Behavior of Slightly Perturbed Lanczos and Conjugate-Gradient Recurrences

A. Greenbaum*
New York University
Courant Institute of Mathematical Sciences
251 Mercer Street
New York, New York 10012

Submitted by Beresford N. Parlett

ABSTRACT

The Lanczos and conjugate-gradient algorithms use a three-term recurrence to generate polynomials orthogonal with respect to a certain set of weights on the eigenvalues of the matrix. The roots of these polynomials are taken as approximate eigenvalues, and the weighted deviation of the polynomials from zero on the set of eigenvalues is a measure of the error in the approximate solution to a linear system. If this recurrence is perturbed slightly, as, for example, by rounding errors, then the polynomials constructed will no longer be orthogonal with respect to the desired measure. It is shown, however, that the computed polynomials are orthogonal with respect to a slightly different measure—one with weights on a larger set of points, all of which lie very near to eigenvalues of the matrix. This analogy is exploited to gain information about approximate eigenvalues generated by a perturbed Lanczos recurrence and about the rate of convergence of the perturbed conjugate-gradient algorithm. In particular, it is shown that the Chebyshev error bound holds (to a close approximation) for slightly perturbed conjugate-gradient recurrences, and that a sharper error bound can be expressed in terms of the minimax polynomial on a set of small intervals about the eigenvalues of the matrix.

1. INTRODUCTION

The Lanczos algorithm [3] for computing eigenvalues and eigenvectors of a symmetric matrix and the conjugate-gradient algorithm [2] for solving

*This work was supported by the Applied Mathematical Sciences Program of the U.S. Department of Energy under contract DE-AC02-76ER03077.

655 Avenue of the Americas, New York, NY 10010 0024-3795/89/$3.50

symmetric positive definite linear systems use a three-term recurrence to generate polynomials orthogonal with respect to a certain set of weights on the eigenvalues of the matrix. The roots of these polynomials are taken as approximate eigenvalues, and the weighted deviation of the polynomials from zero on the set of eigenvalues is a measure of the error in the approximate solution to a linear system. If this recurrence is perturbed slightly, as, for example, by rounding errors, then the polynomials constructed will no longer be orthogonal with respect to the desired measure, and many of the attractive convergence properties proved in exact arithmetic may not hold. In particular, it has been observed that slightly perturbed Lanczos recurrences may generate "multiple copies" of some eigenvalues (which could not occur if the eigenvalue approximations were roots of the desired orthogonal polynomials), and that slightly perturbed conjugate gradient recurrences may fail to solve an N-by-N linear system in N steps (which could not occur if the error were given by the deviation of the Nth orthogonal polynomial from zero on the N eigenvalues).

This observed behavior is, however, consistent with, and to be expected from, that of a slightly different set of orthogonal polynomials—polynomials orthogonal with respect to weights on a larger set of points, all of which lie very near to eigenvalues of the matrix. These are the polynomials constructed by an exact Lanczos or conjugate-gradient recurrence applied to a matrix with more distinct eigenvalues than the given matrix, but with all of its eigenvalues lying within small intervals about the eigenvalues of the given matrix. The roots of such polynomials interlace the weighted points, and hence a polynomial may have several roots near one eigenvalue without having any near another. Convergence to the solution of a linear system is not to be expected in N steps, but in a number of steps equal to the number of distinct weighted points.

It is shown that the polynomials generated by a slightly perturbed Lanczos or conjugate-gradient recurrence are of this form—orthogonal polynomials for a set of weights on points in small intervals about the eigenvalues of the matrix—and that the individual components of the vectors generated represent approximately the weighted values of these polynomials across each interval. This analogy is exploited to gain information about approximate eigenvalues generated by a perturbed Lanczos recurrence and about the rate of convergence of the perturbed conjugate-gradient algorithm.

It is shown that the Chebyshev error bound—based on the square root of the ratio of the largest to the smallest eigenvalue—holds approximately for slightly perturbed conjugate-gradient (CG) recurrences when the matrix is not too ill conditioned, because it holds (using a slightly larger ratio) for exact CG recurrences applied to any matrix whose eigenvalues lie within small intervals about the eigenvalues of the given matrix. Bounds based on approxi-

mation on discrete sets—such as the bound described in [1] based on the minimax polynomial on the discrete set of eigenvalues of the matrix—cannot, in general, be expected to hold for perturbed recurrences. The minimax polynomial on a set of small intervals about the eigenvalues can, however, be used to obtain sharper error bounds for slightly perturbed CG recurrences.

Unfortunately, the method of proof for this backward error analysis is by no means direct. A simpler, more direct proof, especially one establishing smaller bounds on the size of the intervals about the eigenvalues, would be of great interest to the author.

The proof relies heavily on the results of Paige [4, 5], who showed that while orthogonality among the Lanczos vectors is completely lost in a finite-precision Lanczos computation, the latest generated Lanczos vector is approximately orthogonal to unconverged Ritz vectors whose Ritz values are well separated from the others. This result is used to show that a given slightly perturbed Lanczos recurrence can be continued—with small additional perturbation terms required to orthogonalize future vectors against each other and against the unconverged Ritz vectors of the original computation—to produce, at some step $N+m$, a coefficient β_{N+m} equal to zero. Paige's analysis is then used to show that all eigenvalues of the final tridiagonal matrix lie within small intervals about the eigenvalues of A.

In the following section, the Lanczos and conjugate-gradient algorithms and their properties in exact arithmetic are briefly described. A well-known identity is established between Lanczos vectors and normalized conjugate-gradient residuals that will later be used to extend analysis of the perturbed Lanczos algorithm to the perturbed conjugate-gradient algorithm. The next section illustrates the similarity between a slightly perturbed Lanczos recurrence and the exact recurrence for the orthogonal polynomials for weights on points within small intervals about the eigenvalues. Implications of this resemblance are discussed, as are the consequences of a proven identity. In the following section, the proof of such an identity is outlined. Next, the results of Paige are presented and notation is established. Finally, the identity is proved by showing that the original recurrence can be continued in such a way as to produce a tridiagonal matrix whose eigenvalues are all close to eigenvalues of A.

Throughout this paper, capital letters will be used to denote matrices and small letters to denote vectors (matrices whose second dimension is 1). Scalars may be denoted by either capital or small letters. For example, Q_J will denote the N-by-J matrix of Lanczos vectors produced at step J of the Lanczos algorithm, and q^j, $j = 1, \ldots, J$, will be its columns. Subscripts will be used to denote the step at which a matrix or scalar is produced, while subscripts on vectors will denote the individual components. Thus, T_J is the tridiagonal matrix produced at step J, but s_i^j is the ith component of the jth

eigenvector of T_J. All scalar quantities will be assumed to be real, and a superscript T will denote the transpose of a matrix or vector.

2. THE ALGORITHMS AND THEIR PROPERTIES IN EXACT ARITHMETIC

The Lanczos recurrence for an N-by-N symmetric matrix A can be written as follows:

(1) Given an N-vector q^1 with $\|q^1\| = 1$, set $\beta_0 = 0$, and for $j = 1, \ldots, N$:
Compute $v^j = Aq^j - \alpha_j q^j - \beta_{j-1} q^{j-1}$,
where $\alpha_j = \langle Aq^j - \beta_{j-1} q^{j-1}, q^j \rangle$.
Compute $\beta_j = \|v^j\|$, and if $\beta_j \neq 0$,
set $q^{j+1} = v^j / \beta_j$.

If q^1 has nonzero components in the direction of all eigenvectors of A, and if A has n distinct eigenvalues, then the algorithm terminates with $\beta_n = 0$, and the eigenvalues of the tridiagonal matrix

$$T_n = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{n-1} \\ & & \beta_{n-1} & \alpha_n \end{bmatrix}$$

are the eigenvalues of A. Multiple eigenvalues are represented only once. The eigenvectors of A are given by

$$Q_n s^i, \qquad i = 1, \ldots, n,$$

where $Q_n = (q^1, \ldots, q^n)$ is an orthogonal matrix and the vectors s^i, $i = 1, \ldots, n$, are the eigenvectors of T_n.

Eigenvalues of earlier tridiagonal matrices

$$T_J = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{J-1} \\ & & \beta_{J-1} & \alpha_J \end{bmatrix}, \qquad J < n,$$

can be used to approximate some of the eigenvalues of A, though it is shown in [8] that for certain initial vectors the matrices T_J, $J < n$, will have no eigenvalues very close to those of A. In practice, however, it is found that tridiagonal matrices generated early in the algorithm generally provide very accurate approximations to extreme eigenvalues. The Ritz vectors,

$$y^i \equiv Q_J s^i, \qquad Q_J = (q^1, \ldots, q^J),$$

$$s^i, \quad i = 1, \ldots, J, \text{ eigenvectors of } T_J,$$

represent approximate eigenvectors of A. (To be precise, Ritz vectors should be denoted as, say, $y^{i,J}$ to show the dependence on the step J. Since it is usually clear from the context which step is being considered, we will omit the superscript J.)

Recurrence (1) can be written in matrix form as

$$AQ_J = Q_J T_J + v^J e^{J^T}, \tag{2.1}$$

where e^J is the J-vector with Jth component one and all other components zero. This form proved useful for the analysis in [4, 5] and [7].

The conjugate-gradient method for solving a symmetric positive definite linear system $Ax = b$ is as follows:

(2) Given an initial guess x^0, compute $r^0 = b - Ax^0$ and set $p^0 = r^0$.
For $k = 1, \ldots, N$:
Compute $x^k = x^{k-1} + a_{k-1} p^{k-1}$,
where $a_{k-1} = \langle r^{k-1}, r^{k-1} \rangle / \langle r^{k-1}, Ap^{k-1} \rangle$.
Set $r^k = r^{k-1} - a_{k-1} Ap^{k-1}$.
Compute $p^k = r^k - b_{k-1} p^{k-1}$,
where $b_{k-1} = -\langle r^k, r^k \rangle / \langle r^{k-1}, r^{k-1} \rangle$.

The error $e^k \equiv x - x^k$ in the kth approximate solution vector can be written as

$$e^k = P_k(A) e^0, \tag{2.2}$$

where P_k is the kth-degree polynomial with value one at zero, which, among all such polynomials that could be placed in Equation (2.2), minimizes the A-norm of the error,

$$\|e^k\|_A \equiv \langle e^k, Ae^k \rangle^{1/2}.$$

That is, for any kth-degree polynomial p_k with $p_k(0) = 1$, we have

$$\|e^k\|_A \leq \|p_k(A)e^0\|_A. \tag{2.3}$$

From this characterization, various error bounds are obtained by substituting different polynomials in (2.3). Substituting the kth-degree Chebyshev polynomial on the interval $[\lambda_{\min}, \lambda_{\max}]$, where $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and largest eigenvalues of A, respectively, yields the bound

$$\begin{aligned}\frac{\|e^k\|_A}{\|e^0\|_A} &\leq 2\left[\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k + \left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k\right]^{-1} \\ &\leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k, \qquad \kappa = \frac{\lambda_{\max}}{\lambda_{\min}}.\end{aligned} \tag{2.4}$$

A sharp error bound (one that, at any given step, can actually be attained with a certain initial vector) is derived by substituting in (2.3) the kth-degree minimax polynomial on the set $\{\lambda_1, \ldots, \lambda_n\}$ of eigenvalues of A. This bound, derived in [1], can be written as

$$\frac{\|e^k\|_A}{\|e^0\|_A} \leq \frac{1}{\left|\displaystyle\sum_{j=1}^{k+1} (-1)^{j-1} \prod_{\substack{i=1 \\ i \neq j}}^{k+1} \left(\frac{\lambda_{\alpha_i}}{\lambda_{\alpha_i} - \lambda_{\alpha_j}}\right)\right|}, \tag{2.5}$$

where the λ_{α_i}'s are the eigenvalues at which the kth minimax polynomial assumes its maximum absolute value on the set of all eigenvalues.

To see the relation between the conjugate-gradient algorithm and the Lanczos algorithm, note from recurrence (2) that the residual vector r^k can be expressed in terms of two previous residuals as

$$r^k = r^{k-1} - a_{k-1}Ar^{k-1} + \frac{a_{k-1}b_{k-2}}{a_{k-2}}(r^{k-2} - r^{k-1}).$$

Defining a normalized vector z^k by

$$z^k \equiv (-1)^k \frac{r^k}{\|r^k\|},$$

it can be seen that z^k satisfies the recurrence

$$z^k = a_{k-1}\frac{\|r^{k-1}\|}{\|r^k\|}Az^{k-1} - \left(1 - \frac{a_{k-1}b_{k-2}}{a_{k-2}}\right)\frac{\|r^{k-1}\|}{\|r^k\|}z^{k-1}$$

$$+ \frac{a_{k-1}b_{k-2}}{a_{k-2}}\frac{\|r^{k-2}\|}{\|r^k\|}z^{k-2}.$$

Finally, noting that the coefficients in the conjugate-gradient algorithm are chosen so that (1) r^k is orthogonal to r^{k-1} and hence z^k is orthogonal to z^{k-1}, and (2) the above recurrence is symmetric (in the sense that the coefficient for z^k in the equation for Az^{k-1} is equal to the coefficient for z^{k-1} in the equation for Az^k), it is seen that the recurrence for z^k is the same as the Lanczos recurrence for q^{k+1}. Thus, if the initial residual r^0 in the conjugate-gradient algorithm is parallel to the initial vector q^1 in the Lanczos procedure, then all residuals r^k will be parallel to the Lanczos vectors q^{k+1}. The polynomial P_k in (2.2) is the characteristic polynomial of the tridiagonal matrix T_k. Its roots are the approximate eigenvalues generated by the Lanczos algorithm.

In the following sections the perturbed Lanczos algorithm is analyzed. The analogy between normalized conjugate-gradient residuals and Lanczos vectors is then used to extend the results to the perturbed conjugate-gradient algorithm.

3. THE PERTURBED LANCZOS RECURRENCE AND THE EXACT RECURRENCE FOR THE ORTHOGONAL POLYNOMIALS FOR A SLIGHTLY DIFFERENT MEASURE

If recurrence (1) is perturbed slightly, then the perturbed recurrence can be written in the form

$$\begin{aligned} (3)\quad & v^j = Aq^j - \alpha_j q^j - \beta_{j-1}q^{j-1} - f^j, \\ & \alpha_j = \langle Aq^j - \beta_{j-1}q^{j-1}, q^j\rangle, \\ & \|f^j\| \leqslant \epsilon\|A\|,\ \epsilon \ll 1, \\ & q^{j+1} = v^j/\beta_j, \\ & \beta_j = \|v^j\|, \end{aligned}$$

or, in a form analogous to (2.1), as

$$AQ_J = Q_J T_J + v^J e^{J^T} + F_J, \qquad F_J \equiv (f^1, \ldots, f^J). \tag{3.1}$$

If the perturbation is the result of rounding errors in a finite-precision computation, then ϵ will be a small multiple of the unit roundoff of the machine [5]. The computed coefficients α_j and β_j may not satisfy the formulas in (3) exactly, but they can be computed to very nearly satisfy these formulas, and the differences can be included in the perturbation terms f^j. These small differences will not be important to the analysis, since it is the eigenvalues of T_j that are of interest, and the eigenvalues of a slightly perturbed symmetric matrix are very close to the eigenvalues of the unperturbed matrix [6].

