ANALYSIS OF THE SYMMETRIC LANCZOS PROCESS

by

D. St.-C. Scott

Memorandum No. UCB/ERL M78/40

June 1978

(cover)

ANALYSIS OF THE SYMMETRIC LANCZOS PROCESS*

by

David St. Clair Scott

Department of Mathematics
University of California, Berkeley

Ph.D. Dissertation

June 1978

Abstract

The Lanczos algorithm is a powerful method for finding a few eigenvalues and eigenvectors of large sparse symmetric matrices. The quantities actually computed by the Lanczos algorithm diverge completely from their theoretical counterparts. In 1971 C. Paige showed that this "instability" merely resulted in the computation of multiple copies of eigenpairs of the matrix. This work presents and analyzes a new way of implementing the Lanczos algorithm which prevents the computation of redundant copies of eigenpairs and costs little more than simple Lanczos itself.

B.N. Parlett
Professor B.N. Parlett
Chairman of Thesis Committee

*Research sponsored by Office of Naval Research Contract N00014-76-C-0013.

Acknowledgments

My advisor Professor B.N. Parlett suggested that I investigate the new variant of the Lanczos algorithm. The many discussions I held with him were instrumental both in consolidating work already done and in suggesting new lines of research. His many comments and suggestions for the writing of this thesis were also greatly appreciated.

I would like to express my gratitude to C.C. Paige for his pioneering work on the Lanczos algorithm, without which the algorithm might still be mired in obscurity. Thanks are also due to Professors W. Kahan, F.A. Grunbaum, and R.L. Taylor for reading all or part of this dissertation.

I would like to thank Ruth Suzuki for her fast and expert typing of this manuscript.

Finally I wish to thank my wife Sheila for her constant patience and support throughout my graduate career.

Introduction

The Lanczos algorithm is one of the most powerful methods for finding a few eigenvalues (and eigenvectors if desired) from one or both ends of the spectrum of a large sparse symmetric matrix. It has been known since its introduction in 1950 that the Lanczos algorithm is unstable, in that the quantities computed in finite precision arithmetic will diverge completely from their theoretical counterparts.

In his Ph.D. thesis of 1971, C. Paige showed that this "instability" of the Lanczos algorithm merely resulted in the computation of repeated copies of the eigenpairs of the matrix. Despite the importance of Paige's results to understanding the behavior of the Lanczos algorithm in finite precision, many of them have never been published in the open literature.

The main contribution of this thesis is an analysis of Selective Orthogonalization, a new and efficient method of implementing the Lanczos algorithm, based on Paige's analysis, which prevents the appearance of repeated copies of eigenpairs. To help the reader understand the effects of Selective Orthogonalization some necessary background material is given in the first two chapters.

Chapter 1 gives a description of the Lanczos algorithm in the context of exact arithmetic. Included are a derivation of the Kaniel-Paige a priori error bounds on the accuracy of the eigenvalue estimates computed by the algorithm and some new results relating the choice of the starting vector to the convergence of the algorithm.

Chapter 2 describes the surprising behavior of the Lanczos algorithm in finite precision arithmetic. Since Paige's results are rather inaccessible we hope that the derivations presented here will further the appreciation of the importance of his work.

Chapter 3 explores a suggestion of Parlett that nearly converged Ritz vectors be explicitly purged from new Lanczos vectors as a means of maintaining linear independence at little extra cost.

Table of Contents

1. The Lanczos Algorithm in Exact Arithmetic

This chapter will give an overview of the theory of the Lanczos algorithm in the context of exact arithmetic. Section 1 will describe the Rayleigh-Ritz procedure, a general method for choosing approximations to eigenvectors of A from a subspace $\mathcal{W} \subseteq \mathbb{R}^n$. Some of the results in this section are not widely appreciated. Section 2 describes the Lanczos algorithm, a special case of the Rayleigh-Ritz procedure. Sections 3 and 4 investigate various aspects of the convergence of the algorithm. Only Section 4 is new material.

Throughout this thesis, A will be an $n \times n$ (real) symmetric matrix with eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and eigenvectors $z_1, z_2, \ldots, z_n$. Thus $A = Z\Lambda Z^*$ where $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ and $Z = (z_1, z_2, \ldots, z_n)$ with $Z^*Z = 1$. The symbol 1 will stand for the appropriate size identity matrix. $B-\sigma$ will stand for the matrix $B - \sigma I$.

1.1 Approximations from a Subspace -- The Rayleigh-Ritz Procedure

One possible approach for computing approximations to eigenpairs of A is to choose some j dimensional subspace $\mathcal{W}_j \subseteq \mathbb{R}_n$ and compute the j best approximations to eigenvectors of A contained in $\mathcal{W}_j$ and the corresponding best approximate eigenvalues. We refer to an approximate eigenpair (y,θ) as a <u>pair</u>. Of course this approach requires some precise definition of the word "best." To motivate possible definitions of "best" we quote some of the most important theorems for evaluating the accuracy of such pairs (y,θ).

Theorem 1. Let y be any unit vector, let θ be any scalar, and let $\rho = \|Ay-y\theta\|$. Then there exists λ, an eigenvalue of A such that

$$|\lambda-\theta| \leq \rho .$$

Furthermore, for a fixed vector y, ρ is minimized as a function of θ by $\theta = y^*Ay$.

Theorem 1 is best possible in that equality may hold.

Example 1. Let

$$A = \mathrm{diag}(1,-1)$$

and

$$y = (\tfrac{\sqrt{2}}{2},\tfrac{\sqrt{2}}{2})^* .$$

Then

$$\theta = y^*Ay = 0 ,$$

$$r = Ay - y\theta = Ay = (\tfrac{\sqrt{2}}{2},-\tfrac{\sqrt{2}}{2})^* ,$$

$$\rho = \|r\| = 1 ,$$

and

$$\min_{i=1,2} |\lambda_i-0| = 1 . \qquad \square$$

The quantity y^*Ay (or y^*Ay/y^*y for a non-unit vector) is now called the Rayleigh quotient of y (with respect to A). For a given unit vector y, the vector $r = Ay - y\theta$, where $\theta = y^*Ay$, is the residual vector of y and $\rho = \|r\|$ is the residual norm of y.

If $j > 1$, more than one pair must be chosen and we would want each pair to approximate a different eigenpair of A. Unfortunately the obvious generalization of Theorem 1 to more than one vector fails in this respect, even when the chosen vectors are orthogonal. That is, it is possible to have two pairs (y_1,θ_1) and (y_2,θ_2) such that $y_1^*y_2 = 0$ and yet only <u>one</u> eigenvalue of A lies in the union of the intervals around θ_1 and θ_2 given by Theorem 1.

<u>Example 2</u>. Let

$$A = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}, \quad y_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \text{and} \quad y_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

Then $\theta_1 = \theta_2 = 0$ and $\|Ay_1-y_1\theta_1\| = \|Ay_2-y_2\theta_2\| = 1$. The eigenvalues of A are 0 and $\pm\sqrt{2}$ and only 0 lies within $[-1,1]$. □

The best bound for locating the proper number of eigenvalues of A was given by W. Kahan in [Kahan 1967].

<u>Theorem 2 (Kahan)</u>. Let Y be an $n \times j$ orthonormal matrix, let H be any $j \times j$ symmetric matrix, let $\theta_1,\theta_2,\ldots,\theta_j$ be the eigenvalues of H, and let $R = AY - YH$. Then there exist distinct integers $1',2',\ldots,j'$ such that for $i = 1,2,\ldots,j$,

$$|\theta_i-\lambda_{i'}| \leq \|R\| .$$

For a fixed matrix Y, $\|R\|$ is minimized by $H = Y^*AY$.

Note that for Example 2, $Y = (y_1,y_2)$, $Y^*AY = H = 0$, and $\|R\| = \sqrt{2}$, so that Theorem 2 is the best possible in that equality may hold.

A common method of choosing pairs from the subspace $\mathcal{W}_j$ is the Rayleigh-Ritz procedure, defined as follows.

The Rayleigh-Ritz Procedure. Given any subspace $\mathcal{W}_j$ of dimension j and a symmetric matrix A, do steps 1 through 6.

1. Compute Q_j, an orthonormal matrix such that $\text{span}(Q_j) = \mathcal{W}_j$.
2. Form $H = Q_j^*(AQ_j)$.
3. Compute $S\Theta S^*$, the eigensystem of H.
4. Form $Y = (y_1, y_2, \ldots, y_j) = Q_jS$.
5. Compute $R = (r_1, \ldots, r_j) = AY - Y\Theta = (AQ_j)S - Y\Theta$.
6. Compute $\|R\|$ and $\rho_i = \|r_i\| = \|Ay_i - y_i\theta_i\|$, for $i = 1,2,\ldots,j$.

The columns of Y are the Ritz vectors, the eigenvalues of H are the Ritz values, and a pair (y_i, θ_i) is a Ritz pair. The norms computed in step 6 can be used in Theorems 1 and 2 to bound the accuracy of the Ritz values. The Ritz pairs are determined solely by the action of A on the subspace $\mathcal{W}_j$ and they are independent of the particular matrix Q_j used to compute them.

The word optimal is often associated with the Rayleigh-Ritz procedure although there appears to be some confusion as to the sense in which the Ritz pairs are indeed optimal. The following result is hardly new but is included here to clarify this point.

Definitions. Let P_j be the orthogonal projector onto the subspace $\mathcal{W}_j$, that is $P_j\mathbb{R}_n = \mathcal{W}_j$ and $P_j = P_j^*$. For any orthonormal matrix $Y = (y_1, y_2, \ldots, y_j)$ let $R(Y) \equiv AY - Y\Theta$, where $\Theta = \text{diag}(\theta_1, \theta_2, \ldots, \theta_j)$ and $\theta_i = y_i^*Ay_i$ for all i.

Theorem 3. Let $Y_j = (y_1, y_2, \ldots, y_j)$ and $\Theta_j = (\theta_1, \theta_2, \ldots, \theta_j)$ be the Ritz pairs derived from $\mathcal{W}_j$. Then

1. $Y_j^* A Y_j = \Theta_j$.
2. For all i, (y_i, θ_i) is an eigenpair of $P_j A P_j$.
3. Y_j minimizes $\|R(Y)\|$ over all orthonormal matrices whose columns span $\mathcal{W}_j$.

Remarks. As shown by 1, the Ritz vectors are the distinguished basis for $\mathcal{W}_j$ which makes the reduced matrix, $H = Y_j^* A Y_j$, diagonal. Furthermore $Y_j Y_j^*$ is a matrix representation of P_j so that if A is represented in a basis which has the columns of Y_j as its first j elements then the matrix $P_j A P_j$ takes on the simple form,

$$P_j A P_j = \begin{bmatrix} \Theta & 0 \\ 0 & 0 \end{bmatrix} .$$

Proof. Let Q_j be an orthonormal matrix whose columns span $\mathcal{W}_j$ and let $H = Q_j^* A Q_j$. Then $H = S\Theta_j S^*$ and $Y_j = Q_j S$. Therefore

$$\begin{aligned} Y_j^* A Y_j &= (Q_j S)^* A Q_j S , \\ &= S^* Q_j^* A Q_j S , \\ &= S^* H S , \\ &= \Theta_j , \end{aligned}$$

since $S^* S = 1$. This proves 1.

$Q_j Q_j^*$ is the matrix representation of P_j. Therefore

$$\begin{aligned} P_j A P_j &= Q_j Q_j^* A Q_j Q_j^* , \\ &= Q_j H Q_j^* , \end{aligned}$$

and for any $y_i = Q_j s_i$,

$$\begin{aligned} P_j A P_j y_i &= Q_j H Q_j^* Q_j s_i \ , \\ &= Q_j H s_i \ , \\ &= Q_j s_i \theta_i \ , \\ &= y_i \theta_i \ . \end{aligned}$$

This proves 2.

Let $G_j = (g_1, g_2, \ldots, g_j)$ be any orthonormal matrix whose columns span ω_j and let $\Theta_j' = (g_1^* A g_1, \ldots, g_j^* A g_j)$. Since $\text{span}(G_j) = \text{span}(Y_j) = \omega_j$ there exists a $j \times j$ orthogonal matrix L such that $G_j = Y_j L$. Recall from Theorem 2 that $\|R(Y,H)\| = \|AY-YH\|$ is minimized for fixed Y by $H = Y^*AY$. Therefore

$$\begin{aligned} \|R(G_j)\| &= \|AG_j - G_j\Theta_j'\| \ , & & \\ &= \|AY_jL - Y_jL\Theta_j'\| \ , & & \text{definition of } L \\ &= \|(AY_j - Y_jH')L\| \ , & & H' = L\Theta_j'L^* \\ &= \|AY_j - Y_jH'\| \ , & & L \text{ is orthogonal} \\ &\geq \|AY_j - Y_j\Theta_j\| \ , & & \Theta_j = Y_j^*AY_j \\ &= \|R(Y_j)\| \ . & & \end{aligned}$$

This proves 3. □

It is important to realize that in general Ritz vectors are not projections onto ω_j of eigenvectors of A nor are they local minima of the residual norm. Consider the following example.

Example 3. Let

$$A = \begin{bmatrix} 0 & \alpha & 0 \\ \alpha & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

and let

$$Q_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} .$$

Then

$$H \equiv Q_2^* A Q_2 = \begin{bmatrix} 0 & \alpha \\ \alpha & 0 \end{bmatrix} ,$$

$$S = \frac{\sqrt{2}}{2}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} ,$$

$$\Theta = \mathrm{diag}(\alpha,-\alpha) ,$$

$$Y \equiv Q_2 S = \frac{\sqrt{2}}{2}\begin{bmatrix} 1 & -1 \\ 1 & 1 \\ 0 & 0 \end{bmatrix} ,$$

$$R \equiv AY - Y\Theta = \frac{\sqrt{2}}{2}\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 1 \end{bmatrix} ,$$

$$\|R\| = 1 , \text{ and}$$

$$\|r_1\| = \|r_2\| = \frac{\sqrt{2}}{2} .$$

Thus S, Y, and R are independent of the size of α. Do other vectors in the span of Q_2 have smaller residual norms than y_1 and y_2? The answer is yes, regardless of the size of α. To prove this, let

$$y = \begin{bmatrix} \cos\phi \\ \sin\phi \\ 0 \end{bmatrix},$$

which is an arbitrary unit vector in $\text{span}(Q_2)$. Then with some trigonometric manipulation it can be shown that

$$\theta = y^* Ay = \alpha \sin 2\phi$$

and

$$\rho^2 = \|Ay - y\theta\|^2 = \alpha^2 \cos^2 2\phi + \sin^2\phi \quad .$$

The Ritz vectors y_1 and y_2 correspond to $\phi = \pi/4$ and $\phi = 3\pi/4$ which both have $\rho^2 = 1/2$ as expected. The derivative of ρ^2 is

$$\begin{aligned} \frac{d\rho^2}{d\phi} &= -4\alpha^2 \cos 2\phi \sin 2\phi + 2 \sin\phi \cos\phi \\ &= \sin 2\phi - 2\alpha^2 \sin 4\phi \ . \end{aligned}$$

The value of the derivative at $\pi/4$ is 1 and at $3\pi/4$ is -1. This shows that ρ^2 and hence ρ is never minimized by a Ritz vector for any value of α.

The unnormalized eigenvectors of A are the columns of

$$Z = \begin{bmatrix} \alpha & 1 & \alpha \\ \sqrt{1+\alpha^2} & 0 & -\sqrt{1+\alpha^2} \\ 1 & -\alpha & 1 \end{bmatrix} .$$

The projections onto $\text{span}(Q_2)$ are the columns of

$$\begin{bmatrix} \alpha & 1 & \alpha \\ \sqrt{1+\alpha^2} & 0 & -\sqrt{1+\alpha^2} \\ 0 & 0 & 0 \end{bmatrix} .$$

For no value of α are any of these vectors parallel to either of the Ritz vectors. That is, neither Ritz vector is ever a normalized projection onto $\mathcal{W}_j$ of an eigenvector of A. □

Given only the subspaces $\mathcal{W}_j$ and $A\mathcal{W}_j$ is it possible to obtain better approximations to eigenvalues and eigenvectors of A than those given by the Rayleigh-Ritz procedure? The answer is yes and many results along this line have been obtained. Some of the most important results are contained in [N.J. Lehmann 1963] and [Davis and Kahan 1970] but we will not pursue this question any further here.

We finish this section by considering the cost of applying the Rayleigh-Ritz procedure when $j << n$, which will always be the case when A is large. The bulk of the cost of applying Rayleigh-Ritz normally lies in computing the orthonormal basis Q_j and forming the matrix product AQ_j. Of course it may happen that $\mathcal{W}_j$ is specified as the span of an orthonormal matrix which would eliminate the cost of the first step. However for arbitrary subspaces there is no way to avoid the cost of forming AQ_j. However for special subspaces the cost of implementing the Rayleigh-Ritz procedure can be substantially reduced, as is shown in the next section.

1.2 The Lanczos Algorithm

Section 1 described the Rayleigh-Ritz procedure, a method of extracting good approximations to eigenpairs of A from a given subspace $\mathcal{W}_j$. In this section we show that a certain class of subspaces, called Krylov subspaces, are ideally suited for the Rayleigh-Ritz procedure in that the bulk of the cost of using the procedure is eliminated due to the special structure of these subspaces.

For any vector $s \neq 0$, $K_j(s) = (s, As, \ldots, A^{j-1}s)$ is a Krylov matrix and $\mathcal{K}_j(s) = \text{span}(K_j(s))$ is a Krylov subspace. The first step of the Rayleigh-Ritz procedure is to find an orthonormal basis for $\mathcal{K}_j(s)$, so for any j, let $Q_j \equiv (q_1, q_2, \ldots, q_j)$ be the result of orthonormalizing the columns of $K_j(s)$ from left to right. In particular $q_1 = s/\|s\|$. The second step of the Rayleigh-Ritz procedure is to form $Q_j^* A Q_j$. This is made easier by the following.

Theorem 4. $T_j \equiv Q_j^* A Q_j$ is a symmetric tridiagonal matrix.

Proof. By definition, for all i, $A\mathcal{K}_i(s) \subseteq \mathcal{K}_{i+1}(s)$. In particular $Aq_i \in \mathcal{K}_{i+1}(s) = \text{span}(Q_{i+1})$. Hence $q_k^* A q_i = 0$ for all $k \geq i+2$. Since A is symmetric, $q_i^* A q_k = q_k^* A q_i$ and the result follows. □

We label the diagonal elements of T_j as $\alpha_1, \alpha_2, \ldots, \alpha_j$ and the off diagonal elements of T_j as $\beta_1, \beta_2, \ldots, \beta_{j-1}$.

Theorem 4 implies that the columns of Q_j satisfy a three term recurrence. The Lanczos algorithm arises from the observation that the coefficients of the recurrence, the α's and β's can be computed as needed.

There are several mathematically equivalent formulas for computing the α's and β's. We present only the most numerically stable version. A comparison of the different possible methods is made in Section 2.4.

The Lanczos Algorithm. Given q_1 an arbitrary unit vector, define $q_0 = 0$ and $\beta_0 = 0$. For $j = 1,2,\dots$ do 1 through 5.

```
1.  $u_j = Aq_j - q_{j-1}\beta_{j-1}$
2.  $\alpha_j = q_j^* u_j$
3.  $r_j = u_j - q_j\alpha_j$
4.  $\beta_j = \|r_j\|$
5.  If $\beta_j = 0$ stop
    else $q_{j+1} = r_j/\beta_j$
```

One cycle of 1 through 5 is a Lanczos step. Each α_i is chosen to force $q_i^* q_{i+1} = 0$ and each β_i is chosen to normalize q_{i+1} to length 1. That is maintaining local orthonormality is sufficient to guarantee orthonormality of Q_j.

We have introduced the Lanczos algorithm by assuming Q_j was orthonormal and proving the three term recurrence. A more common approach is to define the α's and β's and then prove that Q_j is orthonormal (cf. [Kahan and Parlett 1974]).

If the algorithm is interrupted after j steps, the computed quantities satisfy

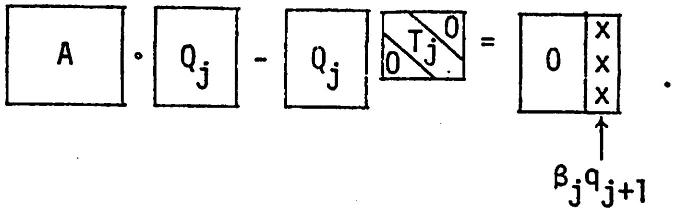
$$AQ_j - Q_jT_j = \beta_j q_{j+1} e_j^* \tag{1}$$

and

$$1 - Q_j^* Q_j = 0 \ , \tag{2}$$

where $e_j^* = (0,0,\dots,0,1)$ has j elements. Rewriting equation (1) with appropriate sized rectangles yields

(3)



Equations (1) and (2) are a compact way of displaying the Lanczos algorithm and most of the analysis presented will start from these equations.

At the j^{th} step of the algorithm, if $S_j \Theta_j S_j^*$ is the spectral decomposition of T_j, where $\Theta_j = (\theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_j^{(j)})$ and $S_j = (s_1^{(j)}, s_2^{(j)}, \ldots, s_j^{(j)})$, and $Y_j = (y_1^{(j)}, y_2^{(j)}, \ldots, y_j^{(j)}) = Q_j S_j$ then $(y_i^{(j)}, \theta_i^{(j)})$, for $i = 1,2,\ldots,j$ are the Ritz pairs obtained from $K_j(q_1)$. We note that normalized eigenvectors of T_j are only determined up to a factor of ±1. In later analysis it will be convenient to assume that s_{ji} the bottom element of $s_i^{(j)}$ is positive for all i. Unless otherwise stated, all Ritz pairs will be taken from $K_j(q_1)$ and the superscript j will be dropped.

The Lanczos algorithm permits the first two steps of the Rayleigh-Ritz procedure to be performed simultaneously at a substantial savings in cost. Only the vectors q_{j-1} and q_j are needed for computing q_{j+1}. The rest of the vectors can be put into secondary store until they are needed for forming Ritz vectors. If only the Ritz values are desired, the Lanczos vectors (the q's) need not be kept at all. This is a very attractive feature with respect to large problems. Also the matrix $H \equiv Q_j^* A Q_j$ is tridiagonal and this substantially lowers the cost of computing its eigensystem. Finally we are going to show that the composite residual norm is computed by the algorithm itself (i.e. β_j) and that the residual norm of each Ritz vector can be computed without forming the Ritz vectors! This obviates the need for performing step 5 of the procedure and reduces the cost of step 6. We now show how this can be accomplished.

Theorem 5. Let (Y_j,Θ_j) be the set of Ritz pairs derived at the j^{th} step of the Lanczos algorithm. Then

$$\|R_j\| = \|AY_j - Y_j\Theta_j\| = \beta_j$$

and

$$\rho_i = \|Ay_i - y_i\theta_i\| = \beta_{ji} ,$$

for $i = 1,2,\ldots,j$, where $\beta_{ji} = \beta_j s_{ji}$ and $s_{ji} > 0$ is the bottom (j^{th}) element of s_i, the i^{th} eigenvector of T_j.

Proof. Multiply equation (1) on the right by S_j to yield

$$AQ_jS_j - Q_jT_jS_j = \beta_j q_{j+1} e_j^* S_j .$$

Since $T_jS_j = S_j\Theta_j$ and $Y_j = Q_jS_j$, this reduces to

(4) $$AY_j - Y_j\Theta_j = \beta_j q_{j+1} e_j^* S_j .$$

Taking the norm of both sides yields

$$\|R_j\| = \|AY_j - Y_j\Theta_j\| = \|\beta_j q_{j+1} e_j^* S_j\| .$$

S_j is orthogonal and for any matrix B, $\|BP\| = \|B\|$ if P is orthogonal so

$$\begin{aligned} \|R_j\| &= \|\beta_j q_{j+1} e_j^*\| , \\ &= \beta_j \|q_{j+1}\| \|e_j^*\| , \\ &= \beta_j , \end{aligned}$$

since $\|uv^*\| = \|u\|\|v^*\|$ for any vectors u and v. Equating the i^{th} column of each side of (4) yields

$$\begin{aligned} Ay_i - y_i\theta_i &= \beta_j q_{j+1} e_j^* s_i \ , \\ &= \beta_j s_{ji} q_{j+1} \ , \end{aligned}$$

and by taking the norm of each side we obtain

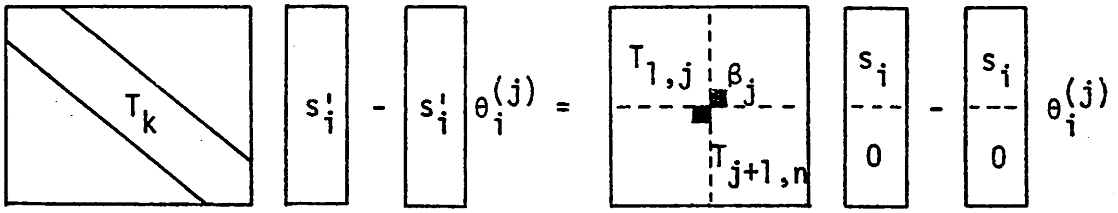
$$\begin{aligned} \rho_i &= \beta_j s_{ji} \|q_{j+1}\| \\ &= \beta_{ji} \ . \end{aligned}$$ □

These numbers β_{ji}, which can be computed without forming the Ritz vectors, explain how it can happen that <u>some</u> of the Ritz values may be very accurate approximations to eigenvalues of A without the appearance of any small off diagonal elements of T (β's). A Ritz value $\theta_i^{(j)}$ with a negligible β_{ji} has <u>converged</u>. This definition is justified by the following result.

