MATHEMATICS OF COMPUTATION
VOLUME 62, NUMBER 205
JANUARY 1994, PAGES 209–226

ERROR ANALYSIS OF THE LANCZOS ALGORITHM FOR THE NONSYMMETRIC EIGENVALUE PROBLEM

ZHAOJUN BAI

ABSTRACT. This paper presents an error analysis of the Lanczos algorithm in finite-precision arithmetic for solving the standard nonsymmetric eigenvalue problem, if no breakdown occurs. An analog of Paige's theory on the relationship between the loss of orthogonality among the Lanczos vectors and the convergence of Ritz values in the symmetric Lanczos algorithm is discussed. The theory developed illustrates that in the nonsymmetric Lanczos scheme, if Ritz values are well conditioned, then the loss of biorthogonality among the computed Lanczos vectors implies the convergence of a group of Ritz triplets in terms of small residuals. Numerical experimental results confirm this observation.

1. INTRODUCTION

This paper is concerned with an error analysis of the Lanczos algorithm for solving the nonsymmetric eigenvalue problem of a given real $n \times n$ matrix A:

$$Ax = \lambda x, \qquad y^H A = \lambda y^H,$$

where the unknown scalar λ is called an eigenvalue of A, and the unknown nonzero vectors x and y are called the right and left eigenvectors of A, respectively. The triplet (λ, x, y) is called eigentriplet of A. In the applications of interest, the matrix A is usually large and sparse, and only a few eigenvalues and eigenvectors of A are wanted. In [2], a collection of such matrices is presented describing their origins in problems of applied sciences and engineering.

The Lanczos algorithm, proposed by Cornelius Lanczos in 1950 [19], is a procedure for successive reduction of a given general matrix to a nonsymmetric tridiagonal matrix. The eigenvalue problem for the latter matrix is then solved. The remarkable feature in practice is that in this procedure a few eigenvalues of A (often the largest ones in algebraic magnitude) appear as the eigenvalues of a smaller reduced tridiagonal matrix. The scheme references the matrix A only

Received by the editor April 7, 1992 and, in revised form, December 3, 1992 and January 12, 1993.

1991 *Mathematics Subject Classification.* Primary 65F15, 65F10.

Key words and phrases. Nonsymmetric matrices, eigenvalue problem, error analysis, Lanczos method.

This work was completed while the author was a visitor at the Institute for Mathematics and its Applications, University of Minnesota. This work was supported in part by NSF grant ASC-9102963 and by the Applied and Computational Mathematics Program, Defense Advanced Research Projects Agency, under contract DM28E04120.

0025-5718/94 $1.00 + $.25 per page

through the matrix-vector products Ax and $A^T x$; hence the structure of the matrix is maintained, which renders the scheme particularly useful for finding a few eigenvalues of a very large and sparse problem.

In the 1970s and 80s, great progress has been made on the Lanczos algorithm for solving a large linear system of equations with symmetric coefficient matrix and the symmetric eigenvalue problem. Paige [20] was the first to give an error analysis of the Lanczos algorithm in finite-precision arithmetic. Later, Parlett, Scott, Grcar, Simon, Greenbaum, Strakos, and many others [23, 11, 30, 15, 37] presented further analyses of the Lanczos scheme and its variants. These analyses conclude that the loss of orthogonality among the computed Lanczos vectors is not necessarily a calamity, since it accompanies the convergence of a group of Ritz values to the eigenvalues of the original matrix. In [8], the standard Lanczos algorithm is extended to solve the symmetric generalized eigenvalue problem $Ax = \lambda Bx$. Today, the Lanczos algorithm is regarded as the most powerful tool for finding a few eigenvalues of a large symmetric eigenvalue problem. Software, developed by Parlett and Scott [23] and Cullum and Willoughby [4], can be accessed via netlib, a software distribution system.

In recent years, there has been considerable interest in the Lanczos algorithm for solving linear systems of equations with nonsymmetric coefficient matrix and the nonsymmetric eigenvalue problem. Parlett, Taylor, and Liu [26], Freund, Gutknecht, and Nachtigal [9] have proposed robust schemes for overcoming possible failure (called *breakdown*), or huge intermediate quantities (called *instability*) in the nonsymmetric Lanczos procedure. A theoretical investigation of the possible breakdown and instability of the nonsymmetric Lanczos procedure is made by Gragg [10], Parlett [27], Gutknecht [16], and Boley et al. [3].

