$, respectively. Now

$$w^T C_{22}^2 w = c_2^T \bar{\Lambda}_2 c_2$$

and, with $\bar{D}_j = \operatorname{diag}(0, \lambda_{j+1} - \lambda_j, \ldots, \lambda_p - \lambda_j)$, we have

$$w^T \Gamma^{(j)} w = c_j^T \bar{D}_j c_j.$$

Since

$$c_2 = U_2^T B^2 U_2 v$$

and $v \neq 0$, then c_2 is nonzero. Using this notation, we may then rewrite r_2 as

$$r_2 = \max_{c_2 \neq 0} \frac{c_j^T \bar{D}_j c_j}{c_j^T \bar{D}_j c_j + c_2^T \bar{\Lambda}_2 c_2}.$$

Now set

$$m = \max_{c_2 \neq 0} \frac{c_j^T \bar{D}_j c_j}{c_2^T \bar{\Lambda}_2 c_2}$$

so that r_2 is bounded from above by $m(m+1)^{-1}$. Define

$$c = \max_{c_2 \neq 0} \frac{\| c_j \|^2}{\| c_2 \|^2}.$$

Then the largest value of m is achieved by setting $c_j = c^{1/2} e_p$ and $c_2 = e_{p+1}$. With $c_1 = (c_{(j-1)}^T, c_j^T)^T$, then

$$\begin{aligned} c &\leqslant \max_{c_2 \neq 0} \frac{\| c_1 \|^2}{\| c_2 \|^2} \\ &= \max_{v \neq 0} \frac{\| U_1^T B^2 U_2 v \|^2}{\| U_2^T B^2 U_2 v \|^2} \\ &= \max_{v \neq 0} \left| \frac{v^T U_2^T B^3 U_2 v}{v^T U_2^T B^2 U_2 U_2^T B^2 U_2 v} - 1 \right| \end{aligned}$$

where the last line follows by noting that

$$U_1 U_1^T = B^{-1} - U_2 U_2^T.$$

Let

$$f_4(v) = \frac{v^T U_2^T B^3 U_2 v}{v^T U_2^T B^2 U_2 U_2^T B^2 U^2 v}.$$

The maximum of $f_4(v)$ is at least unity, so that

$$\max_{v\neq 0} |f_4(v)-1| = \max_{v\neq 0} f_4(v) - 1.$$

In fact,

$$\max_{v\neq 0} f_4(v) = \max_{v^Tv=1} \frac{v^T U_2^T B^3 U_2 v}{\|U_2^T B^2 U_2 v\|^2}$$

$$\leqslant \max_{v^Tv=1} \frac{v^T U_2^T B^3 U_2 v}{\left(v^T U_2^T B^2 U_2 v\right)^2}$$

$$= \max_{q^Tq=1} \frac{q^T B^2 q}{\left(q^T B q\right)^2}$$

where $q = B^{1/2} U_2 v$. Changing the form in order to invoke Lemma 7, we have

$$\max_{q^Tq=1} \frac{q^T B^2 q}{\left(q^T B q\right)^2} = \left[1 - \max_{q^Tq=1}\left(1 - \frac{\left(q^T B q\right)^2}{q^T q q^T B^2 q}\right)\right]^{-1}$$

$$= \left[1 - \left(\frac{K-1}{K+1}\right)^2\right]^{-1}$$

$$= \frac{(K+1)^2}{4K},$$

which is at least unity. Hence,

$$c \leqslant \frac{K^2+2K+1}{4K} - 1$$

$$= \frac{(K-1)^2}{4K}.$$

Thus, $r_2 \leqslant r_2'$ as claimed.