It will be necessary to consider recurrence (3) in a basis for which the matrix A is diagonal. If we write A in the form

$$A = U\Lambda U^T, \qquad \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_N), \quad UU^T = U^TU = I, \quad \lambda_1 \leqslant \cdots \leqslant \lambda_N,$$

then the vectors in (3) can be seen to satisfy

$$\begin{aligned}
(4)\quad & U^Tv^j = \Lambda U^Tq^j - \alpha_j U^Tq^j - \beta_{j-1}U^Tq^{j-1} - U^Tf^j, \\
& \alpha_j = \langle \Lambda U^Tq^j, U^Tq^j \rangle + \langle f^{j-1}, q^{j-1} \rangle, \\
& \|U^Tf^j\| \leqslant \epsilon \|\Lambda\|, \\
& U^Tq^{j+1} = U^Tv^j/\beta_j, \\
& \beta_j = \|U^Tv^j\|.
\end{aligned}$$

Now consider the problem of constructing the orthonormal polynomials for a set of weights on points very near the eigenvalues of A, or, more generally, for a measure $dw(x)$ with points of increase of $w(x)$ lying in small intervals $[\lambda_i - \delta_i, \lambda_i + \Delta_i]$, $i = 1, \ldots, N$, about the eigenvalues. These polynomials satisfy the following recurrence:

$$\begin{aligned}
(5)\quad & \bar{v}_j(x) = x\bar{q}_j(x) - \bar{\alpha}_j\bar{q}_j(x) - \bar{\beta}_{j-1}\bar{q}_{j-1}(x), \\
& \bar{\alpha}_j = \langle x\bar{q}_j(x), \bar{q}_j(x) \rangle_{dw} \equiv \textstyle\sum_{i=1}^{N} \int_{\lambda_i - \delta_i}^{\lambda_i + \Delta_i} x\bar{q}_j^2(x)\, dw(x), \\
& \bar{q}_{j+1}(x) = \bar{v}_j(x)/\bar{\beta}_j, \\
& \bar{\beta}_j = \|\bar{v}_j(x)\|_{dw} \equiv \Big(\textstyle\sum_{i=1}^{N} \int_{\lambda_i - \delta_i}^{\lambda_i + \Delta_i} \bar{v}_j^2(x)\, dw(x)\Big)^{1/2}, \\
& \Big(\bar{q}_0(x) = 0,\ \bar{q}_1(x) = 1/\Big(\textstyle\sum_{i=1}^{N} \int_{\lambda_i - \delta_i}^{\lambda_i + \Delta_i} dw(x)\Big)^{1/2}\Big).
\end{aligned}$$

Let $\lambda_i + \xi_{ij}$ be the point in $(\lambda_i - \delta_i, \lambda_i + \Delta_i)$ at which $\bar{v}_j^2(x)$ attains its mean value with respect to the measure $dw(x)$:

$$\bar{v}_j^2(\lambda_i + \xi_{ij}) \int_{\lambda_i - \delta_i}^{\lambda_i + \Delta_i} dw(x) = \int_{\lambda_i - \delta_i}^{\lambda_i + \Delta_i} \bar{v}_j^2(x)\, dw(x). \tag{3.2}$$

Define N-vectors $\bar{v}^j$ and $\bar{q}^{j+1}$ by

$$\bar{v}_i^j = \bar{v}_j(\lambda_i + \xi_{ij})w_i, \qquad \bar{q}_i^{j+1} = \bar{q}_{j+1}(\lambda_i + \xi_{ij})w_i, \tag{3.3}$$

$$w_i^2 \equiv \int_{\lambda_i - \delta_i}^{\lambda_i + \Delta_i} dw(x), \qquad i = 1, \ldots, N.$$

These vectors satisfy the recurrence

(6) $\bar{v}^j = \Lambda\bar{q}^j - \bar{\alpha}_j\bar{q}^j - \bar{\beta}_{j-1}\bar{q}^{j-1} - \bar{f}^j,$
$\bar{f}_i^j = -\xi_{ij}\bar{q}_j(\lambda_i + \xi_{ij})w_i - (\lambda_i - \bar{\alpha}_j)[\bar{q}_j(\lambda_i + \xi_{ij}) - \bar{q}_j(\lambda_i + \xi_{i,j-1})]w_i$
$+ \bar{\beta}_{j-1}[\bar{q}_{j-1}(\lambda_i + \xi_{ij}) - \bar{q}_{j-1}(\lambda_i + \xi_{i,j-2})]w_i, \quad i = 1, \ldots, N,$
$\bar{q}^{j+1} = \bar{v}^j/\bar{\beta}_j,$

where the coefficients $\bar{\alpha}_j$ and $\bar{\beta}_j$ are given by

$$\bar{\alpha}_j = \sum_{i=1}^{N} (\lambda_i + \eta_{ij}) \int_{\lambda_i - \delta_i}^{\lambda_i + \Delta_i} \bar{q}_j^2(x)\, dw(x) \qquad \text{for some} \quad \eta_{ij} \in (-\delta_i, \Delta_i),$$

$$= \langle \Lambda\bar{q}^j, \bar{q}^j \rangle + \sum_{I=1}^{N} \eta_{ij}(\bar{q}_i^j)^2,$$

$$\bar{\beta}_j = \|\bar{v}^j\|.$$

Comparing (4) with (6), it can be seen that if the initial vector q^1 in a perturbed Lanczos recurrence satisfies

$$(U^T q^1)_i = \frac{w_i}{\left(\sum_{k=1}^{N} w_k^2\right)^{1/2}}, \qquad i = 1, \ldots, N, \tag{3.4}$$

and if the perturbation terms f^j are given by

$$(U^T f^j)_i = -\xi_{ij}\bar{q}_j(\lambda_i + \xi_{ij})w_i - (\lambda_i - \bar{\alpha}_j)\left[\bar{q}_j(\lambda_i + \xi_{ij}) - \bar{q}_j(\lambda_i + \xi_{i,j-1})\right]w_i$$
$$+ \bar{\beta}_{j-1}\left[\bar{q}_{j-1}(\lambda_i + \xi_{ij}) - \bar{q}_{j-1}(\lambda_i + \xi_{i,j-2})\right]w_i$$
$$+ \left(\langle f^{j-1}, q^{j-1} \rangle - \sum_{i=1}^{N} \eta_{ij}(\bar{q}_i^j)^2\right)(U^T q^j)_i, \tag{3.5}$$

then the coefficients α_j, β_j, $j = 1, \ldots,$ generated by the perturbed recurrence are related to the coefficients $\bar{\alpha}_j, \bar{\beta}_j$, $j = 1, \ldots,$ generated by an exact recurrence for the orthonormal polynomials corresponding to the measure $dw(x)$ by

$$\alpha_j = \bar{\alpha}_j + \langle f^{j-1}, q^{j-1} \rangle - \sum_{i=1}^{N} \eta_{ij} (\bar{q}_i^j)^2, \qquad \beta_j = \bar{\beta}_j. \tag{3.6}$$

It follows that the roots of the orthonormal polynomials for $dw(x)$ differ from the eigenvalues generated by the perturbed Lanczos recurrence by no more than

$$\epsilon \|A\| + \max_{i=1,\ldots,N} \left[\max(|\delta_i|, |\Delta_i|) \right]. \tag{3.7}$$

Individual components $(U^T q^j)_i$ of the generated Lanczos vectors satisfy

$$(U^T q^j)_i^2 = \int_{\lambda_i - \delta_i}^{\lambda_i + \Delta_i} \bar{q}_j^2(x)\, dw(x), \qquad i = 1, \ldots, N. \tag{3.8}$$

By considering measures $dw(x)$ with the points of increase of $w(x)$ lying in very tiny intervals $[\lambda_i - \delta_i, \lambda_i + \Delta_i]$ about the eigenvalues of A, one can find orthonormal polynomials for which the terms $\bar{f}^j$ in (6) are arbitrarily small. It follows that if it is known *only* that the perturbation terms in a perturbed Lanczos recurrence are smaller than some given value (e.g., $\epsilon \|A\|$), then it is *possible* that these terms will be of the form (3.5) for some measure $dw(x)$ that weights several points or an entire small interval about each eigenvalue of A. In this case, the perturbed recurrence will generate (approximately) the coefficients of the exact orthonormal polynomials for $dw(x)$. These polynomials may have multiple roots within the intervals $[\lambda_i - \delta_i, \lambda_i + \Delta_i]$, and so this is also to be expected of the characteristic polynomials of the tridiagonal matrices generated by a perturbed Lanczos recurrence. The Nth orthogonal polynomial for the measure $dw(x)$ may not be zero at all weighted points, and so a perturbed conjugate-gradient recurrence cannot be expected to generate a zero residual vector at step N.

It is known that for matrices with one or a few eigenvalues much larger than the others, finite-precision Lanczos computations tend to converge to the large, well-separated eigenvalues quickly and to produce several close approximations to such eigenvalues before producing any approximations to some of the interior eigenvalues. This behavior is also seen in the exact orthogonal polynomials for a set of weights on points in small intervals about

such eigenvalues (i.e., in the exact Lanczos algorithm applied to a matrix with several eigenvalues near each eigenvalue of the given matrix).

In Table 1, the eigenvalues generated by an "exact" (double precision, full reorthogonalization) Lanczos recurrence applied to the matrix

$$\bar{A} = \operatorname{diag}((\lambda_{il},\ l = 1,\ldots,11),\ i = 1,\ldots,10),$$

$$\lambda_{il} = \lambda_i + (l-6)\times 2.\text{E}-9, \quad l = 1,\ldots,11,$$

$$\lambda_i = i,\ i = 1,\ldots,9, \quad \lambda_{10} = 200,$$

are given. The initial vector was taken to have all components of equal size. Also given are the sizes of the perturbation terms that would result in the generation of (approximately) the same eigenvalues by a perturbed Lanczos recurrence applied to the matrix

$$A = \operatorname{diag}(\lambda_i,\ i = 1,\ldots,10).$$

Column 1 shows the sizes of the perturbation terms necessary to produce (in addition to the same eigenvalue approximations) Lanczos vectors satisfying (3.8). These were computed using the formula (3.5). Note that the size of these items varies considerably from step to step. Since roundoff terms in a finite-precision Lanczos computation usually exhibit less variation in size, this suggests that the equality (3.8) might be only roughly approximated in finite-precision computations. If individual components in (3.8) are allowed to differ slightly, we can construct perturbation terms of more uniform size that still result in the generation of the same eigenvalue approximations. The sizes of one such set of perturbations are given in column 2. These perturbations terms were computed by varying individual components of the perturbed Lanczos vectors slightly, while keeping fixed the appropriate sums of components that determine the coefficients in the recurrence.

Table 2 shows the $\bar{A}$-norm of the error in the exact CG recurrence applied to $\bar{A}$, with initial residual having equal components and initial error having A-norm one. Also shown is the Chebyshev bound (2.4) for the error in the exact CG iterates generated by a recurrence with matrix A. The sharp error bound (2.5) for matrix A is also given, as is a "modified" Chebyshev error bound, based on the polynomial

$$p_k(x) = \frac{200 - x}{200} T_{k-1}(x),$$

$$T_{k-1}(x) = (k-1)\text{st Chebyshev polynomial on } [1,9].$$

TABLE 1

EIGENVALUE APPROXIMATIONS GENERATED BY AN EXACT LANCZOS RECURRENCE FOR $\bar{A}$ OR AN EQUIVALENTLY PERTURBED LANCZOS RECURRENCE FOR A

	$\|f^j\|_\infty$		Eigenvalue approximations									
Step	Equiv. perturb. satisfying (3.8)	Equiv. perturb. approx. (3.8)	1	2	3	4	5	6	7	8	9	200
1	1.8E − 15	2.8E − 8										24.50
2	1.0E − 14	8.8E − 8					4.93					199.69
3	1.2E − 11	1.6E − 7		2.39						7.56		200.00
4	6.6E − 06	3.4E − 9		1.55			4.97			8.42		200.00
5	6.4E − 04	2.6E − 7	1.22			3.51		6.45			8.77	200.00
6	3.6E − 06	3.1E − 9	1.08		2.71		4.99		7.26		8.92	200.00
7	4.3E − 12	1.0E − 7	1.03	2.39		4.24		6.23		7.86	9.00	200.00
8	1.6E − 12	1.8E − 8	1.02	2.29		4.01		5.95		7.70	8.98	196.45 200.00
9	6.1E − 13	4.1E − 8	1.00	2.08	3.42		4.99		6.56	7.91	9.00	200.00 200.00
10	2.4E − 06	2.7E − 9	1.00	2.01	3.10	4.34		5.65	6.89	7.99	9.00	200.00 200.00
11	6.1E − 04	7.5E − 7	1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00	9.00	200.00 200.00

TABLE 2

ACTUAL ERROR IN THE EXACT CG ITERATES FOR $\bar{A}$ AND ERROR BOUNDS BASED ON A

Step	Actual error for exact CG on $\bar{A}$, $\|e^k\|_{\bar{A}}$	Chebyshev bound for A or $\bar{A}$	Modified Chebyshev bound for A	Minimax bound for A	Minimax bound for any matrix whose evals. are within 1.E − 8 of those of A
1	.93	.99	1.0	.99	.99
2	.60	.96	.80	.79	.79
3	.36	.92	.47	.46	.46
4	.20	.86	.25	.24	.24
5	.10	.79	.12	.12	.12
6	.047	.72	.062	.058	.058
7	.025	.65	.031	.024[a]	.050
8	.018	.58	.016[a]	.008[a]	.023
9	.006	.52	.008	.002[a]	.008
10	.001	.46	.004	0.[a]	.002
11	.42E − 7	.40	.002	0.[a]	.35E − 6

[a]Step at which error bound is exceeded.

The maximum absolute value of this polynomial on the set of eigenvalues of A is bounded by

$$2\left[\left(\frac{\sqrt{\kappa'}-1}{\sqrt{\kappa'}+1}\right)^{k-1}+\left(\frac{\sqrt{\kappa'}+1}{\sqrt{\kappa'}-1}\right)^{k-1}\right]^{-1} \leqslant \frac{1}{2^{k-2}}, \qquad k \geqslant 2, \quad \kappa' = 9.$$

Note that since the condition number of $\bar{A}$ is approximately the same as that of A, the Chebyshev bound based on A must also hold to a close approximation for $\bar{A}$, and hence for any slightly perturbed recurrence with A that happens to generate iterates having approximately the same size error as the exact iterates for $\bar{A}$. The error at each step shown in Table 2 is indeed less than the Chebyshev bound. At certain steps, however, the error is greater than the bounds based on the minimax polynomial for A or the modified Chebyshev polynomial with a root at $x = 200$. These bounds cannot be expected to hold for the CG algorithm applied to $\bar{A}$, since the polynomials on which they are based may be large at some eigenvalues of $\bar{A}$. Thus, in general, such bounds cannot be expected to hold for perturbed CG recurrences with A. In the last column of Table 2 is shown an error bound based on the minimax polynomial for the union of intervals $\bigcup_{i=1}^{10}[\lambda_i - 1.\text{E} - 8, \lambda_i + 1.\text{E} - 8]$. The minimax polynomial on this union of intervals—i.e., the kth-degree polynomial with value one at zero whose maximum deviation

from zero on the intervals is as small as possible—was computed using the Remes exchange algorithm. Table 2 shows the maximum absolute value of this polynomial over the intervals, and so, by (2.3), this error bound holds for the exact CG algorithm applied to $\bar{A}$, or to A, or to any matrix whose eigenvalues lie within these intervals.

We have shown that the exact Lanczos-CG recurrence for a matrix whose eigenvalues are clustered in small intervals can be thought of as a slightly perturbed recurrence for a matrix with one or a few eigenvalues from each interval. The goal of the remainder of this paper will be to show that *every* slightly perturbed recurrence is equivalent to an exact recurrence for a matrix whose eigenvalues lie in small intervals about the eigenvalues of the given matrix. From this it will follow that the Chebyshev error bound holds approximately for slightly perturbed CG recurrences and that a sharper error bound can be expressed in terms of the maximum absolute value of the minimax polynomial on the union of these small intervals. While a slightly perturbed Lanczos recurrence might generate several eigenvalue approximations within a single interval, it could not generate (at a single step) more than one eigenvalue approximation between a pair of consecutive intervals.

4. MAIN THEOREMS AND OUTLINE OF PROOF

While the details of the theorems and proofs in the following sections are gory, the basic ideas are not. These ideas will be presented here.

To show that the tridiagonal matrix T_J generated at step J of a perturbed Lanczos recurrence applied to matrix A is the same as that generated by an exact Lanczos recurrence applied to a matrix whose eigenvalues are all close to those of A, it is sufficient (and necessary) to show that T_J can be extended to a larger tridiagonal matrix,

$$T_{J+M} = \left[\begin{array}{c|cccccc} T_J & \beta_J & & & & & \\ \hline \beta_J & \alpha_{J+1} & \beta_{J+1} & & & & \\ & \beta_{J+1} & \cdot & \cdot & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & \cdot & \cdot & \cdot & \\ & & & & \cdot & \cdot & \beta_{J+M-1} \\ & & & & & \beta_{J+M-1} & \alpha_{J+M} \end{array}\right],$$

whose eigenvalues are all close to eigenvalues of A. For the characteristic

polynomials of the successive principal submatrices of T_{J+M} are the orthogonal polynomials for a set of weights—given by the squares of the first components of the eigenvectors of T_{J+M}—on the eigenvalues of T_{J+M} [6]. Thus, T_J is the matrix generated at the Jth step of an exact Lanczos recurrence applied to any matrix whose eigenvalues are the same as those of T_{J+M}, with initial vector having components equal to the first components of the eigenvectors of T_{J+M}, in the directions of the eigenvectors of the matrix.

From this it follows that *any perturbed Lanczos recurrence that eventually generates a tridiagonal matrix whose eigenvalues are all close to eigenvalues of* A, *generates the same eigenvalue approximations at each step as an exact recurrence applied to a matrix* (*e.g., the final tridiagonal matrix*), *whose eigenvalues are all close to eigenvalues of* A. *Thus, the Lanczos method with full reorthogonalization* [5], *or that with selective orthogonalization* (*provided the criterion for orthogonalizing is small enough*) [7], *generate the same eigenvalue approximations at each step as an exact Lanczos recurrence applied to a matrix of order* N, *whose eigenvalues are all close to eigenvalues of* A.

Sometimes finite-precision Lanczos recurrences that do not use reorthogonalization generate (often after step N) a tridiagonal matrix whose eigenvalues are all close to those of A. Paige, in fact, showed [4] that the eigenvalues of the successive tridiagonal matrices "stabilize" only to points near eigenvalues of A. There is no guarantee, however, that the recurrence will ever reach a step at which *all* eigenvalues of the tridiagonal matrix are "stabilized." To prove that the tridiagonal matrix generated at any step J of a slightly perturbed Lanczos recurrence can be extended to a larger one whose eigenvalues are all close to those of A, we will demonstrate a method of continuing the recurrence with carefully chosen small perturbations, specifically designed to achieve the desired result.