<u>Theorem 6</u>. Let $\theta_i^{(j)}$ be any eigenvalue of T_j. Then <u>for all</u> $k > j$ there exists an index i_k such that $\theta_{i_k}^{(k)}$ (an eigenvalue of T_k) satisfies

$$|\theta_i^{(j)} - \theta_{i_k}^{(k)}| \leq \beta_{ji} \ .$$

<u>Proof</u>. Let s_i be the eigenvector of $\theta_i^{(j)}$. Let s_i' be a k-vector with $s_i' = \binom{s_i}{0}$. To prove the Theorem we compute $\|T_k s_i' - s_i'\theta_i^{(j)}\|$ and then invoke Theorem 1. In pictures,



$$= \left[\begin{array}{c} T_j s_i - s_i\theta_i^{(j)} \\ \hline \beta_j s_{ji} \\ 0 \\ \vdots \\ 0 \end{array}\right]$$

$$= \left[\begin{array}{c} 0 \\ \hline \beta_j s_{ji} \\ 0 \\ \vdots \\ 0 \end{array}\right] .$$

Hence $\|T_k s'_i - s'_i\theta_i^{(j)}\| = \beta_j s_{ji} = \beta_{ji}$. By Theorem 1 there must be some eigenvalue of T_k, call it $\theta_{i_k}^{(k)}$ such that $|\theta_i^{(j)} - \theta_{i_k}^{(k)}| \le \beta_{ji}$. □

These results show that the converged Ritz values can be identified by inspection of the bottom row of the matrix S_j. They do not address the fundamental questions of whether convergence will occur quickly and to which eigenvalues of A Ritz values are most likely to converge. The examination of these questions is put off until Section 3. In the mean time we derive some useful algebraic properties of the Lanczos algorithm.

The algorithm must terminate ($\beta_j = 0$) at some step $j \le n$, since $n+1$ orthonormal n-vectors cannot exist. The exact number of steps taken before termination depends on the starting vector q_1 as follows.

<u>Theorem 7</u>. Let $\mathcal{W} \subseteq \mathbb{R}^n$ be the smallest A-invariant subspace containing q_1 and let $m = \dim \mathcal{W}$. Then the Lanczos algorithm started with q_1 will terminate at the m^{th} step $(\beta_m = 0)$ with $\mathcal{K}_m(q_1) = \mathcal{W}$.

<u>Proof</u>. Suppose the algorithm terminates at the j^{th} step $(\beta_j = 0)$. Then

$$AQ_j - Q_jT_j = 0 \ ,$$

so $\mathcal{K}_j(q_1) = \mathrm{span}(Q_j)$ is A-invariant. Since $q_1 \in \mathcal{K}_j(q_1)$, $\mathcal{W} \subseteq \mathcal{K}_j(q_1)$ and $m \leq j$.

On the other hand, since $\mathcal{W}$ is A-invariant and $q_1 \in \mathcal{W}$,

$$A^i q_1 \in \mathcal{W} \quad \text{for all} \quad i \ .$$

Hence $\mathcal{K}_j(q_1) = \mathrm{span}(q_1, Aq_1, \ldots, A^{j-1}q_1) \subseteq \mathcal{W}$ and $j \leq m$. □

The Lanczos algorithm is invariant under certain algebraic operations. Since these results are needed in later analysis, they are listed together here for easy reference. To express these identities we use the notation $L^j(A,q_1) = (Q_j,T_j)$ to mean that j steps of the Lanczos algorithm run on A starting with q_1 yield Q_j and T_j.

Theorem 8. If $L^j(A,q_1) = (Q_j,T_j)$ then

1. $L^j(\gamma A,q_1) = (Q_j,\gamma T_j)$ for all $\gamma \in \mathbb{R}$.
2. $L^j(A-\gamma,q_1) = (Q_j,T_j-\gamma)$ for all $\gamma \in \mathbb{R}$.
3. $L^j(P^*AP,P^*q_j) = (P^*Q_j,T_j)$ for all orthogonal P.
4. $L^j(A|_{\mathcal{W}},q_1) = (Q_j,T_j)$, where $A|_{\mathcal{W}}$ is A restricted to $\mathcal{W}$ and $\mathcal{W}$ is the smallest A-invariant subspace containing q_1.

Statement 4, which follows directly from Theorem 7, has an important theoretical consequence. Any eigenvector orthogonal to q_1 will be orthogonal to all of $\mathcal{W}$ and will not be discovered by the algorithm. In particular if A has multiple eigenvalues, only one representative (at most) of the multiplicity will lie in $\mathcal{W}$ and the multiplicity will not be discerned. Therefore in analyzing the algorithm one may always assume that A has no multiple eigenvalues at all. This feature of the Lanczos algorithm must be viewed as a serious drawback in the context of using the algorithm to find a few eigenvalues of a given matrix A.

1.3 The Kaniel-Paige Theory

In 1966 S. Kaniel [Kaniel 1966] computed some bounds on the accuracy of Ritz pairs obtained from Krylov subspaces. Since this paper is difficult to read and contains some significant errors, C. Paige reworked the results in his Ph.D. thesis [Paige 1971]. Unfortunately this material is not readily available so we present the theory here, correcting a minor error in Paige and adding a mild improvement.

The foundation of the theory lies in the characterization

$$K_j(q_1) = \{\phi(A)q_1 | \phi \text{ is a polynomial of degree} < j\} .$$

The point is that using Tchebychev polynomials it is possible to choose a polynomial ϕ which greatly amplifies the q_1-component of one eigenvector while crushing down the components in all the other eigenvectors. The remaining results account for the fact that the vector so constructed is not exactly a Ritz vector. All the results presented bound the accuracy of the algebraically smallest Ritz values. Similar results for the largest Ritz values can be obtained by applying the given results to $-A$. There is a numerical example at the end of the section to illustrate the theorems.

As mentioned at the end of Section 2, we assume that A has no eigenvectors perpendicular to q_1. In particular A has no multiple eigenvalues. Let $\lambda_1 < \lambda_2 < \cdots < \lambda_n$ be the eigenvalues of A and let $z_1, z_2, \ldots, z_n$ be the corresponding (normalized) eigenvectors. Let (y_i, θ_i), for $i = 1,2,\ldots,j$, be the Ritz pairs obtained from $K_j(q_1)$. Then each Ritz vector can be expanded (uniquely) as

$$(1) \qquad y_i = \sum_{k=1}^{n} \gamma_{ki} z_k .$$

For each Ritz vector define

$$(2) \qquad \begin{aligned} \epsilon_i^2 &= \| y_i - \gamma_{ii} z_i \|^2 , \\ &= \sum_{k \neq i} \gamma_{ki}^2 , \\ &= 1 - \gamma_{ii}^2 . \end{aligned}$$

Note that ϵ_i is the norm of the component of y_i orthogonal to z_i rather than $\|y_i - z_i\|$. It is confusion of these two possible definitions of ϵ_i which leads to the errors in Kaniel's paper. The next theorem gives a bound on the ϵ_i^2.

Theorem 9. For $i = 1,2,\ldots,j$

$$\epsilon_i^2 \leq [\theta_i - \lambda_i + \sum_{k=1}^{i-1} \epsilon_k^2(\lambda_{i+1} - \lambda_k)]/(\lambda_{i+1} - \lambda_i) .$$

Proof. By equation (1) and Theorem 3 (in section 1),

$$\begin{aligned} \theta_i &= y_i^* A y_i , \\ &= \sum_{k=1}^{n} \gamma_{ki}^2 \lambda_k , \end{aligned}$$

and since $\sum_{k=1}^{n} \gamma_{ki}^2 = \|y_i\|^2 = 1$,

$$\lambda_i = \sum_{k=1}^{n} \gamma_{ki}^2 \lambda_i .$$

Combining these results and rearranging yields

$$\begin{aligned} \theta_i - \lambda_i + \sum_{k=1}^{i-1} \gamma_{ki}^2 (\lambda_i - \lambda_k) &= \sum_{k=i+1}^{n} \gamma_{ki}^2 (\lambda_k - \lambda_i) , \\ &\geq (\lambda_{i+1} - \lambda_i) \sum_{k=i+1}^{n} \gamma_{ki}^2 . \end{aligned}$$

From (2), $\sum_{k=i+1}^{n} \gamma_{ki}^2 = \epsilon_i^2 - \sum_{k=1}^{i-1} \gamma_{ki}^2$ so

$$\epsilon_i^2 \leq [\theta_i - \lambda_i + \sum_{k=1}^{i-1} \gamma_{ki}^2 (\lambda_{i+1} - \lambda_k)]/(\lambda_{i+1} - \lambda_i) .$$

Finally by (2), $\gamma_{ki}^2 \leq \epsilon_k^2$ for $k \neq i$ and the result follows. □

The vector in $K_j(q_1)$, which we will obtain with an artfully chosen polynomial ϕ, will not be a Ritz vector. Therefore this vector's Rayleigh quotient will not be exactly a Ritz value. We now compute a bound on the error introduced by this discrepancy.

Theorem 10. Let $v \in K_j(q_1)$ with $v^*v = 1$ and $z_k^*v = 0$ for $k < i$. Then

$$\lambda_i \leq \theta_i \leq v^*Av + \sum_{k=1}^{i-1} \epsilon_k^2(\lambda_n - \theta_k) \leq v^*Av + \sum_{k=1}^{i-1} \epsilon_k^2(\lambda_n - \lambda_k) .$$

The first and last inequalities follow directly from the Cauchy interlace theorem. To obtain the middle inequality we first prove the following lemma.

Lemma. Let A be negative semidefinite and let v be as in Theorem 10. Then

$$\theta_i \leq v^*Av - \sum_{k=1}^{i-1} \epsilon_k^2 \theta_k .$$

Proof of Lemma. Resolve v as

$$v = v_1 + \sum_{k=1}^{i-1} \nu_k y_k$$

where $v_1^* y_k = 0$ for $k < i$. By (2) and the Schwartz inequality,

$$\begin{aligned} |\nu_k| &= |v^* y_k| , \\ &= |v^*(y_k - \gamma_{kk} z_k)| , \\ &\leq \epsilon_k . \end{aligned}$$

By Theorem 3 the Ritz vectors are A-orthogonal so

$$\begin{aligned} v^*Av &= v_1^*Av_1 + \sum_{k=1}^{i-1} \nu_k^2\theta_k , \\ &\le v_1^*Av_1/v_1^*v_1 + \sum_{k=1}^{i-1} \epsilon_k^2\theta_k , \end{aligned}$$

since $v_1^*Av_1 \le 0$. Using the Courant-Fischer characterization of eigenvalues of a symmetric matrix it can be shown that

$$\theta_i = \min_{\substack{x \in K_j(q_1) \\ x^*y_k=0 \text{ for } k<i}} (x^*Ax/x^*x) .$$

Since v_1 is a candidate for x, the Lemma follows. Since the Lanczos algorithm is translation-invariant (see Theorem 8), the Theorem follows from the Lemma applied to $A - \lambda_n$. □

We now use the polynomial characterization of $K_j(q_1)$ to obtain a good vector v for use in Theorem 10. Let $q_1 = \sum_{i=1}^{n} \sigma_i z_i$ be the spectral decomposition of q_1. Then $\phi(A)q_1 = \sum_{i=1}^{n} \sigma_i \phi(\lambda_i) z_i$ for any polynomial ϕ. What is needed is a polynomial ϕ such that

$\phi(\lambda_k) = 0$ for $k < i$,

$\phi(\lambda_i)$ is "large", and

$\phi(\lambda_k)$ is "small" for $k > i$.

Ideally ϕ should vanish at all the eigenvalues of A except λ_i, but this takes a polynomial of degree $n-1$ which means $j = n$ and all the "approximations" are exact. For real applications j will be much smaller than n and a different approach is needed.

Partition the set $\{1,2,\ldots,n\}$ into three sets I, T, and K. I will consist of the number i alone, T will be the index set for those eigenvalues of A at which ϕ will be made to vanish, and K will index the rest of the eigenvalues at which ϕ will be made small using a Tchebychev polynomial. By the hypothesis of Theorem 10 T must include $1,2,\ldots,i-1$. It may be advantageous for T to be larger. The larger the set T the lower the degree of the Tchebychev polynomial for K, but the interval containing the eigenvalues indexed by K is also smaller. It is this trade off which determines the optimal size of T. Both Kaniel and Paige observed that it may be advantageous for T to include $i+1,i+2,\ldots,s$, for some s but neither mentions that it may also help to include $n,n-1,\ldots,t$ in T for some number t. Let $|T|$ be the number of elements in T.

Theorem 11. Let I, T, and K be a partition of $\{1,2,\ldots,n\}$ such that $\{1,2,\ldots,i-1\} \subseteq T$ and $I = \{i\}$. Let $m = j-1-|T|$. Then

$$\lambda_i \leq \theta_i \leq \lambda_i + \frac{\sum_{k \in K} [\sigma_k \prod_{j \in T} (\lambda_k - \lambda_j)]^2 (\lambda_k - \lambda_i)}{[\sigma_i T_m(1+2\rho) \prod_{j \in T} (\lambda_i - \lambda_j)]^2} + \sum_{k=1}^{i-1} \epsilon_k^2 (\lambda_n - \lambda_k)$$

where T_m is the m^{th} Tchebychev polynomial (of the first kind), $q_1 = \sum_{k=1}^{n} \sigma_k z_k$ is the spectral decomposition of q_1, and $\rho = (\lambda_{\underline{k}} - \lambda_i)/(\lambda_{\bar{k}} - \lambda_{\underline{k}})$ where $\bar{k}$ and $\underline{k}$ are the largest and smallest elements of K respectively.

Proof. Let $\phi(x) = \hat{T}_m(x) \prod_{j \in T} (x-\lambda_j)$, where $\hat{T}_m(x)$ is the m^{th} Tchebychev polynomial scaled and translated to the interval $[\lambda_{\underline{k}}, \lambda_{\bar{k}}]$. By the definition of m, $\deg \phi = j-1$. If we let $w = \phi(A)q_1$ then

$$\begin{aligned} \|w\|^2 &\geq [\sigma_i \hat{T}_m(\lambda_i) \prod_{j \in T} (\lambda_i-\lambda_j)]^2 \\ &= [\sigma_i T_m(1+2\rho) \prod_{j \in T} (\lambda_i-\lambda_j)]^2 \end{aligned}$$

and for $k \in K$

$$\begin{aligned} |\phi(\lambda_k)| &= |\hat{T}_m(\lambda_k) \prod_{j \in T} (\lambda_k-\lambda_j)| \\ &\leq | \prod_{j \in T} (\lambda_k-\lambda_j)| \ . \end{aligned}$$

Now let $v = w/\|w\|$, apply Theorem 10, and use these bounds to obtain the result. □

To illustrate the bounds obtainable from these theorems we consider the following example.

Example. Let the eigenvalues of A satisfy

$$\begin{aligned} \lambda_1 &= 0 \\ \lambda_2 &= .01 \\ \lambda_3 &= .04 \\ .1 \leq \lambda_k &\leq .9 \quad \text{for} \quad k = 4,5,\dots,n-1 \\ \lambda_n &= 1.0 \ . \end{aligned}$$

Let $q_1^* z_i = .01$ for $i = 1,2,3$. Assume the Lanczos algorithm is interrupted at $j = 53$ and the three smallest Ritz values are computed. To bound the accuracy of each of these Ritz values we choose, for $i = 1,2,3,$

$$K = \{4,5,\ldots,n-1\} ,$$
$$I = \{i\} , \text{ and}$$
$$T = \{1,2,3,n\} \setminus I .$$

For all i we can bound the numerator appearing in Theorem 11 by

$$\sum_{k\in K} [\sigma_k \prod_{j\in T} (\lambda_k-\lambda_j)]^2(\lambda_k-\lambda_i) \le \sum_{k\in K} \sigma_k^2 \le 1 .$$

Also $T_m(\cosh x) = \cosh(mx) > e^{mx}/2$. Using these bounds, Theorem 11 assures that

$$\begin{aligned}\theta_1 - \lambda_1 &\le [\sigma_1 T_{49}(1+2\rho) \prod_{j\in T} (\lambda_1-\lambda_j)]^{-2} ,\\ &\le [.01T_{49}(1+2(.125))(0-.01)(0-.04)(0-1.0)]^{-2} ,\\ &\le 10^{-18} .\end{aligned}$$

Then by Theorem 9,

$$\begin{aligned}e_1^2 &\le 10^{-18}/.01 ,\\ &= 10^{-16} .\end{aligned}$$

By Theorem 11,

$$\begin{aligned}\theta_2 - \lambda_2 &\le [.01 \times T_{49}(1.225) \times .01 \times .03 \times .99]^{-2} + 10^{-16} ,\\ &\le .103 \times 10^{-15} .\end{aligned}$$

By Theorem 9 again,

$$\begin{aligned}e_2^2 &\le (.103\times 10^{-15} + .04\times 10^{-16})/.03 ,\\ &\le .345 \times 10^{-14} .\end{aligned}$$

Finally,

$$\begin{aligned}\theta_3 - \lambda_3 &\leq [.01T_{49}(1.15) \times .04 \times .03 \times .96]^{-2} + .99 \times .345 \times 10^{-14} \ , \\ &< .3 \times 10^{-12}\end{aligned}$$

and

$$\begin{aligned}\epsilon_3^2 &\leq (.3\times10^{-12} + .1\times10^{-16} + .31\times10^{-15})/.06 \ , \\ &\leq .51 \times 10^{-11} \ .\end{aligned}$$

Summarizing:

$\theta_1 - \lambda_1 \leq 10^{-18}$	$\epsilon_1^2 \leq 10^{-16}$
$\theta_2 - \lambda_2 \leq .103 \times 10^{-15}$	$e_2^2 \leq .345 \times 10^{-14}$
$\theta_3 - \lambda_3 \leq .3 \times 10^{-12}$	$e_3^2 \leq .51 \times 10^{-11}$

These bounds are slightly stronger than those obtained by Kaniel and Paige on essentially the same example due to the assumed gap between λ_{n-1} and λ_n. □

The results in this section are not often useful in practical problems since the gaps in the spectrum are not usually known. However these results are important from a theoretical standpoint. Consider Theorem 11 which, in the simplest case of $i = 1$ and $T = \emptyset$, states

$$\begin{aligned}0 \leq \theta_1 - \lambda_1 &\leq \frac{\sum_{k=2}^{n} \sigma_k^2(\lambda_k - \lambda_1)}{[\sigma_1 T_{j-1}(1+2\rho)]^2} \ , \\ &\leq \frac{\tan^2\psi[\lambda_n - \lambda_1]}{T_{j-1}^2(1+2\rho)} \ ,\end{aligned}$$

where ψ is the (acute) angle between z_1 and q_1 and $\rho = (\lambda_2-\lambda_1)/(\lambda_n-\lambda_2)$.

The role of ρ shows clearly that the larger the relative separation of λ_1 and λ_2, the faster the convergence to λ_1. On the other hand if $\psi \approx \pi/2$, that is if q_1 is almost orthogonal to z_1, convergence to λ_1 wil be slow even if λ_1 is well separated. Is it possible to choose q_1 so that no Ritz value converges quickly? This question is examined in the next section.

1.4 Slow Convergence

Let A and $\tau > 0$ be given. Define a Ritz pair (y_i,θ_i) to be converged if $\|Ay_i-y_i\theta_i\| = \beta_{ji} \leq \tau$. Is there a starting vector q_1 such that no Ritz pair converges before $j = n$? For matrices with well separated eigenvalues the answer is yes. Even for matrices with clustered eigenvalues there are starting vectors which delay convergence for a long time. In this section we will derive formulas for these perverse starting vectors.

Of course q_1 and A determine everything in the Lanczos algorithm. The main result of this section is the derivation of a simple relationship between q_1 and the Ritz values at the penultimate step (the step before termination).

Theorem 12. Let $\lambda_1 < \lambda_2 < \cdots < \lambda_n$ be the eigenvalues of A and let $Z = (z_1, z_2, \ldots, z_n)$ be the corresponding normalized eigenvectors. Let $\mu_1, \mu_2, \ldots, \mu_{n-1}$ be any numbers such that

$$\lambda_1 < \mu_1 < \lambda_2 < \cdots < \mu_{n-1} < \lambda_n .$$

If the Lanczos algorithm is run on (A, q_1) for $n-1$ steps, then $\mu_1, \mu_2, \ldots, \mu_{n-1}$ are the eigenvalues of T_{n-1} iff $q_1 = Zp_1$ where

$$p_{i1}^2 = \pi_n^2 [\prod_{\substack{j=1 \\ j \neq i}}^{n} (\lambda_i - \lambda_j) \prod_{j=1}^{n-1} (\lambda_i - \mu_j)]^{-1}$$

and the constant π_n^2 can always be determined.

Note that the formula for p_{i1}^2 can be rewritten as

$$p_{i1}^2 = \pi_n^2 [\chi_n'(\lambda_i) \chi_{n-1}(\lambda_i)]^{-1}$$

where $\chi_j(\xi)$ is the characteristic polynomial of T_j. In particular $\chi_n(\xi)$ is also the characteristic polynomial of A since T_n is similar to A.

Results similar to Theorem 12 have been used by D. Boley and G.H. Golub [Boley and Golub 1977] and C. de Boor and G.H. Golub [de Boor and Golub] in the context of inverse eigenvalue problems for banded matrices.

To prove the theorem we first prove two lemmas.

Definition. Let adj(R) be the transpose of the matrix of co-factors of R. This is usually called the adjugate or classical adjoint of R. By the Cauchy-Binet Theorem we have

$$R\,\mathrm{adj}(R) = \mathrm{adj}(R)R = \det(R)I \ .$$

Lemma 1 (Thompson and McEnteggert). Let $A = Z\Lambda Z^*$ be the spectral decomposition of A, where $\Lambda = \mathrm{diag}(\lambda_1,\lambda_2,\dots,\lambda_n)$ and $Z = (z_1,z_2,\dots,z_n)$. Then for $i = 1,2,\dots,n$

$$\begin{aligned}\mathrm{adj}(\lambda_i - A) &= \prod_{\substack{j=1\\ j\neq i}}^{n} (\lambda_i-\lambda_j)z_iz_i^* \\ &= \chi_A'(\lambda_i)z_iz_i^* \ ,\end{aligned}$$

where $\chi_A'(\xi)$ is the derivative of the characteristic polynomial of A.

Note that if λ_i is a multiple eigenvalue of A then $\chi_A'(\lambda_i) = 0$, so that the ambiguity in the choice of eigenvectors is unimportant.

Proof of Lemma 1. Let $\mu \neq \lambda_i$ for all i so that $(\mu-A)^{-1}$ exists and

$$\begin{aligned}\mathrm{adj}(\mu-A) &= [\det(\mu-A)](\mu-A)^{-1} \\ &= \chi_A(\mu)Z(\mu-\Lambda)^{-1}Z^* \\ &= Z\Delta Z^* \qquad (1)\end{aligned}$$

where $\Delta = \mathrm{diag}(\delta_1,\delta_2,\dots,\delta_n)$ with

$$\begin{aligned} \delta_k &= \chi_A(\mu)/(\mu-\lambda_k) \\ &= \prod_{\substack{j=1 \\ j\neq k}}^{n} (\mu-\lambda_j) \ . \end{aligned}$$

Since computing cofactors does not involve division, $\operatorname{adj}(R)$ is a continuous function of R. Therefore by continuity, equation (1) must hold even for $\mu = \lambda_i$. Setting $\mu = \lambda_i$ yields

$$\operatorname{adj}(\lambda_i - A) = Z\Delta Z^*$$

with

$$\begin{aligned} \delta_k &= \prod_{\substack{j=1 \\ j\neq k}}^{n} (\lambda_i-\lambda_j) \\ &= \begin{matrix} 0 & \text{for } k \neq i \\ \chi_A'(\lambda_i) & \text{for } k = i \end{matrix} \end{aligned}$$

and the result follows. □

Thompson and McEnteggert were working with general Hermitian matrices. The application of their result to tridiagonal matrices was made by Paige [Paige 1971].