Compared to the existing sophisticated error analysis of the Lanczos algorithm for the symmetric eigenvalue problem, much less progress has been made on error analysis of the nonsymmetric Lanczos algorithm. In this paper, we give an error analysis for the simple nonsymmetric Lanczos algorithm and study the effects of finite-precision arithmetic. In the spirit of Paige's floating-point error analysis for the symmetric Lanczos algorithm [20], based on the rounding error model of the basic sparse linear algebra operations, such as saxpy, inner product, and matrix-vector multiplication, we present a set of matrix equations which govern all computed quantities of the simple nonsymmetric Lanczos algorithm in finite-precision arithmetic. An analogy of Paige's theory on the relationship between the loss of orthogonality among the computed Lanczos vectors and the convergence of a Ritz value for the symmetric eigenvalue problem is also discussed in this paper. We conclude that if Ritz values are well conditioned, then the loss of biorthogonality among the computed Lanczos vectors implies the convergence of a group of Ritz triplets in terms of small residuals. The error analysis results developed in this paper also provide insight into the need for robustness schemes, such as look-ahead strategies [26, 9], to avoid potential breakdown and instability in the nonsymmetric Lanczos algorithm.

Other competitive numerical techniques for solving large nonsymmetric eigenvalue problems are the subspace iteration method [35, 36, 6, 7] and Arnoldi's method [31, 32, 28, 34]. The reader is referred to [33] for a more complete and elegant treatment of all these methods.

Throughout this paper we shall use the notational conventions in [14]. Specifically, matrices are denoted by upper-case italic and Greek letters, vectors by

lower-case italic letters, and scalars by lower-case Greek letters or lower-case italic if there is no confusion. The (i, j) entry of a matrix A is denoted by a_{ij}. The symbol $\mathbb{R}$ denotes the set of real numbers, $\mathbb{R}^n$ the set of real n-vectors, and $\mathbb{R}^{m\times n}$ the set of real $m \times n$ matrices. The matrix A^T is the transpose of A. By $|A|$ we denote the matrix $|A| = (|a_{ij}|)$, and $|A| \leq |B|$ means $|a_{ij}| \leq |b_{ij}|$ for any i, j. By $\|\cdot\|_2$ and $\|\cdot\|_F$ we denote the 2-norm and Frobenius norm, respectively, of a vector or matrix.

The rest of this paper is organized as follows. Section 2 recalls the nonsymmetric Lanczos scheme and reviews its properties. Section 3 presents a rounding error analysis of the Lanczos scheme in finite-precision arithmetic. Section 4 discusses the effects of rounding errors and the loss of biorthogonality in the Lanczos algorithm. Section 5 gives some numerical results to support the theoretical analysis of the previous sections.

2. Lanczos algorithm and its properties in exact arithmetic

In this section, we recall the standard nonsymmetric Lanczos scheme for the reduction of a general matrix to tridiagonal form and review some of its important properties in connection with the nonsymmetric eigenvalue problem. This sets up a framework for the following discussion on the behavior of the Lanczos scheme in finite-precision arithmetic.

Given any two starting vectors $u_1, v_1 \in \mathbb{R}^n$ such that $\omega_1 = u_1^T v_1 \neq 0$, the standard nonsymmetric Lanczos algorithm can be viewed as biorthonormalizing, via a two-sided Gram-Schmidt procedure, the two Krylov sequences

$$\mathscr{K}_j(u_1, A) = \{u_1, Au_1, A^2u_1, \ldots, A^{j-1}u_1\},$$
$$\mathscr{K}_j(v_1, A^T) = \{v_1, A^Tv_1, (A^T)^2v_1, \ldots, (A^T)^{j-1}v_1\}.$$

Specifically, the algorithm can be described as follows, where $\operatorname{sign}(\omega)$ denotes the sign of ω.

Lanczos algorithm.