To examine the bound r_1', rewrite (11) as

$$U_2^T B^2 U_2 v = \mu \bar{\Lambda}_2^{-1} v$$

where $v=\bar{\Lambda}_2^{-1/2}w$. Then

$$\max_{v^Tv=1} v^TU_2^TB^2U_2v=\max_{q^Tq=1} q^TBq$$

$$\leqslant b_n,$$

where b_n is the largest eigenvalue of B. In a similar manner, we have that

$$\min_{v^Tv=1} v^TU_2^TB^2U_2v\geqslant b_1,$$

where b_1 is the smallest eigenvalue of B. We can thus bound the eigenvalues for (11) according to

$$\mu_n\leqslant b_n(\lambda_n-\lambda_j)$$

and

$$\mu_{p+1}\geqslant b_1(\lambda_{p+1}-\lambda_j).$$

Substitution of these bounds into the expression for r_1 yields the bound r_1', and the corollary is proved. ■

V. THE CONJUGATE-GRADIENT VERSION

The method of conjugate gradients applied to the solution of linear equations involving positive definite symmetric matrices was first described by Hestenes and Stiefel [15]. It is based upon successive correction vectors that are conjugate with respect to the associated matrix, that is, the directions are orthogonal in the inner product induced by this matrix. Bradbury and Fletcher [7] used a nonlinear version of the conjugate-gradient method to solve the algebraic eigenproblem. A summary of further developments for eigenproblems can be found in [16] and [17]. We propose the use of this method in conjunction with simultaneous iteration which we briefly describe below.

The conjugate gradient version of SIRQIT treated here, which we call SIRQIT-CG, is a simple generalization of conjugate-gradient minimization of the Rayleigh quotient. In particular, our approach is to perform several steps of the iterative process on each of the individual columns of X, perform a

Ritz acceleration on X, and restart the entire procedure. Thus, Ritz acceleration is used precisely when the conjugate gradient iteration is to be restarted. This is critical, since interaction among the columns of X undermines the motivation for using columnwise conjugate directions.

Conjugacy of the direction vectors can be attempted by explicit use of the Hessian (or bigradient) of $R(x)$ given by

$$H(x)=\frac{2}{x^TBx}\left[A-R(x)B-xg(x)^T-g(x)x^T\right]. \tag{18}$$

An asymptotically equivalent approach (cf. [16]) is to modify the current gradient by a properly chosen scalar multiple of the previous direction. For the kth column of the ith iterate $Y^{(i)}$, one such choice is given by

$$\beta_k^{(i)}=\frac{\|g(y_k^{(i)})\|^2}{\|g(y_k^{(i-1)})\|^2}, \qquad i\leqslant k\leqslant p. \tag{19}$$

Numerical experience suggests that this is probably the best approach in terms of efficiency.

SIRQIT-CG is outlined as follows:

(i), (ii), and (iii) are the same as in SIRQIT as described in section II.

(iv) Compute $\Psi^{(i)}=G(Y^{(i)})-\Psi^{(i-1)}\mathcal{B}(i)$, where $\mathcal{B}^{(i)}$ is a diagonal matrix of dimension p with diagonal entries $\beta_k^{(i)}$ given in (19).

(v) Set $Z^{(i)}=Y^{(i)}-\Psi^{(i)}S^{(i)}$, where $S^{(i)}$ is a diagonal matrix determined to minimize $R(z_k^{(i)})$ for each $k=1,2,\ldots,p$. (See Remark 8 below.)

(vi) Decide whether or not to restart (presumably by fixing the number of conjugate-gradient iterations at the start of SIRQIT-CG):

(a) If so, make the replacement $X^{(i+1)}=Z^{(i)}M^{(i)}$, where $M^{(i)}$ is determined (in effect by the Gram-Schmidt procedure) as an upper triangular matrix so that $X^{(i+1)T}BX^{(i+1)}=I$. Increase i by 1 and go to step (ii).
(b) If not, make the replacement $Y^{(i+1)}=Z^{(i)}M^{(i)}$, increase i by 1, and go to step (iii).