<u>Notation</u>. Let

$$T_{r,t} = \begin{bmatrix} \alpha_r & \beta_r & & & \\ \beta_r & \alpha_{r+1} & \beta_{r+1} & & \bigcirc \\ & \beta_{r+1} & \ddots & \ddots & \\ & \bigcirc & \ddots & & \beta_{t-1} \\ & & & \beta_{t-1} & \alpha_t \end{bmatrix},$$

let $T_t = T_{1,t}$ as before, let $\chi_{r,t}(\xi) = \det(\xi - T_{r,t})$, the characteristic polynomial of $T_{r,t}$, let $\chi_t(\xi) = \chi_{1,t}(\xi)$, and let $\chi_{r,r-1}(\xi) \equiv 1$, for all r.

<u>Lemma 2</u> (Paige). Let $T_n = S\Theta S^*$ be the spectral decomposition of T_n with $\Theta = \mathrm{diag}(\theta_1,\theta_2,\ldots,\theta_n)$ and $S = (s_1,s_2,\ldots,s_n)$. Then for $r \leq t$ and all i

$$\chi_n'(\theta_i)s_{ri}s_{ti} = \chi_{1,r-1}(\theta_i)\beta_r\beta_{r+1}\cdots\beta_{t-1}\chi_{t+1,n}(\theta_i) \ .$$

<u>Proof of Lemma 2</u>. By Lemma 1,

$$\mathrm{adj}(\theta_i - T_n) = \chi_n'(\theta_i)s_is_i^* \ . \tag{2}$$

The (r,t) element of the R.H.S. of (2) is $\chi_n'(\theta_i)s_{ri}s_{ti}$. Because of the tridiagonal form of T_n, the (r,t) element of the L.H.S. of (2) is $\chi_{1,r-1}(\theta_i)\beta_r\beta_{r+1}\cdots\beta_{t-1}\chi_{t+1,n}(\theta_i)$. For example

$$\theta_i - T_6 = \begin{bmatrix} \boxed{\theta_i-\alpha_1} & -\beta_1 & & & \\ -\beta_1 & \theta_i-\alpha_2 & -\beta_2 & & \bigcirc \\ & \boxed{-\beta_2} & \theta_i-\alpha_3 & -\beta_3 & \\ \bigcirc & & -\beta_3 & \boxed{\begin{matrix} \theta_i-\alpha_4 & -\beta_4 & \\ -\beta_4 & \theta_i-\alpha_5 & -\beta_5 \\ & -\beta_5 & \theta_i-\alpha_6 \end{matrix}} \end{bmatrix} .$$

The circled elements contribute to the $(2,3)$ cofactor. The minus signs associated with the β's cancel with the alternating signs assigned to the cofactors. □

This lemma gives many relationships among the elements of S. We will need two of them for proving Theorem 12, namely for $i - 1,2,\ldots,n$,

(3) $$s_{i1}s_{ni}\chi_n'(\theta_i) = \beta_1\beta_2\cdots\beta_{n-1} \equiv \pi_n \text{ , a constant,}$$

and

(4) $$s_{ni}^2\chi_n'(\theta_i) = \chi_{n-1}(\theta_i) \text{ .}$$

Proof of Theorem 12. By the invariance properties given in Theorem 8, if $L^j(A,q_1) = (Q_j,T_j)$ then $L^j(Z^*AZ,Z^*q_1) = (Z^*Q_j,T_j)$. Since $Z^*AZ = \Lambda$ and $Z^*q_1 = p_1$, $L^j(\Lambda,p_1) = (P_j,T_j)$, where $P_j = Z^*Q_j$. The Lanczos algorithm will terminate at the n^{th} step with

$$\Lambda P_n = P_nT_n \text{ .}$$

Thus $P_n = S^*$, the transpose of the matrix of eigenvectors of T_n. Equation (3) can now be interpreted as relating the first Lanczos vector p_1 to the last Lanczos vector p_n. Equation (4) relates p_n to the eigenvalues of T_{n-1} and the eigenvalues of T_n. The eigenvalues of T_n are just $\lambda_1,\lambda_2,\ldots,\lambda_n$, since T_n, A, and Λ are all similar. Combining equations (3) and (4) and changing to the P notation yields

(5) $$p_{i1}^2\chi_n'(\lambda_i)\chi_{n-1}(\lambda_i) = \pi_n^2 \text{ ,}$$

for any starting vector p_1.

If p_1 is given, then (5) gives the values of $\chi_{n-1}(\lambda_i)$, for $i = 1,2,\ldots,n$ in terms of the constant π_n^2. $\chi_{n-1}(\xi)$ is a polynomial of degree n-1. By choosing an arbitrary value for π_n^2 (say 1) the roots of $\chi_{n-1}(\xi)$ can be found by interpolation. The value of π_n^2 can then be found from the fact that $\chi_{n-1}(\xi)$ is a monic polynomial.

If $\mu_1,\mu_2,\ldots,\mu_{n-1}$ are specified then by choosing an arbitrary value for π_n^2 (say 1), tentative values $\hat{p}_{i1}^2$ can be calculated for all i. Since $\sum p_{i1}^2 = \|p_1\|^2 = 1$, $\pi_n^2 = (\sum\hat{p}_{i1}^2)^{-1}$ and the tentative values can be correctly normalized.

The choice in signs of the elements of p_1 merely reflect the choice of signs for the eigenvectors of T_n. All choices yield the same T_n and hence the same μ's. □

For the original matrix A, the specified starting vector q_1 depends on both the eigenvalues and eigenvectors of A. The expression $q_1 = Zp_1$ clarifies their roles; Z is independent of Λ and p_1 is independent of Z.

Example. Let $\Lambda = \text{diag}(1,3,5,7,9)$ and let $\mu_i = 2i$ for $i = 1,2,3,4$.

$$\begin{aligned}
\chi'(1)\chi_\mu(1) &= (1-3)(1-5)(1-7)(1-9)(1-2)(1-4)(1-6)(1-8) = 40320\\
\chi'(3)\chi_\mu(3) &= (3-1)(3-5)(3-7)(3-9)(3-2)(3-4)(3-6)(3-8) = 1440\\
\chi'(5)\chi_\mu(5) &= (5-1)(5-3)(5-7)(5-9)(5-2)(5-4)(5-6)(5-8) = 576\\
\chi'(7)\chi_\mu(7) &= (7-1)(7-3)(7-5)(7-9)(7-2)(7-4)(7-6)(7-8) = 1440\\
\chi'(9)\chi_\mu(9) &= (9-1)(9-3)(9-5)(9-7)(9-2)(9-4)(9-6)(9-8) = 40320
\end{aligned}$$

$$\begin{aligned}
p_{11} = p_{51} &= \pi_5/\sqrt{40320} = .00498\pi_5\\
p_{21} = p_{41} &= \pi_5/\sqrt{1440} = .02635\pi_5\\
p_{31} &= \pi_5/\sqrt{576} = .04167\pi_5
\end{aligned}$$

By normalization, $\pi_5 = 17.749$ and

$$p_1 = (.0880,\ .4677,\ .7396,\ .4677,\ .0880)^* .$$

The Lanczos algorithm run on (Λ,p_1) yielded a T_4 with eigenvalues 2, 4, 6, and 8 correct to the precision of the machine used. □

Theorem 12 shows that an appropriate choice of the starting vector can place the eigenvalues of T_{n-1} anywhere between the λ's. Let $\mu_1,\mu_2,\ldots,\mu_{n-1}$ be fixed and let q_1 be chosen. What can be said about convergence in this case? The following result gives a lower bound on all the β_{ji} for $j < n$.

<u>Theorem 13</u>. Let the Lanczos algorithm on (A,q_1) yield $\mu_1;\mu_2,\ldots,\mu_{n-1}$ as the eigenvalues of T_{n-1}. Then for all $j < n$ and all $i \leq j$

$$\beta_{ji} \geq \delta_\mu/2 \ ,$$

where $\delta_\mu = \min |\mu_j-\lambda_i|$.

<u>Proof</u>. Let (y_i,θ_i) be a Ritz pair with residual norm β_{ji}. By Theorem 5 there must be a λ such that

$$|\theta_i-\lambda| \leq \beta_{ji} \ .$$

By Theorem 6 (with $k=n-1$) there must be a μ such that

$$|\theta_i-\mu| \leq \beta_{ji} \ .$$

The smallest value of β_{ji} which can satisfy both inequalities is $\beta_{ji} = \delta_\mu/2$. □

<u>Corollary</u>. Let $\delta_A = \min_{i \neq j} |\lambda_i - \lambda_j|$. If τ (the given convergence tolerance) satisfies $\tau \leq \delta_A/4$, then there exists a starting vector q_1 such that the Lanczos algorithm run on (A, q_1) will have

$$\beta_{ji} \geq \tau$$

for all $j < r$ and $i \leq j$.

<u>Proof</u>. Let $\mu_i = (\lambda_i + \lambda_{i+1})/2$ for $i = 1, 2, \ldots, n-1$. By Theorem 12 there is a q_1 such that the μ's are the eigenvalues of T_{n-1}. Then $\delta_\mu = \delta_A/2$ and the result follows from Theorem 13. □

This result does not imply that no Ritz value will be accurate enough. It only guarantees that the corresponding β_{ji} will not reveal such accuracy. In the previous example of $\Lambda = \mathrm{diag}(1,3,5,7,9)$ and $\mu_i = 2i$, for $i = 1,2,3,4$, $\theta_2^{(3)}$, the middle eigenvalue of T_3 is 5, correct to working accuracy. $\beta_{32} = 1.25$ which shows that this fortuitous accuracy is due to the symmetry of the example, rather than the accuracy of the Ritz vector.

If $\delta_A/4 < \tau$ then the corollary does not apply, but it is still possible to find perverse starting vectors which delay convergence for a long time.

Theorem 14. Let $\mathcal{W}$ be an A-invariant subspace of maximal dimension such that $\delta_{\bar{A}}/4 > \tau$, where $\bar{A}$ is A restricted to $\mathcal{W}$. Let $m = \dim \mathcal{W}$. Then there exists a starting vector for A which delays convergence until $j = m$.

Proof. Apply the corollary to $\bar{A}$ to obtain a starting vector q_1 for $\bar{A}$ which delays convergence until $j = m$. By Theorem 8 the algorithm run on $(\bar{A},q_1)$ and (A,q_1) produces the same T_j for all j. Hence this q_1 will delay convergence for A until $j = m$. □

2. The Lanczos Algorithm in Finite Precision

Chapter 1 paints a very rosy picture of the Lanczos algorithm. If these theoretical results were closely approximated in practice, the Lanczos algorithm would be the preferred method of tridiagonalizing _any_ symmetric matrix. However as was known to Lanczos when he introduced the algorithm the computed quantities can deviate greatly from their theoretical counterparts.

At this point we make an important change in notation. From now on Q_j and T_j will represent the quantities actually computed with finite precision arithmetic. However the spectral decomposition $T_j = S_j\Theta_jS_j^*$, with Θ_j diagonal and S_j orthogonal, will be assumed to hold exactly. High quality subroutines exist for accurately computing Θ_j and S_j and the small roundoff errors committed therein are always dominated by the errors inherent in T_j. Recall that the columns of S are normalized so that the bottom row of S is all positive.

The eigenvalues of T_j will still be called Ritz values and the columns of $Y_j = Q_jS_j$ will still be called Ritz vectors even when Q_j is not even close to orthonormal. In principal it is possible to compute the true Ritz vectors from $\text{span}(Q_j)$ but this would be very expensive and nullify the advantages of the Lanczos algorithm. We will never consider the true Ritz pairs so no confusion should arise.

2.1 Description and Example

In exact arithmetic the quantities computed at the j^{th} step of the Lanczos algorithm satisfy the equations

(1) $$\boxed{\quad A \quad}\,\boxed{Q_j} - \boxed{Q_j}\,\boxed{T_j} = \boxed{0 \,\big|\, r_j}$$

and

(2) $$1 - Q_j^* Q_j = 0 \ .$$

Equation (1) can be written compactly as

(3) $$AQ_j - Q_j T_j = r_j e_j^*$$

where $e_j^* = (0,0,\ldots,0,1)$ has j elements. In finite precision arithmetic neither (2) nor (3) will be satisfied exactly. Instead they must be replaced by

(4) $$AQ_j - Q_j T_j = r_j e_j^* + F_j$$

and

(5) $$1 - Q_j^* Q_j = G_j$$

where F_j and G_j account for the rounding errors. Bounds on $\|F_j\|$ and $\|G_j\|$ depend on the specific implementation of the algorithm but the surprising fact is that while any reasonable implementation will keep $\|F_j\|$ tiny ($\|F_j\| \doteq \epsilon\|A\|$, where ϵ is the relative machine precision), no implementation of the simple Lanczos algorithm (the three term recurrence) yields a small a priori bound on $\|G_j\|$. This "loss of orthogonality" among the Lanczos vectors (columns of Q_j) is the infamous instability of the algorithm.

W. Kahan has shown that the bound given in Theorem 2 of Chapter 1 for an orthonormal matrix Q_j fails gracefully as Q_j loses orthogonality. The number $\sigma_1(Q_j) = (\lambda_1[Q_j^*Q_j])^{1/2}$, the smallest singular value of Q_j, appears as a measure of the loss of orthogonality in Q_j.

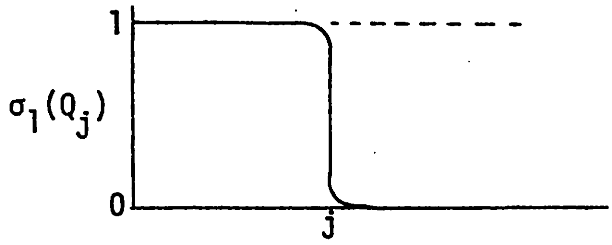
<u>Theorem 1</u> (Kahan). Let Q_j be any $n \times j$ matrix, let H be a $j \times j$ symmetric matrix with eigenvalues $\theta_1, \theta_2, \ldots, \theta_j$, and let $R = AQ_j - A_jH$. Then there exists $1', 2', \ldots, j'$ distinct integers such that for all i

$$|\theta_i - \lambda_{i'}| \leq \sqrt{2}\|R\|/\sigma_1(Q_j) ,$$

where $\sigma_1(Q_j)$ is the smallest singular value of Q_j.

The proof is in [Kahan 1967].

Unfortunately computational experience indicates that the graph of $\sigma_1(Q_j)$ looks qualitatively like:



That is, $\sigma_1(Q_j)$ rapidly approaches 0 once it has moved away from 1. We give an example of this phenomenon.

Example 1.

$n = 10$

$\epsilon = .6 \times 10^{-7}$ (relative machine precision)

$A = \mathrm{diag}(0,.01,.02,\ldots,.08,1.0)$

$q_1 = u/\|u\|$, where $u = (1,1,\ldots,1)^*$

Step j	$\sigma_1(Q_j)$
1	1.00000
2	1.00000
3	1.00000
4	1.00000
5	.99997
6	.99852
7	.92250
8	.08190
9	.00104
10	.000001

Parlett and Kahan also give graphs of $\sigma_1(Q_j)$ in [Kahan and Parlett 1976].

In theory $\beta_{10} = 0$ since 11 orthonormal 10-vectors cannot exist. In Example 1 orthogonality has been lost completely by step 10 and there is no compelling reason why β_{10} should be tiny. Indeed for Example 1 $\beta_{10} = .01033$ which is not small at all compared to $\|A\| = 1$.

The Ritz values at step $j = 10$ are

i	$\theta_i^{(10)}$
1	.000002
2	.010116
3	.020998
4	.033318
5	.045409
6	.058857
7	.069857
8	.079996
9	1.000000
10	1.000000

and these values shed some light on what has happened. A spurious multiplicity has appeared at 1.0 where A has only a simple eigenvalue. The two Ritz vectors, y_9 and y_{10}, are <u>both</u> good approximations to the one eigenvector z_{10}. Therefore y_9 and y_{10} are essentially parallel and $\sigma_1(Q_{10})$ must be tiny. □

This example will be examined more closely in the light of later results.

2.2 A Misleading Example

Theorem 5 of Chapter 1 shows that the residual norm of a Ritz vector y_i can be computed without computing y_i. Namely

$$\|Ay_i - y_i\theta_i\| = \beta_{ji}$$

where $\beta_{ji} = \beta_j s_{ji}$. Therefore it is possible to bound the accuracy of θ_i without computing y_i and if only eigenvalues are desired the

Lnaczos vectors need not be saved at all.

In finite precision arithmetic we obtain the following analog of Theorem 5.

<u>Theorem 2</u>. Let $AQ_j - Q_jT_j = \beta_j q_{j+1} e_j^* + F_j$. Let $y_i = Q_j s_i$ with $T_j s_i = s_i\theta_i$ and $\|s_i\| = 1$. Then

$$\|Ay_i - y_i\theta_i\| \leq \beta_{ji} + \|F_j s_i\| \leq \beta_{ji} + \|F_j\| .$$

<u>Proof</u>. The second inequality follows from $\|s_i\| = 1$. To obtain the first inequality multiply the matrix equation on the right by s_i to find

$$AQ_j s_i - Q_j T_j s_i = \beta_j q_{j+1} e_j^* s_i + F_j s_i \ ,$$

which simplifies to

$$Ay_i - y_i\theta_i = \beta_i s_{ji} q_{j+1} + F_j s_i \ .$$

Since $\|q_{j+1}\| = 1$ and $\beta_{ji} = \beta_j s_{ji}$, the result follows from taking the norm of each side. □

Theorem 2 shows that β_{ji} is a good estimate of the residual norm of y_i provided that $\beta_{ji} \geq \|F_j\|$. For the simple Lanczos algorithm $\|F_j\|$ is always tiny, like roundoff in A (see Appendix 1), so $\beta_{ji} \geq \|F_j\|$ will hold in almost all cases and β_{ji} can be used as a good estimate of the residual norm of y_i.

Unfortunately this need not lead to a good estimate of the accuracy of θ_i. Since Q_j is not orthonormal, $y_i = Q_j s_i$ need not have length 1. The best obtainable bound for the accuracy of θ_i in the absence of further information is

$$\min_k |\theta_i-\lambda_k| \le \|Ay_i-y_i\theta_i\|/\|y_i\| . \tag{1}$$

A lower bound for $\|y_i\|$ is given by the following lemma in which the smallest singular value of Q_j again appears.

<u>Lemma 1</u>. Let $y_i = Q_j s_i$ with $\|s_i\| = 1$. Then

$$\|y_i\| \ge \sigma_1(Q_j) ,$$

where $\sigma_1(Q_j) = (\lambda_1[Q_j^*Q_j])^{1/2}$, the smallest singular value of Q_j.

<u>Proof</u>.

$$\begin{aligned} \|y_i\|^2 &= y_i^*y_i \\ &= s_i^*Q_j^*Q_js_i \\ &\ge \lambda_1[Q_j^*Q_j] \quad (Q_j^*Q_j \text{ is symmetric}) . \end{aligned}$$

Since $Q_j^*Q_j$ is nonnegative definite, the square root of both sides can be taken. □

Thus it would appear that it is necessary to either calculate $\|y_i\|$ or estimate $\sigma_1(Q_j^*Q_j)$ to obtain an error bound for θ_i. This suspicion is further strengthened by the following perverse example.

<u>Example 2</u>. Take as an instance of $AQ_2 = Q_2T_2 + r_2e_2^* + F_2$,

$$[1][1\ -1] = [1\ -1]\begin{bmatrix} 1001 & 1000 \\ 1000 & 1001 \end{bmatrix}$$

where both F_2 and r_2 equal zero. Observe that T_2 is similar to

$$\Theta_2 = \mathrm{diag}(1,2001)$$

and since $\beta_2 = \|r_2\| = 0$, both β_{21} and β_{22} are zero. Furthermore, since $F_2 = 0$ as well the residual norms of both y_1 and y_2 are zero by Theorem 2. If A were hidden from us we might be led to believe that 2001 was an eigenvalue of A. This paradox is resolved by computing the Ritz vectors. We have

$$S_2 = (\sqrt{2}/2)\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

and

$$Y_2 = Q_2 S_2 = [\sqrt{2} \quad 0] \ .$$

Thus $y_2 = 0$ and $\|Ay_2 - y_2\theta_2\| = 0$ even though $\theta_2 = 2001$ is <u>not</u> an eigenvalue of A. The vanishing of y_2 is possible because the columns of Q_2 are linearly dependent and so $\sigma_1(Q_2) = 0$. □

Example 2 itself is <u>not</u> an example of the Lanczos algorithm since α_1, the (1,1) element of T_2, is 1001 whereas α_1 for the Lanczos algorithm would be $q_1^* A q_1 = 1$. In order to analyze the Lanczos algorithm it is necessary to find a characterization of it which distinguishes the true examples (like Example 1) from the spurious ones (like Example 2). To do this we must investigate the specific manner in which orthogonality is lost in the course of the algorithm.

2.3 Loss of Orthogonality

In his Ph.D. thesis, C. Paige derived a powerful characterization of the manner in which orthogonality is lost in the course of the simple Lanczos algorithm. This result is central to understanding the behavior of the algorithm in the context of finite precision arithmetic. Since it has never been published in the open literature its full derivation will be given here.

It is the vector $Q_j^* q_{j+1}$, which would be zero in exact arithmetic, which displays how well orthogonality is preserved at the j^{th} step. However the elements of this vector are difficult to analyze and little insight can be gained from them. A change of basis is needed to clarify the situation. It is the vector $S_j^* Q_j^* q_{j+1} = Y_j^* q_{j+1}$ which can be easily described.

Theorem 3 (Paige). At any step j of the simple Lanczos algorithm and for any Ritz vector $y_i = Q_j s_i$,

$$|y_i^* q_{j+1}| = \epsilon \|A\| \gamma_{ji} / \beta_{ji}$$

where $\beta_{ji} = \beta_j s_{ji}$, ϵ is the relative machine precision, and $\gamma_{ji} \doteq 1$.

Remarks. By Theorem 2, β_{ji} is essentially the residual norm of y_i so Theorem 3 states that the smaller the residual norm of y_i the greater the loss of orthogonality of q_{j+1} in the direction of y_i. This can be stated as

$$\boxed{\text{loss of orthogonality} \Leftrightarrow \text{convergence}} \quad .$$

The proof of Theorem 3 is rather long so we break it into several lemmas starting from the basic equation

$$AQ_j = Q_jT_j + r_je_j^* + F_j \ . \tag{1}$$

Lemma 1. If (1) holds then

$$Q_j^*r_je_j^* = (1-Q_j^*Q_j)T_j - T_j(1-Q_j^*Q_j) + F_j^*Q_j - Q_j^*F_j + e_jr_j^*Q_j \ .$$

Proof of Lemma 1. Multiplication of (1) on the left by Q_j^* yields

$$Q_j^*AQ_j = Q_j^*Q_jT_j + Q_j^*r_je_j^* + Q_j^*F_j \ . \tag{2}$$

Since $Q_j^*AQ_j$ is symmetric we may subtract (2) from the transpose of (2) and rearrange to obtain

$$Q_j^*r_je_j^* = (1-Q_j^*Q_j)T_j - T_j(1-Q_j^*Q_j) + F_j^*Q_j - Q_j^*F_j + e_jr_j^*Q_j \ . \quad \square \tag{3}$$

Notation. Let $\nabla(R)$ be the upper triangular part of the matrix R (including the diagonal). For any conformal R and S, $\nabla(R+S) = \nabla(R) + \nabla(S)$ but $\nabla(RS) \neq \nabla(R)\nabla(S)$ in general. Let $1 - Q_j^*Q_j = C_j^* + \Delta_j + C_j$ with Δ_j diagonal and C_j strictly upper triangular.

Lemma 2. If (1) holds then

$$Q_j^*r_je_j^* = C_jT_j - T_jC_j + B_j + E_j \ ,$$

where $E_j = \nabla(F_j^*Q_j - Q_j^*F_j)$ and $B_j = \nabla(\Delta_jT_j - T_j\Delta_j) + \nabla(C_j^*T_j - T_jC_j^*) + \nabla(e_jr_j^*Q_j)$.