1. Choose two starting vectors u_1, v_1 such that $\omega_1 = u_1^T v_1 \neq 0$. Define
   $\beta_1 = \sqrt{|\omega_1|}$;
   $\gamma_1 = \operatorname{sign}(\omega_1)\beta_1$;
   $q_1 = u_1/\beta_1$;
   $p_1 = v_1/\gamma_1$;
2. for $j = 1, 2, \ldots,$ do
   $\alpha_j = p_j^T A q_j$;
   $u_j = Aq_j - \alpha_j q_j - \gamma_j q_{j-1}$;
   $v_j = A^T p_j - \alpha_j p_j - \beta_j p_{j-1}$;
   $\omega_j = u_j^T v_j$;
   $\beta_{j+1} = \sqrt{|\omega_j|}$;
   $\gamma_{j+1} = \operatorname{sign}(\omega_j)\beta_{j+1}$;
   $q_{j+1} = u_j/\beta_{j+1}$;
   $p_{j+1} = v_j/\gamma_{j+1}$.

One pass through loop 2 is called a *Lanczos step*. The two sequences of vectors $\{q_i\}$ and $\{p_i\}$ are called *Lanczos vectors*. In matrix notation, in the jth step, assuming that $\omega_j \neq 0$, the Lanczos algorithm generates two $n \times j$ matrices Q_j and P_j,

$$Q_j = (q_1, q_2, \ldots, q_j), \qquad P_j = (p_1, p_2, \ldots, p_j),$$

which satisfy

$$P_j^T Q_j = I_j \tag{2.1}$$

and

$$AQ_j = Q_j T_j + \beta_{j+1} q_{j+1} e_j^T, \tag{2.2}$$

$$A^T P_j = P_j T_j^T + \gamma_{j+1} p_{j+1} e_j^T, \tag{2.3}$$

where $e_j = (0, 0, \ldots, 0, 1)^T \in \mathbb{R}^j$ and T_j is the tridiagonal matrix

$$T_j = \begin{pmatrix} \alpha_1 & \gamma_2 & & \\ \beta_2 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \gamma_j \\ & & \beta_j & \alpha_j \end{pmatrix}, \qquad \beta_i = \pm\gamma_i.$$

Relation (2.1) is called the *biorthonormality condition* for the Lanczos vectors. In exact arithmetic, the above procedure must stop at the nth step with $\omega_{n+1} = 0$. However, it may terminate early whenever $\omega_i = 0$. This is the so-called *breakdown* of the procedure, which has been discussed extensively; see, for example, [38, 27, 9]. In this paper, we assume that breakdown will not occur during the procedure.

We note that if A is a symmetric matrix, then the above Lanczos algorithm with the same starting vectors generates $Q_j = P_j$ and a symmetric tridiagonal matrix T_j. Therefore, when A is symmetric, all the results we shall present in this paper reduce to those obtained by Paige [20, 21] for the symmetric Lanczos algorithm.

We also note that there are infinitely many ways of choosing the scalars β_{j+1} and γ_{j+1} in the Lanczos algorithm, as long as they satisfy the equality

$$\omega_j = \beta_{j+1}\gamma_{j+1}.$$

For example, in [5], the choice $\beta_{j+1} = \gamma_{j+1} = \sqrt{\omega_j}$ is made, which may lead to a complex symmetric tridiagonal matrix T_j. In [9], β_{j+1} and γ_{j+1} are chosen so that the condition $(p_i, q_i) = 1$ for $i = 1, \ldots, j$, is replaced by $\|q_i\|_2 = \|p_i\|_2 = 1$ for all i. There are certain tradeoffs among these choices. We will not go into the details of these choices.