REMARK 8. Let $\psi_1,\psi_2,\ldots,\psi_p$ and $y_1,y_2,\ldots,y_p$ represent the columns of Ψ and Y, respectively. Each diagonal element, s_k, of $S=S^{(i)}$ is found by minimizing $R(y_k-s\psi_k)$ with respect to s. Equations (4) through (8) are used with $\hat{g}$ replaced by ψ_k for each $k=1,2,\ldots,p$. Note that $c=-\psi_k^Tg(y_k)$ and $b=\psi_k^TB\psi_k[R(\psi_k)-R(y_k)]$, so that the choice s_+ in either (8) or (8′) again assures minimization of $R(y_k-s\psi_k)$.

Conjugate-gradient techniques tend to accelerate gradient methods, so that SIRQIT-CG is very useful in a neighborhood of the solution. Near solutions the Y iterates are nearly A- and B-orthogonal, and hence so are the columns

of Z. Steps (ii) and (vi) have little effect on this orthogonality. This is fortunate, since otherwise it may be necessary to add a B-orthogonalization step in (iv). This would amount to a mixing of the columns of Z and destruction of the underlying motivation for conjugacy, thus undermining the method itself.

There are two other natural ways to develop a conjugate-gradient scheme that is properly integrated with simultaneous iteration for solving (1). Both involve extending the Rayleigh-quotient concept to $R^{n\times p}$ by defining

$$R_G(X)=\operatorname{tr}\left[(X^TBX)^{-1/2}(X^TAX)(X^TBX)^{-1/2}\right],$$

where $R_G: R^{n\times p}\to R$. Viewing $R^{n\times p}$ as an np-dimensional vector space, a nonlinear version of conjugate gradients can then be applied to the problem of minimizing $R_G(X)$ subject to the constraint that $X^TBX=I$, X in $R^{n\times p}$. Ignoring B-orthonormalization, this yields a method of the form

$$X^{(i+1)}=X^{(i)}-s_1\Psi^{(i)},$$

where $\Psi^{(i)}$ is computed in terms of $\Psi^{(i-1)}$ and the gradient of $R_G(X^{(i)})$ in $R^{n\times p}$ and in accordance with the conjugacy requirement. The step size is a scalar determined to minimize $R_G(X^{(i)}-s\Psi^{(i)})$ over s. The difficulty with this first approach is that, roughly speaking, the iteration is extended in scope (to $R^{n\times p}$) without a commensurate increase in power (e.g., the step size is only a scalar quantity).

A second approach involves using the same direction, $\Psi^{(i)}$, but computing the step size implicitly by determining $X^{(i+1)}$ as the first p eigenvectors of (1) as it is projected onto $\operatorname{span}(X^{(i)},\Psi^{(i)})$. However, although this extension appears to have the necessary power, it is unfortunately an improper extension of Rayleigh-quotient minimization by conjugate gradients. (The Hessian given in (18) operates columnwise on X. Thus, conjugacy of $n\times p$ matrices is not a well-defined concept.)

We have confirmed the poor behavior of both of these alternative extensions in numerical tests.

VI. NUMERICAL RESULTS

In this section, we content ourselves with illustrating the convergence properties of SIRQIT-G and SIRQIT-CG by reporting on numerical tests involving four artificially constructed eigenvalue problems. Each problem is constructed with $n=10$, $B=\operatorname{diag}(1,2,\ldots,10)$, U generated by applying the

Gram-Schmidt procedure in the B inner product to a random set of n vectors, and $A = BUU^TB$. (Although these methods are intended for problems with very large n, the purpose of this section is to illustrate simply the accuracy and sharpness of the rates developed in Sec. IV. Extensive numerical tests and comparisons with the Lanczos–conjugate-gradient method are left to a later paper.) The problems differ according to choices for the entries of the diagonal matrix of eigenvalues, Λ. (Note that there is no loss of generality by restricting our attention to diagonal B, since SIRQIT is invariant under orthogonal transformations.) We also assume that $p = 4$ and that only the first three eigenvalues and their eigenvectors are desired.