It will be shown that if the original recurrence is (hypothetically) continued—making small additional perturbations in order to orthogonalize future vectors against each other and against the unconverged Ritz vectors of the original computation—then the recurrence will reach a step (at or before step $N+m$, if there are m converged Ritz vectors) at which the coefficient β will be zero. Assuming, for simplicity, that this occurs at step $N+m$, the resulting recurrence can be written in matrix form as

$$AQ_{N+m} = Q_{N+m}T_{N+m} + \tilde{F}_{N+m}, \qquad \tilde{F}_{N+m} \equiv (f^1, \ldots, f^{J-1}, \tilde{f}^J, \ldots, \tilde{f}^{N+m}) \tag{4.1}$$

$f^1, \ldots, f^{J-1}$ perturbation terms in original recurrence,
$\tilde{f}^J, \ldots, \tilde{f}^{N+m}$ perturbation terms resulting from reorthogonalizing.

Letting S_{N+m} denote the matrix of eigenvectors of T_{N+m} and Θ_{N+m} the diagonal matrix of eigenvalues, and multiplying (4.1) by S_{N+m} on the right, we have

$$AY_{N+m} = Y_{N+m}\Theta_{N+m} + \tilde{F}_{N+m}S_{N+m}, \qquad Y_{N+m} \equiv Q_{N+m}S_{N+m}. \quad (4.2)$$

For any Ritz vector y^j, $j = 1, \ldots, N+m$, this implies

$$\|Ay^j - y^j\theta_j\| \leq \|\tilde{F}_{N+m}s^j\| \leq \|\tilde{F}_{N+m}\| \leq \left[(J-1)(\epsilon\|A\|)^2 + \sum_{j=J}^{N+m} \|\tilde{f}^j\|^2\right]^{1/2}. \quad (4.3)$$

(The vectors y^j in a perturbed Lanczos recurrence are not the true Ritz vectors, but we will refer to them as such and to the corresponding eigenvalues θ_j as Ritz values.) If the norm of y^j is not too small, this implies that θ_j is close to some eigenvalue λ_i of A:

$$|\theta_j - \lambda_i| \leq \frac{1}{\|y^j\|}\left[(J-1)(\epsilon\|A\|)^2 + \sum_{j=J}^{N+m} \|\tilde{f}^j\|^2\right]^{1/2}. \quad (4.4)$$

Paige showed [4] that Ritz vectors y^j with well-separated Ritz value θ_j have norm close to one. More generally, Paige proved the following theorem relating the eigenvalues of T_{N+m} to those of A:

THEOREM (Paige). *Let T_{N+m} be the tridiagonal matrix generated at step $N+m$ of a slightly perturbed Lanczos recurrence of the form* (4.1), *with $\beta_{N+m} = 0$. Then every eigenvalue of T_{N+m} lies within*

$$\sigma(N+m)^3\xi\|A\| \quad (4.5)$$

of an eigenvalue of A, if the perturbation terms f^j, $j = 1, \ldots, J-1$, and $\tilde{f}^j$, $j = J, \ldots, N+m$ are bounded by

$$\max\{\|f^1\|, \ldots, \|f^{J-1}\|, \|\tilde{f}^J\|, \ldots, \|\tilde{f}^{N+m}\|\} \leq \xi\|A\|. \quad (4.6)$$

Here $\sigma = O(1)$, independent of ξ, N, m, and $\|A\|$.

While the expression (4.5) is usually a large overestimate of the distance between eigenvalues of T_{N+m} and eigenvalues of A, it does prove that if the original perturbation terms, $f^1,\ldots,f^{J-1}$, and the additional terms, $\tilde{f}^J,\ldots,\tilde{f}^{N+m}$, are small enough ($\ll 1/(N+m)^3\|A\|$), then the eigenvalues of T_{N+m} are close to those of A.

Combining Paige's theorem with the previous arguments, we obtain the following theorem relating the eigenvalue approximations generated by a slightly perturbed recurrence to those generated by a certain exact Lanczos recurrence.

THEOREM 1. *The tridiagonal matrix T_J generated at any step J of a perturbed Lanczos recurrence of the form* (3) *is equal to that generated by an exact Lanczos recurrence applied to a matrix whose eigenvalues lie within*

$$\sigma(N+m)^3\max\{\epsilon\|A\|,\|\tilde{f}^J\|,\ldots,\|\tilde{f}^{N+m}\|\} \tag{4.7}$$

of eigenvalues of A, where $\tilde{f}^J,\ldots,\tilde{f}^{N+m}$ are the smallest perturbation terms that will cause the procedure to generate a coefficient β equal to zero at or before step $N+m$. Thus, the eigenvalue approximations generated at each step by a perturbed Lanczos recurrence for A are equal to those generated by an exact Lanczos recurrence for a matrix whose eigenvalues lie within intervals of the size (4.7) *about the eigenvalues of A.*

As argued in Section 2, the conjugate-gradient recurrence for the normalized "residual" vectors,

$$z^k \equiv \frac{(-1)^k r^k}{\|r^k\|},$$

is identical to the Lanczos recurrence. The coefficient formulas of recurrence (2) are not quite equivalent to those of recurrence (1), if the recurrences are perturbed slightly, but they do force approximate orthogonality between z^k and z^{k-1}, and they do generate a symmetric tridiagonal matrix.

A slightly perturbed CG recurrence might be written in the form

(7)
$$\begin{aligned} x^k &= x^{k-1} + a_{k-1}p^{k-1} + \delta_{x^k},\\ r^k &= r^{k-1} - a_{k-1}Ap^{k-1} + \delta_{r^k},\quad k=1,2,\ldots\\ p^k &= r^k - b_{k-1}p^{k-1} + \delta_{p^k},\\ &(x^0 \text{ given},\ r^0 = b - Ax^0 + \delta_{r^0}), \end{aligned}$$

so that the vectors $z^0,\ldots,z^{K-1}$ satisfy the equation

$$AZ_K = Z_K T_K + \beta_K z^K e^{K^T} + G_K,$$

$$Z_K \equiv (z^0,\ldots,z^{K-1}),$$

$$(T_K)_{k,k} \equiv \alpha_k = \frac{1}{a_{k-1}} - \frac{b_{k-2}}{a_{k-2}}$$

$$(T_K)_{k,k+1} = (T_K)_{k+1,k} \equiv \beta_k = \frac{\|r^k\|}{a_{k-1}\|r^{k-1}\|},$$

$$G_K \equiv (g^0,\ldots,g^{K-1}),$$

$$g^{k-1} = \frac{(-1)^k}{\|r^{k-1}\|}\left[\frac{b_{k-2}}{a_{k-2}}\delta_{r^{k-1}} - A\delta_{p^{k-1}} + \frac{1}{a_{k-1}}\delta_{r^k}\right]. \tag{4.8}$$

The inner product of $\beta_k z^k$ with z^{k-1} satisfies

$$\left|\langle \beta_k z^k, z^{k-1}\rangle\right| = \frac{-1}{a_{k-1}\|r^{k-1}\|}\left\langle \delta_{r^k}, \frac{r^{k-1}}{\|r^{k-1}\|}\right\rangle,$$

$$\left|\langle \beta_k z^k, z^{k-1}\rangle\right| \leq \frac{1}{|a_{k-1}|\|r^{k-1}\|}\|\delta_{r^k}\|.$$

If the perturbations are small enough so that $\|g^{k-1}\|$ and $|\langle \beta_k z^k, z^{k-1}\rangle|$ satisfy

$$\|g^{k-1}\| \leq \epsilon\|A\|, \quad \left|\langle \beta_k z^k, z^{k-1}\rangle\right| \leq \epsilon\|A\|, \qquad k = 1,\ldots,K-1, \quad \epsilon \ll 1, \tag{4.9}$$

then the analysis in this and the following sections will apply also to the perturbed CG algorithm. It will establish that the tridiagonal matrix T_K generated by a perturbed CG recurrence for the matrix A is equal to that generated by an exact CG recurrence for a matrix whose eigenvalues lie within intervals of size (4.7) about the eigenvalues of A.

Now, the 2-norm of the "residual" vectors, r^k, $k = 1,\ldots,K$, can be expressed in terms of the 2-norm of the initial residual and the elements of

T_K. Using (4.8) and the coefficient formulas in (2), we can write

$$\|r^k\| = \frac{\beta_k \|r^{k-1}\| \, \|r^{k-2}\|}{\alpha_k \|r^{k-2}\| - \beta_{k-1} \|r^{k-1}\|}, \qquad 2 \leqslant k \leqslant K-1,$$
$$\|r^1\| = \frac{\beta_1}{\alpha_1} \|r^0\|. \tag{4.10}$$

Thus, it will follow that the 2-norms of the vectors r^k, $k = 1, \ldots,$ generated by a perturbed CG recurrence are equal to the 2-norms of the residual vectors generated by an exact CG recurrence, with initial residual of size $\|r^0\|$, for the matrix described in Theorem 1.

In order for this result to be of interest, we need to know that the vectors r^k are closely related to the true residuals, $b - Ax^k$. The following theorem establishes such a relationship.

THEOREM 2. *For $k = 0, 1, \ldots,$ let x^k and r^k satisfy recurrence* (7). *Then the difference between r^k and the true residual, $b - Ax^k$, obeys*

$$\left\| r^k - (b - Ax^k) \right\| \leqslant \|\delta_{r^0}\| + \sum_{j=1}^{k} (\|\delta_{r^j}\| + \|A\delta_{x^j}\|).$$

Proof. Comparing the expressions in (7) for r^k and for $b - Ax^k$, we can write

$$\begin{aligned} r^k - (b - Ax^k) &= r^{k-1} - (b - Ax^{k-1}) + \delta_{r^k} + A\delta_{x^k} \\ &\vdots \\ &= r^0 - (b - Ax^0) + \sum_{j=1}^{k} (\delta_{r^j} + A\delta_{x^j}) \\ &= \delta_{r^0} + \sum_{j=1}^{k} (\delta_{r^j} + A\delta_{x^j}), \end{aligned}$$

from which the desired result follows, upon taking norms on both sides. ∎

Theorem 2 shows that as long as the vectors r^k are larger than the (simple) sum of perturbation terms at steps 0 through k, they approximate

the true residuals. Beyond this point, they may bear no resemblance to the true residuals, and, in fact, finite-precision CG computations continue to reduce the size of the vectors r^k (but not $b - Ax^k$), long after the obtainable precision of the machine has been met. In this paper, we will be concerned not with how accurate a solution can ultimately be obtained, but with the *rate* of convergence while the ultimately obtainable accuracy is still far away. Therefore, we will regard the vectors r^k as true residuals and prove results about the rate at which they are reduced in size.

As argued previously, the fact that the tridiagonal matrix generated by a perturbed CG recurrence is the same as that generated by a certain exact CG recurrence implies that the 2-norms of the residuals in the perturbed recurrence are the same as those of this exact recurrence. It is not the 2-norm of the residual, however, that is minimized by the conjugate-gradient algorithm. It is the A-norm of the error, where A is the matrix to which the algorithm is applied. Thus, we would like to show that the A-norm of the error, $\|e^k\|_A \approx \|A^{-1/2}r^k\|_2$, in a slightly perturbed CG recurrence is approximately equal to the $\bar{A}$-norm of the error in the corresponding exact recurrence for the matrix $\bar{A}$ described in Theorem 1.

To this end, we will establish an approximate identity analogous to (3.8) in Section 3. That is, writing A and $\bar{A}$ (the matrix of Theorem 1) in the form

$$A = U\Lambda U^T, \qquad UU^T = U^TU = I, \quad \Lambda = \operatorname{diag}(\lambda_i,\ i = 1,\ldots,N),$$

$$\bar{A} = \bar{U}\bar{\Lambda}\bar{U}^T, \qquad \bar{U}\bar{U}^T = \bar{U}^T\bar{U} = I, \quad \bar{\Lambda} = \operatorname{diag}(\bar{\lambda}_l,\ l = 1,\ldots,N+m)$$

and defining the eigenvalue clusters C_i, $i = 1,\ldots,N$, of $\bar{A}$ by

$$C_i = \left\{ l \,\middle|\, \bar{\lambda}_l \in \left[\lambda_i - \sigma(N+m)^3\xi\|A\|, \lambda_i + \sigma(N+m)^3\xi\|A\|\right]\right\},$$

it will be shown that individual components of the normalized CG residuals z^k satisfy

$$(U^Tz^k)_i^2 \approx \sum_{l \in C_i} (\bar{U}^T\bar{z}^k)_l^2, \qquad i = 1,\ldots,N, \quad k = 1,2,\ldots,$$

where the vectors $\bar{z}^k$ are the normalized residuals of the corresponding exact CG recurrence.

The following theorem will then establish approximate equality between $\|e^k\|_A$ and $\|\bar{e}^k\|_{\bar{A}}$.

THEOREM 3. *Let r^k be the kth "residual" vector and $e^k \equiv A^{-1}r^k$ the kth "error" vector, in a perturbed CG recurrence for matrix A, satisfying*

(4.9). *Let $\bar{r}^k$ be the kth residual vector and $\bar{e}^k \equiv \bar{A}^{-1}\bar{r}^k$ the kth error vector, in an exact CG recurrence for the matrix $\bar{A}$ described in Theorem* 1, *with initial error $\bar{e}^0$ satisfying $\|\bar{e}^0\|_{\bar{A}} = \|e^0\|_A$. Let $\tau \leqslant \sigma(N+m)^3\xi\|A\|$ be the greatest distance from an eigenvalue of $\bar{A}$ to the nearest eigenvalue of A, and assume that τ is small enough so that*

$$\lambda_1 - \tau > 0.$$

Then $\|e^k\|_A$ is related to $\|\bar{e}^k\|_{\bar{A}}$ by

$$\left|1 - \frac{\|\bar{e}^k\|_{\bar{A}}^2}{\|e^k\|_A^2}\right| \leqslant (\bar{\kappa}^3 + \bar{\kappa}^2)\frac{\tau}{\|A\|} + \bar{\kappa}\sum_{i=1}^{N}\left|(U^T z^k)_i^2 - \sum_{l\in C_i}(\bar{U}^T\bar{z}^k)_l^2\right|$$

$$+ \bar{\kappa}^2\sum_{i=1}^{N}\left|(U^T z^0)_i^2 - \sum_{l\in C_i}(\bar{U}^T z^0)_l^2\right|, \qquad \bar{\kappa} = \frac{\lambda_N + \tau}{\lambda_1 - \tau}.$$

Proof. From the definitions of $\|e^k\|_A$ and $\|\bar{e}^k\|_{\bar{A}}$, we have

$$\|e^k\|_A^2 - \|\bar{e}^k\|_{\bar{A}}^2 = \sum_{i=1}^{N}\frac{1}{\lambda_i}(U^T r^k)_i^2 - \sum_{i=1}^{N}\sum_{l\in C_i}\frac{1}{\bar{\lambda}_l}(\bar{U}^T\bar{r}^k)_l^2$$

$$= \sum_{i=1}^{N}\frac{1}{\lambda_i}(U^T r^k)_i^2 - \sum_{i=1}^{N}\frac{1}{\bar{\lambda}_{C_i}}\sum_{l\in C_i}(\bar{U}^T\bar{r}^k)_l^2,$$

$$\bar{\lambda}_{C_i} \in [\lambda_i - \tau, \lambda_i + \tau].$$

Now expressing r^k as $\|r^k\|z^k$ and $\bar{r}^k$ as $\|\bar{r}^k\|\bar{z}^k$, this becomes

$$\|e^k\|_A^2 - \|\bar{e}^k\|_{\bar{A}}^2 = \|r^k\|^2\sum_{i=1}^{N}\frac{1}{\lambda_i}(U^T z^k)_i^2 - \|\bar{r}^k\|^2\sum_{i=1}^{N}\frac{1}{\bar{\lambda}_{C_i}}\sum_{l\in C_i}(\bar{U}^T\bar{z}^k)_l^2$$

$$= \|r^k\|^2\left[\sum_{i=1}^{N}\left(\frac{1}{\lambda_i} - \frac{1}{\bar{\lambda}_{C_i}}\right)(U^T z^k)_i^2\right.$$

$$\left. + \sum_{i=1}^{N}\frac{1}{\bar{\lambda}_{C_i}}\left[(U^T z^k)_i^2 - \sum_{l\in C_i}(\bar{U}^T\bar{z}^k)_l^2\right]\right]$$

$$+ (\|r^k\|^2 - \|\bar{r}^k\|^2)\sum_{i=1}^{N}\frac{1}{\bar{\lambda}_{C_i}}\sum_{\bar{\lambda}_{C_i}}(\bar{U}^T\bar{z}^k)_l^2, \tag{4.11}$$

and taking absolute values on both sides gives

$$
\begin{aligned}
&\left| \|e^k\|_A^2 - \|\bar{e}^k\|_{\bar{A}}^2 \right| \\
&\quad \leq \|r^k\|^2 \max_{i=1,\ldots,N} \left| \frac{1}{\lambda_i} - \frac{1}{\bar{\lambda}_{C_i}} \right| \\
&\qquad + \|r^k\|^2 \max_{i=1,\ldots,N} \left| \frac{1}{\bar{\lambda}_{C_i}} \right| \sum_{i=1}^{N} \left| (U^T z^k)_i^2 - \sum_{l \in C_i} (\bar{U}^T \bar{z}^k)_l^2 \right| \\
&\qquad + \left| \|r^k\|^2 - \|\bar{r}^k\|^2 \right| \max_{i=1,\ldots,N} \left| \frac{1}{\bar{\lambda}_{C_i}} \right| \\
&\quad \leq \|r^k\|^2 \left| \frac{\tau}{\lambda_1(\lambda_1 - \tau)} \right| + \|r^k\|^2 \frac{1}{\lambda_1 - \tau} \sum_{i=1}^{N} \left| (U^T z^k)_i^2 - \sum_{l \in C_i} (\bar{U}^T \bar{z}^k)_l^2 \right| \\
&\qquad + \left| \|r^k\|^2 - \|\bar{r}^k\|^2 \right| \frac{1}{\lambda_1 - \tau}. \qquad (4.12)
\end{aligned}
$$