Proof of Lemma 2. Taking the upper triangular part of each side of (3) we find

$$(4)\quad \nabla(Q_j^*r_je_j^*) = \nabla((1-Q_j^*Q_j)T_j - T_j(1-Q_j^*Q_j)) + \nabla(F_j^*Q_j - Q_j^*F_j) + \nabla(e_jr_j^*Q_j)\ .$$

Substituting $1 - Q_j^*Q_j = C_j^* + \Delta_j + C_j$ and using the definition of ϵ_j we find

$$(5)\quad \nabla(Q_jr_je_j) = \nabla(C_jT_j - T_jC_j) + \nabla(\Delta_jT_j - T_j\Delta_j) + \nabla(C_j^*T_j - T_jC_j^*) + E_j + \nabla(e_jr_j^*Q_j)\ .$$

The matrix $Q_j^*r_je_j^*$ is zero except for the last column which is $Q_j^*r_j$. Therefore $\nabla(Q_j^*r_je_j^*) = Q_j^*r_je_j^*$. Furthermore C_j is strictly upper triangular and T_j is tridiagonal. Hence $\nabla(C_jT_j - T_jC_j) = C_jT_j - T_jC_j$ and the result follows. □

Lemma 3. If $y_i = Q_js_i$ with $T_js_i = s_i\theta_i$ then

$$|y_i^*q_{j+1}| = \epsilon\|A\|\gamma_{ji}/\beta_{ji}$$

with $\gamma_{ji} = |s_i^*(B_j+E_j)s_i|/\epsilon\|A\|$ and $\beta_{ji} = \beta_js_{ji}$.

Proof of Lemma 3. Multiplying the assertion of Lemma 2 on the left by s_i^* and on the right by s_i we obtain

$$(6)\qquad s_i^*Q_j^*r_je_j^*s_i = s_i^*(C_jT_j - T_jC_j)s_i + s_i^*(B_j+E_j)s_i\ ,$$

which simplifies to

$$(7)\qquad y_i^*r_js_{ji} = (s_i^*C_js_i)\theta_i - \theta_i(s_i^*C_js_i) + s_i^*(B_j+E_j)s_i\ .$$

Since $r_j = q_{j+1}\beta_j$, (7) becomes

(8) $$\beta_j s_{ji} y_i^* q_{j+1} = s_i^*(B_j+E_j)s_i$$

and the result follows. □

Lemma 4. B_j (of Lemma 3) is a bidiagonal matrix with

$$b_{11} = q_1^* r_1$$

$$b_{ii} = q_i^* r_i - q_{i-1}^* r_{i-1}$$

$$b_{i-1,i} = \beta_{i-1}(q_i^* q_i - q_{i-1}^* q_{i-1}) \qquad \text{for } i = 2,3,\ldots,j .$$

Proof of Lemma 4. From Lemma 2

(9) $$B_j = \nabla(\Delta_j T_j - T_j \Delta_j) + \nabla(C_j^* T_j - T_j C_j^*) + \nabla(e_j r_j^* Q_j) .$$

Δ_j is diagonal so that $\Delta_j T_j - T_j \Delta_j$ is tridiagonal with zero diagonal. Hence $\nabla(\Delta_j T_j - T_j \Delta_j)$ is super diagonal.

C_j^* is strictly lower triangular so that $C_j^* T_j - T_j C_j^*$ is lower triangular and $\nabla(C_j^* T_j - T_j C_j^*)$ is diagonal. Finally $er_j^* Q_j$ is zero except for the bottom row which is $r_j^* Q_j$. Thus $\nabla(e_j r_j^* Q_j)$ is zero except for the (j,j) element which is $r_j^* q_j = q_j^* r_j$.

Therefore B_j is bidiagonal and the formulas for its elements can be derived from the fact that $r_i = \beta_i q_{i+1}$ for all $i \leq j$. □

Logically it only remains to prove that $\|B_j\| \doteq \epsilon\|A\|$ and $\|E_j\| \doteq \epsilon\|A\|$. We will not do this for two reasons. The bounds for $\|B_j\|$ and $\|E_j\|$ depend on the specific implementation of the algorithm which we have not yet discussed. Furthermore these bounds depend on several characteristics of the matrix A which obscures the basic simplicity of the result. Therefore the statement and proof of the final lemma are relegated to Appendix 1 which completes the proof of Theorem 3. □

Corollary. For all $i \leq j$

$$\beta_{ji} \geq \epsilon \|A\| \gamma_{ji} / \|y_i\|$$

Proof. By Theorem 3 we have

$$\beta_{ji} |y_i^* q_{j+1}| = \epsilon \|A\| \gamma_{ji}$$

and so by the Schwartz inequality

$$\beta_{ji} \|y_i\| \|q_{j+1}\| \geq \epsilon \|A\| \gamma_{ji} \ .$$

Since $\|q_{j+1}\| = 1$ the result follows. □

The Corollary indicates the importance of the γ_{ji}. No β_{ji} can be much smaller than $\epsilon \|A\| \gamma_{ji}$. Thus the smaller the values of the γ_{ji}, the smaller the β_{ji} can be, and the better the accuracy which can be obtained in the Ritz values.

2.4 Implementation

The Corollary at the end of Section 3 provides a natural way to discriminate between various implementations of the simple Lanczos algorithm. The smaller the bounds on $\|E_j\|$ and $\|B_j\|$ the better the implementation.

All four versions of the algorithm which we consider keep $\|F_j\|$ and hence $\|E_j\|$ tiny. Therefore we analyze these choices to determine which yields the smallest bounds for the elements of B_j.

Lanczos Algorithms. Start with $\|q_1\| = 1$ and $u_1 = Aq_1$. For $j = 1,2,\ldots$ repeat

```
(1a)  $\alpha_j = q_j^* A q_j$  or  (1b)  $\alpha_j = q_j^* u_j$
(2)   $r_j = u_j - q_j \alpha_j$
(3)   $\beta_j = \|r_j\|$
(4)   if $\beta_j = 0$ STOP else $q_{j+1} = r_j / \beta_j$
(5a)  $\eta_j = q_{j+1}^* A q_j$  or  (5b)  $\eta_j = \beta_j$
(6)   $u_{j+1} = Aq_{j+1} - q_j \eta_j$
```

These same implementations were analyzed by C. Paige [Paige 1972]. The conclusion is the same but the approach used here is different from that of Paige.

Recall Lemma 4 of Section 3 which gives formulas for the elements of B_j, namely

$$\left.\begin{aligned} b_{11} &= q_1^* r_1 \\ b_{ii} &= q_1^* r_1 - q_{i-1}^* r_{i-1} \\ b_{i-1,i} &= \beta_{i-1}(q_i^* q_i - q_{i-1}^* q_{i-1}) \end{aligned}\right. \quad \text{for } 2 \le i \le j \;.$$

We now give an informal analysis of the choices between (1a)-(1b) and (5a)-(5b) in the light of these formulas.

Remark 1. (1b) is slightly better than (1a).

If (1a) is used, α_j is chosen to force $q_j^*(Aq_j - q_j\alpha_j) \doteq 0$ and thus

$$\begin{aligned} q_j^* r_j &\doteq q_j^*(Aq_j - q_j\alpha_j - q_{j-1}\beta_{j-1}) \\ &\doteq q_j^*(Aq_j - q_j\alpha_j) - q_j^* q_{j-1}\beta_{j-1} \\ &\doteq -q_j^* q_{j-1}\beta_{j-1} \end{aligned}$$

$$\doteq -q_{j-1}^{*}q_j\beta_{j-1}$$
$$\doteq -q_{j-1}^{*}r_{j-1} \ .$$

Thus the size of $|q_j^{*}r_j|$ depends on $q_{j-1}^{*}r_{j-1}$ and may grow (slowly) as j increases.

If (1b) is used α_j is chosen to force $q_j^{*}r_j \doteq 0$ and there is no dependence on earlier steps.

Thus the diagonal elements of B_j are kept smaller by (1b) than (1a) while both (1a) and (1b) keep the off diagonal elements tiny. □

The more interesting choice is between (5a) and (5b). Historically (5a) was recommended (cf. [Wilkinson 1965], p. 395) despite the extra computation involved. This was unfortunate since

Remark 2. (5b) is better than (5a).

Recall that (5a) explicitly computes $\eta_j = q_j^{*}Aq_j$ as the coefficient of q_j in the formula for u_{j+1}, namely

$$u_{j+1} = Aq_{j+1} - q_j\eta_j \ ,$$

while (5b) merely sets $\eta_j = \beta_j = \|r_j\|$. If (5a) is used the resulting tridiagonal matrix has η's on the subdiagonal and β's on the super diagonal. Thus it is not symmetric and we denote it T_j'.

Since T_j' is asymmetric the analysis of Section 3 does not apply directly to

$$(7) \qquad AQ_j - Q_jT_j' = r_je_j^{*} + F_j \ .$$

If any of the η's are negative then T_j' will have complex eigenvalues which is clearly wrong. If all the η's are positive we need the following standard result to finish the justification of Remark 2.

Lemma. If all the η's in T_j' are positive then there exists a diagonal matrix $\Xi_j = \text{diag}(\xi_1, \xi_2, \ldots, \xi_j)$ such that $\xi_j = 1$ and $\Xi^{-1}T'\Xi$ is symmetric.

Proof of Lemma 3. Let $\xi_j = 1$ and for $i = j-1, j-2, \ldots, 1$, let $\xi_i = \xi_{i+1}(\beta_i/\eta_i)^{1/2}$. Then $\Xi_j^{-1}T_j'\Xi_j = T_j$ with

$$T_j = \begin{bmatrix} \alpha_1 & \zeta_1 & & & \\ \zeta_1 & \alpha_2 & \zeta_2. & & \\ & \zeta_2. & \ddots & \ddots & \\ & & \ddots & \ddots & \zeta_{j-1} \\ & & & \zeta_{j-1} & \alpha_j \end{bmatrix}$$

and $\zeta_i = \sqrt{\beta_i \eta_i}$, for $i \leq j-1$. □

We note that the greater the asymmetry of T_j', the greater the ratio between successive ξ's.

Now let Ξ_j be the diagonal matrix that symmetrizes T_j'. Then the quantities computed using (5a) satisfy

$$(8) \qquad \begin{aligned} A(Q_j\Xi_j) &= (Q_j\Xi_j)(\Xi_j^{-1}T_j'\Xi_j) + r_je_j^*\Xi_j + F_j\Xi_j \\ &= (Q_j\Xi_j)T_j + r_je_j^* + F_j\Xi_j \end{aligned}$$

since $\xi_j = 1$. Now the lemmas of Section 3 can be applied to equation (8) but the result is not encouraging. The lengths of the columns of $(Q_j\Xi_j)$ are not 1 but are $\xi_1, \xi_2, \ldots, \xi_j$ instead.

Therefore the off diagonal elements of B_j are of the form

$$b_{i-1,i} = \zeta_{i-1}(\xi_i^2-\xi_{i-1}^2) \ . \tag{9}$$

Hence the greater the asymmetry in T_j', the greater the variation in the ξ's, the larger the elements of B_j, and the larger the γ_{ji} will be.

On the other hand (5b) always maintains symmetry of T and always normalizes the q_i to have length 1. Hence the off diagonal elements of B_j are always tiny when (5b) is used and (5b) is better than (5a). □

In the rest of the thesis we will consider only the most stable version of the Lanczos algorithm which uses (1b) and (5b).

2.5 Distinguishing the Lanczos Algorithm

To analyze the Lanczos algorithm in the context of finite precision arithmetic it is necessary to have a tractable definition of what constitutes an instance of the algorithm.

Definition. The matrix equation

$$AQ_j - Q_jT_j = \beta_j q_{j+1} e_j^* + F_j \tag{1}$$

is an instance of the Lanczos algorithm if $\|F_j\| \doteq \epsilon\|A\|$ and $\gamma_{ki} \doteq 1$ for $k \leq j$ and all $i \leq k$, where $\gamma_{ki} = \beta_{ki}|y_i^* q_{k+1}|/\epsilon\|A\|$.

We note that this definition is justified, for the most stable implementation of the Lanczos algorithm, by Theorem 3 and the analysis given in Appendix 1.

Recall the following perverse Example.

Example 3.

$$[1]\,[1\ \ -1] = [1\ \ -1]\begin{bmatrix} 1001 & 1000 \\ 1000 & 1001 \end{bmatrix}$$

Since $\beta_2 = 0$, both β_{21} and β_{22} are zero and hence γ_{21} and γ_{22} are zero. On the other hand $\beta_1 = 1000$, $s_{11} = 1$, $y_1 = q_1$, and $\|A\| = 1$. Hence $y_1^* q_2 = -1$, $\beta_{11} = 1000$ and $\gamma_{11} = \beta_{11}|y_1^* q_2|/\epsilon\|A\| =$ $= 1000/\epsilon$. Thus Example 3 is not an instance of the Lanczos algorithm. □

To show the power of this definition we prove a theorem which gives a bound for $|y_i^* y_k|$, the inner product for two different Ritz vectors at step j.

Theorem 4 (Paige). Let y_i and y_k be two Ritz vectors at the j^{th} step of the Lanczos algorithm. Then

$$|\theta_i - \theta_k|\,|y_i^* y_k| \le [\gamma_{ji}(\beta_{jk}/\beta_{ji}) + \gamma_{jk}(\beta_{ji}/\beta_{jk}) + \nu_{ik}]\epsilon\|A\| ,$$

where $\nu_{ik} = |s_i^*(Q_j^* F_j - F_j^* Q_j)s_i|/\epsilon\|A\| \doteq 1$.

Proof. Recall that $y_i = Q_j s_i$ and $T_j s_i = s_i \theta_i$. Multiply the basic equation (1) on the left by y_i^* and on the right by s_k to obtain

$$(2)\qquad y_i^* A Q_j s_k - y_i^* Q_j T_j s_k = y_i^* \beta_j q_{j+1} e_j^* s_k + s_i^* Q_j^* F_j s_k ,$$

which simplifies to

$$(3) \qquad y_i^* A y_k - y_1^* y_k \theta_k = \beta_j s_{jk} y_i^* q_{j+1} + s_i^* Q_j^* F_j s_k \ .$$

On the other hand if (1) is multiplied on the left by y_k^* and on the right by s_i the result is

$$(4) \qquad y_k^* A y_1 - y_k^* y_i \theta_i = \beta_j s_{ji} y_k^* q_{j+1} + s_k^* Q_j^* F_j s_i \ .$$

Subtract (4) from (3) to discover

$$(5) \quad (\theta_i - \theta_k)(y_i^* y_k) = \beta_j s_{jk} y_i^* q_{j+1} - \beta_j s_{ji} y_k^* q_{j+1} + s_i^* Q_j^* F_j s_k - s_k^* Q_j^* F_j s_i \ .$$

Taking the absolute value of each side of (5) and observing that $s_k^* Q_j^* F_j s_i = s_i^* F_j^* Q_j s_k$ we obtain

$$(6) \quad |\theta_i - \theta_k| \, |y_1^* y_k| \le \beta_{jk} |y_i^* q_{j+1}| + \beta_{ji} |y_k^* q_{j+1}| + |s_i^* (Q_j^* F_j - F_j^* Q_j) s_k| \ .$$

By Theorem 3, $|y_i^* q_{j+1}| = \gamma_{ji} \epsilon \|A\| / \beta_{ji}$, for all i. By definition of an instance of the Lanczos algorithm $\|F_j\| \doteq \epsilon \|A\|$ so $\nu_{ik} \doteq 1$ and the result follows. □

Theorem 4 shows how the loss of orthogonality in the matrix Q_j is manifested in the matrix $Y_j = Q_j S_j$. If two Ritz vectors have equal β_{ji} then they will be orthogonal unless their Ritz values are almost equal. If two Ritz vectors have very different values of β_{ji} (that is, one is well converged and the other is not) then they will never be orthogonal.

2.6 Behavior of the Lanczos Algorithm

With Theorem 3 and Theorem 4 in hand it is possible to give a detailed description of what occurs during a Lanczos run. To illustrate the various stages of the process we will intersperse the verbal description with selected output from Example 1, which was first discussed in Section 1.

Example 1.

$$n = 10$$

$$\epsilon = .6 \times 10^{-7}$$

$$A = \mathrm{diag}(0,.01,.02,\ldots,.08,1.0)$$

$$q_1 = u/\|u\| \ , \quad u = (1,1,\ldots,1)^*$$

Note the large gap between .08 and 1.0. Thus by the results in Section 3 of Chapter 1 we would expect the Ritz value with the smallest β_{ji} at any step j to approximate 1.0.

The algorithm was arbitrarily terminated at $j = 11$. The elements of T_{11} are

j	alpha	beta
1	.1360	.2890
2	.8964	.0810
3	.0465	.0227
4	.0401	.0215
5	.0401	.0203
6	.0401	.0190
7	.0614	.1423
8	.9653	.1114
9	.0532	.0143
10	.0404	.0103
11	.0405	.0032

As noted before β_{10}, which would be zero in exact arithmetic, is not tiny at all. This behavior is quite common. Rarely, if ever, are tiny off diagonal elements encountered in large problems even for $j > n$. This phenomenon will be explained later.

In the early steps of the algorithm ($j \leq 3$ for the example) no β_{ji} is small and orthogonality among the columns of both Q_j and Y_j is well maintained.

$j = 3 \quad (\epsilon = .6 \times 10^{-7})$

i	Ritz value	β_{ji}	$\|y_i^* q_4\|$	$\gamma_{ji} = \beta_{ji} \|y_i^* q_4\| / (\epsilon \\|A\\|)$
1	.01366	$.158 \times 10^{-1}$	$.358 \times 10^{-7}$	.0094
2	.06527	$.163 \times 10^{-1}$	$.116 \times 10^{-6}$	.0313
3	1.00000	$.183 \times 10^{-2}$	$.140 \times 10^{-5}$	.0425

$$|1 - Y_3^* Y_3|$$

$$\boxed{\begin{array}{ccc} .48 \times 10^{-6} & .29 \times 10^{-7} & .88 \times 10^{-7} \\ .29 \times 10^{-7} & .24 \times 10^{-6} & .48 \times 10^{-8} \\ .88 \times 10^{-7} & .48 \times 10^{-8} & .24 \times 10^{-6} \end{array}}$$

After a while ($j = 6$) some Ritz value begins to converge. The smaller the corresponding β_{ji} the greater the loss of orthogonality. Since Q_j is no longer orthonormal, neither is Y_j ($= Q_j S_j$). This loss of orthogonality does not affect the converging Ritz vector y. Instead each of the other Ritz vectors is contaminated by a spurious component in the direction of y. The greater the convergence the greater the contamination. Note that the β_{ji} of the unconverged Ritz vectors are larger at $j = 7$ than $j = 6$ because of the greater contamination.

$j = 6$

i	Ritz value	β_{ji}	$\lvert y_i^* q_7 \rvert$	γ_{ji}
1	$.744 \times 10^{-3}$	$.405 \times 10^{-2}$	$.308 \times 10^{-6}$	.027
2	.0170	$.959 \times 10^{-2}$	$.157 \times 10^{-6}$	.025
3	.0371	$.118 \times 10^{-1}$	$.672 \times 10^{-7}$	.013
4	.0625	$.976 \times 10^{-2}$	$.173 \times 10^{-6}$	.028
5	.0719	$.421 \times 10^{-2}$	$.800 \times 10^{-7}$	.006
6	1.00000	$.164 \times 10^{-7}$	.149 ×	.041

$|1-\gamma_6^*\gamma_6|$

$.12 \times 10^{-6}$	$.74 \times 10^{-7}$	$.23 \times 10^{-6}$	$.16 \times 10^{-6}$	$.49 \times 10^{-7}$	$.60 \times 10^{-3}$
$.74 \times 10^{-7}$	$.95 \times 10^{-6}$	$.12 \times 10^{-6}$	$.62 \times 10^{-7}$	$.11 \times 10^{-7}$	$.15 \times 10^{-2}$
$.23 \times 10^{-6}$	$.12 \times 10^{-6}$	0.	$.12 \times 10^{-6}$	$.91 \times 10^{-7}$	$.18 \times 10^{-2}$
$.16 \times 10^{-6}$	$.62 \times 10^{-7}$	$.12 \times 10^{-6}$	$.72 \times 10^{-6}$	$.29 \times 10^{-6}$	$.16 \times 10^{-2}$
$.49 \times 10^{-7}$	$.11 \times 10^{-7}$	$.91 \times 10^{-7}$	$.29 \times 10^{-6}$	$.24 \times 10^{-6}$	$.68 \times 10^{-3}$
$.60 \times 10^{-3}$	$.15 \times 10^{-2}$	$.18 \times 10^{-2}$	$.16 \times 10^{-2}$	$.68 \times 10^{-3}$	$.12 \times 10^{-6}$

$j = 7$

i	Ritz value	β_{ji}	$y_i^* q_8$	γ_{ji}
1	$.417 \times 10^{-3}$	$.114 \times 10^{-1}$	$.121 \times 10^{-7}$	.002
2	.0148	$.329 \times 10^{-1}$	$.335 \times 10^{-7}$	.018
3	.0348	$.529 \times 10^{-1}$	$.410 \times 10^{-7}$	.036
4	.0564	$.704 \times 10^{-1}$	$.894 \times 10^{-7}$	.104
5	.0733	$.902 \times 10^{-1}$	$.224 \times 10^{-7}$	.034
6	.0811	$.561 \times 10^{-1}$	$.521 \times 10^{-7}$	.049
7	1.0000000	$.516 \times 10^{-8}$	.982	.085

$|1 - Y_7^* Y_7|$

$.12\times10^{-6}$	$.17\times10^{-6}$	$.25\times10^{-6}$	$.17\times10^{-6}$	$.31\times10^{-6}$	$.11\times10^{-6}$	$.11\times10^{-1}$
$.17\times10^{-6}$	$.12\times10^{-5}$	$.24\times10^{-6}$	$.31\times10^{-6}$	$.68\times10^{-8}$	$.48\times10^{-7}$	$.33\times10^{-1}$
$.25\times10^{-6}$	$.24\times10^{-6}$	$.48\times10^{-6}$	$.67\times10^{-6}$	$.23\times10^{-6}$	$.46\times10^{-7}$	$.54\times10^{-1}$
$.17\times10^{-6}$	$.31\times10^{-6}$	$.67\times10^{-7}$	$.24\times10^{-6}$	$.34\times10^{-6}$	$.76\times10^{-7}$	$.73\times10^{-1}$
$.31\times10^{-6}$	$.68\times10^{-8}$	$.23\times10^{-6}$	$.34\times10^{-6}$	$.18\times10^{-6}$	$.70\times10^{-7}$	$.96\times10^{-1}$
$.11\times10^{-6}$	$.48\times10^{-7}$	$.46\times10^{-7}$	$.76\times10^{-7}$	$.70\times10^{-7}$	$.18\times10^{-6}$	$.60\times10^{-1}$
$.11\times10^{-1}$	$.33\times10^{-1}$	$.54\times10^{-1}$	$.73\times10^{-1}$	$.96\times10^{-1}$	$.60\times10^{-1}$	$.24\times10^{-6}$

Suddenly at some step ($j = 8$) the contamination of the unconverged Ritz vectors decreases and is transformed into a second copy of y. When it first appears the second copy is much less accurate than the first. Note the improvement in the unconverged Ritz vectors with the lessening of the contamination.

$j = 8$

i	Ritz value	β_{ji}	$y_i^* q_9$	γ_{ji}
1	$.188\ \ 10^{-3}$	$.223 \times 10^{-2}$	$.576 \times 10^{-6}$	.021
2	.0128	$.662 \times 10^{-2}$	$.455 \times 10^{-6}$	.050
3	.0299	$.954 \times 10^{-2}$	$.291 \times 10^{-6}$	.046
4	.0493	$.967 \times 10^{-2}$	$.151 \times 10^{-6}$	.024
5	.0669	$.692 \times 10^{-2}$	$.245 \times 10^{-6}$	.028
6	.0797	$.224 \times 10^{-2}$	$.263 \times 10^{-6}$	.011
7	.987	.110	$.536 \times 10^{-6}$	.985
8	1.000000	$.159 \times 10^{-8}$	.116	.003

The bottom row of $1 - Y_8^* Y_8$ is

$$\boxed{.26\times10^{-3}\ \ .77\times10^{-3}\ \ .11\times10^{-2}\ \ .12\times10^{-2}\ \ .86\times10^{-3}\ \ .31\times10^{-3}\ \ .99\ \ .24\times10^{-6}}$$

which shows that y_7 and y_8 are almost identical.

On succeeding steps the poorer copy gets better while the more accurate Ritz vector gets worse. That is, the appearance of the second coyp perturbs the first copy away from z, the eigenvector of A, until both are equally accurate (β_{ji}'s approximately equal). This usually occurs at about $\beta_{ji} = \sqrt{\varepsilon}\|A\|$.