The first column of Table 1 specifies the problem by listing the diagonal entries of Λ in order, while the second indicates to which column of X the rates refer. The third, fourth, and fifth columns depict the rates $r = r_1 + r_2 - r_1 r_2$ predicted by Theorem 2, with the more pessimistic rates $r' = r_1' + r_2' - r_1' r_2'$ of Corollary 1 appearing in parentheses. The last column contains a selected average of rates r observed from numerical experiments with SIRQIT-G, while SIRQIT-CG is included for comparison. (SIRQIT-CG for these examples was restarted every third iteration, since longer restart periods were generally less efficient.)

We finish with an example that helps to clarify the statements of Sec. IV concerning the eigenproblems (11) and (12). Specifically, the effect of the condition number of B on r_1 and r_2 of Theorem 2 depends upon the subspaces span(U_1) and span(U_2). However, the pessimistic values r_1' and r_2' of Corollary 1 are determined independently of these subspaces. As this example illustrates, although these bounds can sometimes be attained, it is also possible that the condition of B has no effect at all.

EXAMPLE. Suppose $\frac{1}{3} > \varepsilon > 0$, and let $V = (v_1\ v_2\ v_3\ v_4\ v_5)$, where $v_1 = (1+\varepsilon)^{-1/2}(e_1 + e_5)$, $v_2 = (1+\varepsilon^{-1})^{-1/2}(e_1 - \varepsilon^{-1}e_5)$, $v_3 = e_2$, $v_4 = e_3$, and $v_5 = e_4$. Note that $V^TBV = I$. Choosing U by permuting columns of V in some way, let $B = \text{diag}(1,1,1,1,\varepsilon)$, $\Lambda = \text{diag}(0,1,2,3,4)$, and $A = BU\Lambda U^TB$. Fixing $p = 2$ and $j = 1$, then Corollary 1 yields the estimates

$$r_1' = \left(\frac{2-\varepsilon}{2+\varepsilon}\right)^2 \quad \text{and} \quad r_2' = \frac{(1-\varepsilon)^2}{8\varepsilon + (1-\varepsilon)^2}.$$

However, depending upon the permutation used to define U, Theorem 2 provides the sharper estimates as follows:

(i) If $U = (v_1\ v_3\ v_4\ v_5\ v_2)$, then

$$r_1 = \left(\frac{5\varepsilon - 3}{11\varepsilon + 3}\right)^2 \quad \text{and} \quad r_2 = 0.$$

TABLE 1

Eigenvalues Λ	Column	Predicted $r_1(r_1')$	Predicted $r_2(r_2')$	Predicted by Theorem 2 (Corollary 1)	Observed: SIRQIT-G	Observed: SIRQIT-CG
0,1,2,...,9	1	.55(.85)	.20(.60)	.64(.94)	.41	.29
	2	.58(.88)	.12(.57)	.63(.95)	.44	.26
	3	.62(.90)	.03(.50)	.63(.95)	.61	.33
0,1,2,3,10,11,...,15	1	.51(.77)	.11(.38)	.56(.86)	.49	.20
	2	.52(.77)	.06(.31)	.55(.84)	.37	.22
	3	.52(.78)	.01(.20)	.52(.82)	.51	.17
0,10,20,30,31,32,...,36	1	.51(.71)	.32(.66)	.67(.90)	.37	.27
	2	.51(.72)	.22(.66)	.62(.90)	.46	.23
	3	.51(.76)	.08(.65)	.55(.92)	.50	.11
0,0,1,1,2,3,...,7	1	.62(.89)	.13(.50)	.67(.95)	.30	.11
	2	.62(.89)	.13(.50)	.67(.95)	.35	.33
	3	.71(.94)	0(0)	.71(.94)	.36	.29

(ii) If $U=(v_3\ v_1\ v_4\ v_5\ v_2)$, then

$$r_1=\left(\frac{5\varepsilon-3}{11\varepsilon+3}\right)^2 \quad \text{and} \quad r_2=\frac{(1-\varepsilon)^2}{16\varepsilon+(1-\varepsilon)^2}.$$

(iii) If $U=V$, then $r_1=\frac{1}{9}$ and $r_2=0$.