From (4.10) it follows that $\|r^k\|$ and $\|\bar{r}^k\|$ satisfy

$$
\frac{\|r^k\|}{\|r^0\|} = \frac{\|\bar{r}^k\|}{\|\bar{r}^0\|}.
$$

Using (4.11) with $k = 0$ gives

$$
\begin{aligned}
&\|r^0\|^2 - \|\bar{r}^0\|^2 \\
&\quad = \frac{-1}{\sum_{i=1}^{N} \frac{1}{\bar{\lambda}_{C_i}} \sum_{l \in C_i} (\bar{U}^T \bar{z}^k)_l^2} \|r^0\|^2 \\
&\qquad \times \left[\sum_{i=1}^{N} \left(\frac{1}{\lambda_i} - \frac{1}{\bar{\lambda}_{C_i}} \right) (U^T z^0)_i^2 + \sum_{i=1}^{N} \frac{1}{\bar{\lambda}_{C_i}} \left[(U^T z^0)_i^2 - \sum_{l \in C_1} (\bar{U}^T \bar{z}^0)_l^2 \right] \right],
\end{aligned}
$$

$$
\begin{aligned}
&\left| \|r^0\|^2 - \|\bar{r}^0\|^2 \right| \leq (\lambda_N + \tau) \|r^0\|^2 \\
&\qquad \times \left[\frac{\tau}{\lambda_1(\lambda_1 - \tau)} + \frac{1}{\lambda_1 - \tau} \sum_{i=1}^{N} \left| (U^T z^0)_i^2 - \sum_{l \in C_i} (\bar{U}^T \bar{z}^0)_l^2 \right| \right]
\end{aligned}
$$

and hence we can write

$$\left|1-\frac{\|\bar{r}^k\|^2}{\|r^k\|^2}\right|$$

$$\leqslant(\lambda_N+\tau)\left[\frac{\tau}{\lambda_1(\lambda_1-\tau)}+\frac{1}{\lambda_1-\tau}\sum_{i=1}^{N}\left|(U^Tz^0)_i^2-\sum_{l\in C_i}(\bar{U}^T\bar{z}^0)_l^2\right|\right]. \quad (4.13)$$

Combining (4.12) and (4.13) gives the desired result:

$$\left|\|e^k\|_A^2-\|\bar{e}^k\|_A^2\right|$$

$$\leqslant\|r^k\|^2\left[\frac{\tau}{\lambda_1(\lambda_1-\tau)}+\frac{1}{\lambda_1-\tau}\sum_{i=1}^{N}\left|(U^Tz^k)_i^2-\sum_{l\in C_i}(\bar{U}^T\bar{z}^k)_l^2\right|\right.$$

$$\left.+\frac{(\lambda_N+\tau)\tau}{\lambda_1(\lambda_1-\tau)^2}+\frac{\lambda_N+\tau}{(\lambda_1-\tau)^2}\sum_{i=1}^{N}\left|(U^Tz^0)_i^2-\sum_{l\in C_i}(\bar{U}^T\bar{z}^0)_l^2\right|\right]$$

$$\leqslant\|e^k\|_A^2\lambda_N\left[\frac{\tau(\lambda_1+\lambda_N)}{\lambda_1(\lambda_1-\tau)^2}+\frac{1}{\lambda_1-\tau}\sum_{i=1}^{N}\left|(U^Tz^k)_i^2-\sum_{l\in C_i}(\bar{U}^T\bar{z}^k)_l^2\right|\right.$$

$$\left.+\frac{\lambda_N+\tau}{(\lambda_1-\tau)^2}\sum_{i=1}^{N}\left|(U^Tz^0)_i^2-\sum_{l\in C_i}(\bar{U}^T\bar{z}^0)_l^2\right|\right]$$

$$\leqslant\|e^k\|_A^2\left[(\bar{\kappa}^3+\bar{\kappa}^2)\frac{\tau}{\|A\|}+\bar{\kappa}\sum_{i=1}^{N}\left|(U^Tz^k)_i^2-\sum_{l\in C_i}(\bar{U}^T\bar{z}^k)_l^2\right|\right.$$

$$\left.+\bar{\kappa}^2\sum_{i=1}^{N}\left|(U^Tz^0)_i^2-\sum_{l\in C_i}(\bar{U}^T\bar{z}^0)_l^2\right|\right]. \quad \blacksquare$$

With Theorems 1–3 established, the goal of the following sections will be to prove that a given slightly perturbed Lanczos recurrence can be continued

with small additional perturbation terms to generate a coefficient β that is zero, and that individual components of the perturbed Lanczos vectors

$$\left(U^T q^{k+1}\right)_i \equiv \left(U^T z^k\right)_i$$

satisfy an approximate identity of the form

$$\left(U^T q^{k+1}\right)_i^2 \approx \sum_{l \in C_i} \left(\bar{U}^T \bar{q}^{k+1}\right)_l^2, \qquad i = 1,\ldots,N, \quad k = 0,1,\ldots,J-1,$$

where $\bar{U}$ and $\bar{q}^{k+1} \equiv \bar{z}^k$ are as defined previously. It will then follow that the eigenvalue approximations generated by the perturbed Lanczos recurrence coincide with those generated by an exact recurrence applied to a "nearby" matrix $\bar{A}$, and that the A-norm of the error in the equivalently perturbed CG recurrence is reduced at approximately the same rate as the $\bar{A}$-norm of the error in the exact recurrence.

5. PAIGE'S THEORY

Paige showed [5] that while orthogonality among the Lanczos vectors is completely lost in a finite-precision computation, the latest-generated Lanczos vector is approximately orthogonal to unconverged Ritz vectors whose Ritz values are well separated from the others. More precisely, let q^{J+1} be the $(J+1)$st Lanczos vector in a recurrence of the form (3) of Section 3. Let T_J be the tridiagonal matrix generated at step J, and assume that T_J satisfies

$$T_J S_J = S_J \Theta_J, \qquad S_J S_J^T = S_J^T S_J = I,$$

$$\Theta_J = \operatorname{diag}(\theta_1,\ldots,\theta_J), \qquad S_J \equiv \left(s^1,\ldots,s^J\right), \qquad \theta_1 \leqslant \cdots \leqslant \theta_J.$$

The number β_{Ji}, defined as β_J times the Jth component of s^i,

$$\beta_{Ji} \equiv \beta_J s_J^i,$$

is approximately equal to the residual norm of $y^i \equiv Q_J s^i$:

$$Ay^i - y^i\theta_i = \beta_J s_J^i q^{J+1} + F_J s^i,$$

$$\|Ay^i - y^i\theta_i\| \approx |\beta_J s_J^i|.$$

(Since the vectors $s^1,\ldots,s^J$ are determined only up to a multiple of plus or minus one, we will assume from here on that the sign is chosen so that each vector has a positive Jth component.) Paige showed that if β_{Ji} is not too small, then the inner product of q^{J+1} with y^i is tiny.

Since q^{J+1} has norm one, this implies that q^{J+1} is approximately orthogonal to the direction of y^i, provided the norm of y^i is not too small. Paige showed that if θ_i is well separated from the other Ritz values, then y^i has norm approximately one. The case when θ_i is not well separated from the other Ritz values but is part of a cluster, and the cluster is well separated from the others, was also considered by Paige. He showed that the sum of squares of the norms of the clustered Ritz vectors is approximately equal to the number of vectors in the cluster. The following theorem summarizes these results:

THEOREM (Paige). *At any step J of a perturbed Lanczos computation satisfying the equations in* (3) *of Section* 3, *the inner product of q^{J+1} with any Ritz vector y^i is given by*

$$\langle q^{J+1}, y^i \rangle = \frac{\epsilon \|A\| \gamma_{Ji}}{\beta_{Ji}}, \qquad \gamma_{Ji} \leqslant O(J). \tag{5.1}$$

The norm of y^i satisfies

$$\left|1 - \|y^i\|^2\right| \leqslant \frac{J(J-1)\gamma\epsilon}{\mu_i}, \qquad \mu_i = \min_{k \neq i} \frac{|\theta_i - \theta_k|}{\|A\|}, \quad \gamma \leqslant O(J). \tag{5.2}$$

The sum of squares of norms of several consecutive Ritz vectors, $y^{i-p},\ldots,y^{i+q}$ obeys

$$\left| p + q + 1 - \sum_{l=i-p}^{i+q} \|y^l\|^2 \right| \leqslant (p+q+1)\frac{J(J-1)\gamma\epsilon}{\mu},$$

$$\mu = \min_{\substack{l \in \{i-p,\ldots,i+q\} \\ k \notin \{i-p,\ldots,i+q\}}} \frac{|\theta_l - \theta_k|}{\|A\|}. \tag{5.3}$$

Here $O(\ )$ denotes a constant independent of J, ϵ, and $\|A\|$ times the quantity inside parentheses, plus higher-order terms in ϵ.

Using this theorem, it is not hard to show that the inner product of two unconverged Ritz vectors with well-separated Ritz values is approximately zero:

THEOREM (Paige). *If (y^i, θ_i) and (y^k, θ_k) are two different Ritz pairs at step J, then the inner product $\langle y^i, y^k \rangle$ is bounded by*

$$|\langle y^i, y^k \rangle| \leq \frac{|\gamma_{Ji}(\beta_{Jk}/\beta_{Ji}) + \gamma_{Jk}(\beta_{Ji}/\beta_{Jk}) + \nu_{ik}|\epsilon}{|\theta_i - \theta_k|/\|A\|},$$

$$\nu_{ik} = \frac{|s^{iT}(Q_J^T F_J - F_J^T Q_J)s^k|}{\epsilon\|A\|} \leq O(J). \tag{5.4}$$

Paige's analysis of the case with clustered Ritz values is somewhat incomplete, in the sense that it does not make clear just which part of the space spanned by the vectors in the cluster, is approximately orthogonal to q^{J+1}. The result (5.3) implies that at least one vector in the cluster has norm greater than or equal to one, and if this vector is unconverged, then (5.1) implies that it is approximately orthogonal to q^{J+1}. But the vector with largest norm could be converged, and there could be unconverged Ritz vectors in the cluster having tiny norms. Then it is unclear whether q^{J+1} is approximately orthogonal to all or any part of the space spanned by these vectors. For our analysis, we would like to find some "representative" vector in the span of the cluster that has norm approximately one and is approximately orthogonal to q^{J+1}, if *any* of the Ritz vectors in the cluster are unconverged.

In proving (5.3), Paige actually established a somewhat stronger result that will be of use in our analysis. This result is stated in the following lemma:

LEMMA (Paige). *Let P_J be the strictly upper triangular part of $Q_J^T Q_J$. For each $t \in \{1, \ldots, J-1\}$, the following inequality holds:*

$$\left| \sum_{l=i-p}^{i+q} \left(s^{lT} P_J\right)_{t+1} s_{t+1}^l \right| \leq t(p+q+1)\frac{\gamma\epsilon}{\mu}, \qquad \gamma, \mu \text{ as in (5.2)–(5.3)}. \tag{5.5}$$

Method of proof: In proving (5.3), Paige showed

$$\left| \sum_{l=i-p}^{i+q} s^{lT} P_J s^l \right| = \left| \sum_{t=1}^{J-1} \sum_{l=i-p}^{i+q} \left(s^{lT} P_J\right)_{t+1} s_{t+1}^l \right|$$

$$\leq (p+q+1)\frac{J(J-1)}{2}\frac{\gamma\epsilon}{\mu}.$$

He proved this by showing that each term inside the sum over t is bounded as in (5.5). ∎

Let C be a subset of consecutive indices from $\{1,\dots,J\}$. Define a vector y^C by

$$y^C = \frac{1}{w_C} \sum_{l \in C} s_J^l y^l, \qquad w_C = \sqrt{\sum_{l \in C} \left(s_J^l\right)^2}. \tag{5.6}$$

From (5.1), it can be seen that if $\beta_J w_C$ is not too small (i.e., if at least one of the Ritz vectors y^l, $l \in C$, is unconverged), then the inner product of q^{J+1} and y^C is tiny. The following theorem shows that if w_C is not too small, then the norm of y^C is approximately one.

THEOREM 4. *The linear combination of Ritz vectors* $\sum_{l \in C} s_J^l y^l$, *satisfies*

$$\left| \sum_{l \in C} \left(s_J^l\right)^2 - \left\| \sum_{l \in C} s_J^l y^l \right\|^2 \right| \leqslant \frac{J(\gamma+\nu)\epsilon}{\mu} + |C| \frac{(J-1)\gamma\epsilon}{\mu}, \tag{5.7}$$

$$\mu = \max_{\substack{l \in C \\ l' \notin C}} \frac{\|\theta_l - \theta_{l'}\|}{\|A\|}, \qquad |C| = \textit{number of elements in } C,$$

$$\gamma = \max_{i=1,\dots,J} |\gamma_{Ji}|, \qquad \nu = \max_{\substack{l \in C \\ l' \notin C}} |\nu_{l',l}|.$$

Proof. By definition, the matrix Q_J of Lanczos vectors satisfies $Q_J = Y_J S_J^T$, where Y_J is the matrix of Ritz vectors at step J. Hence q^J can be expressed as

$$q^J = \sum_{l'=1}^{J} s_J^{l'} y^{l'}.$$

Taking the inner product of this sum with $\sum_{l \in C} s_J^l y^l$ gives

$$\begin{aligned}
\left\langle \sum_{l'=1}^{J} s_J^{l'} y^{l'}, \sum_{l \in C} s_J^l y^l \right\rangle &= \left\| \sum_{l \in C} s_J^l y^l \right\|^2 + \left\langle \sum_{l' \notin C} s_J^{l'} y^{l'}, \sum_{l \in C} s_J^l y^l \right\rangle \\
&= \left\langle q^J, \sum_{l \in C} s_J^l y^l \right\rangle = \left\langle q^J, \sum_{l \in C} s_J^l \sum_{r=1}^{J} q^r s_r^l \right\rangle \\
&= \sum_{l \in C} \left(s_J^l\right)^2 + \sum_{l \in C} s_J^l \sum_{r=1}^{J-1} \langle q^J, q^r \rangle s_r^l.
\end{aligned}$$

Hence we have

$$\sum_{l\in C}\left(s_J^l\right)^2-\left\|\sum_{l\in C}s_J^l y^l\right\|^2=\left\langle\sum_{l'\notin C}s_J^{l'}y^{l'},\sum_{l\in C}s_J^l y^l\right\rangle-\sum_{l\in C}s_J^l\sum_{r=1}^{J-1}\langle q^J,q^r\rangle s_r^l. \tag{5.8}$$

Now, the second term on the right-hand side of (5.8) is just $\Sigma_{l\in C}s_J^l(s^{lT}P_J)_J$. By Paige's lemma (5.5), this is bounded by

$$\left|\sum_{l\in C}s_J^l\left(s^{lT}P_J\right)_J\right|\leqslant(J-1)|C|\frac{\gamma\epsilon}{\mu}. \tag{5.9}$$

The first term on the right-hand side of (5.8) can be bounded using (5.4):

$$\begin{aligned}\left|\left\langle\sum_{l'\notin C}s_J^{l'}y^{l'},\sum_{l\in C}s_J^l y^l\right\rangle\right|&\leqslant\sum_{l'\notin C}\sum_{l\in C}\left|s_J^{l'}s_J^l\langle y^{l'},y^l\rangle\right|\\&\leqslant\sum_{l'\notin C}\sum_{l\in C}\frac{\left[\left|\gamma_{Jl'}\left(s_J^l\right)^2\right|+\left|\gamma_{Jl}\left(s_J^{l'}\right)^2\right|+\left|\nu_{l',l}s_J^{l'}s_J^l\right|\right]\epsilon}{\mu}\\&\leqslant\frac{\epsilon}{\mu}J\left[\max_{i=1,\ldots,J}|\gamma_{Ji}|+\max_{\substack{l\in C\\l'\notin C}}|\nu_{l',l}|\right].\end{aligned} \tag{5.10}$$

Substituting (5.9) and (5.10) into (5.8) gives the desired result. ■

6. NOTATION

A Ritz value θ_i at step J will be said to be *well separated* from the other Ritz values if

$$\min_{k\neq i}\frac{|\theta_k-\theta_i|}{\|A\|}>\mu. \tag{6.1}$$

It will be said to be part of a *cluster* if

$$\min_{k \neq i} \frac{|\theta_k - \theta_i|}{\|A\|} \leqslant \Delta. \tag{6.2}$$

The numbers μ and Δ will later be defined in terms of ϵ (the size of the perturbation terms), and will be chosen so that all Ritz values are either well separated or part of a cluster.