Then both Ritz vectors will improve and spurious components of z in the other Ritz vectors will grow again until a third copy makes its appearance. Thus the algorithm grinds out more and more copies of z. The other Ritz pairs will continue to improve despite the appearance of repeated copies of z.

Most computational examples will be more complicated than the one given here. This example was chosen so that only one Ritz vector converged quickly. In general, at any one step, there will be Ritz vectors at all different levels of accuracy. However the same basic cycle can be discerned for each individual eigenvector.

The cycle time for a particular eigenvector, that is, the number of steps between the appearance of one copy and the next, seems to be fairly constant. Of course the cycle times are different for different eigenvectors and depend on the locations of the corresponding eigenvalues in the spectrum of A. Predicting the average cycle times from the spectrum of A alone appears to be quite difficult.

The concept of cycles and cycle times does explain why tiny off diagonal elements are rarely ever encountered in the Lanczos algorithm even for $j > n$. A tiny β can occur only if a very large fraction of the cycles condense on the same step. Since this is statistically unlikely, small β's are rarely seen.

2.7 The Lengths of Ritz Vectors

In exact arithmetic, Q_j is orthonormal and so $Y_j = Q_j S_j$ is also orthonormal and all the Ritz vectors have length 1. However once Q_j has lost orthogonality the Ritz vectors need not have length 1. In Lemma 1 of Section 2 we established a lower bound for the lengths of Ritz vectors. Namely

$$\|y_i\| \geq \sigma_1(Q_j) \tag{1}$$

where $\sigma_1(Q_j) = (\lambda_1[Q_j^* Q_j])^{1/2}$ is the smallest singular value of Q_j.

In Example 2 we found that this maximal shrinkage of Ritz vectors can occur:

(2) $$[1][1\ \ -1] = [1\ \ -1]\begin{bmatrix} 1001 & 1000 \\ 1000 & 1001 \end{bmatrix},$$

$Q_2 = [1\ \ -1]$, $\sigma_1(Q_2) = 0$, and $\|y_2\| = 0$. As shown in Section 5, Example 2 is not an instance of the Lanczos algorithm ($\gamma_{11} = 1000/\epsilon$). Can this same maximal shrinkage occur in a Lanczos example? The answer is yes. Consider

<u>Example 3</u>.

$$[1][1\ \ -1] = [1\ \ -1]\begin{bmatrix} 1 & \epsilon \\ \epsilon & 1 \end{bmatrix} + [\epsilon][0\ \ 1] + [\epsilon\ \ 0]\ .$$

Here $r_2 = [\epsilon]$ and $F_2 = [\epsilon\ \ 0]$. It can be verified that $\|F_2\| = \epsilon$, $\gamma_{11} = 1$, $\gamma_{21} = \sqrt{2}$, and $\gamma_{22} = 0$. Therefore since $\|A\| = 1$, Example 3 is an instance of the Lanczos algorithm. $T_2 = S_2\Theta_2S_2^*$ with

$$\Theta_2 = \mathrm{diag}(1-\epsilon, 1+\epsilon)$$
$$S_2 = (\sqrt{2}/2)\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

and so $Y_2 = Q_2S_2 = [\sqrt{2}\ \ 0]$. □

There is a great difference between Example 2 and Example 3. In Example 2 the zero Ritz vector y_2 is associated with the spurious Ritz value 2001 and is completely misleading. In Example 3 y_2 is associated with $1+\epsilon$ which <u>is</u> an eigenvalue of A correct to working accuracy. Furthermore there is another Ritz value close to the same eigenvalue of A whose Ritz vector has length at last 1, namely $\theta_1 = 1-\epsilon$ and $y_1 = [\sqrt{2}]$.

Both of these facts are characteristic of the Lanczos algorithm. In the next two sections we will establish bounds which show that $\|y_i\|^2 \doteq 1$ holds for an isolated Ritz vector and that $\sum\|y_i\|^2 \doteq m$ for the Ritz vectors associated with a cluster of m Ritz values.

We first give an intuitive explanation of why it is impossible to establish a robust lower bound for the length of a single Ritz vector whose Ritz value is not well separated from the rest of the Ritz values.

Consider the simplest case of two copies of a single eigenpair of A where there will exist two orthogonal eigenvectors of T_j, call them s_1 and s_2, such that $y_1 = Q_j s_1$ and $y_2 = Q_j s_2$ have length 1 and are equal to working accuracy. The corresponding Ritz values are also equal to working accuracy and eigenvectors of very close eigenvalues are not well determined. All that can be computed in practice are two orthonormal vectors s_1' and s_2' which span the same space as s_1 and s_2.

Thus there exists a 2×2 orthogonal matrix P such that $S = S'P$, where $S' = (s_1', s_2')$ and $S = (s_1, s_2)$. If $y_1' = Q_j s_1'$ and $y_2' = Q_j s_2'$ then

$$\begin{aligned}\|y_1'\|^2 + \|y_2'\|^2 &= \operatorname{trace}(S'^* Q_j^* Q_j S') \\ &= \operatorname{trace}(P^* S'^* Q_j^* Q_j S' P) \\ &= \operatorname{trace}(S^* Q_j^* Q_j S) \\ &= \|y_1\|^2 + \|y_2\|^2 \\ &= 2 ,\end{aligned}$$

but since Q_j has lost linear independence we need not have $\|y_2'\|^2 = 1$. If $s_2' = (s_1 - s_2)/\sqrt{2}$ then

$$\begin{aligned} y_2' &= Q_j s_2' \\ &= (Q_j s_1 - Q_j s_2)/\sqrt{2} \\ &= (y_1 - y_2)/\sqrt{2} \\ &\doteq 0 \end{aligned}$$

since $y_1 \doteq y_2$. This is exactly what happens in Example 3.

2.8 Bounding the Length of an Isolated Ritz Vector

Let (y_i, θ_i) be the i^{th} Ritz vector at the j^{th} step of the Lanczos algorithm. We wish to bound the departure of $\|y_i\|^2$ from the expected value of 1. The obtainable bound depends on $\mu_i = \min_{k \neq i} |\theta_i - \theta_k| / \|A\|$, the relative separation of θ_i from the rest of the eigenvalues of T_j. The following result shows that if μ_i is not too small then $\|y_i\|^2$ cannot be too small either.

Theorem 5 (Paige). Let (y_i, θ_i) be the i^{th} Ritz pair at the j^{th} step of the Lanczos algorithm. Let $1 - Q_j^* Q_j =$ $= C_j^* + \Delta_j + C_j$, where Δ_j is diagonal and C_j is strictly upper triangular. Let $\mu_i = \min_{k \neq i} |\theta_i - \theta_k| / \|A\|$. Then

$$|1 - \|y_i\|^2| \leq \|\Delta_j\| + \zeta_i$$

where $\zeta_i = j(j-1)\gamma\epsilon/\mu_i$ and γ is a number such that $\gamma_{tr} \leq \gamma$ for all $t \leq j$ and $r \leq t$.

Remarks. Since each Lanczos vector is normalized $\|\Delta_j\|$ will always be tiny and by Theorem 3 γ will not be large. Therefore Theorem 5 shows that an isolated Ritz vector (a Ritz vector whose Ritz value is isolated) will never be small. For instance if

$$\mu_i \geq 10j(j-1)\gamma\epsilon$$

then

$$\begin{aligned} |1 - \|y_i\|^2| &\leq \|\Delta_j\| + 0.1 \\ &\doteq 0.1 \end{aligned}$$

and

$$\|y_i\|^2 \geq .9 .$$

Furthermore

$$\begin{aligned} |1 - \|y_i\|^2| &= |1 - y_i^* y_i| \\ &= |1 - s_i^* Q_j^* Q_j s_i| \\ &= |s_i^*(1 - Q_j^* Q_j) s_i| \\ &= |s_i^*(C_j^* + \Delta_j + C_j) s_i| \\ &\leq |s_i^* \Delta_j s_i| + 2|s_i^* C_j s_i| \\ &\leq \|\Delta_j\| + 2|s_i^* C_j s_i| \end{aligned}$$

and so it remains to show that $2|s_i^* C_j s_i| \leq \zeta_i$. The proof of this inequality is rather long so we break it into a series of lemmas.

The first lemma gives a formula for evaluating the "inner product" of eigenvectors of T_j and T_t for $t < j$ at the t^{th} step.

<u>Lemma 1</u>. Let $(s_i^{(j)}, \theta_i^{(j)})$ be an eigenpair of T_j and let $(s_r^{(t)}, \theta_r^{(t)})$ be an eigenpair of T_t for $t < j$. Associated with $s_r^{(t)}$ is the j-vector $s_r' = \begin{pmatrix} s_r^{(t)} \\ 0 \end{pmatrix}$. Then for all $r \leq t$

$$(\theta_i^{(j)} - \theta_r^{(t)}) s_i^{(j)*} s_r' = \beta_t s_{tr}^{(t)} s_{t+1,i}^{(j)} .$$

Proof of Lemma 1. As in the proof of Theorem 6 of Chapter 1

$$(1) \qquad T_j s'_r - s'_r \theta_r^{(t)} = \begin{bmatrix} 0 \\ - - - - \\ \beta_t s_{tr}^{(t)} \\ 0 \end{bmatrix} .$$

Multiply (1) on the right by $s_i^{(j)*}$ to obtain

$$(2) \qquad \theta_j^{(j)} s_i^{(j)*} s'_r - s_i^{(j)*} s'_r \theta_r^{(t)} = \beta_t s_{tr}^{(t)} s_{t+1,i}^{(j)} . \qquad \square$$

Lemma 2. If $\theta_i^{(j)} = \theta_m^{(t)}$ for some $m \leq t$ in Lemma 1 then $s_{t+1,i}^{(j)} = 0$ and, for $r = 1,\ldots,t$,

$$(s_i^{(j)*} s'_r)^2 = \begin{cases} 0 & , \text{ if } r \neq m , \\ \sum_{k=1}^{t} s_{ki}^{(j)^2} & , \text{ if } r = m . \end{cases}$$

Proof of Lemma 2. For $r = m$ in Lemma 1,

$$\begin{aligned} \beta_t s_{tm}^{(t)} s_{t+1,i}^{(j)} &= (\theta_i^{(j)} - \theta_m^{(t)}) s_i^{(j)*} s'_m , \\ &= 0 . \end{aligned}$$

In the Lanczos algorithm T_j and T_t are unreduced (no β vanishes) and so $\beta_t \neq 0$. $s_{tm}^{(t)}$, the bottom element of an eigenvector of T_t, cannot be zero either so we must have $s_{t+1,i}^{(j)} = 0$.

Thus by Lemma 1, for all $r \leq t$

$$(3) \qquad (\theta_i^{(j)} - \theta_r^{(t)}) s_i^{(j)*} s'_r = 0 .$$

For $r \neq m$, $\theta_r^{(t)} \neq \theta_m^{(t)} = \theta_i^{(j)}$ since T_t, an *unreduced* tridiagonal matrix has distinct eigenvalues, and so by (3) for $r \neq m$

(4) $$s_i^{(j)*} s_r' = 0 \; .$$

Let s_i' be the t-vector of the first t elements of s_i. Thus prime shortens vectors of length j and lengthens vectors of length t. Then for all $r \leq t$

(5) $$s_r^* s_i' = s_i^* s_r' \; ,$$

and since S_t, the matrix of eigenvectors of T_t is orthogonal,

$$\begin{aligned} \|s_i'\|^2 &= \|S_t^* s_i'\|^2 \\ &= \sum_{r=1}^{t} (s_r^* s_i')^2 \\ &= \sum_{r=1}^{t} (s_i^* s_r')^2 \; , \quad \text{by (5)} \\ &= (s_i^* s_m')^2 \; , \qquad \text{by (4)} \; . \end{aligned}$$

Since $\|s_i'\|^2 = \sum_{k=1}^{t} s_{ki}^{(j)2}$ the result follows. □

Lemma 3. Let $1 - Q_j^* Q_j = C_j^* + \Delta_j + C_j$ as in Theorem 5. Then

$$C_j = (0, S_1' v_1, S_2' v_2, \ldots, S_{j-1}' v_{j-1}) \; ,$$

where $S_k' = \binom{S_k}{0}$ is a $j \times k$ matrix, $v_k = \epsilon \|A\| (\gamma_{k1}/\beta_{k1}, \gamma_{k2}/\beta_{k2}, \ldots, \gamma_{kk}/\beta_{kk})^*$ is a k-vector, and γ_{ki} is defined in Theorem 3.

Proof of Lemma 3. By definition C_j is the upper triangular part of $Q_j^*Q_j$ and hence the $(i+1)^{st}$ column of C_j is

$$\begin{aligned}
\begin{pmatrix} Q_i^* q_{i+1} \\ 0 \end{pmatrix} &= \begin{pmatrix} S_i S_i^* Q_i^* q_{i+1} \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} S_i Y_i^* q_{i+1} \\ 0 \end{pmatrix} \\
&= S_i' Y_i^* q_{i+1} \\
&= S_i' v_i
\end{aligned}$$

by Theorem 3. □

Lemma 4. The following equality holds:

$$s_i^* C_j s_i = \sum_{t=1}^{j-1} s_{t+1,i}^{(j)2} \omega_{it}^{(j)}$$

where

$$\omega_{it}^{(j)} = \begin{cases} 0 & , \text{ if } \theta_m^{(t)} = \theta_i^{(j)} \text{ for some } m \\ \epsilon \sum_{r=1}^{t} \gamma_{tr} \|A\| / (\theta_i^{(j)} - \theta_r^{(t)}) & , \text{ otherwise} \end{cases}$$

Proof of Lemma 4. By Lemma 3

$$\begin{aligned}
(6) \qquad s_i^* C_j s_i &= s_i^* (0, S_1' v_1, S_2' v_2, \ldots, S_{j-1}' v_{j-1}) s_i \\
&= \sum_{t=1}^{j-1} (s_i^* S_t' v_t) s_{t+1,i}^{(j)} \\
&= \sum_{t=1}^{j-1} s_{t+1,i}^{(j)} \sum_{r=1}^{t} (s_i^* s_r^{(t)'}) v_{rt} \\
&= \sum_{t=1}^{j-1} s_{t+1,i}^{(j)} \sum_{r=1}^{t} (s_i^* s_r^{(t)'}) \epsilon \|A\| \gamma_{tr} / \beta_{tr}
\end{aligned}$$

where $s_r^{(t)'}$ is the r^{th} column of S_t'.

If $\theta_i^{(j)} = \theta_m^{(t)}$ for some m then by Lemma 2 $s_i^* s_r^{(t)}$ is always bounded and $s_{t+1,i}^{(j)} = 0$. Hence $\omega_{it}^{(j)}$ can be set to zero.

If $\theta_i^{(j)} \neq \theta_m^{(t)}$ for all $m \leq t$ then by Lemma 1

$$(7) \qquad \begin{aligned} s_i^* s_r^{(t)'} &= \beta_t s_{tr}^{(t)} s_{t+1,i}^{(j)} / (\theta_i^{(j)} - \theta_r^{(t)}) \ , \\ &= \beta_{tr} s_{t+1,i}^{(j)} / (\theta_i^{(j)} - \theta_r^{(t)}) \ . \end{aligned}$$

Substituting (7) into (6) we find

$$\begin{aligned} s_i^* C_j s_i &\leq \epsilon \|A\| \sum_{t=1}^{j-1} s_{t+1,i}^{(j)} \sum_{r=1}^{t} \left(\beta_{tr} s_{t+1,i}^{(j)} / (\theta_i^{(j)} - \theta_r^{(t)})\right)(\gamma_{tr}/\beta_{tr}) \\ &= \epsilon \|A\| \sum_{t=1}^{j-1} s_{t+1,i}^{(j)^2} \sum_{r=1}^{t} \gamma_{tr} / (\theta_i^{(j)} - \theta_r^{(t)}) \ . \qquad \Box \end{aligned}$$

To make use of Lemma 3 we need an expression for $s_{t+1,i}^{(j)^2}$ involving the eigenvalues of T_j. Such a formula was given in Section 4 of Chapter 1 in terms of the following definition.

Let $\chi_{r,t}(\xi)$ be the characteristic polynomial of $T_{r,t}$, where

$$T_{r,t} = \begin{bmatrix} \alpha_r & \beta_r & & & \\ \beta_r & \alpha_{r+1} & \beta_{r+1} & & \\ & \beta_{r+1} & \ddots & \ddots & \\ & & \ddots & \ddots & \beta_{t-1} \\ & & & \beta_{t-1} & \alpha_t \end{bmatrix} .$$

Lemma 2 in Section 1.4 states

$$(8) \qquad s_{t+1,i}^{(j)^2} = \chi_{1,t}(\theta_i^{(j)}) \chi_{t+2,j}(\theta_i^{(j)}) / \chi'_{1,j}(\theta_i^{(j)}) \ .$$

Observe that if the $(t+1)^{st}$ row and column of T_j are deleted the remaining matrix is a direct sum of two smaller matrices namely $T_{1,t}$ and $T_{t+2,j}$. For instance if $j = 4$ and $t = 2$ we have

$$T_{1,4} = \left[\begin{array}{cccc} \alpha_1 & \beta_1 & & \\ \beta_1 & \alpha_2 & \beta_1 & \\ & \beta_1 & \alpha_3 & \beta_1 \\ & & \beta_1 & \boxed{\alpha_4} \end{array}\right]$$

We label the union of the eigenvalues of $T_{1,t}$ and $T_{t+2,j}$ as

$$\nu_1^{(t)} \leq \nu_2^{(t)} \leq \cdots \leq \nu_{j-1}^{(t)} .$$

Equation (8) can now be rewritten as

$$(9)\quad s_{t+1,i}^{(j)^2} = \prod_{k=1}^{i-1}[(\theta_i^{(j)}-\nu_k^{(t)})/(\theta_i^{(j)}-\theta_k^{(j)})]\prod_{k=1}^{j-1}[(\theta_i^{(j)}-\nu_k^{(t)})/(\theta_i^{(j)}-\theta_{k+1}^{(j)})] .$$

Note that by the Cauchy Interlace Theorem each factor in the R.H.S. of (9) lies between 0 and 1.

$\theta_r^{(t)}$ is an eigenvalue of $T_{1,t}$ and so there is a subscript, call it m such that $\nu_m^{(t)} = \theta_r^{(t)}$.

Lemma 5. The following equality holds:

$$s_i^* C_j s_i = \epsilon \sum_{t=1}^{j-1} \sum_{r=1}^{t} \gamma_{tr} \|A\| \pi_r^{(t)} / (\theta_i^{(j)} - \theta_{m'}^{(j)})$$

where

$$\pi_r^{(t)} = \prod_{\substack{k=1 \\ k\neq m}}^{i-1} [(\theta_i^{(j)}-\nu_k^{(t)})/(\theta_i^{(j)}-\theta_k^{(j)})] \cdot \prod_{\substack{k=i \\ k\neq m}}^{j-1} [(\theta_i^{(j)}-\nu_k^{(t)})/(\theta_i^{(j)}-\theta_{k+1}^{(j)})]$$

and

$$m' = \begin{cases} m & , \text{ if } m < i , \\ m+1 & , \text{ if } m \geq i . \end{cases}$$

We note that the exclusion $k \neq m$ can only be relevant for one of the two products in the expression for $\pi_r^{(t)}$. However since the value of m is unknown neither can be omitted.

Proof of Lemma 5. Substitute (9) into Lemma 4 and recall that $\nu_m^{(t)} = \theta_r^{(t)}$. □

To complete the proof of Theorem 5 from Lemma 5 we observe that $0 \leq \pi_r^{(t)} \leq 1$ by the Cauchy Interlace Theorem, $\gamma_{tr}\|A\|/(\theta_i^{(j)}-\theta_{m'}^{(j)}) \leq$ $\leq \gamma/\mu_i$ by hypothesis, and that there are exactly $j(j-1)/2$ terms in the double summation. □

Remarks. The bound established in Theorem 5 is unrealistic for two reasons. In most examples $\pi_r^{(t)}$ will be much smaller than 1. Also most of the eigenvalues of T_j will be much farther away from θ_i than the minimum separation μ_i. In practice the dependence on j of the variation in lengths of the computed Ritz vectors is much less than quadratic. Indeed in most practical examples the lengths of the Ritz vectors seems to be almost independent of j.

2.9 Clustered Ritz Vectors

We now extend the bound established in Theorem 5 to clusters of Ritz vectors.

Theorem 6 (Paige). Let $y_p, y_{p+1}, \ldots, y_{p+h}$ be a cluster of Ritz vectors at the j^{th} step of the Lanczos algorithm. Let $\mu = \min(\theta_p^{(j)}-\theta_{p-1}^{(j)}, \theta_{p+n+1}^{(j)}-\theta_{p+h}^{(j)})/\|A\|$, the relative separation of the cluster. Let Δ_j be the diagonal of $1 - Q_j^* Q_j$. Then

$$|h+1-\sum_{i=p}^{p+h} y_i^* y_i| \le (h+1)(\|\Delta_j\|+\zeta)$$

where $\zeta = j(j-1)\gamma\varepsilon/\mu$ and $\gamma_{tr} \le \gamma$ for all $t \le j$ and $r \le t$.

Remarks. As in Theorem 5, $\|\Delta_j\|$ and $\gamma\varepsilon$ are always tiny. Theorem 6 shows that a cluster of Ritz vectors, which is well separated from the rest of the Ritz vectors, will always contain at least one Ritz vector which is not small. Note that the clustered Ritz values may be arbitrarily close. Theorem 6 depends only on the separation of the cluster from the rest of the spectrum of T_j.

Proof. Since $y_i = Q_j s_i$ for all i,

$$\begin{aligned} (1) \qquad |h+1-\sum_{i=p}^{p+h} y_i^* y_i| &= |h+1-\sum_{i=p}^{p+h} s_i^* Q_j^* Q_j s_i| \\ &= |\sum_{i=p}^{p+h} s_i^*(1-Q_j^*Q_j)s_i| \\ &= |\sum_{i=p}^{p+h} s_i^*(C_j^*+\Delta_j+C_j)s_i| \\ &\le (h+1)\|\Delta_j\| + 2|\sum_{i=p}^{p+h} s_i^* C_j s_i| \ . \end{aligned}$$

It remains to show that $2|\sum_{i=p}^{p+h} s_i^* C_j s_i| \le (h+1)\zeta$. We establish this inequality with a sequence of lemmas.

Definition. As before let the union of the eigenvalues of $T_{1,t}$ and $T_{t+2,j}$ be denoted

$$\nu_1^{(t)} \le \nu_2^{(t)} \le \cdots \le \nu_{j-1}^{(t)} \ ;$$

and as before, by the Cauchy Interlace Theorem, Lemma 1. For all $0 \le t \le j-1$

$$\theta_1^{(j)} \le \nu_1^{(t)} \le \theta_2^{(j)} \le \cdots \le \nu_{j-1}^{(t)} \le \theta_j^{(j)} .$$

For any $r \le t$, $\theta_r^{(t)}$ is an eigenvalue of $T_{1,t}$. Hence there exists a subscript m such that $\nu_m^{(t)} = \theta_r^{(t)}$.