Let us examine the eigenvalue problem of the $j \times j$ tridiagonal matrix T_j:

$$T_j z_i = z_i \theta_i, \tag{2.4}$$

$$w_i^H T_j = \theta_i w_i^H, \tag{2.5}$$

for $i = 1, \ldots, j$, where z_i and w_i are normalized so that $w_i^H z_i = 1$. We define the *Ritz triplets* (θ_i, x_i, y_i) for $i = 1, \ldots, j$ by

$$(\theta_i, x_i, y_i) \equiv (\theta_i, Q_j z_i, P_j w_i),$$

where for ease of notation, the index of the Ritz triplets corresponding to the Lanczos step j is omitted. If we consider a Ritz triplet (θ_i, x_i, y_i) as an approximate eigentriplet of the large matrix A, and let r_i and s_i define the corresponding residual vectors of the right and left Ritz vectors, respectively, then we have for $i = 1, \ldots, j$, using (2.2) and (2.3),

$$r_i = Ax_i - x_i\theta_i = \beta_{j+1}(e_j^T z_i)q_{j+1} \equiv \beta_{ji}q_{j+1}, \tag{2.6}$$

$$s_i^H = y_i^H A - \theta_i y_i^H = \gamma_{j+1}(w_i^H e_j)p_{j+1}^T \equiv \gamma_{ji}p_{j+1}^T. \tag{2.7}$$

Moreover, from the biorthogonality property (2.1), we know that the Ritz vectors x_i and y_i satisfy

$$p_{j+1}^T x_i = 0, \tag{2.8}$$

$$y_i^H q_{j+1} = 0. \tag{2.9}$$

Here is another way to describe the biorthogonality of the Lanczos vectors q_i and p_i. From the biorthogonality condition, we have the following equalities, which measure the backward error for the Ritz triplet (θ_i, x_i, y_i):

$$(A - E_i)x_i = \theta_i x_i, \tag{2.10}$$

$$y_i^H(A - E_i) = \theta_i y_i^H, \tag{2.11}$$

where the backward error matrix E_i is

$$E_i = \frac{r_i x_i^H}{\|x_i\|_2^2} + \frac{y_i s_i^H}{\|y_i\|_2^2}.$$

It is easy to show that the Frobenius norm of E_i is

$$\|E_i\|_F^2 = |\beta_{ji}|^2 \frac{\|q_{j+1}\|_2^2}{\|x_i\|_2^2} + |\gamma_{ji}|^2 \frac{\|p_{j+1}\|_2^2}{\|y_i\|_2^2}. \tag{2.12}$$

In [18], it has been shown that the E_i is a perturbation of A satisfying (2.10) and (2.11) with minimal Frobenius norm. If we are interested in the perturbation E of A satisfying (2.10) and (2.11) with minimal 2-norm, it is also shown in [18] that

$$\min_E \|E\|_2 = \max\left\{\frac{\|r_i\|_2}{\|x_i\|_2}, \frac{\|s_i\|_2}{\|y_i\|_2}\right\} = \max\left\{|\beta_{ji}|\frac{\|q_{j+1}\|_2}{\|x_i\|_2}, |\gamma_{ji}|\frac{\|p_{j+1}\|_2}{\|y_i\|_2}\right\}.$$

If $\|E_i\|$ is sufficiently small, then (2.10) and (2.11) tell us that the Ritz triplet (θ_i, x_i, y_i) is the exact eigentriplet of a slightly perturbed matrix of the original matrix A. For measuring the absolute accuracy of the Ritz value θ_i to some simple eigenvalue λ of A, it is well known (see, for example, [38]) that when $\|E_i\|$ is sufficiently small, we have, up to first order,

$$|\lambda - \theta_i| \lesssim \mathrm{cond}(\lambda)\|E_i\|,$$

where $\mathrm{cond}(\lambda) = \|x\|_2\|y\|_2$ is the condition number of the eigenvalue λ, with x and y the right and left eigenvectors corresponding to λ. The vectors x and y are normalized so that $y^H x = 1$. Obviously, we cannot estimate $\mathrm{cond}(\lambda)$ without knowing x and y. In practice, we may replace this unknown condition number by the computable approximate condition number

$$\mathrm{cond}(\theta_i) = \|Q_j\|_F\|P_j\|_F\|z_i\|_2\|w_i\|_2. \tag{2.13}$$

The quantity $\mathrm{cond}(\theta_i)$ is therefore called the condition number of the Ritz value θ_i. The quantities $\|Q_j\|_F^2 = \sum_{i=1}^{j} \|q_i\|_2^2$ and $\|P_j\|_F^2 = \sum_{i=1}^{j} \|p_i\|_2^2$ can be accumulated during execution of the Lanczos steps. Consequently, $\|E_i\|_F$ and $\mathrm{cond}(\theta_i)$ can be used as stopping criteria for the Lanczos procedure. We should note that the above discussion is under the assumption of the biorthogonality of the Lanczos vectors. This turns out to be much more involved in the presence of roundoff error; see [18, 5] for more details.