A Ritz vector y^i corresponding to a well-separated Ritz value θ_i will be considered *converged* if β_{J_i} satisfies

$$\beta_{J_i} \leqslant \delta, \tag{6.3}$$

and *unconverged* if β_{J_i} satisfies

$$\beta_{J_i} > \delta. \tag{6.4}$$

A linear combination,

$$y^C = \frac{1}{w_C} \sum_{l \in C} s_J^l y^l, \qquad w_C = \sqrt{\sum_{l \in C} \left(s_J^l\right)^2}, \tag{6.5}$$

of Ritz vectors corresponding to a cluster of Ritz values (which is well separated from the other Ritz values) will be referred to as a *cluster vector*. It will be considered *converged* if

$$\beta_J w_C \leqslant \delta, \tag{6.6}$$

and *unconverged* if

$$\beta_J w_C > \delta. \tag{6.7}$$

The parameter δ will also be defined later in terms of ϵ. The value

$$\theta_C = \frac{1}{2}\left(\max_{l \in C} \theta_l + \min_{l \in C} \theta_l\right) \tag{6.8}$$

will be called the *cluster value*.

7. PROOF THAT A PERTURBED LANCZOS RECURRENCE CAN BE CONTINUED WITH SMALL ADDITIONAL PERTURBATIONS TO GENERATE A ZERO COEFFICIENT β

Using the notation of Sections 5 and 6, assume that J steps of a perturbed Lanczos recurrence have been executed, producing a tridiagonal matrix $T_J = S_J \Theta_J S_J^T$ and Lanczos vectors $q^1, \ldots, q^J$, and v^J. Following the arguments of Section 4, we would like to show that the recurrence can be continued, with small additional perturbations, to generate a coefficient β equal to zero.

Define $\hat{\Theta}_{J-m}$ to be a diagonal matrix whose diagonal entries consist of the well-separated Ritz values that correspond to *unconverged Ritz vectors* and the cluster values that correspond to *unconverged cluster vectors.* Thus m is the number of converged Ritz vectors whose Ritz values are well separated, plus the number of Ritz vectors whose Ritz values are clustered, minus one for each unconverged cluster vector. Define $\hat{Y}_{J-m}$ to be the matrix whose columns are these unconverged Ritz vectors and unconverged cluster vectors, and $\hat{S}_{J-m}$ to be the matrix whose columns are the corresponding *eigenvectors* or *linear combinations of eigenvectors* $[s^C \equiv (1/w_C)\Sigma_{l \in C} s_J^l s^l]$ of T_J. Multiplying the basic recurrence (3.1) by $\hat{S}_{J-m}$ on the right gives

$$\begin{aligned} A\hat{Y}_{J-m} &= \hat{Y}_{J-m}\hat{\Theta}_{J-m} + v^J e^{JT}\hat{S}_{J-m} + F_J\hat{S}_{J-m} + G, \\ G &= Q_J\left(T_J\hat{S}_{J-m} - \hat{S}_{J-m}\hat{\Theta}_{J-m}\right). \end{aligned} \tag{7.1}$$

There are three important properties to observe about the matrix $\hat{Y}_{J-m}$: First, according to Paige's theorem (5.1), v^J is approximately orthogonal to the columns of $\hat{Y}_{J-m}$. Hence, orthogonalizing v^J against the span of these columns to obtain a new vector $v^{J'}$ will result in a small perturbation to the three-term Lanczos recurrence. Second, according to (7.1), $\hat{Y}_{J-m}$ satisfies an approximate identity of the form

$$A\hat{Y}_{J-m} \approx \hat{Y}_{J-m}\hat{\Theta}_{J-m} + v^J e^{JT}\hat{S}_{J-m}.$$

It follows that if successive vectors $q^{J+1} = v^{J'}/\|v^{J'}\|$, $q^{J+2}, \ldots, q^{J+k}$ are constructed to be orthogonal to each other and to the columns of $\hat{Y}_{J-m}$, then the next vector in the recurrence,

$$v^{J+k} = Aq^{J+k} - \alpha_{J+k}q^{J+k} - \beta_{J+k-1}q^{J+k-1},$$

will be approximately orthogonal to the columns of $\hat{Y}_{J-m}$:

$$v^{J+k^T}\hat{Y}_{J-m} = q^{J+k^T}A\hat{Y}_{J-m} \approx q^{J+k^T}v^J e^{J^T}\hat{S}_{J-m} \approx q^{J+k^T}v^{J'}e^{J^T}\hat{S}_{J-m} = 0.$$

Third, it will be shown that, because the only columns of Y_J not present in $\hat{Y}_{J-m}$ are ones for which $\beta_J s_J^i$ or $\beta_J w_C$ is small, the vector $\beta_J q^J$ can be written approximately as a linear combination of the columns of $\hat{Y}_{J-m}$. It follows that if successive vectors are constructed as above, then v^{J+k} is also approximately orthogonal to q^{J+1}:

$$v^{J+k^T}q^{J+1} = q^{J+k^T}Aq^{J+1} = \beta_J q^{J+k^T}q^J \approx q^{J+k^T}\hat{Y}_{J-m}w = 0.$$

The result is that if the original Lanczos recurrence is continued—perturbing the recurrence only to orthogonalize future vectors against each other and against the columns of $\hat{Y}_{J-m}$—then the additional perturbation terms will be small, and the (hypothetically) continued recurrence will reach a step, at or before step $N+m$, at which the coefficient β will be zero.

The following lemmas quantify these results. Here and in the following section, only terms of the lowest order in ϵ will be retained. Essentially the same conclusions hold when the higher-order terms are included.

LEMMA 1. *Let $\hat{Y}_{J-m}$ and G be as defined above. Then the columns of G satisfy*

$$g^j = 0 \qquad \text{if } \hat{s}^j \text{ is an eigenvector of } T_J, \tag{7.2}$$

$$\|g^j\| \leqslant \frac{(c_{\max}-1)\Delta\|A\|}{2}\sqrt{c_{\max}}$$

if $\hat{s}^j$ is a linear combination of eigenvectors of T_J,

$c_{\max} =$ max *number of elements in an unconverged cluster, or* 1 *if there are no unconverged clusters.*

The vector v^J satisfies

$$\|\hat{Y}_{J-m}^T v^J\| \leqslant \|v^J\|\frac{J\gamma\epsilon}{\delta/\|A\|}, \tag{7.3}$$

and $\beta_J q^J$ *satisfies*

$$\left\| \beta_J q^J - \sum_{t=1}^{J-m} \beta_J w_t \hat{y}^t \right\| \leq m\delta, \tag{7.4}$$

$$w_t = \begin{cases} s_J^t & \text{if } \hat{y}^t \text{ is a Ritz vector,} \\ w_C & \text{if } \hat{y}^t \text{ is a cluster vector corresponding to cluster } C. \end{cases}$$

Proof. From the definition of G, we have

$$G \equiv (g^1, \ldots, g^{J-m}),$$

$$g^j = 0 \quad \text{if } \hat{s}^j \text{ is an eigenvector of } T_J,$$

$$g^j = \frac{1}{w_C} \sum_{l \in C} s_J^l y^l (\theta_l - \theta_C)$$

if $\hat{s}^j$ is a linear combination of eigenvectors corresponding to cluster C. Using the definition (6.2) of a cluster, the norm of g^j can be bounded by

$$\|g^j\| \leq \max_{C \text{ an unconverged cluster}} \left\{ \frac{1}{w_C} \sum_{l \in C} |s_J^l| |\theta_l - \theta_C| \|y^l\| \right\}$$

$$\leq \max_{C \text{ an unconverged cluster}} \left\{ \frac{(|C|-1)\Delta\|A\|}{2} \frac{1}{w_C} \sqrt{\left(\sum_{l \in C} |s_J^l|^2 \right) \left(\sum_{l \in C} \|y^l\|^2 \right)} \right\}$$

$$\leq \frac{(c_{\max}-1)\Delta\|A\|}{2} \max_{C \text{ an unconverged cluster}} \sqrt{\sum_{l \in C} \|y^l\|^2},$$

and this last expression can be bounded using (5.3):

$$\|g^j\| \leq \frac{(c_{\max}-1)\Delta\|A\|}{2} \sqrt{c_{\max}\left(1 + \frac{J(J-1)\gamma\epsilon}{\mu}\right)}.$$

Ignoring the term of order ϵ times the others gives the result (7.2).

From Paige's theorem (5.1), the inner product of v^J with an unconverged Ritz vector y^i satisfies

$$|\langle v^J, y^i \rangle| \leq \|v^J\| \frac{\epsilon \|A\| \gamma_{Ji}}{\delta},$$

and the inner product of v^J with an unconverged cluster vector $y^C = (1/w_C)\Sigma_{l \in C} s_J^l y^l$ satisfies

$$\left| \left\langle v^J, \frac{1}{w_C} \sum_{l \in C} s_J^l y^l \right\rangle \right| \leq \|v^J\| \sum_{l \in C} \frac{\epsilon \|A\| \gamma_{Jl}}{\delta}.$$

Hence $\|\hat{Y}_{J-m}^T v^J\|$ is bounded by

$$\|\hat{Y}_{J-m}^T v^J\| \leq \|v^J\| \frac{J\gamma\epsilon}{\delta/\|A\|}.$$

The vector q^J can be written in the form

$$q^J = \sum_{i=1}^{J} s_J^i y^i = \sum_{t=1}^{J-m} w_t \hat{y}^t + \sum_{i \in \Phi} s_J^i y^i,$$

$\Phi = \{$indices of converged Ritz vectors and

Ritz vectors that are part of converged clusters$\}$.

Since for each converged Ritz vector y^i we have

$$\beta_J s_J^i \leq \delta,$$

and for each converged cluster vector $y^C = (1/w_C)\Sigma_{l \in C} s_J^l y^l$ we have

$$\beta_J w_C \leq \delta,$$

we can write

$$\left\| \sum_{i \in \Phi} \beta_J s_J^i y^i \right\| \leq \delta \left[\sum_{y^i \text{ a converged Ritz vector}} \|y^i\| + \sum_{y^C \text{ a converged cluster vector}} \|y^C\| \right].$$

From (5.2), the norm of a Ritz vector with a well-separated Ritz value is approximately one, and from (5.3), the norm of a cluster vector is no greater than about $\sqrt{|C|}$. Hence a conservative bound on the quantity above is

$$\left\| \sum_{i \in \Phi} \beta_J s_J^i y^i \right\| \leq m\delta.$$

From this the result (7.4) follows. ∎

The following lemma shows that the columns of $\hat{Y}_{J-m}$ are approximately orthonormal.

LEMMA 2. *The matrix* $\hat{Y}_{J-m}^T \hat{Y}_{J-m}$ *satisfies*

$$\|I - \hat{Y}_{J-m}^T \hat{Y}_{J-m}\| \leq \max\left\{ \frac{J(J-1)\gamma\epsilon}{\mu}, \frac{(c_{\max}-1)J\phi_1\epsilon}{(\delta/\|A\|)^2\mu} \right\}$$

$$+ (J-m-1)\max\left\{ \frac{\phi_2\epsilon}{(\delta/\|A\|)\mu}, \frac{(c_{\max}-1)\phi_3\epsilon}{(\delta/\|A\|)\mu} \right\}, \quad (7.5)$$

$$\phi_1 = 3\gamma + \nu, \qquad \phi_2 = 2\gamma + \frac{\delta}{\|A\|}\nu, \qquad \phi_3 = 4\gamma + \sqrt{2}\,\frac{\delta}{\|A\|}\nu.$$

Proof. The norm of the matrix $I - \hat{Y}_{J-m}^T \hat{Y}_{J-m}$ is bounded by

$$\|I - \hat{Y}_{J-m}^T \hat{Y}_{J-m}\| \leq \max_{t=1,\ldots,J-m} |1 - \|\hat{y}^t\|^2| + (J-m-1)$$

$$\times \max_{\substack{t=1,\ldots,J-m \\ r \neq t}} |\langle \hat{y}^t, \hat{y}^r \rangle|. \quad (7.6)$$

According to Paige's theorem (5.2), the norm of a well-separated Ritz vector y^i satisfies

$$|1 - \|y^i\|^2| \leq \frac{J(J-1)\gamma\epsilon}{\mu}.$$

From (5.7) in Theorem 4, the norm of an unconverged cluster vector satisfies

$$\left|1-\frac{1}{w_C^2}\left\|\sum_{l\in C} s_j^l y^l\right\|^2\right| \leqslant \frac{1}{w_C^2}\frac{(J(\gamma+\nu)+|C|(J-1)\gamma)\epsilon}{\mu}$$

$$\leqslant \frac{1}{(\delta/\beta_J)^2}\frac{(J(\gamma+\nu)+|C|(J-1)\gamma)\epsilon}{\mu}$$

$$\leqslant \frac{1}{(\delta/\|A\|)^2}\frac{(|C|-1)J\phi_1\epsilon}{\mu}$$

$$\text{if} \quad \phi_1 \geqslant \frac{(1+|C|)\gamma+\nu}{|C|-1}, \quad |C|>1.$$

Hence, whether $\hat{y}^t$ is a Ritz vector or a cluster vector, we can write

$$|1-\|\hat{y}^t\|^2| \leqslant \max\left\{\frac{J(J-1)\gamma\epsilon}{\mu}, \frac{(c_{\max}-1)J\phi_1\epsilon}{(\delta/\|A\|)^2\mu}\right\},$$

$$\phi_1 = 3\gamma+\nu \geqslant \frac{(1+|C|)\gamma+\nu}{|C|-1} \qquad \text{if} \quad |C|>1. \tag{7.7}$$

From Paige's second theorem (5.4), the inner product of two unconverged, well-separated Ritz vectors y^i and y^k obeys

$$|\langle y^i, y^k\rangle| \leqslant \frac{|\gamma_{Ji}(\beta_{Jk}/\beta_{Ji})+\gamma_{Jk}(\beta_{Ji}/\beta_{Jk})+\nu_{ik}|\epsilon}{\mu}$$

$$\leqslant \frac{[|\gamma_{Ji}|(\|A\|/\delta)+|\gamma_{Jk}|(\|A\|/\delta)+|\nu_{ik}|]\epsilon}{\mu}$$

$$\leqslant \frac{2\gamma\epsilon}{(\delta/\|A\|)\mu}+\frac{\nu\epsilon}{\mu}$$

$$\leqslant \frac{\phi_2\epsilon}{(\delta/\|A\|)\mu} \qquad \text{if} \quad \phi_2 \geqslant 2\gamma+\frac{\delta}{\|A\|}\nu.$$

Again using (5.4), we can bound the inner product of an unconverged Ritz

vector and an unconverged cluster vector:

$$\left|\left\langle y^i, \frac{1}{w_C}\sum_{l\in C} s_J^l y^l\right\rangle\right| \leq \frac{1}{w_C}\sum_{l\in C}\left|s_J^l\langle y^i, y^l\rangle\right|$$

$$\leq \frac{1}{w_C}\sum_{l\in C}\frac{\left|\gamma_{Ji}\beta_J(s_J^l)^2/\beta_{Ji} + \gamma_{Jl}\beta_{Ji}/\beta_J + s_J^l\nu_{il}\right|\epsilon}{\mu}$$

$$\leq \frac{1}{w_C}\frac{\epsilon}{\mu}\left|\gamma_{Ji}\beta_J w_C^2/\beta_{Ji} + |C|\gamma\beta_{Ji}/\beta_J + w_C\sqrt{|C|}\,\nu\right|$$

$$\leq \frac{\epsilon}{\mu}\left|\frac{\gamma\|A\|}{\delta} + \frac{|C|\gamma\|A\|}{\delta} + \sqrt{|C|}\,\nu\right| \leq \frac{(|C|-1)\phi_3\epsilon}{(\delta/\|A\|)\mu}$$

$$\text{if} \quad \phi_3 \geq \frac{|C|+1}{|C|-1}\gamma + \frac{(\delta/\|A\|)\sqrt{|C|}\,\nu}{|C|-1}, \quad |C|>1.$$

Finally, using (5.4) to bound the inner product of two unconverged cluster vectors gives

$$\left|\left\langle \frac{1}{w_{C_i}}\sum_{l\in C_i} s_J^l y^l, \frac{1}{w_{C_k}}\sum_{l'\in C_k} s_J^{l'} y^{l'}\right\rangle\right| \leq \frac{1}{w_{C_i}w_{C_k}}\sum_{l\in C_i}\sum_{l'\in C_k}\left|s_J^l s_J^{l'}\langle y^l, y^{l'}\rangle\right|$$

$$\leq \frac{1}{w_{C_i}w_{C_k}}\sum_{l\in C_i}\sum_{l'\in C_k}\frac{\left|\gamma_{Jl}(s_J^{l'})^2 + \gamma_{Jl'}(s_J^l)^2 + s_J^l s_J^{l'}\nu_{ll'}\right|\epsilon}{\mu}$$

$$\leq \frac{1}{w_{C_i}w_{C_k}}\frac{\epsilon}{\mu}\left[|C_i|\gamma w_{C_k}^2 + |C_k|\gamma w_{C_i}^2 + \nu w_{C_i}w_{C_k}\right]$$

$$\leq \frac{\epsilon}{\mu}\left[|C_i|\gamma\frac{w_{C_k}}{w_{C_i}} + |C_k|\gamma\frac{w_{C_i}}{w_{C_k}} + \nu\right]$$

$$\leq \frac{2\gamma c_{\max}\epsilon}{(\delta/\|A\|)\mu} + \frac{\nu\epsilon}{\mu}$$

$$\leq \frac{(c_{\max}-1)\phi_3\epsilon}{(\delta/\|A\|)\mu}$$

$$\text{if} \quad \phi_3 \geq \frac{2c_{\max}\gamma}{c_{\max}-1} + \frac{(\delta/\|A\|)\nu}{c_{\max}-1}, \quad c_{\max}>1.$$

Thus, the inner product of any two vectors $\hat{y}^t$ and $\hat{y}^r$, $r \neq t$, is bounded by

$$|\langle \hat{y}^t, \hat{y}^r \rangle| \leq \max\left\{ \frac{\phi_2 \epsilon}{(\delta/\|A\|)\mu}, \frac{(c_{\max}-1)\phi_3 \epsilon}{(\delta/\|A\|)\mu} \right\}, \tag{7.8}$$

$$\phi_2 = 2\gamma + (\delta/\|A\|)\nu, \qquad \phi_3 = 4\gamma + \sqrt{2}\,(\delta/\|A\|)\nu.$$

Substituting (7.8) and (7.7) into (7.6) gives the desired result (7.5). ■

In order for Lemmas 1 and 2 to establish the desired approximate relations, it is necessary that ϵ be small enough and that μ and δ be defined in such a way that the bounds (7.2)–(7.4) are small, and the bound (7.5) is significantly less than one.