Lemma 2. The following equality holds:

$$\sum_{i=p}^{p+h} s_i^* C_j s_i = \epsilon\|A\| \sum_{t=1}^{j-1} \sum_{r=1}^{t} \gamma_{tr}\omega_{tr}$$

where

$$\omega_{tr} = \sum_{i=p}^{p+h} \Big(\prod_{\substack{k=1 \\ k\neq m}}^{j-1} (\theta_i^{(j)}-\nu_k^{(t)}) / \prod_{\substack{k=1 \\ k\neq i}}^{j} (\theta_i^{(j)}-\theta_k^{(j)})\Big)$$

Proof of Lemma 2. By Lemma 5 of Section 7,

$$\sum_{i=p}^{p+h} s_i^* C_j s_i = \epsilon\|A\| \sum_{i=p}^{p+h} \sum_{t=1}^{j-1} \sum_{r=1}^{t} \left[\gamma_{tr} \prod_{\substack{k=1 \\ k\neq m}}^{j-1} (\theta_i^{(j)}-\nu_k^{(t)}) / \prod_{\substack{k=1 \\ k\neq i}}^{j} (\theta_i^{(j)}-\theta_k^{(j)})\right] .$$

The result follows from exchanging the order of the summations. □

Using equation (1) and Lemma 2, Theorem 6 will be proved if we can establish that $|\omega_{tr}| \le (h+1)/(\mu\|A\|)$. Therefore from now on we consider t and r fixed, we drop the superscripts (j) and (t), and we define m' as a function of i by

$$m' = \begin{cases} m \ , & \text{if } m < i \\ m+1 \ , & \text{if } m \ge i \ . \end{cases}$$

Lemma 3. If $m < p$ or $m \geq p+h$ then

$$|\omega_{tr}| \leq (h+1)/(\mu\|A\|) .$$

Proof of Lemma 3. By Lemma 2

$$|\omega_{tr}| \leq \left|\sum_{i=p}^{p+h} \prod_{\substack{k=1\\k\neq m}}^{j-1} (\theta_i-\nu_k)/ \prod_{\substack{k=1\\k\neq i}}^{j} (\theta_i-\theta_k)\right|$$

$$= \sum_{i=p}^{p+h} \pi_i/|\theta_i-\theta_{m'}|$$

where

$$\pi_i = \prod_{\substack{k=1\\k\neq m}}^{j-1} (\theta_i-\nu_k)/ \prod_{\substack{k=1\\k\neq i,m'}}^{j} (\theta_i-\theta_k) .$$

By Lemma 1, the ratio of successive factors in the numerator and the denominator lie between 0 and 1 and so $0 \leq \pi_i \leq 1$. By hypothesis on m we have $m' < p$ or $m' > p+h$ and so $|\theta_i-\theta_{m'}| \geq \mu\|A\|$ and the result follows. □

The case of $p \leq m < p+h$ requires a longer chain of reasoning. Let $\rho(\xi)$ be a rational function with

$$\rho(\xi) = \prod_{k=1}^{p-1} ((\xi-\nu_k)/(\xi-\theta_k)) \prod_{k=p+h+1}^{j} ((\xi-\nu_{k-1})/(\xi-\theta_k)) ,$$

and let $\rho_i = \rho(\theta_i)$. Note that by Lemma 1

$$0 \leq \rho_i \leq 1 \quad \text{for} \quad p \leq i \leq p+h .$$

<u>Lemma 4</u>. If $p \leq m \leq p+h-1$ then

$$|\omega_{tr}| \leq \sum_{\substack{i=p \\ i \neq m}}^{p+h} |(\rho_m - \rho_i)/(\theta_m - \theta_i)| \ .$$

<u>Proof of Lemma 4</u>. By definition of ρ, the restriction on m, and Lemma 2,

$$(2) \qquad \omega_{tr} = \sum_{i=p}^{p+h} \rho_i \sigma_i$$

where

$$(3) \qquad \sigma_i = \prod_{\substack{k=p \\ k \neq m}}^{p+h-1} (\theta_i - \nu_k) / \prod_{\substack{k=p \\ k \neq i}}^{p+h} (\theta_i - \theta_k) \ .$$

By the restriction on m, $i = m$ for some i and we consider σ_m in more detail. If we expand the formula for σ_m in partial fractions it can be shown that

$$(4) \qquad \sigma_m = \sum_{\substack{i=p \\ i \neq m}}^{p+h} \tau_i / (\theta_m - \theta_i)$$

where

$$\tau_i = \prod_{\substack{k=p \\ k \neq m}}^{p+h-1} (\theta_i - \nu_k) / \prod_{\substack{k=p \\ k \neq i,m}}^{p+h} (\theta_i - \theta_k) \ .$$

On the other hand from (3), for $i \neq m$,

$$\sigma_i = \tau_i / (\theta_i - \theta_m) \ .$$

Hence using (4), equation (3) can be rewritten as

$$\omega_{tr} = \sum_{\substack{i=p\\ i\neq m}}^{p+h} \tau_i(\rho_m-\rho_i)/(\theta_m-\theta_i)$$

and so

$$|\omega_{tr}| \leq \sum_{\substack{i=p\\ i\neq m}}^{p+h} |\tau_i||(\rho_m-\rho_i)/(\theta_m-\theta_i)| \ .$$

Finally by Lemma 1 $|\tau_i| \leq 1$ for each i and the result follows. □

<u>Lemma 5</u>. If $p \leq m < p+h$ then

$$|\omega_{tr}| \leq (h+1)/(\mu\|A\|) \ .$$

<u>Proof of Lemma 5</u>. By definition of ρ,

$$(5) \qquad \rho(\xi) = \{\frac{\xi-\nu_1}{\xi-\theta_1}\}\cdots\{\frac{\xi-\nu_{p-1}}{\xi-\theta_{p-1}}\}\{\frac{\xi-\nu_{p+h}}{\xi-\theta_{p+h+1}}\}\cdots\{\frac{\xi-\nu_{j-1}}{\xi-\theta_j}\} \ .$$

By Lemma 1, if $\nu_{p-1} \leq \xi \leq \nu_{p+h}$ then each factor of $\rho(\xi)$ lies between 0 and 1 and in particular $0 \leq \rho(\xi) \leq 1$. Furthermore ρ is differentiable in this interval and so by the mean value theorem, for $i \neq m$

$$(\rho_m-\rho_i)/(\theta_m-\theta_i) = \rho'(\xi_i)$$

for some ξ_i between θ_m and θ_i.

The derivative of ρ satisfies,

$$\begin{aligned}\rho'(\xi) &= \rho(\xi)(d/d\xi)(\ln(\rho(\xi)))\\ &= \rho(\xi)[\eta_1-\eta_2]\end{aligned}$$

where

$$\eta_1 = \sum_{k=1}^{p-1}[(\xi-\nu_k)^{-1}-(\xi-\theta_k)^{-1}]$$

and

$$\eta_2 = \sum_{k=p+h+1}^{j}[(\nu_{k-1}-\xi)^{-1}-(\theta_k-\xi)^{-1}] \ .$$

If we rearrange the formula for η_1 as

$$\eta_1 = (\xi-\nu_{p-1})^{-1} - \sum_{k=1}^{p-2}[(\xi-\nu_k)^{-1}-(\xi-\theta_{k+1})^{-1}] - (\xi-\theta_1)^{-1} \ ,$$

we find by Lemma 1, for $\nu_{p-1} < \xi < \nu_{p+h}$,

$$0 \leq \eta_1 \leq (\xi-\nu_{p-1})^{-1} \ .$$

Similarly

$$0 \leq \eta_2 \leq (\nu_{p+h}-\xi)^{-1}$$

and so, since $\rho(\xi) \geq 0$,

$$|\rho'(\xi)| \leq \rho(\xi)\max[(\xi-\nu_{p-1})^{-1},(\nu_{p+h}-\xi)^{-1}] \ .$$

Both terms in the brackets appear as numerators in the formula $\rho(\xi)$ given in (5). Using the fact that each factor in (5) lies between 0 and 1 we find for $\theta_p \leq \xi \leq \theta_{p+h}$

$$\begin{aligned}|\rho'(\xi)| &\leq \max[(\xi-\theta_{p-1})^{-1},(\theta_{p+h+1}-\xi)^{-1}] \\ &\leq 1/\mu\|A\| \ .\end{aligned}$$

Hence for all i

$$|(\rho_m-\rho_i)/(\theta_m-\theta_i)| \leq 1/\mu\|A\|$$

and the result follows from Lemma 4. □

This completes the proof of Theorem 6. □

Remarks. In practical examples the bounds given by Theorem 6 are unrealistic. As with Theorem 5 the behavior of the lengths of Ritz vectors appears to be almost independent of j.

2.10 Corrective Measures

Lanczos himself was aware of the inevitable loss of linear independence among the columns of Q_j. When he introduced the algorithm in 1950 he suggested that each newly computed Lanczos vector q_i be explicitly orthogonalized against all the preceding vectors. This is called reorthogonalization and despite the great cost of this device (about $j/3$ times the cost of the simple Lanczos algorithm both in time and storage) the Lanczos algorithm with reorthogonalization was the standard method of reducing a symmetric matrix to tridiagonal form (1950-54) until the advent of explicit orthogonal transformations.

In current usage A is a large matrix ($n \geq 1000$) and the cost of using reorthogonalization, even for $j \doteq \sqrt{n}$, is prohibitive. This poses a serious dilemma. Reorthogonalization is too expensive but independence will surely be lost without it.

C. Paige has suggested that no corrective action be taken. The loss of linear independence among the Lanczos vectors merely results in the appearance of multiple copies of the converged Ritz vectors. The rest of the Ritz pairs continue to improve as the algorithm proceeds. This approach was used by J. Lewis [Lewis 1977] on a difficult interior eigenvalue problem. There are two possible drawbacks to this approach. It is necessary for the user to distinguish which Ritz pairs are copies and which are distinct. This is not usually too difficult for the user but it is not clear how to automate such a decision procedure. Another drawback is that the algorithm may compute many copies of some Ritz pairs before the desired Ritz pairs are found. See [Lewis 1977] for a striking example of this phenomenon.

Another possible approach is to use the Lanczos algorithm iteratively. The basic idea is to stop at some step, compute the best approximation to a desired eigenvector, and use it or some modification of it as a new starting vector. If reorthogonalization is used the step at which the algorithm is iterated is determined by storage and cost considerations. If reorthogonalization is not used it is necessary to monitor the loss of orthogonality and iterate whenever significant loss of orthogonality is detected.

It is important to realize that iterative use of the Lanczos algorithm is theoretically unfortunate. Information is always lost when the algorithm is restarted. In exact arithmetic, Ritz pairs obtained by iteration are always inferior to the Ritz pairs which would be obtained if the algorithm were carried on for the same number of total steps. This occurs because the Krylov subspace computed in the last iteration is strictly contained in the Krylov subspace which would be obtained by going on. For a striking example of this phenomenon, consider the problem of finding the smallest eigenvalue of a 6×6 matrix. Of course the Lanczos algorithm will find all six eigenvalues in six steps, but if we are forced to iterate after five steps, it will take several iterations to obtain good accuracy. Iteration is forced on us only because of the problems associated with loss of orthogonality.

A practical difficulty in using iterative Lanczos arises in choosing the restarting vector. If more than one eigenpair is desired how can they all be represented in one vector? This problem, along with the theoretical difficulties in finding multiple eigenvalues, led several researchers to investigate block generalizations of the Lanczos algorithm, now called simply block Lanczos. The block Lanczos algorithm replaces each q vector by an $n \times p$ orthonormal matrix. The resulting T is block tridiagonal with block size p. One of the unsolved problems in using block Lanczos is the a priori determination of the optimal block size. Costs increase sharply if the optimum is missed.

Both J. Cullum and W.E. Donath [Cullum and Donath 1974] and R. Underwood [Underwood 1975] have implemented block Lanczos programs. Underwood uses full reorthogonalization while Cullum and Donath do not. In any case the block version does solve many of the problems associated with iterating the simple Lanczos algorithm. In particular, since the starting block contains more than one vector, more information may be saved when the algorithm restarts. Furthermore multiple eigenvalues (up to the size of the block) can be found simultaneously.

On the other hand, if an efficient method existed for preventing the loss of independence among the columns of Q_j the need for iteration would be eliminated. In Chapter 3 we analyze Selective Orthogonalization a new and efficient method for maintaining independence. As a byproduct, Selective Orthogonalization also allows multiple eigenvalues to be found without using blocks and without iterating.

3. The Lanczos Algorithm with Selective Orthogonalization

Selective Orthogonalization, hereafter referred to as SO, is a variant of the Lanczos algorithm which interpolates between the simple Lanczos algorithm and Lanczos with full reorthogonalization in an attempt to obtain the best of both worlds.

The simple Lanczos algorithm is very cheap but suffers from the inevitable loss of linear independence among the Lanczos vectors (columns of Q_j). This loss of independence is manifested in the Ritz vectors by the appearance of repeated copies of converged eigenvectors, as detailed in Chapter 2. Full reorthogonalization, in which each newly computed Lanczos vector, q_{j+1}, is explicitly orthogonalized against all preceding Lanczos vectors, cures the instability[†] of the simple Lanczos algorithm but is ruinously expensive in both time and storage.

SO attempts to obtain the stability of full reorthogonalization (no redundant copies of eigenvectors computed) at a cost which is close to that of the simple Lanczos algorithm.

3.1 Motivation for Selective Orthogonalization

To motivate SO we consider two thought experiments on possible variants of full reorthogonalization. If e-orthogonality of the Lanczos vectors ($|q_i^* q_k| \doteq e$, for $i \neq k$) is desired then no substantial improvement over full reorthogonalization can be achieved. Instead we relax our standards and concentrate on maintaining robust linear independence among the Lanczos vectors. That is we insure that

[†] An orthogonalization must be repeated if cancellation occurs.

$|q_i^* q_k| \leq \tau$, for $i \neq k$, for some given number τ, which may be much larger than ϵ.

<u>Scheme 1</u>. One way to insure that $|q_i^* q_k| \leq \tau$, for $i \neq k$ is as follows. As each new Lanczos vector q_{j+1} is computed, merely compute $q_i^* q_{j+1}$, for $i \leq j$, and orthogonalize q_{j+1} against q_i whenever $|q_i^* q_{j+1}| > \tau$. (We note that since the Lanczos vectors are not orthogonal, orthogonalization of q_{j+1} against q_i may increase $|q_k^* q_{j+1}|$ for some other k. Discussion of this second order effect is delayed until Section 9.)

Scheme 1 was implemented on the following test problem.

<u>Test 1</u>.

$$\epsilon = .16 \times 10^{-16}$$

$$n = 20$$

$$\lambda_i = 1/i \ , \quad \text{for} \quad i = 1,2,\ldots,20$$

$$q_1 = u/\|u\| \ , \quad u = (1,1,\ldots,1)^*$$

The following results were obtained for various values of τ. No orthogonalizations were performed at step $j = n = 20$.

Scheme 1 on Test 1

	τ	Number of orthogonalizations	$\|1 - Q_{20}^* Q_{20}\|$	max error $\max \|\lambda_i - \theta_i^{(20)}\|$
full reorth.	10^{-18}	190	$.62 \times 10^{-16}$	$.90 \times 10^{-16}$
	10^{-17}	148	$.66 \times 10^{-16}$	$.29 \times 10^{-16}$
	10^{-16}	98	$.32 \times 10^{-15}$	$.97 \times 10^{-16}$
	10^{-15}	90	$.31 \times 10^{-14}$	$.55 \times 10^{-16}$
	10^{-14}	79	$.24 \times 10^{-13}$	$.14 \times 10^{-15}$
	10^{-13}	72	$.28 \times 10^{-12}$	$.69 \times 10^{-16}$
	10^{-12}	66	$.27 \times 10^{-11}$	$.56 \times 10^{-16}$
	10^{-11}	70	$.31 \times 10^{-10}$	$.76 \times 10^{-16}$
	10^{-10}	55	$.21 \times 10^{-9}$	$.97 \times 10^{-16}$
	10^{-9}	55	$.25 \times 10^{-8}$	$.11 \times 10^{-15}$
	10^{-8}	51	$.21 \times 10^{-7}$	$.83 \times 10^{-16}$
	10^{-7}	47	$.22 \times 10^{-6}$	$.14 \times 10^{-14}$
	10^{-6}	39	$.15 \times 10^{-5}$	$.36 \times 10^{-12}$
	10^{-5}	46	$.16 \times 10^{-4}$	$.20 \times 10^{-10}$
	10^{-4}	37	$.14 \times 10^{-3}$	$.25 \times 10^{-9}$
	10^{-3}	36	$.16 \times 10^{-2}$	$.54 \times 10^{-6}$
	10^{-2}	39	$.24 \times 10^{-1}$	$.19 \times 10^{-4}$
	10^{-1}	35	.20	$.12 \times 10^{-2}$
simple L.	1.	0	1.0	.50

The number of orthogonalizations decreases as τ increases yet the Ritz values at step 20 are correct to working accuracy until $\tau = 10^{-7} < \sqrt{\varepsilon}$. At $\tau = 10^{-8}$ only 51 orthogonalizations were required compared to 190 orthogonalizations required by full reorthogonalization.

Scheme 1 with $\tau = 10^{-8}$ is an improvement over full reorthogonalization but it does have two drawbacks. At each step it is necessary to compute $q_i^* q_{j+1}$ for each $i \leq j$ to determine whether q_{j+1} should be orthogonalized against q_i. This requires that the q_i be kept in fast store so they are available at each step. Furthermore the cost of computing $q_i^* q_{j+1}$ is equal to the cost of the vector subtraction so little is saved by omitting the orthogonalization.

Scheme 2. We now consider what appears, at first sight, to be an even sillier method of maintaining robust independence. At each step j, compute each Ritz vector $y_i = Q_j s_i$ (where s_i is the i^{th} eigenvector of T_j), then compute $y_i^* q_{j+1}$ and orthogonalize q_{j+1} against y_i whenever $|y_i^* q_{j+1}| > \tau$.

Scheme 2 was implemented on Test 1 with the following results.

Scheme 2 on Test 1

	τ	Number of orthogonalizations	$\|1 - Q_{20}^*Q_{20}\|$	max error max $\|\lambda_i - \theta_i^{(20)}\|$
full reorth.	10^{-18}	190	$.71 \times 10^{-16}$	$.83 \times 10^{-16}$
	10^{-17}	159	$.80 \times 10^{-16}$	$.11 \times 10^{-15}$
	10^{-16}	89	$.37 \times 10^{-15}$	$.83 \times 10^{-16}$
	10^{-15}	48	$.28 \times 10^{-14}$	$.42 \times 10^{-16}$
	10^{-14}	37	$.25 \times 10^{-13}$	$.48 \times 10^{-16}$
	10^{-13}	31	$.20 \times 10^{-12}$	$.11 \times 10^{-15}$
	10^{-12}	22	$.18 \times 10^{-11}$	$.56 \times 10^{-16}$
	10^{-11}	22	$.13 \times 10^{-10}$	$.62 \times 10^{-16}$
	10^{-10}	19	$.11 \times 10^{-9}$	$.69 \times 10^{-16}$
	10^{-9}	17	$.11 \times 10^{-8}$	$.83 \times 10^{-16}$
	10^{-8}	14	$.18 \times 10^{-7}$	$.45 \times 10^{-16}$
	10^{-7}	12	$.11 \times 10^{-6}$	$.61 \times 10^{-15}$
	10^{-6}	10	$.12 \times 10^{-5}$	$.12 \times 10^{-12}$
	10^{-5}	12	$.13 \times 10^{-4}$	$.18 \times 10^{-10}$
	10^{-4}	8	$.65 \times 10^{-4}$	$.19 \times 10^{-9}$
	10^{-3}	11	$.13 \times 10^{-2}$	$.57 \times 10^{-6}$
	10^{-2}	6	$.11 \times 10^{-1}$	$.70 \times 10^{-5}$
	10^{-1}	9	.30	$.23 \times 10^{-2}$
simple Lanczos	1.	0	1.00	.50

Again we find that the Ritz values at step 20 are correct to working accuracy until $\tau < \sqrt{\epsilon}$. However the number of orthogonalizations required by Scheme 2 for values of τ near $\sqrt{\epsilon}$ is much smaller than for Scheme 1. Furthermore the drawbacks associated with Scheme 1 can be avoided in implementing Scheme 2 as we show in the next section.

3.2 Implementation of Selective Orthogonalization

Section 1 shows that τ-orthogonality can be maintained by Scheme 2 with many fewer orthogonalizations than Scheme 1, for values of τ near $\sqrt{\epsilon}$. This effect is symptomatic of the Lanczos algorithm and does not depend on the particular test problem chosen.

An intuitive explanation of the success of Scheme 2 can be obtained from Theorem 3 of Chapter 2 which states that for all $i \leq j$ and all j,

$$(1) \qquad |y_i^{(j)*}q_{j+1}| = \epsilon\|A\|\gamma_{ji}/\beta_{ji} \text{ ,}$$

where $\gamma_{ji} \doteq 1$. Equation (1) shows that orthogonality is only lost in the direction of Ritz vectors with small β_{ji}. Since β_{ji} is a good estimate of the residual norm of $y_i^{(j)}$, serious loss of orthogonality occurs only in the direction of converged Ritz vectors.

Only a few of the Ritz vectors will have converged at any step j and so only a few orthogonalizations are needed. Furthermore since $\gamma_{ji} \doteq 1$, it is possible to estimate $|y_i^{(j)*}q_{j+1}|$ using equation (1) by

$$(2) \qquad |y_i^{(j)*}q_{j+1}| \doteq \epsilon\|A\|/\beta_{ji} \text{ .}$$

Thus it is possible to determine which Ritz vectors should be used for orthogonalization of q_{j+1} before computing them, namely any Ritz vector which satisfies

$$(3) \qquad \beta_{ji} \leq \epsilon\|A\|/\tau \text{ .}$$

Any Ritz vector which satisfies equation (3) will be called a good Ritz vector, while the remaining Ritz vectors will be called bad Ritz vectors.

$\beta_{ji} = \beta_j s_{ji}$ can be computed from the spectral decomposition of T_j. In practical problems $j << n$ and β_{ji} can be computed much more quickly than first computing $y_i^{(j)} = Q_j s_i^{(j)}$ and then computing $y_i^{(j)*} q_{j+1}$. In general $\|A\|$ is not known but $\|T_j\|$ is known from the spectral decomposition of T_j and $\|T_j\| \doteq \|A\|$ will hold even for $j << n$. Thus in practice $\|T_j\|$ replaces $\|A\|$ in (2). Nevertheless the computation of even a few Ritz vectors is time consuming. It takes jn multiplications to compute a single Ritz vector. In theory the good Ritz vectors are different at each step and must be recomputed, which destroys the efficiency of SO.

Fortunately, for practical values of τ (τ near $\sqrt{\epsilon}$) the good Ritz vectors change very little from step to step and a good Ritz vector computed at one step may be safely used for orthogonalization at later steps. However to avoid complicating the analysis, we will consider only the model in which the good Ritz vectors are recomputed at each step.

Equation (1) is the basis for using equation (2) to determine the good Ritz vectors. Unfortunately once orthogonalization begins in SO the computed quantities no longer satisfy the fundamental equation of the simple Lanczos algorithm. Therefore the conclusions of Chapter 2 may no longer be valid. In particular γ_{ji} defined by equation (1) may be much larger than 1. The rest of this chapter will examine the relationship between the choice of τ and γ growth in SO. In Section 4 we give a numerical example which shows that large γ's can occur for values of τ near 1.

3.3 Governing Equations for SO

We first observe that it is unnecessary to normalize r_j (to become q_{j+1}) before computing the good Ritz vectors, which are determined by $\beta_j = \|r_j\|$ and the bottom row of S_j, the eigenvector matrix of T_j. r_j itself is orthogonalized against the good Ritz vectors and this new r_j is normalized (by dividing by a new β_j) to become q_{j+1}. In order to distinguish the new quantities from the old we use the following notation.

Let r_j, β_j, q_{j+1}, and γ_{ji}, for $i \leq j$, be the quantities computed by the j^{th} step of SO. Let r_j', β_j', q_{j+1}', and γ_{ji}', for $i \leq j$, be the quantities which would be computed if the orthogonalizations of the j^{th} step (only) were omitted.

Remarks. If there are no good Ritz vectors at step j of SO then no orthogonalizations are performed and $r_j = r_j'$, $\beta_j = \beta_j'$, $q_{j+1} = q_{j+1}'$, and $\gamma_{ji} = \gamma_{ji}'$, for $i \leq j$. Furthermore SO never modifies T_j or Q_j and so the Ritz pairs computed at step j are unchanged by the orthogonalizations. Finally $\beta_j = \|r_j\| \leq \|r_j'\| = \beta_j'$ since orthogonalization always shortens the vector r_j'.

As mentioned at the end of Section 2, once orthogonalization begins the computed quantities no longer satisfy

$$(1) \qquad AQ_j - Q_jT_j = r_je_j^* + F_j$$

with $\|F_j\| \doteq \varepsilon\|A\|$, which is the fundamental equation of the simple Lanczos algorithm.

Consider the step j at which a single Ritz vector, y_i, first becomes good. To orthogonalize r_j' against y_i, SO computes

$\xi_{ji} = y_i^* r_j' / y_i^* y_i$ and $r_j = r_j' - y_i \xi_{ji}$. We note that the division by $y_i^* y_i$ in the definition of ξ_{ji} is necessary because y_i will not have length exactly 1. Before the orthogonalization of r_j' against y_i the computed quantities satisfy

(2) $$AQ_j - Q_j T_j = r_j' e_j^* + F_j \ .$$

Since $r_j = r_j' - y_i \xi_{ji}$, equation (2) becomes

$$\begin{aligned} AQ_j - Q_j T_j &= (r_j + y_i \xi_{ji}) e_j^* + F_j \ , \\ &= r_j e_j^* + D_j + F_j \ , \end{aligned}$$

where $D_j = d_j e_j^* = (0,0,\ldots,0,d_j)$ and $d_j = y_i \xi_{ji}$.

For subsequent j, $D_j = (d_1, d_2, \ldots, d_j)$ and d_k is the accumulation of the orthogonalizations performed at step k. That is, if G_k is the index set of the good Ritz vectors at step k, then for $k \leq j$,

$$d_k = \sum_{i \in G_k} y_i^{(k)} \xi_{ki} \ ,$$

where $\xi_{ki} = y_i^{(k)*} r_k' / y_i^{(k)*} y_i^{(k)}$.