Observe that the only difference between the first two examples is that the first two columns of U are interchanged. Further, in (iii), the estimates are independent of the condition, ε^{-1}, of B. The impact of B, therefore, in distorting not only the orthogonality of $B^{1/2}U_1$ and $B^{1/2}U_2$, but also the orthonormality of $B^{1/2}U_2$ with itself, is of more consequence than the condition number of B alone. In (i), the presence of $\bar{\Lambda}_1$ serves to dampen the effects of $U_1^TB^2U_2$. Finally, (ii) provides an example for which r_1 and r_2 are of comparable size.

As a final comment, we wish to again emphasize that this paper reflects more of an attempt to analyze some theoretical aspects of the simultaneous gradient-type methods (with numerical illustrations) than to provide specific guidance to the proper choice of those techniques. However, it should be noted that of the gradient-type methods, SIRQIT-G2 is probably the most robust, but its greater computational cost suggests that it is perhaps best used to start SIRQIT-G (using perhaps G in place of $\hat{G}$). In terms of efficiency, however, SIRQIT-CG should prove to be the best choice in general.

REFERENCES

1 K. J. Bathe and E. L. Wilson, Solution methods for eigenvalue problems in structural mechanics, *Internat. J. Numer. Methods Engrg.* 6:213–226 (1973).

2 R. Gruber, Finite hybrid elements to compute the ideal magnetohydrodynamic spectrum of an axisymmetric plasma, *J. Computational Phys.* 26:379–389 (1978).

3 C. Lanczos, An iteration method for the solution of eigenvalue problems of linear differential and integral operators, *J. Res. Nat. Bur. Standards* 45:255–282 (1950).

4 J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford U. P. (Clarendon), New York, 1965.

5 H. R. Schwarz, The eigenvalue problem $(A-\lambda B)x=0$ for symmetric matrices of high order, *Comput. Methods Appl. Mech. Engrg.* 3:11–28 (1974).

6 M. R. Hestenes and W. Karush, Solutions of $Ax=\lambda Bx$, *J. Res. Nat. Bur. Standards* 49:409–436 (1952).

7 W. W. Bradbury and R. Fletcher, New iterative methods for solution of the eigenproblem, *Numer. Math.* 9:259–267 (1966).

8 P. Laasonen, A Ritz method for simultaneous determination of several eigenvalues and eigenvectors of a big matrix, *Ann. Acad. Sci. Fenn. Ser. A I* 265:3–16 (1959).
9 R. Rutishauser, Computational aspects of F. L. Bauer's simultaneous iteration method, *Numer. Math.* 13:205–223 (1969).
10 S. F. McCormick and T. Noe, Simultaneous iteration for the matrix eigenvalue problem, *J. Linear Algebra Appl.* 16:43–56 (1977).
11 H. R. Schwarz, Two algorithms for treating $Ax=\lambda Bx$, *Comput. Methods Appl. Mech. Engrg.* 12:181–199 (1977).
12 *I.M.S.L. Library 3 Reference Manual*, International Mathematical and Statistical Libraries, Houston, Texas, 1975.
13 B. S. Garbow, J. M. Boyle, J. J. Dongarra and C. B. Moler, *Matrix Eigensystem Routines*—EISPACK *Guide Extension*, Springer, New York, 1977.
14 D. K. Faddeev and V. N. Faddeeva, *Computational Methods of Linear Algebra*, Freeman, San Francisco, 1960.
15 M. R. Hestenes and E. Stiefel, Methods of conjugate gradients for solving linear systems, *J. Res. Nat. Bur. Standards*, 49:409–436 (1952).
16 A. Ruhe, Iterative eigenvalue problems for large symmetric matrices, Report UMINF-31.72, Department of Information Processing, Univ. Of Umeå, 15 Nov. 1973.
17 G. W. Stewart, A bibliographical tour of the large, sparse generalized eigenvalue problem, *Sparse Matrix Computations* (J. R. Bunch and D. J. Rose, Eds.), Academic, New York, 1976, pp. 113–130.

Received 2 April 1979; revised 10 February 1980