Consider defining a converged Ritz vector y^i to be one for which

$$\beta_{Ji} \leq \sqrt{\epsilon}\,\|A\| \equiv \delta. \tag{7.9}$$

This is the definition of a converged Ritz vector adopted in [7]. Suppose, with this definition, that all unconverged Ritz vectors of the original perturbed recurrence correspond to Ritz values that are separated from each other by a distance much greater than $\sqrt{\epsilon}\,\|A\|$. (That is, if two Ritz values are separated by a distance close to or less than $\sqrt{\epsilon}\,\|A\|$, then the corresponding Ritz values will be part of a converged cluster, if Δ—the maximum relative distance between consecutive Ritz values in a cluster—is defined to be approximately $\sqrt{\epsilon}$.) Then the parameter $c_{\max}$ in Lemmas 1 and 2 is one. If $\mu \gg \sqrt{\epsilon}$ is taken to be the minimum relative separation between Ritz values corresponding to unconverged Ritz vectors, then the bounds in Lemmas 1 and 2 are small:

$$g^j = 0, \qquad j = 1, \ldots, J-m,$$

$$\|\hat{Y}_{J-m}^T v^J\| \leq \|v^J\| J\gamma\sqrt{\epsilon}\,, \qquad \left\| \beta_J q^J - \sum_{t=1}^{J-m} \beta_J w_t \hat{y}^t \right\| \leq m\sqrt{\epsilon}\,\|A\|,$$

$$\|I - \hat{Y}_{J-m}^T \hat{Y}_{J-m}\| \leq \frac{J(J-1)\gamma\epsilon}{\mu} + (J-m-1)\frac{\phi_2\sqrt{\epsilon}}{\mu}.$$

If defining δ by (7.9) and taking μ to be much greater than $\sqrt{\epsilon}$ results in $c_{\max}$ being equal to one, then this is a near-optimal way of defining δ in order to obtain small bounds on both $\|\hat{Y}_{J-m}^T v^J\|$ and $\|\beta_J q^J - \Sigma_{t=1}^{J-m} \beta_J w_t \hat{y}^t\|$. For with this definition, both bounds are of order $\sqrt{\epsilon}$, whereas a smaller δ gives a larger bound on the first quantity, and a larger δ gives a larger bound on the

second. Also, for small enough ϵ, the bound on $\|I - \hat{Y}_{J-m}^T \hat{Y}_{J-m}\|$ will be significantly less than one.

On the other hand, if defining δ by (7.9) and taking μ to be much greater than $\sqrt{\epsilon}$ results in there being unconverged Ritz vectors whose Ritz values are not well separated, then $c_{\max}$ is not one, and the term

$$\frac{(c_{\max}-1)J\phi_1\epsilon}{(\delta/\|A\|)^2\mu} = \frac{(c_{\max}-1)J\phi_1}{\mu}$$

in the bound (7.5) will be large. In this case, a more liberal definition of a converged Ritz vector is needed, say,

$$\delta \equiv \epsilon^{1/4}\|A\|. \tag{7.10}$$

If Ritz values are defined to be part of a *cluster* if their relative separation is less than or equal to $\Delta \leq \epsilon^{1/4}$ and *well separated* if their relative separation is greater than $\mu \geq \epsilon^{1/4}$, then the bounds from Lemmas 1 and 2 can be expressed as

$$\|g^j\| \leq \frac{(c_{\max}-1)\epsilon^{1/4}\|A\|}{2}\sqrt{c_{\max}},$$

$$\|\hat{Y}_{J-m}^T v^J\| \leq |v^J| J\gamma\epsilon^{3/4}, \qquad \left\| \beta_J q^J - \sum_{t=1}^{J-m} \beta_J w_t \hat{y}^t \right\| \leq m\epsilon^{1/4},$$

$$\begin{aligned}\|I - \hat{Y}_{J-m}^T \hat{Y}_{J-m}\| &\leq \max\left\{ \frac{J(J-1)\gamma\epsilon}{\mu}, \frac{(c_{\max}-1)J\phi_1\epsilon^{1/2}}{\mu} \right\} \\ &\quad + (J-m-1)\max\left\{ \frac{\phi_2\epsilon^{3/4}}{\mu}, \frac{(c_{\max}-1)\phi_3\epsilon^{3/4}}{\mu} \right\} \\ &\leq \max\left\{ J(J-1)\gamma\epsilon^{3/4}, (c_{\max}-1)J\phi_1\epsilon^{1/4} \right\} \\ &\quad + (J-m-1)\max\left\{ \phi_2\epsilon^{1/2}, (c_{\max}-1)\phi_3\epsilon^{1/2} \right\}.\end{aligned}$$

In this case, the bounds are of order $\epsilon^{1/4}$, but for small enough ϵ, they will still be small.

In practice, the former case is much more common than the latter. Also, in practice, the eigenvalues of T_J generally occur either in tight groups or as

well-spread-out individuals. Thus, there is a natural choice for the parameters Δ and μ, and μ can be taken to be much larger than Δ, while still including all Ritz values as either clustered (spacing less than or equal to Δ) or well separated (spacing greater than μ).

With Lemmas 1 and 2 established, we can now show that the original Lanczos recurrence can be continued by orthogonalizing future vectors against each other and against the space spanned by the columns of $\hat{Y}_{J-m}$, and that the resulting perturbations to the recurrence will be small. The hypothetically continued recurrence will generate a coefficient β equal to zero at or before step $N+m$, and so it will follow from Theorem 1 that the eigenvalues of the final tridiagonal matrix will all be close to eigenvalues of A.

Assume that ϵ is small enough so that the bound (7.5) on $\|I - \hat{Y}_{J-m}^T \hat{Y}_{J-m}\|$ is significantly less than one, say,

$$\|I - \hat{Y}_{J-m}^T \hat{Y}_{J-m}\| \leq \tfrac{3}{4}. \tag{7.11}$$

Then $\hat{Y}_{J-m}^T \hat{Y}_{J-m}$ is nonsingular, and hence $\hat{Y}_{J-m}$ can be written in the form

$$\hat{Y}_{J-m} = \bar{Y}_{J-m} R,$$

where the columns of $\bar{Y}_{J-m}$ are orthonormal and where R is an upper triangular matrix (the Cholesky factor of $\hat{Y}_{J-m}^T \hat{Y}_{J-m}$).

As a first step in the continuation, replace v^J by

$$v^{J'} = v^J - \bar{Y}_{J-m} \bar{Y}_{J-m}^T v^J$$

to obtain a vector that is orthogonal to all columns of $\bar{Y}_{J-m}$. Note that $v^{J'}$ satisfies

$$v^{J'} = Aq^J - \alpha_J q^J - \beta_{J-1} q^{J-1} - \tilde{f}^J, \qquad \tilde{f}^J = f^J + \bar{Y}_{J-m} \bar{Y}_{J-m}^T v^J, \tag{7.12}$$

where α_J and β_{J-1} are the usual Lanczos coefficients defined in recurrence (3). Next, compute q^{J+1} by

$$q^{J+1} = \frac{v^{J'}}{\beta_{J'}}, \qquad \beta_{J'} = \|v^{J'}\|,$$

so that q^{J+1} satisfies

$$\bar{Y}_{J-m}^T q^{J+1} = 0, \qquad \|q^{J+1}\| = 1.$$

Successive vectors v^{J+k}, $k = 1, \ldots, N+m-J$, are then formed according to

$$v^{J+k} = Aq^{J+k} - \alpha_{J+k} q^{J+k} - \beta_{J+k-1} q^{J+k-1} - \tilde{f}^{J+k},$$

$$\tilde{f}^{J+k} = \left(\bar{Y}_{J-m}, q^{J+1}, \ldots, q^{J+k}\right)\left(\bar{Y}_{J-m}, q^{J+1}, \ldots, q^{J+k}\right)^T \times \left(Aq^{J+k} - \alpha_{J+k} q^{J+k} - \beta_{J+k-1} q^{J+k-1}\right), \tag{7.13}$$

where the normalized vectors q^{J+k+1}, $k = 1, \ldots, N+m-J-1$, satisfy

$$q^{J+k+1} = v^{J+k}/\beta_{J+k}, \qquad \beta_{J+k} = \|v^{J+k}\|.$$

The following lemma bounds the size of the perturbation terms $\tilde{f}^J, \ldots, \tilde{f}^{N+m}$:

LEMMA 3. *Let $\tilde{f}^J$ and $\tilde{f}^{J+k}$, $k = 1, \ldots, N+m-J$, be as defined in* (7.12) *and* (7.13). *The norms of these vectors satisfy*

$$\|\tilde{f}^J\| \leq \epsilon\|A\| + \frac{2J\gamma\epsilon\|A\|}{\delta/\|A\|}, \tag{7.14}$$

$$\|\tilde{f}^{J+1}\| \leq 2\frac{J(J-1)\gamma\epsilon\|A\|}{(\delta/\|A\|)\mu} + 2\sqrt{J}\,\epsilon\|A\| + (c_{\max} - 1)\sqrt{c_{\max}}\sqrt{J-m}\,\Delta\|A\|, \tag{7.15}$$

$$\|\tilde{f}^{J+k}\| \leq 2\sqrt{J}\,\epsilon\|A\| + (c_{\max} - 1)\sqrt{c_{\max}}\,\Delta\|A\| + m\delta, \qquad k = 2, \ldots, N+m-j. \tag{7.16}$$

Proof. Taking norms on each side in the expression (7.12) for $\tilde{f}^J$, we can write

$$\|\tilde{f}^J\| \leq \|f^J\| + \|\bar{Y}_{J-m}\|\|\bar{Y}_{J-m}^T v^J\| \leq \epsilon\|A\| + \|\bar{Y}_{J-m}^T v^J\|. \tag{7.17}$$

From the definition of $\bar{Y}_{J-m}$ and the bound (7.11), we have

$$\|\bar{Y}_{J-m}^T v^J\| \leq \|R^{-1^T}\| \|\hat{Y}_{J-m}^T v^J\| \leq 2\|\hat{Y}_{J-m}^T v^J\|,$$

and using the bound (7.3) from Lemma 1, this becomes

$$\|\bar{Y}_{J-m}^T v^J\| \leq 2\|v^J\| \frac{J\gamma\epsilon}{\delta/\|A\|}. \tag{7.18}$$

Substituting (7.18) into (7.17) and noting that $\|v^J\|$ is bounded by $\|A\|$ gives the desired result (7.14).

When k is 1, (7.13) becomes

$$\tilde{f}^{J+1} = \left(\bar{Y}_{J-m}, q^{J+1}\right)\left(\bar{Y}_{J-m}, q^{J+1}\right)^T\left(Aq^{J+1} - \alpha_{J+1}q^{J+1} - \beta_J q^J\right).$$

By definition of the coefficient α_{J+1}, the expression in parentheses is orthogonal to q^{J+1}, and by construction, q^{J+1} is orthogonal to all columns of $\bar{Y}_{J-m}$. Hence we can write

$$\tilde{f}^{J+1} = \left(\bar{Y}_{J-m}, q^{J+1}\right)\begin{pmatrix} \bar{Y}_{J-m}^T Aq^{J+1} - \beta_J \bar{Y}_{J-m}^T q^J \\ 0 \end{pmatrix}. \tag{7.19}$$

Using (7.1) and the definition of $\bar{Y}_{J-m}$, we can write

$$\begin{aligned}\bar{Y}_{J-m}^T Aq^{J+1} &= R^{-1^T}\left(\hat{\Theta}_{J-m}\hat{Y}_{J-m}^T q^{J+1} + \hat{S}_{J-m}^T e^J v^{J^T} q^{J+1} + \hat{S}_{J-m}^T F_J^T q^{J+1} + G^T q^{J+1}\right) \\ &= R^{-1^T}\left(\hat{S}_{J-m}^T e^J v^{J^T} q^{J+1} + \hat{S}_{J-m}^T F_J^T q^{J+1} + G^T q^{J+1}\right),\end{aligned}$$

while $\bar{Y}_{J-m}^T q^J$ can be written as

$$\bar{Y}_{J-m}^T q^J = R^{-1^T}\hat{Y}_{J-m}^T q^J = R^{-1^T}\hat{S}_{J-m}^T Q_J^T q^J.$$

Hence $\|\tilde{f}^{J+1}\|$ satisfies

$$\begin{aligned}\|\tilde{f}^{J+1}\| &\leq \|R^{-1^T}\|\left[\|\hat{S}_{J-m}^T e^J v^{J^T} q^{J+1} - \beta_J \hat{S}_{J-m}^T Q_J^T q^J\| + \|F_J^T\| + \|G^T\|\right] \\ &\leq 2\|\hat{S}_{J-m}^T e^J v^{J^T} q^{J+1} - \beta_J \hat{S}_{J-m}^T Q_J^T q^J\| + 2\sqrt{J}\,\epsilon\|A\| \\ &\quad + \sqrt{J-m}\,(c_{\max} - 1)\Delta\|A\|\sqrt{c_{\max}},\end{aligned} \tag{7.20}$$

with the latter inequality making use of (7.11) and (7.2). To bound the first term on the right-hand side in (7.20), first note that $v^{J^T} q^{J+1}$ is related to β_J by

$$v^{J^T} q^{J+1} = \frac{1}{\beta_{J'}} v^{J^T} \left(v^J - \bar{Y}_{J-m} \bar{Y}_{J-m}^T v^J \right) = \frac{1}{\beta_{J'}} \left(\beta_J^2 - \| \bar{Y}_{J-m}^T v^J \|^2 \right) = \sqrt{\beta_J^2 - \| \bar{Y}_{J-m}^T v^J \|^2}$$

$$= \beta_J \left(1 - \frac{1}{2} \left(\frac{\| \bar{Y}_{J-m}^T v^J \|}{\beta_J} \right)^2 + O\left(\left(\frac{\| \bar{Y}_{J-m}^T v^J}{\beta_J} \right)^4 \right) \right).$$

Hence, this term is bounded by

$$\| \hat{S}_{J-m}^T e^J v^{J^T} q^{J+1} - \beta_J \hat{S}_{J-m}^T Q_J^T q^J \|$$

$$\leqslant \beta_J \| \hat{S}_{J-m}^T e^J - \hat{S}_{J-m}^T Q_J^T q^J \| + 2\beta_J \left(\frac{J\gamma\epsilon}{\delta / \|A\|} \right)^2. \quad (7.21)$$

The term $\hat{S}_{J-m}^T Q_J^T q^J$ can be written as

$$\hat{S}_{J-m}^T Q_J^T q^J = \hat{S}_{J-m}^T e^J + \begin{pmatrix} \left(\hat{S}^{1^T} P_J \right)_J \\ \vdots \\ \left(\hat{S}^{J-m^T} P_J \right)_J \end{pmatrix},$$

where, as in Paige's lemma (5.5), P_J is the strictly upper triangular part of $Q_J^T Q_J$. For an eigenvector s^i of T_J, corresponding to an unconverged Ritz vector, Paige's lemma implies

$$\left| \left(s^{i^T} P_J \right)_J \right| \leqslant \frac{1}{s_J^i} (J-1) \frac{\gamma\epsilon}{\mu} \leqslant \frac{\beta_J}{\delta} (J-1) \frac{\gamma\epsilon}{\mu},$$

and for a linear combination of eigenvectors s^C corresponding to an unconverged cluster vector, we have

$$\left| \frac{1}{w_C} \sum_{l \in C} s_J^l \left(s^{l^T} P_J \right)_J \right| \leqslant \frac{1}{w_C} (J-1)|C| \frac{\gamma\epsilon}{\mu} \leqslant \frac{\beta_J}{\delta} (J-1)|C| \frac{\gamma\epsilon}{\mu}.$$

Using these results in (7.21) gives

$$\|\hat{S}_{J-m}^T e^J v^{J^T} q^{J+1} - \beta_J \hat{S}_{J-m}^T Q_J^T q^J\| \leqslant \frac{\beta_J^2}{\delta}(J-1)J\frac{\gamma\epsilon}{\mu} + 2\beta_J\left(\frac{J\gamma\epsilon}{(\delta/\|A\|)}\right)^2.$$

Noting that β_J is bounded by $\|A\|$, and ignoring higher powers of ϵ, this becomes

$$\|\hat{S}_{J-m}^T e^J v^{J^T} q^{J+1} - \beta_J \hat{S}_{J-m}^T Q_J^T q^J\| \leqslant \frac{(J-1)J\gamma\epsilon\|A\|}{(\delta/\|A\|)\mu}. \tag{7.22}$$

Substituting this result in (7.20) gives the desired relation (7.15).