Thus $D + F$ must replace F in applying the results of Chapter 2. In particular, Theorem 3 of Chapter 2 states that for the simple Lanczos algorithm,

$$\beta_{ji} | y_i^* q_{j+1} | = \epsilon \| A \| \gamma_{ji}$$

with $\gamma_{ji} = 1$, for all i and j. The introduction of the matrix D into the analysis may cause γ_{ji} to be much larger than 1. In the next section we show by example that large γ's can occur for values of τ near 1.

3.4 An Extreme Example

In this section we give an example which shows that if orthogonalization is delayed long enough (τ near 1) it is possible for the γ's to grow enormously. Consider Example 1 of Chapter 2, namely,

$$e = .6 \times 10^{-7}$$

$$n = 10$$

$$A = \mathrm{diag}(0,.01,.02,\ldots,.08,1.0)$$

$$q_1 = u/\|u\| \ , \quad u = (1,1,\ldots,1)^*$$

At $j = 7$, y_7 is a very good approximation to Z_{10}, the eigenvector of 1.0. We repeat the results obtained by simple Lanczos at $j = 7$ but including two extra columns to show the effects or orthogonalizing r_7' against y_7 and then normalizing.

$j = 7$

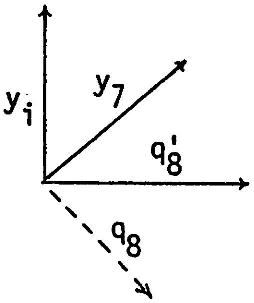
i	Ritz value	β'_{ji}	$\lvert y_i^* q_8' \rvert$	γ'_{ji}	$\lvert y_i^* q_8 \rvert$	γ_{ji}
1	$.417 \times 10^{-3}$	$.114 \times 10^{-1}$	$.121 \times 10^{-7}$	.002	.0585	2100
2	.0147	$.329 \times 10^{-1}$	$.335 \times 10^{-7}$	.018	.171	17667
3	.0348	$.529 \times 10^{-1}$	$.410 \times 10^{-7}$	.036	.280	46667
4	.0564	$.704 \times 10^{-1}$	$.894 \times 10^{-7}$	.105	.381	84500
5	.0733	$.902 \times 10^{-1}$	$.224 \times 10^{-7}$	.034	.492	141333
6	.0811	$.561 \times 10^{-1}$	$.521 \times 10^{-7}$	.049	.312	55167
7	1.0000	$.516 \times 10^{-8}$	.982	.085	$.668 \times 10^{-6}$	1.08×10^{-8}

How is it possible for orthogonalization of q_8' against y_7 to make $|y_i^* q_8|$ seven orders of magnitude greater than $|y_i^* q_8'|$, for $i \neq 7$? The main cause of the lack of orthogonality between q_8 and y_i, $i \neq 7$, is that since Q_7 is already far from orthonormal, y_7

is far from orthogonal to $y_1, y_2, \ldots, y_6$. In fact the bottom row of $|1-Y_7^*Y_7|$ is

.011	.038	.054	.073	.096	.060	$.24 \times 10^{-6}$

Thus both q_8' and y_i have spurious components in the direction of y_7. These components are correlated so that $y_i^* q_8'$ is tiny. Removal of the component of y_7 from q_8' destroys this correlation as shown by the following picture.



We now pause to put this example into perspective. The purpose of SO is to insure that

$$|y_i^* q_{j+1}| \le \tau \tag{1}$$

for all the Ritz vectors at step j. To avoid computing all the Ritz vectors SO uses the equation

$$|y_i^* q_{j+1}'| = \epsilon \|A\| \gamma_{ji}' / \beta_{ji}' \tag{2}$$

to determine which Ritz should be labelled good. SO then computes these Ritz vectors and explicitly orthogonalizes r_j' against them.

The resulting r_j is normalized to become q_{j+1} which satisfies

(3) $$|y_i^* q_{j+1}| = \epsilon \|A\| \gamma_{ji} / \beta_{ji} \ , \quad \text{for all} \quad i \leq j \ .$$

Comparing equation (2) and (3) we see that SO will fail to maintain equation (1), if $\gamma_{ji} >> \gamma'_{ji}$ and Example 1 shows that this behavior is possible for values of τ near 1. Therefore the optimal choice of τ is determined by two competing factors. If τ is chosen too small then many Ritz vectors will be declared good and many unnecessary orthogonalizations will be performed. If τ is chosen too large then the orthogonalizations will be delayed too long and large γ's will appear causing serious loss of orthogonality, as shown by equation (3).

3.5 The Effects of Orthogonalization

We now establish a bound on the difference between γ'_{jk}, the quantity before the orthogonalization of the j^{th} step, and γ_{jk}, the corresponding quantity after r'_j is orthogonalized against the good Ritz vectors.

<u>Theorem 1.</u> Let G_j be the index set of the good Ritz vectors at step j. Then for all $k \leq j$

$$\gamma_{jk} \leq \gamma'_{jk} + \rho_{jk} \ ,$$

where $\rho_{jk} = s_{jk} \sum_{i \in G_j} |(y_k^* y_i) \xi_{ji}| / \epsilon \|A\|$, and $\xi_{ji} = y_i^* r'_j / y_i^* y_i$.

Proof. Recall that γ_{jk} is defined by the equation

$$\beta_{jk}|y_k^* q_{j+1}| = \epsilon\|A\|\gamma_{jk} .$$

Since $\beta_{jk} = \beta_j s_{jk}$ and $\beta_j q_{j+1} = r_j$, we find

(1) $$\epsilon\|A\|\gamma_{jk} = |y_k^* r_j| s_{jk} .$$

Similarly, γ'_{jk} satisfies

(2) $$\epsilon\|A\|\gamma'_{jk} = |y_k^* r'_j| s_{jk} .$$

By definition of r_j (see Section 3) $r_j = r'_j - d_j$ with $d_j = \sum_{i \in G_j} y_i \xi_{ji}$ and $\xi_{ji} = y_i^* r'_j / y_i^* y_i$. Substituting this formula for r_j into equation (1) we obtain

(3) $$\begin{aligned} \epsilon\|A\|\gamma_{jk} &= |y_k^* r'_j - \sum_{i \in G_j} (y_k^* y_i)\xi_{ji}| s_{jk} , \\ &\leq |y_k^* r'_j| s_{jk} + \sum_{i \in G_j} |(y_k^* y_i)\xi_{ji}| s_{jk} . \end{aligned}$$

The result now follows from equation (2). □

Remarks. Theorem 1 is not a realistic bound for a good Ritz vector. If $i \in G_j$ then r_j has been explicitly orthogonalized against $y_i^{(j)}$ which makes $|y_i^{(j)*} r_j| \doteq \epsilon\|r_j\| = \epsilon\beta_j$. Thus from equation (1)

$$\begin{aligned} \gamma_{ji} &= \epsilon\beta_j s_{ji}/\epsilon\|A\| \\ &= \beta_{ji}/\|A\| \\ &\leq \tau \end{aligned}$$

by definition of a good Ritz vector. On the other hand Theorem 1 is a realistic bound on γ_{jk} for a bad Ritz vector y_k. Insight into γ

growth in SO can be gained from analyzing the quantity ρ_{jk}.

The formula for ρ_{jk} in Theorem 1 does indicate why the γ's should remain near 1 for values of τ rather larger than ϵ. In exact arithmetic, for a bad Ritz vector y_k, both $y_i^* y_k$ and $\xi_{ji} = y_i^* r_j' / y_i^* y_i$ are zero. In practice it is unnecessary that each factor be tiny. It is only required that their product be less than $\epsilon \|A\|$ to insure that γ_{jk} is not much larger than γ_{jk}'.

3.6 The First Orthogonalization

We now use Theorem 1 to analyze a simple case, namely the step j at which the first good Ritz vector y_i is found. This is precisely the situation in the numerical example of Section 4.

Since j is the first step at which a good Ritz vector is found no orthogonalizations have occurred at earlier steps. Therefore the quantities r_j', Q_j, T_j, and Y_j have been computed by the simple Lanczos algorithm and the results of Chapter 2 are valid.

In particular by Theorem 3 (Paige) of Chapter 2,

$$(1) \qquad |y_i^* r_j'| = \epsilon \|A\| \gamma_{ji}' / s_{ji} \ ,$$

with $\gamma_{ji}' \doteq 1$. Furthermore by Theorem 4 (Paige) of Chapter 2, since $s_{ji} < s_{jk}$ (y_i is good and y_k is not),

$$(2) \qquad |y_i^* y_k| \doteq \epsilon \|A\| \gamma_{ji}' (s_{jk}/s_{ji}) / |\theta_i - \theta_k| \ .$$

Substituting (1) and (2) into the formula for ρ_{jk} in Theorem 1 we obtain

$$(3) \qquad \rho_{jk} \doteq s_{jk}^2 \gamma_{ji}'^2 (\epsilon / s_{ji}^2)(\|A\| / |\theta_i - \theta_k|) / y_i^* y_i \ ,$$

since G_j contains the single number i. Equation (3) displays the various contributions to the size of ρ_{jk}. s_{jk} the bottom element of s_k, which corresponds to a bad Ritz vector y_k, may not be small at all but $s_{jk}^2 \leq 1$ holds a priori. Furthermore $\gamma'_{ji} \doteq 1$ so that γ'^2_{ji} cannot be large either. Finally for any reasonable choice of τ robust linear independence will be maintained and $y_i^* y_i$ will be close to 1.

The factor $\|A\|/|\theta_i - \theta_k|$ is large only when θ_i and θ_k are close together, that is when θ_i and θ_k are two copies of the same eigenvalue of A. However two copies of an eigenvalue can occur only if one good copy already existed at an earlier step which contradicts the assumption that y_i is the first good Ritz vector found.

It is the factor, ϵ/s_{ji}^2, which can cause γ-growth if s_{ji} is tiny. A good Ritz vector satisfies

$$\beta_{ji} \leq \epsilon\|A\|/\tau \quad ,$$

or equivalently

$$s_{ji} \leq \epsilon\|A\|/\tau\beta_j \quad ,$$

since $\beta_{ji} = \beta_j s_{ji}$. If β_j is not much smaller than $\|A\|$ and τ is near 1, then s_{ji} will be near ϵ and ρ_{jk} will be large.

In Example 1 of Section 4 where $i = j = 7$, $s_{77} \doteq \epsilon$ so that ϵ/s_{ji}^2, the third factor in equation (8) is quite large. This causes γ_{7k}, for $k \neq 7$, to be large as illustrated by the numerical results.

Equation (3) does give an indication of the optimal choice of τ. To prevent γ-growth it is necessary that

$$\gamma'^2_{ji}\epsilon/s_{ji}^2 \leq 1 \quad . \tag{4}$$

Using equation (1), equation (4) can be rearranged as

$$(y_i^* r_j' / \|A\|)^2 \leq \epsilon$$

or equivalently

(5) $$|y_i^* q_{j+1}'| \leq \sqrt{\epsilon}\|A\|/\beta_j' \ ,$$

since $r_j' = \beta_j' q_{j+1}$. Equation (5) indicates that τ should satisfy

(6) $$\tau \leq \sqrt{\epsilon}\|A\|/\beta_j' \ .$$

Inevitably $\beta_j' \leq \|A\|$ so equation (6) will certainly be satisfied if $\tau = \sqrt{\epsilon}$. Since β_j' need not be much smaller than $\|A\|$, a choice of τ larger than $\sqrt{\epsilon}$ may lead to γ-growth and so the optimal choice of τ is $\sqrt{\epsilon}$. If τ is chosen much smaller than $\sqrt{\epsilon}$ then many unnecessary orthogonalizations are performed while if τ is chosen much larger than $\sqrt{\epsilon}$ then large γ's will occur.

However equation (5) suggests that the definition of a good Ritz vector should depend on the ratio $\|A\|/\beta_j'$. We investigate this variant of SO in the next section.

3.7 A Variant of SO

Equation (5) of Section 6 indicates that stability of SO will be assured if

$$|y_i^* q_{j+1}'| \leq \sqrt{\epsilon}\|A\|/\beta_j' \ .$$

Since $|y_i^* q_{j+1}'| = \epsilon\|A\|\gamma_{ji}'/\beta_{ji}'$, this suggests that a Ritz vector should be declared good whenever

(1) $$\epsilon\|A\|\gamma_{ji}'/\beta_{ji}' \geq \sqrt{\epsilon}\|A\|/\beta_j' \ .$$

Assuming that γ-growth has not occurred at earlier steps, so that $\gamma'_{ji} \leq 1$, equation (1) can be rearranged to obtain

(2) $$s_{ji} \leq \sqrt{\epsilon} .$$

We call the variant of SO based on equation (2) <u>SO2</u> to distinguish it from the original SO, in which a Ritz vector y_i is declared good whenever

(3) $$\beta_{ji} \leq \sqrt{\epsilon}\|A\| .$$

A Ritz vector which is declared good by SO will be called an SO-vector and similarly for SO2. Since $\beta'_j \leq \|A\|$, any SO2-vector is also SO-vector and so SO2 will always require fewer orthogonalizations than SO. However on most examples the cost of SO and SO2 are quite close. Furthermore on occasions when the costs of the two schemes are disparate, SO2 may suffer from γ-growth as illustrated by the following example.

<u>Test 2</u>.
$$\begin{aligned} \epsilon &= .16 \times 10^{-16} \\ n &= 20 \\ \lambda_i &= (.2)^{i-1}, \quad \text{for } i = 1,2,\ldots,n \\ q_i &= u/\|u\|, \quad u = (1,1,\ldots,1)^* . \end{aligned}$$

Test 2 is a difficult problem in that most of the eigenvalues are clustered near zero ($\lambda_{19} - \lambda_{20} = \lambda_{20} = .53 \times 10^{-13}$).

SO2 was run on Test 2 and the following results were obtained.

j	beta(j)	# good	max γ_{ji}	$\|1-Q_j^*Q_j\|$
1	.13	0	.65	$.25\times 10^{-16}$
2	.17	0	.70	$.42\times 10^{-16}$
3	$.33\times 10^{-1}$	0	.85	$.10\times 10^{-15}$
4	$.80\times 10^{-2}$	0	.85	$.19\times 10^{-14}$
5	$.32\times 10^{-2}$	0	.85	$.22\times 10^{-12}$
6	$.86\times 10^{-4}$	0	.85	$.68\times 10^{-10}$
7	$.13\times 10^{-3}$	1	.85	$.80\times 10^{-6}$
8	$.23\times 10^{-4}$	2	$.39\times 10^{4}$	$.11\times 10^{-5}$
9	$.28\times 10^{-5}$	3	$.12\times 10^{5}$	$.11\times 10^{-5}$
10	$.15\times 10^{-6}$	4	$.97\times 10^{4}$	$.31\times 10^{-4}$
11	$.11\times 10^{-4}$	5	$.45\times 10^{6}$	$.82\times 10^{-1}$
12	$.68\times 10^{-5}$	5	$.27\times 10^{10}$	1.00
13	$.86\times 10^{-5}$	5	$.48\times 10^{9}$	1.00
14	$.55\times 10^{-5}$	5	$.63\times 10^{9}$	1.00
15	$.23\times 10^{-4}$	5	$.90\times 10^{9}$	1.03
16	$.41\times 10^{-4}$	6	$.98\times 10^{9}$	1.37
17	$.24\times 10^{-4}$	6	$.58\times 10^{12}$	1.87
18	$.18\times 10^{-4}$	6	$.30\times 10^{12}$	1.97
19	$.62\times 10^{-5}$	6	$.22\times 10^{12}$	2.01
20	$.25\times 10^{-4}$	0	$.17\times 10^{12}$	2.82

The maximum absolute error at step j = 20 was $.26\times 10^{-3}$ which shows that serious loss of accuracy had occurred. This loss of accuracy was caused by the large γ's in SO2 which first appeared at step j = 8 with a $\gamma_{ji} = .39\times 10^{4}$. The large γ's in turn led to a complete breakdown of orthogonality among the columns of Q_j as shown by $\|1-Q_{20}^*Q_{20}\| = 2.82$.

We also ran SO on Test 2 for comparison against SO2. SO required 148 orthogonalizations compared to 59 for SO2, but for SO the maximum error at j = 20 was $.14\times 10^{-15}$, while $\|1-Q_{20}^*Q_{20}\| = .10\times 10^{-8}$ and the maximum $\gamma_{ji} = .85$. Thus SO suffered no γ-growth and resolved all the eigenvalues to working accuracy.

What causes the γ-growth in SO2? We note that large γ's appeared only when the ratio $\beta_j/\|A\|$ was quite small. It is small off diagonal elements (β's) which cause γ-growth in SO2 as we show in the next section.

3.8 γ-Growth in SO2

The γ-growth in SO2 is caused by the fact that in the face of a tiny β_{j-1} there is no a prior upper bound on the ratio $\min_{i\leq j-1} s^{(j-1)}_{j-1,i}/\min_{k\leq j} s^{(j)}_{ji}$. That is, it is possible for a tiny s_{ji} to appear out of the blue, with no advance warning from any of the s_{ki} for $k < j$.

Consider the following example.

Example 1. Let $\omega = (1+\zeta^2)^{-1/2}$ and let

$$T_2 = \omega^2\begin{bmatrix} 1 & \zeta \\ \zeta & \zeta^2 \end{bmatrix} .$$

By the choice of ω, the eigenvalues of T_2 are 0 and 1 and the matrix of eigenvectors is

$$S_2 = \omega\begin{bmatrix} -\zeta & 1 \\ 1 & \zeta \end{bmatrix} .$$

Note that for all values of ζ, $\omega \leq 1$ and so $\beta_1 \leq \zeta$ and $s_{22} = \omega\zeta$. If $\zeta << 1$ then β_1 is tiny (compared to $\|T_2\| = 1$) and $s_{22} < \zeta$ is tiny as well. On the other hand $s_{11} = 1$ and so $\min_{i\leq 1} s_{1i}/\min_{i\leq 2} s_{2i} = 1/\omega\zeta$ can be arbitrarily large. □

The sudden appearance of a tiny s_{ji} causes γ-growth in SO2. In Section 5 (equation (3)) we showed that one of the factors in the growth of γ_{ji} is $\gamma'^{2}_{ji}\epsilon/s^{2}_{ji}$ which will be quite large if s_{ji} is tiny for the first good Ritz vector.

Another way to understand the γ-growth in SO2 is from the point of view of loss of orthogonality. By definition of γ_{ji},

$$|y_i^* q_{j+1}| = \gamma_{ji}/\beta_{ji} \ .$$

If β_j is tiny then all of the β_{ji} will be tiny and serious loss of orthogonality will occur unless r_j' is orthogonalized against all the y_i before being normalized to become q_{j+1}. SO, which examines the β_{ji}, will perform these needed orthogonalizations. On the other hand most of the s_{ji} will not be tiny and SO2 will fail to perform some needed orthogonalizations. Thus a rather poor q_{j+1} is accepted and the damage is done. At step j+1 SO2 tries to correct the errors in r'_{j+1} inherited from the poor q_{j+1}, which is a hopeless task.

3.9 Loss of Orthogonality of Good Ritz Vectors

At step j, SO computes the good Ritz vectors and orthogonalizes r_j' against them. The good Ritz vectors are some of the columns of

$$Y_j = Q_j S_j \ .$$

S_j, the eigenvector matrix of T_j is assumed to be orthogonal but Q_j is not orthonormal. Therefore Y_j is not orthogonal either and the good Ritz vectors (if there is more than 1) will not be orthogonal.

In order to successfully orthogonalize r_j' against the good Ritz vectors it is necessary to orthogonormalize the good Ritz vectors first. There are many ways to orthogonormalize a set of vectors. The simplest method is to apply the Gram-Schmidt procedure to some ordering of the good Ritz vectors. In this context it is best to order the good Ritz vectors by increasing β_{ji} as shown by the following analysis.

In Section 6 of Chapter 2 it was shown that the loss of orthogonality in the matrix Q_j was manifested in the matrix Y_j by the contamination of the unconverged Ritz vectors by components in the direction of the converging Ritz vector. In simple Lanczos this contamination grows until a second copy of the converged Ritz vector appears. In SO the onset of orthogonalization stifles the growth of the contamination and thus prevents the appearance of repeated copies of eigenvectors of A.

However the orthogonalizations do not purge the contamination that was present before the Ritz vector became good. It is this residual contamination which prevents the second good Ritz vector found from being orthogonal to the first. This contamination can be removed by simply orthonormalizing the good Ritz vectors in the order of increasing β_{ji}.

The required orthonormalization of the good Ritz vectors gives another indication that $\sqrt{\epsilon}$ is the proper value for τ. The Ritz values at step j are the Rayleigh quotients of the Ritz vectors. Do these Rayleigh quotients change when the Ritz vectors are orthonormalized? If $\tau \leq \sqrt{\epsilon}$ the Ritz values remain the same to working accuracy. We first prove a simple result about approximate eigenpairs in general and then apply it in the context of SO.

Theorem 2. Let y_1 and y_2 be two unit vectors and let $\theta_i = y_i^*Ay_i$, for $i = 1,2$. Let $\eta = y_1^*y_2$ and $v = y_2 - y_1\eta$. Then

(1) $1 - v^*v = \eta^2$

(2) $|\theta_2 - v^*Av| \le 2\eta\|Ay_1 - y_1\theta_1\| + \eta^2|\theta_1|$.

Proof. To prove (1),

$$\begin{aligned} v^*v &= (y_2-y_1\eta)^*(y_2-y_1\eta) \ , \\ &= y_2^*y_2 - 2y_1^*y_2\eta + y_1^*y_1\eta^2 \ , \\ &= 1 - 2\eta^2 + \eta^2 \ , \quad (\eta = y_1^*y_2) \\ &= 1 - \eta^2 \ . \end{aligned}$$

To prove (2),

$$\begin{aligned} v^*Av &= (y_2-y_1\eta)^*A(y_2-y_1\eta) \ , \\ &= y_2^*Ay_2 - \eta y_1^*Ay_2 - y_2^*Ay_1\eta + y_1^*Ay_1\eta^2 \ , \\ &= \theta_2 - 2\eta y_2^*Ay_1 + \eta^2\theta_1 \ , \\ &= \theta_2 - 2\eta y_2^*(Ay_1-y_1\theta_1) - \eta^2\theta_1 \ . \end{aligned}$$

Therefore

$$\begin{aligned} \|\theta_2 - v^*Av\| &\le 2\eta\|y_2\|\|Ay_1-y_1\theta_1\| + \eta^2|\theta_1| \ , \\ &= 2\eta\|Ay_1-y_1\theta_1\| + \eta^2|\theta_1| \ . \end{aligned}$$ □

We now apply Theorem 2 in the context of SO. In SO τ-orthogonality is maintained and $\eta \le \tau$ will hold. Furthermore (y_1,θ_1) is a good Ritz pair so

$$\|Ay_1 - y_1\theta_1\| \leq \epsilon\|A\|/\tau \ .$$

Finally $|\theta_1| \leq \|A\|$. Combining these inequalities we find

<u>Corollary</u>. $$1 - v^*v \leq \tau^2 \ , \text{ and}$$

$$|\theta_2 - v^*Av| \leq 2\epsilon\|A\| + \tau^2\|A\| \ .$$

Thus if $\tau \leq \sqrt{\epsilon}$, v^*Av/v^*v will be perturbed from θ_2 only by a term of order $\epsilon\|A\|$. Since θ_2 already has an error of this order of magnitude there is no need to recompute the Ritz value.

3.10 Further Analysis of SO

One of the technical difficulties in analyzing SO is the identification problem for Ritz vectors at different steps of the algorithm. In theory the Ritz vectors at step j are different from those at step j-1 and it may be impossible to identify a particular Ritz vector at step j as the successor of a given Ritz vector at step j-1. Indeed since there is one more Ritz vector at step j a complete one-to-one identification is impossible.

In practice however it is always possible to identify the successor of a good Ritz vector ($\tau = \sqrt{\epsilon}$). A good Ritz vector at step j-1 is a good approximation to an eigenvector of A. The Ritz vectors at step j are chosen from a larger subspace and an even better approximation to the eigenvector will be found.

As a notational convenience we use the symbol $y_\lambda^{(j)}$ to stand for the Ritz vector at step j which is closest to the eigenvector of A associated with the eigenvalue λ. Thus if $y_\lambda^{(j-1)}$ is a good Ritz vector then $y_\lambda^{(j)}$ is the successor of $y_\lambda^{(j-1)}$ and the two vectors will be almost parallel. We define $\beta_{j\lambda}$, $s_{j\lambda}$, $\gamma_{j\lambda}$, etc. to be the quantities corresponding to $y_\lambda^{(j)}$.