3. Lanczos algorithm in finite-precision arithmetic

In this section, we present a rounding error analysis of the nonsymmetric Lanczos algorithm in finite-precision arithmetic. Our analysis is in the same spirit as Paige's one for the symmetric Lanczos algorithm [20], except that we carry out the analysis componentwise rather than normwise.

We use the usual model of floating-point arithmetic:

$$\mathrm{fl}(x \circ y) = (x \circ y)(1 + \tau),$$

barring overflow and underflow, where $\circ$ is one of the basic operations $\{+, -, \times, \div, \sqrt{}\}$ and $|\tau| \le \varepsilon_M$, where ε_M is the machine precision. A quantity with a hat (like $\hat{\alpha}$) denotes the computed quantity. With this floating-point arithmetic model, it is well known [14, pp. 63–67] that the rounding error for some basic linear algebra operations of sparse vectors and/or matrices can be expressed as follows:

Saxpy operation:

$$\mathrm{fl}(\alpha x + y) = \alpha x + y + e, \qquad |e| \le \varepsilon_M(2|\alpha x| + |y|) + O(\varepsilon_M^2).$$

Inner product:

$$\mathrm{fl}(x^T y) = x^T y + e, \qquad |e| \le k\varepsilon_M |x|^T |y| + O(\varepsilon_M^2),$$

where k is the number of overlapping nonzero components in vectors x and y.

Matrix-vector multiplication:

$$\mathrm{fl}(Ax) = Ax + e, \qquad |e| \le m\varepsilon_M |A|\,|x| + O(\varepsilon_M^2),$$

where m is the maximal number of nonzero elements of the matrix A in any row.

We are now in a position to present a full rounding error analysis of the nonsymmetric Lanczos procedure. We examine one Lanczos step to see the effects of the finite-precision arithmetic in the algorithm. At the jth Lanczos step, suppose that the quantities $\hat{\beta}_j$, $\hat{\gamma}_j$, $\hat{q}_{j-1}$, $\hat{q}_j$, $\hat{p}_{j-1}$, and $\hat{p}_j$ are computed; we want to compute scalars $\hat{\alpha}_j$, $\hat{\beta}_{j+1}$, and $\hat{\gamma}_{j+1}$, and Lanczos vectors $\hat{q}_{j+1}$ and $\hat{p}_{j+1}$.

We first need to compute $\alpha_j = p_j^T A q_j$ in the Lanczos algorithm. Let A have at most m nonzero entries in any row or column; then for matrix-vector multiplication Aq_j, we have

$$\hat{s}_1 = \mathrm{fl}(A\hat{q}_j) = A\hat{q}_j + \delta\hat{s}_1, \tag{3.1}$$

where

$$|\delta\hat{s}_1| \le m\varepsilon_M |A|\,|\hat{q}_j| + O(\varepsilon_M^2).$$

Then α_j is computed by an inner product,

$$\hat{\alpha}_j = \mathrm{fl}(\hat{p}_j^T \hat{s}_1) = \hat{p}_j^T \hat{s}_1 + \delta\hat{\alpha}_j,$$

with

$$|\delta\hat{\alpha}_j| \le n\varepsilon_M |\hat{p}_j|^T |\hat{s}_1| + O(\varepsilon_M^2).$$