For $k > 1$, the expression inside parentheses in (7.13) is again orthogonal to q^{J+k}, and, by construction, q^{J+k} is orthogonal to $\bar{Y}_{J-m}, q^{J+1}, \ldots, q^{J+k-1}$. Thus we can write

$$\tilde{f}^{J+k} = \left(\bar{Y}_{J-m}, q^{J+1}, \ldots, q^{J+k}\right)\begin{pmatrix} \bar{Y}_{J-m}^T A q^{J+k} \\ q^{J+1^T} A q^{J+k} \\ \vdots \\ q^{J+k-1^T} A q^{J+k} - \beta_{J+k-1} \\ 0 \end{pmatrix}. \tag{7.23}$$

Using (7.1) and the definition of $\bar{Y}_{J-m}$, we can write

$$\begin{aligned}
&\bar{Y}_{J-m}^T A q^{J+k} \\
&\quad = R^{-1^T}\left(\hat{\Theta}_{J-m}\hat{Y}_{J-m}^T q^{J+k} + \hat{S}_{J-m}^T e^J v^{J^T} q^{J+k} + \hat{S}_{J-m}^T F_J^T q^{J+k} + G^T q^{J+k}\right) \\
&\quad = R^{-1^T}\left(\hat{S}_{J-m}^T e^J v^{J^T} q^{J+k} + \hat{S}_{J-m}^T F_J^T q^{J+k} + G^T q^{J+k}\right) \\
&\quad = R^{-1^T}\left(\hat{S}_{J-m}^T e^J \left(v^{J'} + \bar{Y}_{J-m}\bar{Y}_{J-m}^T v^J\right)^T q^{J+k} + \hat{S}_{J-m}^T F_J^T q^{J+k} + G^T q^{J+k}\right),
\end{aligned}$$

and so we have

$$\|\bar{Y}_{J-m}^T A q^{J+k}\| \leqslant \|R^{-1^T}\|\left(\|F_J^T\| + \|G^T\|\right) \tag{7.24}$$

$$\leqslant 2\sqrt{J}\,\epsilon\|A\| + (c_{\max} - 1)\sqrt{c_{\max}}\,\Delta\|A\|. \tag{7.25}$$

Using the recurrence (7.13) for Aq^{J+1}, the inner product $q^{J+1^T}Aq^{J+k}$ can be expressed as

$$\begin{aligned} q^{J+1^T}Aq^{J+k} &= v^{J+1^T}q^{J+k} + \alpha_{J+1}q^{J+1^T}q^{J+k} + \beta_J q^{J^T}q^{J+k} \\ &\quad + \left(Aq^{J+1} - \alpha_{J+1}q^{J+1} - \beta_J q^J\right)^T \left(\bar{Y}_{J-m}, q^{J+1}\right) \\ &\quad \times \left(\bar{Y}_{J-m}, q^{J+1}\right)^T q^{J+k} \\ &= \beta_J q^{J^T}q^{J+k}, \end{aligned} \tag{7.26}$$

with the latter equality holding because q^{J+k} is orthogonal to all the other terms in the expression for Aq^{J+1}. But from (7.4) of Lemma 1, $\beta_J q^J$ can be written approximately as a linear combination of the vectors $\hat{y}^1, \ldots, \hat{y}^{J-m}$, to which q^{J+k} is orthogonal. Therefore we can write

$$|q^{J+1^T}Aq^{J+k}| \leqslant m\delta. \tag{7.27}$$

Finally, by construction, we have

$$q^{J+k-1^T}Aq^{J+k} - \beta_{J+k-1} = 0, \qquad k > 1. \tag{7.28}$$

Using (7.26)–(7.28) with (7.25) to bound $\|\tilde{f}^{J+k}\|$ gives the desired result (7.16). ■

If taking δ to satisfy (7.9), Δ to be approximately $\sqrt{\epsilon}$, and μ to be much greater than $\sqrt{\epsilon}$ results in the inequality (7.11) being satisfied, then the bounds of Lemma 3 are of order

$$J^2\gamma\sqrt{\epsilon}\|A\|, \tag{7.29}$$

meaning that they can be expressed as a constant, independent of J and ϵ, times the expression (7.29), plus terms of higher order in ϵ. If defining δ and μ as above results in (7.11) not being satisfied, then, as before, we will define δ to be $\epsilon^{1/4}\|A\|$, and choose $\Delta \leqslant \epsilon^{1/4}$, $\mu \geqslant \epsilon^{1/4}$. Then the bounds of Lemma 3 are of order

$$c_{\max}^{3/2}J\epsilon^{1/4}\|A\|. \tag{7.30}$$

Using the results of Lemma 3, then, and the estimates (7.29)–(7.30), Theorem 1 of Section 4 can be restated as follows:

THEOREM 1′. *The tridiagonal matrix T_J generated at any step J of a perturbed Lanczos recurrence of the form* (3) *is equal to that generated by an exact Lanczos recurrence applied to a matrix whose eigenvalues lie within*

$$\sigma(N+m)^3 O\big(J^2\gamma\sqrt{\epsilon}\,\|A\|\big) \quad \text{or} \quad \sigma(N+m)^3 O\big(c_{\max}^{3/2} J\epsilon^{1/4}\|A\|\big) \tag{7.31}$$

of eigenvalues of A, where $c_{\max}$ and $O(\ )$ are as defined above.

8. PROOF THAT INDIVIDUAL COMPONENTS OF THE PERTURBED LANCZOS VECTORS RESEMBLE CLUSTERED COMPONENTS OF THE CORRESPONDING EXACT VECTORS

When a given perturbed Lanczos recurrence is continued in the manner described in Section 7 to generate a zero coefficient β, the resulting equations can be written in the form (4.1). Translating to a basis in which $A \equiv U\Lambda U^T$ is diagonal, the expression (4.1) can be written as

$$\Lambda\big(U^T Q_{N+m}\big) = \big(U^T Q_{N+m}\big) T_{N+m} + U^T \tilde{F}_{N+m}. \tag{8.1}$$

If T_{N+m} is written in the form

$$T_{N+m} = S_{N+m}\Theta_{N+m}S_{N+m}^T,$$

$$S_{N+m}^T S_{N+m} = S_{N+m}S_{N+m}^T = I, \qquad \Theta_{N+m} = \operatorname{diag}(\theta_1,\ldots,\theta_{N+m}), \tag{8.2}$$

then the exact Lanczos algorithm applied to Θ_{N+m}, with starting vector equal to the first column of S_{N+m}^T, will also generate the tridiagonal matrix T_{N+m}:

$$\Theta_{N+m}S_{N+m}^T = S_{N+m}^T T_{N+m}. \tag{8.3}$$

We would like to establish approximate equality between individual components

$$\big(u^{i^T} q^k\big)^2, \qquad i = 1,\ldots,N, \quad k = 1,\ldots,J,$$

of the original perturbed Lanczos vectors and clustered components

$$\sum_{l \in C_i} \left(S_{N+m}^T\right)_{lk}^2 = \sum_{l \in C_i} \left(s_k^l\right)^2,$$

$$C_i = \{\text{indices of eigenvalues of } \Theta_{N+m} \text{ that are within } \tau \text{ of } \lambda_i\},$$

of the corresponding exact Lanczos vectors, where τ is as defined in Theorem 3. It would then follow from Theorem 3 that the A-norm of the error at each step in the equivalently perturbed CG recurrence for A is approximately equal to the Θ_{N+m}-norm of the error at each step in the exact CG recurrence applied to Θ_{N+m}, provided the A-norm of the initial error in the perturbed recurrence is equal to the Θ_{N+m}-norm of the initial error in the exact recurrence.

For simplicity, we will assume that the eigenvalues of A are well separated:

$$\frac{|\lambda_i - \lambda_j|}{\|A\|} > \rho \gg \frac{2\tau}{\|A\|}, \qquad i \neq j. \tag{8.4}$$

The modifications to Theorem 3 and to the lemmas in this section to relate clustered components

$$\sum_{j \in \chi_i} \left(u^{jT} q^k\right)^2, \qquad \chi_i = \{\text{indices of eigenvalues of } A \text{ that are close to } \lambda_i\}$$

of the original perturbed Lanczos vectors to the corresponding clusters

$$\sum_{l \in \cup_{j \in \chi_i} C_j} \left(s_k^l\right)^2$$

in the exact recurrence are straightforward.

With the usual notation, define the matrix of Ritz vectors Y_{N+m} by

$$Y_{N+m} = Q_{N+m} S_{N+m}.$$

The following lemma establishes approximate equality between individual components $(u^{i^T}2q^k)^2$ and the squared norms of the vectors $\Sigma_{l\in C_i}s_k^l y^l$.

LEMMA 4. *Using the above notation, individual components $u^{i^T}q^k$ of the original perturbed Lanczos vectors satisfy*

$$\left|\left(u^{i^T}q^k\right)^2 - \left\|\sum_{l\in C_i} s_k^l y^l\right\|^2\right| \leqslant \frac{\sqrt{|C_i|}}{\rho\|A\| - \tau}\|\tilde{F}_{N+m}\| + \frac{N-1}{(\rho\|A\| - \tau)^2}\|\tilde{F}_{N+m}\|^2,$$

$$i = 1,\ldots,N, \quad k = 1,\ldots,J. \quad (8.5)$$

Proof. Multiplying (8.1) by S_{N+m} on the right gives

$$\Lambda\left(U^T Y_{N+m}\right) = \left(U^T Y_{N+m}\right)\Theta_{N+m} + U^T\tilde{F}_{N+m}S_{N+m},$$

or

$$(\lambda_i - \theta_l)\left(u^{i^T}y^l\right) = u^{i^T}\tilde{F}_{N+m}s^l, \qquad i = 1,\ldots,N, \quad l = 1,\ldots,N+m.$$

If l' is not in C_i, then from the assumption (8.4) and the definitions of τ, the distance between λ_i and $\theta_{l'}$ is greater than $\rho\|A\| - \tau > 0$, and hence $|u^{i^T}y^{l'}|$ satisfies

$$|u^{i^T}y^{l'}| \leqslant \frac{|u^{i^T}\tilde{F}_{N+m}s^{l'}|}{\rho\|A\| - \tau}, \qquad l' \notin C_i. \quad (8.6)$$

Now, q^k can be written in the form

$$q^k = \sum_{l=1}^{N+m} y^l s_k^l,$$

and so $u^{i^T} q^k$ is given by

$$u^{i^T} q^k = \sum_{l=1}^{N+m} s_k^l \left(u^{i^T} y^l \right) = \sum_{l \in C_i} s_k^l \left(u^{i^T} y^l \right) + \sum_{l' \notin C_i} s_k^{l'} \left(u^{i^T} y^{l'} \right),$$

and the square of this component satisfies

$$\left(u^{i^T} q^k \right)^2 = \left[\sum_{l \in C_i} s_k^l \left(u^{i^T} y^l \right) \right]^2 + 2 \left[\sum_{l \in C_i} s_k^l \left(u^{i^T} y^l \right) \right] \left[\sum_{l' \notin C_i} s_k^{l'} \left(u^{i^T} y^{l'} \right) \right] + \left[\sum_{l' \notin C_i} s_k^{l'} \left(u^{i^T} y^{l'} \right) \right]^2. \tag{8.7}$$

The squared norm of the vector $\Sigma_{l \in C_i} s_k^l y^l$ can be written as

$$\left\| \sum_{l \in C_i} s_k^l y^l \right\|^2 = \sum_{p=1}^{N} \left(\sum_{l \in C_i} s_k^l \left(y^{l^T} u^p \right) \right)^2 = \left(\sum_{l \in C_i} s_k^l \left(y^{l^T} u^i \right) \right)^2 + \sum_{p \neq i} \left(\sum_{l \in C_i} s_k^l \left(y^{l^T} u^p \right) \right)^2. \tag{8.8}$$

Subtracting (8.8) from (8.7) gives

$$\left(u^{i^T} q^k \right)^2 - \left\| \sum_{l \in C_i} s_k^l y^l \right\|^2 = - \sum_{p \neq i} \left(\sum_{l \in C_i} s_k^l \left(y^{l^T} u^p \right) \right)^2 + 2 \left[\sum_{l \in C_i} s_k^l \left(u^{i^T} y^l \right) \right] \left[\sum_{l' \notin C_i} s_k^{l'} \left(u^{i^T} y^{l'} \right) \right] + \left[\sum_{l' \notin C_i} s_k^{l'} \left(u^{i^T} y^{l'} \right) \right]^2. \tag{8.9}$$

Taking absolute values on both sides in (8.9) and using (8.6) to bound the terms on the right-hand side gives the desired result:

$$
\begin{aligned}
&\left|\left(u^{i^T}q^k\right)^2-\left\|\sum_{l\in C_i}s_k^l y^l\right\|^2\right| \\
&\quad\leqslant \sum_{p\neq i}\left(\sum_{l\in C_i}|s_k^l|\frac{|u^{p^T}\tilde{F}_{N+m}s^l|}{\rho\|A\|-\tau}\right)^2 \\
&\qquad+2\left(\sum_{l\in C_i}\left(s_k^l\right)^2\right)^{1/2}\left(\sum_{l\in C_i}\left(u^{i^T}y^l\right)^2\right)^{1/2}\left(\sum_{l'\notin C_i}|s_k^{l'}|\frac{|u^{i^T}\tilde{F}_{N+m}s^{l'}|}{\rho\|A\|-\tau}\right) \\
&\qquad+\left(\sum_{l'\notin C_i}|s_k^{l'}|\frac{|u^{i^T}\tilde{F}_{N+m}s|}{\rho\|A\|-\tau}\right) \\
&\quad\leqslant \frac{1}{(\rho\|A\|-\tau)^2}\sum_{p\neq i}\left(\sum_{l\in C_i}|s_k^l||u^{p^T}\tilde{F}_{N+m}s^l|\right)^2 \\
&\qquad+2\left(\sum_{l\in C_i}\left(s_k^l\right)^2\right)^{1/2}\sqrt{|C_i|}\,\frac{1}{\rho\|A\|-\tau}\sum_{l'\notin C_i}|s_k^{l'}||u^{i^T}\tilde{F}_{N+m}s^{l'}| \\
&\qquad+\frac{1}{(\rho\|A\|-\tau)^2}\left(\sum_{l'\notin C_i}|s_k^{l'}||u^{i^T}\tilde{F}_{N+m}s^{l'}|\right)^2 \\
&\quad\leqslant \frac{1}{(\rho\|A\|-\tau)^2}(N-1)\|\tilde{F}_{N+m}\|^2\sum_{l\in C_i}\left(s_k^l\right)^2 \\
&\qquad+2\left(\sum_{l\in C_i}\left(s_k^l\right)^2\right)^{1/2}\sqrt{|C_i|}\,\frac{1}{\rho\|A\|-\tau}\|\tilde{F}_{N+m}\|\left(\sum_{l'\notin C_i}\left(s_k^{l'}\right)^2\right)^{1/2} \\
&\qquad+\frac{1}{(\rho\|A\|-\tau)^2}\|\tilde{F}_{N+m}\|^2\sum_{l'\notin C_i}\left(s_k^{l'}\right)^2 \\
&\quad\leqslant \frac{N-1}{(\rho\|A\|-\tau)^2}\|\tilde{F}_{N+m}\|^2+\frac{\sqrt{|C_i|}}{\rho\|A\|-\tau}\|\tilde{F}_{N+m}\|. \qquad \blacksquare
\end{aligned}
$$

If, for some index i, the matrix Θ_{N+m} has no eigenvalues near λ_i, then C_i is empty, and Lemma 4 implies that each component $u^{i^T}q^k$, $k=1,\ldots,J$, approximates zero (which can be thought of as the corresponding component of the exact arithmetic vectors).