In this section we discuss the effects of the hereditary nature of good Ritz vectors on the behavior of SO. Selected output from SO (with $\tau = \sqrt{\epsilon}$) applied to the following example will be used for illustration.

<u>Example</u>. $\epsilon = .16 \times 10^{-16}$

$n = 20$

$\lambda_i = 1/i \ , \quad i = 1,2,\ldots,n$

$q_1 = u/\|u\|, \quad u = (1,1,\ldots,1)^*$

$\tau = \sqrt{\epsilon} = .4 \times 10^{-8}$

We first give a summary of the output.

j	# good	$\|1-Q_j^*Q_j\|$
1	0	$.28 \times 10^{-16}$
2	0	$.37 \times 10^{-16}$
3	0	$.10 \times 10^{-15}$
4	0	$.15 \times 10^{-15}$
5	0	$.44 \times 10^{-15}$
6	0	$.30 \times 10^{-14}$
7	0	$.36 \times 10^{-13}$
8	0	$.48 \times 10^{-12}$
9	0	$.84 \times 10^{-11}$
10	1	$.23 \times 10^{-9}$
11	1	$.23 \times 10^{-9}$
12	1	$.23 \times 10^{-9}$
13	2	$.23 \times 10^{-9}$
14	2	$.23 \times 10^{-9}$
15	3	$.23 \times 10^{-9}$
16	3	$.23 \times 10^{-9}$
17	4	$.23 \times 10^{-9}$
18	5	$.83 \times 10^{-9}$
19	5	$.83 \times 10^{-9}$
20	0	$.83 \times 10^{-9}$

Note that $\|1-Q_j^*Q_j\|$ increases smoothly until orthogonalization starts. Thereafter $\|1-Q_j^*Q_j\|$ is almost constant which is what is expected since the purpose of SO is to inhibit the further decay of orthogonality. No orthogonalizations were performed at step j = n = 20. In general there is no need to orthogonalize q_{j+1} in the last step of the algorithm since q_{j+1} does not contribute to the Ritz pairs at step j.

By the Kaniel-Paige theory (Section 1.3) the first eigenvalue to converge should be 1.0 since it is both extreme and well separated. We now give a complete history of $\theta_{1.0}$, the Ritz value closest to 1.0 for each step j.

j	$1-\theta_{1.0}^{(j)}$	$\beta_{j,1.0}$	$\gamma'_{j,1.0}$	$\lvert\xi_{j,1.0}\rvert$
1	.87	.142	.87	--
2	.20	.31	.94	--
3	$.15\times10^{-1}$	$.93\times10^{-1}$	.49	--
4	$.32\times10^{-3}$	$.15\times10^{-1}$	.37	--
5	$.53\times10^{-5}$	$.2\times10^{-2}$	.36	--
6	$.31\times10^{-7}$	$.16\times10^{-3}$	.36	--
7	$.18\times10^{-9}$	$.12\times10^{-4}$	.36	--
8	$.55\times10^{-12}$	$.69\times10^{-6}$	.36	--
9	$.73\times10^{-15}$	$.25\times10^{-7}$	.36	--
10	$.28\times10^{-16}$	$.74\times10^{-9}$	.36	$.21\times10^{-9}$
11	$.83\times10^{-16}$	$.26\times10^{-10}$	$.32\times10^{-3}$	$.61\times10^{-11}$
12	$.14\times10^{-16}$	$.10\times10^{-11}$	$.60\times10^{-11}$	$.33\times10^{-17}$
13	$.28\times10^{-16}$	$.22\times10^{-13}$	$.16\times10^{-12}$	$.22\times10^{-17}$
14	$.42\times10^{-16}$	$.32\times10^{-15}$	$.34\times10^{-14}$	$.23\times10^{-17}$
15	$.56\times10^{-16}$	$.43\times10^{-17}$	$.23\times10^{-16}$	$.10\times10^{-17}$
16	$.42\times10^{-16}$	$.53\times10^{-19}$	$.88\times10^{-18}$	$.34\times10^{-17}$
17	$.56\times10^{-16}$	$.11\times10^{-18}$	$.35\times10^{-17}$	$.44\times10^{-17}$
18	$.56\times10^{-16}$	$.18\times10^{-19}$	$.34\times10^{-18}$	$.33\times10^{-17}$
19	.0	$.19\times10^{-19}$	$.18\times10^{-18}$	$.46\times10^{-18}$
20	$.28\times10^{-16}$	$.65\times10^{-28}$	$.48\times10^{-20}$	--

The vector $y_{1.0}$ goes through three distinct phases as SO progresses. At first (j < 10) $y_{1.0}$ is not good (technically), $\gamma'_{j,1.0}$ is near 1.0, and $\xi_{j,1.0}$, the orthogonalization coefficient, is not defined. At j = 10 $y_{1.0}$ becomes good and $\xi_{j,1.0} = y_{1.0}^{(j)*}r'_j/y^*_{1.0}y_{1.0} \doteq y_{1.0}^{(j)*}r'_j$ is computed to be $.21\times10^{-9}$, which is smaller than $\tau\|A\| = .4\times10^{-8}$ as it should be. At j = 11, $\xi_{j,1.0} = .61\times10^{-11}$ which is rather less than $\tau\|A\|$. For all $j > 11$, $|\xi_{j,1.0}| < \epsilon\|A\|$ which is the magnitude of the rounding errors themselves.

This behavior is completely typical of good Ritz vectors. For a given λ, $\xi_{j\lambda}$ is larger than $\epsilon\|A\|$ only in the first two steps at which y_λ is good. This is due to the hereditary nature of good Ritz vectors. We use the following technical result to illuminate this phenomenon.

Theorem 3. For any Ritz vector y_i at step j,

$$y_i^* r_j' = q_j^*(r_j' s_{ji} + F_j s_i + D_j' s_i) - y_i^* q_j(\alpha_j - \theta_i) - y_i^* q_{j-1}\beta_{j-1} - y_i^* f_j \ ,$$

where $D_j' = (d_1, d_2, \ldots, d_{j-1}, 0)$ and $f_j = F_j e_j$ is the last column of F_j.

Proof. (See Section 3 for notation.) Before the orthogonalizations at step j of SO, the computed quantities satisfy

$$(1) \qquad AQ_j - Q_j T_j = r_j' e_j^* + F_j + D_j' \ ,$$

where $D_j' = (d_1, d_2, \ldots, d_{j-1}, 0)$. Multiply equation (1) on the right by e_j to find,

$$(2) \qquad Aq_j - q_j \alpha_j - q_{j-1}\beta_{j-1} = r_j' + f_j \ ,$$

since $D_j' e_j = 0$. Multiply (2) on the left by y_i^* and rearrange to obtain

$$(3) \qquad \begin{aligned} y_i^* r_j' &= y_i^* A q_j - y_i^* q_j \alpha_j - y_i^* q_{j-1}\beta_{j-1} - y_i^* f_j \ , \\ &= q_j^* A y_i - y_i^* q_j \alpha_j - y_i^* q_{j-1}\beta_{j-1} - y_i^* f_j \ , \end{aligned}$$

since A is symmetric. To obtain a formula for Ay_i multiply (1) on the right by s_i to find

$$(4) \qquad AQ_js_i - Q_jT_js_i = r_j'e_j^*s_i + F_js_i + D_j's_i \ .$$

Since $y_i = Q_js_i$, $T_js_i = s_i\theta_i$, and $e_j^*s_i = s_{ji}$, (4) can be rearranged to obtain

$$(5) \qquad Ay_i = y_i\theta_i + r_j's_{ji} + F_js_i + D_j's_i \ .$$

Substituting (5) into (3) and rearranging yields the result. □

We consider the various contributions to $y_i^*r_j'$ given by Theorem 3. (I) The matrix F_j is just the accumulation of the local roundoff errors and $\|F_j\| \doteq \epsilon\|A\|$ will hold. In particular

$$\|q_j^*F_js_i\| \doteq \epsilon\|A\| \ , \quad \text{and}$$
$$\|y_i^*f_j\| \doteq \epsilon\|A\| \ .$$

(II) r_j' is explicitly orthogonalized against q_j by the choice of α_j. Due to rounding errors $q_j^*r_j'$ is not zero but

$$|q_j^*r_j| \doteq \epsilon\|A\|$$

will hold instead.

(III) For $k < j$,

$$d_k = \sum_{t\in G_k} y_t^{(k)}\xi_{kt} \ .$$

Each ξ_{kt} will be less than $\tau\|A\|$ in magnitude while $|q_j^*y_i^{(k)}| \leq \tau$ as well since $k < j$. Therefore

$$\|q_j^*D_j's_i\| \doteq \tau^2\|A\| \ ,$$
$$= \epsilon\|A\| \ ,$$

for $\tau = \sqrt{\epsilon}$. Bringing these observations together,

$$(6) \qquad \begin{aligned} y_i^* r_j' &\le -y_i^* q_j(\alpha_j - \theta_i) - y_i^* q_{j-1}\beta_{j-1} + O(\epsilon\|A\|) , \\ &\doteq -y_i^* q_j(\alpha_j - \theta_i) - y_i^* q_{j-1}\beta_{j-1} . \end{aligned}$$

If $y_\lambda^{(j-2)}$ and $y_\lambda^{(j-1)}$ were good then equation (6) shows that $y_\lambda^{(j)*} r_j'$ will be tiny because both q_j and q_{j-1} were orthogonalized against vectors which are almost parallel to $y_\lambda^{(j)}$.

In the numerical example $\lambda = .5$ was good at step $j = 13$ and the following values for $y_{.5}^{(j)*} r_j'$ were computed.

j	$y_{.5}^{(j)*} r_j'$
13	$.46 \times 10^{-10}$
14	$-.21 \times 10^{-11}$
15	$.39 \times 10^{-18}$
16	$-.37 \times 10^{-18}$
17	$.10 \times 10^{-17}$
18	$-.37 \times 10^{-18}$
19	$.24 \times 10^{-18}$

The other good Ritz vectors behaved similarly. This phenomenon is an important contributant to the success of SO. Each good Ritz vector makes a significant contribution to D only the first two times it appears. Thereafter the corresponding eigenvector is essentially deflated from the system.

3.11 Conclusions

Selective Orthogonalization with $\tau = \sqrt{\epsilon}$ is an effective means of implementing the Lanczos algorithm. If the good Ritz vectors are only computed occasionally (as described in more detail in [Parlett and

Scott]) SO is very efficient as well. SO permits the Lanczos algorithm to be run as originally intended. Since loss of orthogonality is controlled without the expense of full reorthogonalization, there is no need to iterate the algorithm.

As an added bonus SO is capable of finding clustered or multiple eigenvalues without the added complications of using block Lanczos. This is in marked contrast to the simple Lanczos algorithm in exact arithmetic which can compute only one representative of a multiple eigenvalue and also to simple Lanczos in finite precision which computes multiple copies of all eigenvalues regardless of whether they are truly multiple or not.

This phenomenon is due to rounding errors which introduce to q_{j+1} small components in all directions. After one eigendirection of a multiple eigenvalue has been found, components in the orthogonal direction persist after orthogonalization. These components will grow as the algorithm continues until a second eigenvector, orthogonal to the first, is found. For illustration SO was run on the following example with different values of ω.

$$\begin{aligned}\epsilon &= .16\times 10^{-16}\\ n &= 20\\ \lambda_i &= 1/i \quad \text{for} \quad i \neq 2,4\\ \lambda_2 &= \lambda_1 - \omega\\ \lambda_4 &= \lambda_3 - \omega\end{aligned}$$

20 steps were taken and the largest absolute error in the Ritz values was measured.

ω	max error
10^{-1}	$.83\times10^{-16}$
10^{-3}	$.12\times10^{-15}$
10^{-5}	$.11\times10^{-15}$
10^{-7}	$.69\times10^{-16}$
10^{-9}	$.15\times10^{-15}$
10^{-11}	$.83\times10^{-16}$
10^{-13}	$.11\times10^{-15}$
10^{-15}	$.11\times10^{-15}$
10^{-17}	$.38\times10^{-16}$
0	$.11\times10^{-15}$

This shows that SO can resolve clustered eigenvalues to full working accuracy ($n\epsilon\|A\| = .32\times10^{-15}$).

Can large γ's occur for $\tau = \sqrt{\epsilon}$? Heuristically the answer is no. No examples of γ-growth with $\tau = \sqrt{\epsilon}$ are known. Furthermore it is easy to monitor the size of the γ's as the algorithm proceeds. For each good Ritz vector $y_i^{(j)}$, SO computes $y_i^{(j)*}r_j'$ in the orthogonalization process. It is only necessary to compute

$$\gamma_{ji}' = |y_i^{(j)*}r_j'|s_{ji}/\epsilon\|A\|$$

to observe whether γ-growth has occurred.

In conclusion SO is an efficient method of implementing the Lanczos algorithm which points the way towards a subroutine package which could be used off the shelf for large sparse symmetric eigenvalue problems.

Appendix A

To finish the proof of Theorem 3 of Chapter 2, we first quote the main theorem of [Paige 1976] which asserts that for the Lanczos algorithm, the matrices B_j and F_j are tiny, like roundoff in $\|A\|$.

Theorem (Paige). Let A have at most m non-zeros per row. Let $\||A|\| = \nu\|A\|$, where $|A|$ is the matrix with elements $|a_{ij}|$. Let ϵ be the relative machine precision, let $\epsilon_0 = (n+4)\epsilon$, and let $\epsilon_1 = (7+m\nu)\epsilon$. Assume $4j(3\epsilon_0+\epsilon_1) << 1$ and ignore ϵ^2 terms. Then

$$\begin{aligned}
|b_{11}| &\leq 2\epsilon_0\|A\| , \\
|b_{ii}| &\leq 4\epsilon_0\|A\| , \quad \text{for } i = 2,3,\dots,j, \\
|b_{i-1,i}| &\leq 4\epsilon_0\|A\| , \quad \text{for } i = 2,3,\dots,j, \\
\|f_i\| &\leq \epsilon_1\|A\| , \quad \text{for } i = 1,2,\dots,j,
\end{aligned}$$

where f_i is the i^{th} column of F_j.

The proof of Theorem 3 is completed by the following result.

Lemma. For the Lanczos algorithm

$$|\gamma_{ji}| \leq (2j\epsilon_1+8\epsilon_0)/\epsilon .$$

Proof. Recall from Lemma 3 (of Section 2.3) that

$$(1) \qquad \begin{aligned}
\gamma_{ji} &= |s_i^*B_js_i+s_i^*E_js_i|/\epsilon\|A\| \\
&\leq (|s_i^*B_js_i| + |s_i^*E_js_i|)/\epsilon\|A\|
\end{aligned}$$

$$\leq (\|B_j\| + \|E_j\|)/\epsilon\|A\|$$

and it remains to bound each term separately.

Bounding $\|B_j\|$

Let B_j' be the diagonal of B_j and let B_j'' be the superdiagonal of B_j. Since B_j is bidiagonal, $B_j = B_j' + B_j''$ and

(2) $$\begin{aligned}\|B_j\| &= \|B_j' + B_j''\| \\ &\leq \|B_j'\| + \|B_j''\| .\end{aligned}$$

Any matrix C which has only one nonzero element in each row and column satisfies

(3) $$\|C\| = \max_{i,j} |C_{ij}| .$$

Since both B_j' and B_j'' have this property, by using Paige's Theorem,

(4) $$\begin{aligned}\|B_j'\| &= \max_i |b_{ii}| \\ &\leq 4e_0\|A\|\end{aligned}$$

and

(5) $$\begin{aligned}\|B_j''\| &= \max_i |b_{i-1,i}| \\ &\leq 4e_0\|A\| .\end{aligned}$$

Thus by (2)

(6) $$\|B_j\| \leq 8e_0\|A\| .$$

Bounding $\|E_j\|$

Recall that E_j is the upper triangular part of $F_j^*Q_j - Q_j^*F_j$. Therefore

$$(7) \qquad \|E_j\| \leq 2\|Q_j\|\|F_j\| \quad .$$

For any matrix $G_j = (g_1, g_2, \ldots, g_j)$, $\|G_j\| \leq \sqrt{j} \max_i \|g_i\|$ and so

$$(8) \qquad \begin{aligned} \|F_j\| &\leq \sqrt{j} \max_i \|f_i\| \quad , \\ &\leq \sqrt{j}\, e_1 \|A\| \quad , \end{aligned}$$

by Paige's Theorem and

$$(9) \qquad \begin{aligned} \|Q_j\| &\leq \sqrt{j} \max_i \|q_i\| \\ &\leq \sqrt{j} \ . \end{aligned}$$

Hence from (7),

$$(10) \qquad \|E_j\| \leq 2j\epsilon_1\|A\|$$

and the Lemma follows from (1), (6), and (10). □

The bound given in the Lemma for γ_{ji} is likely to be a large overbound for several reasons. For large problems, the greatest over estimate is concealed in ϵ_0 which is a bound on the maximum error committed in normalizing an n-vector. Only for specially chosen vectors will the factor of n be realistic. Also $\|Q_j\| \doteq \sqrt{k+1}$, where k is the maximum number of copies of any one eigenvalue to have appeared, is much more realistic than even $\sqrt{j}$.

In fact, in Example 1 of Section 6 in Chapter 2 the average value of γ_{ji} is about .03. In all examples we have investigated the dependence of γ_{ji} on both n and j is much less than linear. This explains the computational success of the algorithm even when both n and j are large, see [Lewis 1977] for example.

Appendix B

In this appendix we establish a bound on $\|1-Q_j^*Q_j\|$ in terms of the parameter τ.

Theorem. Let $\tau > 0$ be such that

(1) $|1-q_i^*q_i| \leq \tau$ for all $i \leq j$.

(2) $|y_i^{(k)*}q_{k+1}| \leq \tau$ for all $k < j$ and $i \leq k$.

Then for all $k \leq j$

$$\|1-Q_k^*Q_k\| \leq k\tau \quad .$$

Remarks. SO is designed to insure that hypothesis (2) holds for any given τ. In practice we set $\tau = \sqrt{\epsilon}$ and hypothesis (1) will be satisfied easily. Before proving the Theorem we first prove two lemmas. Lemma 1 first appeared in [Kahan and Parlett 1974].

Lemma 1 (Kahan and Parlett). Let $|1-q_j^*q_j| \leq \kappa_1$, let $\|Q_{j-1}^*q_j\| \leq \zeta_{j-1}$, let $\|1-Q_{j-1}^*Q_{j-1}\| \leq \kappa_{j-1}$, and let

$$\kappa_j = (\kappa_1 + \kappa_{j-1} + \sqrt{(\kappa_{j-1}-\kappa_1)^2+4\zeta_{j-1}^2})/2 \ .$$

Then

$$\|1-Q_j^*Q_j\| \leq \kappa_j \ .$$

Proof.

$$\|1-Q_j^*Q_j\| = \left\| \begin{matrix} 1-Q_{j-1}^*Q_{j-1} & -Q_{j-1}^*q_j \\ -q_j^*Q_{j-1} & 1-q_j^*q_j \end{matrix} \right\|$$

$$\leq \left\| \begin{matrix} \|1-Q_{j-1}^{*}Q_{j-1}\| & \|Q_{j-1}^{*}q_j\| \\ \|Q_{j-1}^{*}q_j\| & \|1-q_j^{*}q_j\| \end{matrix} \right\|$$

$$\leq \left\| \begin{matrix} \kappa_{j-1} & \zeta_{j-1} \\ \zeta_{j-1} & \kappa_1 \end{matrix} \right\|$$

$$= \left(\kappa_1 + \kappa_{j-1} + \sqrt{(\kappa_{j-1}-\kappa_1)^2+4\zeta_{j-1}^2}\right)/2 \ . \qquad \square$$

To apply Lemma 1 it is necessary to have a value for ζ_{j-1} in terms of τ.

Lemma 2. Let $|y_i^{(k)*}q_{k+1}| \leq \tau$, for all $i \leq k$, and let $\zeta_k = \sqrt{k}\,\tau$. Then

$$\|Q_k^{*}q_{k+1}\| \leq \zeta_k \ .$$

Proof. Since S_j is an orthogonal matrix,

$$\begin{aligned} \|Q_k^{*}q_{k+1}\|^2 &= \|S_k^{*}Q_k^{*}q_{k+1}\|^2 \ , \\ &= \|Y_k^{*}q_{k+1}\|^2 \ , \\ &= \sum_{i=1}^{k} (y_i^{(k)*}q_{k+1})^2 \\ &\leq \sum_{i=1}^{k} \tau^2 \\ &= k\tau^2 \ . \end{aligned} \qquad \square$$

Proof of Theorem. Induction on k. If $k = 1$ then

$$\begin{aligned} \|1-Q_1^{*}Q_1\| &= \|1-q_1^{*}q_1\| \ , \\ &\leq \tau \end{aligned}$$

by hypothesis (1) so the Theorem is true. Assume that the Theorem is true for all $t \leq k < j$. By Lemma 1, $\|1-Q_{k+1}^*Q_{k+1}\| \leq \kappa_{k+1}$ where

$$(1) \qquad \kappa_{k+1} = [\kappa_1 + \kappa_k + \sqrt{(\kappa_k-\kappa_1)^2+4\zeta_k^2}]/2 \ .$$

By hypothesis (1), $\kappa_1 \leq \tau$, by induction, $\kappa_k \leq k\tau$, and by Lemma 2 $\zeta_k \leq \sqrt{k}\ \tau$.

Combining these inequalities in (1) we obtain

$$\begin{aligned} \kappa_{k+1} &\leq [\tau + k\tau + \sqrt{(k\tau-\tau)^2+4k\tau^2}]/2 \\ &= [(k+1)\tau + \sqrt{(k-1)^2\tau^2+4k\tau^2}]/2 \\ &= [(k+1)\tau + (k+1)\tau]/2 \\ &= (k+1)\tau \quad . \end{aligned}$$

□

References

D. Boley and G.H. Golub, "Inverse Eigenvalue Problems for Band Matrices," Technical Report STAN-CS-77-623, Computer Science Department, Stanford University (1977).

J. Cullum and W.E. Donath, "A Block Generalization of the Symmetric s-step Lanczos Algorithm," Report #RC 4845 (#21570), IBM Thomas J. Watson Research Center, Yorktown Heights, New York (1974).

C. Davis and W. Kahan, "The Rotation of Eigenvectors by a Perturbation --III," SIAM Journal of Numerical Analysis 7, 1 (March 1970).

C. de Boor and G.H. Golub, "The Numerically Stable Reconstruction of a Jacobi Matrix from Spectral Data," to appear in Linear Algebra and Its Applications.

W. Kahan, "Inclusion Theorems for Clusters of Eigenvalues of Hermitian Matrices," Computer Science Report, University of Toronto, Canada (1967).

W. Kahan and B. Parlett, "An Analysis of Lanczos Algorithms for Symmetric Matrices," Electronics Research Memorandum ERL-M467, University of California (September 1974).

W. Kahan and B. Parlett, "How Far Should You Go with the Lanczos Algorithm?," in Sparse Matrix Computations (eds. J. Bunch and D. Rose), Academic Press, New York, 1976.

S. Kaniel, "Estimates for Some Computational Techniques in Linear Algebra," Mathematics of Computation 20, 95 (July 1966), 369-378.

C. Lanczos, "An Iteration Method for the Solution of the Eigenvalue Problem of Linear Differential and Integral Operators," J. Res. Nat. Bur. Stand. 45 (1950), 255-282.

N.J. Lehmann, "Optimale Eigenwerteinschließungen," Numerische Mathematik 5 (September 1963), 246-272.

J. Lewis, "Algorithms for Sparse Matrix Eigenvalue Problems," Technical Report STAN-CS-77-595, Computer Science Department, Stanford University (1977).

C.C. Paige, "The Computation of Eigenvalues and Eigenvectors of Very Large Sparse Matrices," Ph.D. Thesis, University of London (1971).

C.C. Paige, "Computational Variants of the Lanczos Method for the Eigenproblem," J. Inst. Math. Applics. 10 (1972), 373-381.

C.C. Paige, "Error Analysis of the Lanczos Algorithm for Tridiagonalizing a Symmetric Matrix," J. Inst. Math. Applics. 18 (1976), 341-349.

R.C. Thompson and P. McEnteggert, "Principal Submatrices II: The Upper and Lower Quadratic Inequalities," *Linear Algebra and Its Applications 1* (1968), 211-243.

R. Underwood, "An Iterative Block Lanczos Method for the Solution of Large Sparse Symmetric Eigenproblems," Ph.D. Thesis, Stanford University, STAN-CS-75-496 (1975).

J.H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.