If the set C_i contains only one element—that is, if only one eigenvalue of Θ_{N+m} approximates λ_i—then $\|\sum_{l\in C_i} s_k^l y^l\|^2$ satisfies

$$\left\| \sum_{l\in C_i} s_k^l y^l \right\|^2 = \left(s_k^{l(i)}\right)^2 \|y^{l(i)}\|^2, \qquad C_i = \{l(i)\}. \tag{8.10}$$

By Paige's theorem (5.2), $\|y^{l(i)}\|^2$ is approximately one:

$$\left|1 - \|y^{l(i)}\|^2\right| \leq \frac{J(J-1)\gamma\xi}{\rho\|A\| - 2\tau}, \tag{8.11}$$

where $\xi\|A\|$ is a bound on the size of the perturbation terms, $f^1,\ldots, f^{J-1}, \tilde{f}^J,\ldots,\tilde{f}^{N+m}$. Using (8.10) and (8.11) with (8.5) gives the desired approximate equality:

$$\begin{aligned}\left|\left(s^{i^T}q^k\right)^2 - \left(s_k^{l(i)}\right)^2\right| &\leq \left|\left(s^{i^T}q^k\right)^2 - \left(s_k^{l(i)}\right)^2\|y^{l(i)}\|^2\right| + \left|\left(s_k^{l(i)}\right)^2\|y^{l(i)}\|^2 - \left(s_k^{l(i)}\right)^2\right| \\ &\leq \frac{\sqrt{|C_i|}}{\rho\|A\| - \tau}\|\tilde{F}_{N+m}\| + \frac{N-1}{(\rho\|A\| - \tau)^2}\|\tilde{F}_{N+m}\|^2 \\ &\quad + \left(s_k^{l(i)}\right)^2 \frac{J(J-1)\gamma\xi}{\rho\|A\| - 2\tau}.\end{aligned} \tag{8.12}$$

Suppose the set C_i contains more than one element. Define vectors $s^{C_i,k}$ and $y^{C_i,k}$ by

$$s^{C_i,k} \equiv \sum_{l\in C_i} s_k^l s^l, \quad y^{C_i,k} \equiv Q_{N+m}s^{C_i,k} = \sum_{l\in C_i} s_k^l y^l, \qquad k=1,\ldots,N+m. \tag{8.13}$$

We would like to show that $\|y^{C_i,k}\|$ and $\|s^{C_i,k}\|$ are approximately equal; for

it would then follow from Lemma 4 that $u^{i^T} q^k$ satisfies

$$\left(u^{i^T} q^k\right)^2 \approx \|y^{C_i k}\|^2 \approx \|s^{C_i, k}\|^2 = \sum_{l \in C_i} \left(s_k^l\right)^2.$$

Defining P_{N+m} to be the strictly upper triangular part of $Q_{N+m}^T Q_{N+m}$, we can write

$$\begin{aligned} \|y^{C_i, k}\|^2 &= s^{C_i, k^T} Q_{N+m}^T Q_{N+m} s^{C_i, k} = s^{C_i, k^T} \left(I + P_{N+m} + P_{N+m}^T\right) s^{C_i, k} \\ &= \|s^{C_i, k}\|^2 + 2 s^{C_i, k^T} P_{N+m} s^{C_i, k}. \end{aligned} \tag{8.14}$$

We will use Paige's lemma (5.5) to shows that $|s^{C_i, k^T} P_{N+m} s^{C_i, k}|$ is small for all k.

First note that multiplying Equation (4.1) on the left by Q_{N+m}^T and equating the right-hand side with its transpose gives

$$\begin{aligned} Q_{N+m}^T A Q_{N+m} &= Q_{N+m}^T Q_{N+m} T_{N+m} + Q_{N+m}^T \tilde{F}_{N+m} \\ &= T_{N+m} Q_{N+m}^T Q_{N+m} + \tilde{F}_{N+m} Q_{N+m}. \end{aligned}$$

Writing $Q_{N+m}^T Q_{N+m}$ in the form

$$Q_{N+m}^T Q_{N+m} = I + P_{N+m} + P_{N+m}^T,$$

this becomes

$$\left(P_{N+m} + P_{N+m}^T\right) T_{N+m} - T_{N+m} \left(P_{N+m} + P_{N+m}^T\right) = \tilde{F}_{N+m}^T Q_{N+m} - Q_{N+m}^T \tilde{F}_{N+m}.$$

Equating the strictly upper triangular parts of each side gives a form of a useful identity established by Paige [5]:

$$P_{N+m} T_{N+m} - T_{N+m} P_{N+m} = \nabla\left(\tilde{F}_{N+m}^T Q_{N+m} - Q_{N+m}^T \tilde{F}_{N+m}\right),$$

$$\nabla = \text{strict upper triangle.} \tag{8.15}$$

Using this identity, we can prove the following lemma:

LEMMA 5. *Let* $s^{C_i,k}$ *and* P_{N+m} *be as defined above. The quantity* $|s^{C_i,k^T}P_{N+m}s^{C_i,k}|$ *satisfies*

$$|s^{C_i,k^T}P_{N+m}s^{C_i,k}| \leq \left[(k-1)|C_i|\gamma + \frac{(N+m)^2}{2}\right]\frac{\xi\|A\|}{\rho\|A\| - 2\tau}, \quad (8.16)$$

where each column of $\tilde{F}_{N+m}$ *is bounded by* $\xi\|A\|$.

Proof. Multiplying equation (8.15) by S_{N+m}^T on the left and S_{N+m} on the right gives

$$S_{N+m}^T P_{N+m}S_{N+m}\Theta_{N+m} - \Theta_{N+m}S_{N+m}^T P_{N+m}S_{N+m}$$
$$= S_{N+m}^T \nabla\left(\tilde{F}_{N+m}^T Q_{N+m} - Q_{N+m}^T \tilde{F}_{N+m}\right)S_{N+m},$$

or

$$\left(s^{l^T}P_{N+m}s^{l'}\right)(\theta_{l'} - \theta_l) = s^{l^T}\nabla\left(\tilde{F}_{N+m}^T Q_{N+m} - Q_{N+m}^T \tilde{F}_{N+m}\right)s^{l'},$$
$$l, l' = 1, \ldots, N+m.$$

If l is in C_i and l' is not in C_i, then $|s^{l^T}P_{N+m}s^{l'}|$ can be bounded by

$$|s^{l^T}P_{N+m}s^{l'}| \leq \frac{\left|s^{l^T}\nabla\left(\tilde{F}_{N+m}^T Q_{N+m} - Q_{N+m}^T \tilde{F}_{N+m}\right)s^{l'}\right|}{\rho\|A\| - 2\tau}. \quad (8.17)$$

Now, $P_{N+m}^T s^{C_i,k}$ can be expressed as

$$P_{N+m}^T s^{C_i,k} = \sum_{l \in C_i}\left(s^{C_i,k^T}P_{N+m}s^l\right)s^l + \sum_{l' \notin C_i}\left(s^{C_i,k^T}P_{N+m}s^{l'}\right)s^{l'}.$$

Taking the inner product of each side with $s^{C_i,k}$ gives

$$s^{C_i,k^T}P_{N+m}s^{C_i,k} = \sum_{l \in C_i}\left(s^{C_i,k^T}P_{N+m}s^l\right)s_k^l,$$

which can be written as

$$s^{C_i,k^T}P_{N+m}s^{C_i,k} = \left(P_{N+m}^T s^{C_i,k}\right)_k - \sum_{l' \notin C_i} \left(s^{C_i,k^T}P_{N+m}s^{l'}\right)s_k^{l'}. \quad (8.18)$$

The second term on the right-hand side in (8.18) can be bounded using (8.17):

$$\begin{aligned}
&\left|\sum_{l' \notin C_i} \left(s^{C_i,k^T}P_{N+m}s^{l'}\right)s_k^{l'}\right| \\
&\quad = \left|\left(\sum_{l \in C_i} s_k^l s^l\right)^T P_{N+m}\left(\sum_{l' \notin C_i} s_k^{l'} s^{l'}\right)\right| \\
&\quad \leqslant \frac{\displaystyle\sum_{l \in C_i}\sum_{l' \notin C_i} |s_k^l s_k^{l'}|\left|s^{l^T}\left(\tilde{F}_{N+m}^T Q_{N+m} - Q_{N+m}^T \tilde{F}_{N+m}\right)s^{l'}\right|}{\rho\|A\| - 2\tau} \\
&\quad \leqslant \frac{N+m}{4}\,\frac{\|\tilde{F}_{N+m}^T Q_{N+m} - Q_{N+m}^T \tilde{F}_{N+m}\|}{\rho\|A\| - 2\tau} \\
&\quad \leqslant \frac{(N+m)^2}{2}\,\frac{\xi\|A\|}{\rho\|A\| - 2\tau}.
\end{aligned} \quad (8.19)$$

The first term on the right-hand side in (8.18) is just

$$\left[P_{N+m}^T \sum_{l \in C_i} s_k^l s^l\right]_k = \sum_{l \in C_i} s_k^l \left(s^{l^T}P_{N+m}\right)_k,$$

and by Paige's lemma (5.5), this is bounded by

$$\left|\sum_{l \in C_i} s_k^l \left(s^{l^T}P_{N+m}\right)_k\right| \leqslant (k-1)|C_i|\frac{\gamma\xi}{\rho - 2\tau/\|A\|}. \quad (8.20)$$

Substituting (8.19) and (8.20) into (8.18) gives the desired result. ■

Combining the results of Lemmas 4 and 5, we can write

$$\left|\left(u^{i^T}q^k\right)^2 - \sum_{l\in C_i}\left(s_k^l\right)^2\right| \leqslant O\left((N+m)^2\xi\right), \tag{8.21}$$

where $O(\ )$ denotes a constant independent of ξ, N, m, and $\|A\|$ times the quantity inside parentheses, plus higher-order terms in ξ. Recall from Theorem 1′ that ξ is of order

$$J^2\gamma\sqrt{\epsilon} \quad \text{or} \quad c_{\max}^{3/2}J\epsilon^{1/4},$$

depending on the distribution of Ritz values corresponding to unconverged Ritz vectors at step J. Using (8.21), Theorem 3 can be restated as follows:

THEOREM 3′. *Let r^k be the kth "residual" vector and $e^k \equiv A^{-1}r^k$ the kth "error" vector in a perturbed CG recurrence for matrix A, satisfying (4.9). Let $\bar{r}^k$ be the kth residual vector and $\bar{e}^k \equiv \Theta_{N+m}^{-1}\bar{r}^k$ the kth error vector in an exact CG recurrence for matrix Θ_{N+m}, with initial residual parallel to the first column of S_{N+m}^T and satisfying $\|\bar{e}^0\|_{\Theta_{N+m}} = \|e^0\|_A$. Let $\tau \leqslant \sigma(N+m)^3\xi\|A\|$ be the greatest distance from an eigenvalue of Θ_{N+m} to the nearest eigenvalue of A, and assume that τ is small enough so that*

$$\lambda_1 - \tau > 0.$$

Then $\|e^k\|_A$ is related to $\|\bar{e}^k\|_{\Theta_{N+m}}$ by

$$\left|1 - \frac{\|\bar{e}^k\|_{\Theta_{N+m}}}{\|e^k\|_A}\right| \leqslant \left(\bar{\kappa}^3 + \bar{\kappa}^2\right)\frac{\tau}{\|A\|} + \left(\bar{\kappa}^2 + \bar{\kappa}\right)O\left((N+m)^2\xi\right). \tag{8.22}$$

9. CONCLUSIONS AND FURTHER REMARKS

The analysis in this paper shows that for *very* small values of ϵ, the eigenvalues generated over a fixed number of perturbed Lanczos steps and the A-norm of the error vectors generated over a fixed number of perturbed conjugate-gradient steps are approximately the same as those quantities generated by the exact recurrences applied to a "nearby" matrix. While the

theorems require excessively small ϵ,

$$(N+m)^3 J^2 \gamma \sqrt{\epsilon} \ll 1,$$

it is observed in practice that such results seem to hold even when ϵ is significantly larger. A precise bound on the size of ϵ necessary for such results to hold is not known.

The relationship between a slightly perturbed CG recurrence and the exact CG recurrence for a larger matrix with nearby eigenvalues yields much information about the convergence rate of the perturbed CG algorithm. In particular, to predict whether some form of reorthogonalization could significantly reduce the number of iterations required by a finite-precision CG implementation, one might compare the sharp error bound (2.5) for the exact algorithm with the bound in terms of the minimax polynomial on the appropriate union of intervals, which holds for the finite-precision implementation. We have made such a comparison for the five-point Laplace operator (assuming intervals of width 1.E − 6) and found that the two bounds differ very little until both have become quite small. For such problems, the desired level of accuracy would probably be achieved before the effects of roundoff caused a significant delay in convergence. On the other hand, matrices can be constructed for which the two error bounds differ significantly. The matrix

$$A = \text{diag}(1/i), \qquad I = 1, \ldots, N,$$

is one such example, and, as expected, a finite-precision CG recurrence for this matrix requires more iterations to achieve even a modest level of accuracy than would be predicted by the exact arithmetic error bound (2.5).

An open question concerning finite-precision Lanczos computations is the frequency at which multiple copies of eigenvalues are generated. The analysis in this paper reduces this problem to one of determining the frequency at which the exact Lanczos algorithm, applied to a matrix with many eigenvalues distributed in tight clusters, picks out different individuals from the same cluster. This depends very much on the spacing between the clusters. Using the analogy between the characteristic polynomials of the tridiagonal matrices in the Lanczos algorithm and the weighted least-squares polynomials of the CG algorithm, one can obtain rough estimates of this frequency. For example, consider the problem of Section 3:

$$\bar{A} = \text{diag}((\lambda_{il},\ l = 1, \ldots, 11),\ i = 1, \ldots, 10),$$

$$\lambda_{il} = \lambda_i + (l-6) \times 2.\text{E} - 9, \quad l = 1, \ldots, 11,$$

$$\lambda_i = i,\ i = 1, \ldots, 9, \quad \lambda_{10} = 200.$$

A possible polynomial p_k to substitute in (2.3) to obtain an error bound for the exact CG algorithm applied to this matrix is one that resembles a $(k-k')$th-degree Chebyshev polynomial on the interval $[1-1.\text{E}-8, 9+1.\text{E}-8] \approx [1,9]$ and a (k')th-degree Chebyshev polynomial on the interval $[200-1.\text{E}-8, 200+1.\text{E}-8]$:

$$p_k(x) = T_{k-k'}^{[1,9]}(x)T_{k'}^{[200+-]}(x).$$

The maximum absolute value of this polynomial on the interval $[1,9]$ is approximately

$$2^{-(k-k')+1} \tag{9.1}$$

and on the interval $[200-1.\text{E}-8, 200+1.\text{E}-8]$ is between approximately

$$2(2.5\text{E}-9)^{k'}(199)^{k-k'} \quad \text{and} \quad 2(2.5\text{E}-9)^{k'}\left(\tfrac{191}{9}\right)^{k-k'}. \tag{9.2}$$

[The latter estimates hold because the roots of $T_{k-k'}^{[1,9]}$ lie between 1 and 9. The upper bound is the maximum value of the polynomial $(1-x)^{k-k'}T_{k'}^{[200+-]}(x)$ on the interval about 200, and the lower bound is the maximum value of the polynomial $[(9-x)/9]^{k-k'}T_{k'}^{[200+-]}(x)$ on the interval about 200.] In order that the values (9.1) and (9.2) be approximately equal, k' should satisfy

$$\tfrac{1}{4}k \geqslant k' \geqslant \tfrac{1}{7}k.$$

A polynomial with more than about $\frac{6}{7}$ of its roots in the interval $[1,9]$ will be larger than this polynomial in the interval $[200-1.\text{E}-8, 200+1.\text{E}-8]$, and one with fewer than about $\frac{3}{4}$ of its roots in the interval $[1,9]$ will be larger than this polynomial in the interval $[1,9]$. Hence if each interval has a significant weight, one would expect the exact weighted least-squares polynomial to have about $\frac{1}{4}$ to $\frac{1}{7}$ of its roots in the interval about 200. The exact Lanczos algorithm applied to this problem, then, or the equivalent perturbed algorithm applied to A, can be expected to generate approximations to the eigenvalue 200 about once every four to seven steps.

REFERENCES

1 A. Greenbaum, Comparison of splittings used with the conjugate gradient algorithm, *Numer. Math.* 33:181–194 (1979).

2 M. R. Hestenes and E. Stiefel, Methods of conjugate gradients for solving linear systems, *J. Res. Nat. Bur. Standards* 49:409–436 (1952).

3 C. Lanczos, An iteration method for the solution of the eigenvalue problem of linear differential and integral operators, *J. Res. Nat. Bur. Standards* 45:225–280 (1950).
4 C. Paige, Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem, *Linear Algebra Appl.* 33:235–258 (1980).
5 C. Paige, The Computation of Eigenvalues and Eigenvectors of Very Large Sparse Matrices, Ph.D. Thesis, Univ. of London, London, 1971.
6 B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, N.J., 1980.
7 B. N. Parlett and D. S. Scott, The Lanczos algorithm with selective orthogonalization, *Math. Comp.* 33:217–238 (1979).
8 D. S. Scott, How to make the Lanczos algorithm converge slowly, *Math. Comp.* 33:239–246 (1979).

Received 1 December 1986; final manuscript accepted 24 September 1987