By (3.1) and two saxpy operations, the computed vector $\hat{u}_j$ of $u_j = Aq_j - \alpha_j q_j - \gamma_j q_{j-1}$ is obtained as

$$\hat{s}_2 = \mathrm{fl}(\hat{s}_1 - \hat{\alpha}_j \hat{q}_j) = \hat{s}_1 - \hat{\alpha}_j \hat{q}_j + \delta\hat{s}_2,$$
$$\hat{u}_j = \mathrm{fl}(\hat{s}_2 - \hat{\gamma}_j \hat{q}_{j-1}) = \hat{s}_2 - \hat{\gamma}_j \hat{q}_{j-1} + \delta t_1,$$

where the roundoff errors $\delta\hat{s}_2$ and δt_1 are bounded as follows:

$$|\delta\hat{s}_2| \le \varepsilon_M (2|\hat{\alpha}_j \hat{q}_j| + |\hat{s}_1|) + O(\varepsilon_M^2),$$
$$|\delta t_1| \le \varepsilon_M (2|\hat{\gamma}_j \hat{q}_{j-1}| + |\hat{s}_2|) + O(\varepsilon_M^2).$$

Thus, overall we have

$$\hat{u}_j = A\hat{q}_j - \hat{\alpha}_j \hat{q}_j - \hat{\gamma}_j \hat{q}_{j-1} + \delta\hat{u}_j, \tag{3.2}$$

where

$$\begin{aligned} |\delta\hat{u}_j| &\le |\delta\hat{s}_1| + |\delta\hat{s}_2| + |\delta t_1| \\ &\le m\varepsilon_M |A|\,|\hat{q}_j| + 2\varepsilon_M |\hat{\alpha}_j|\,|\hat{q}_j| + 2\varepsilon_M |\hat{\gamma}_j|\,|\hat{q}_{j-1}| + \varepsilon_M |\hat{s}_1| + \varepsilon_M |\hat{s}_2| + O(\varepsilon_M^2) \\ &\le (2+m)\varepsilon_M |A|\,|\hat{q}_j| + 3\varepsilon_M |\hat{\alpha}_j|\,|\hat{q}_j| + 2\varepsilon_M |\hat{\gamma}_j|\,|\hat{q}_{j-1}| + O(\varepsilon_M^2). \end{aligned}$$

The analysis of the computation of $v_j = A^T p_j - \alpha_j p_j - \gamma_j p_{j-1}$ is entirely analogous. We get

$$\hat{v}_j = A^T \hat{p}_j - \hat{\alpha}_j \hat{p}_j - \hat{\beta}_j \hat{p}_{j-1} + \delta\hat{v}_j,$$

where

$$|\delta\hat{v}_j| \le (2+m)\varepsilon_M |A|^T |\hat{p}_j| + 3\varepsilon_M |\hat{\alpha}_j|\,|\hat{p}_j| + 2\varepsilon_M |\hat{\beta}_j|\,|\hat{p}_{j-1}| + O(\varepsilon_M^2).$$

With $\hat{u}_j$ and $\hat{v}_j$ at hand, the scalars ω_j, β_{j+1}, and γ_{j+1} are computed as

$$\hat{\omega}_j = \mathrm{fl}(\hat{u}_j^T \hat{v}_j) = \hat{u}_j^T \hat{v}_j + \delta\hat{\omega}_j, \tag{3.3}$$

$$\hat{\beta}_{j+1} = \mathrm{fl}(\sqrt{|\hat{\omega}_j|}) = \sqrt{|\hat{\omega}_j|} + \delta\hat{\beta}_{j+1}, \qquad \hat{\gamma}_{j+1} = \mathrm{sign}(\hat{\omega}_j)\hat{\beta}_{j+1}, \tag{3.4}$$

where

$$|\delta\hat{\omega}_j| \le n\varepsilon_M |\hat{u}_j|^T |\hat{v}_j| + O(\varepsilon_M^2),$$
$$|\delta\hat{\beta}_{j+1}| \le \varepsilon_M \sqrt{|\hat{\omega}_j|} \le \varepsilon_M (|\hat{u}_j|^T |\hat{v}_j|)^{1/2} + O(\varepsilon_M^2).$$

Finally, the new Lanczos vectors q_{j+1} and p_{j+1} are computed by

$$\hat{q}_{j+1} = \mathrm{fl}(\hat{u}_j / \hat{\beta}_{j+1}) = \hat{u}_j / \hat{\beta}_{j+1} + \delta\hat{q}_{j+1}, \tag{3.5}$$

and

$$\hat{p}_{j+1} = \mathrm{fl}(\hat{v}_j / \hat{\gamma}_{j+1}) = \hat{v}_j / \hat{\gamma}_{j+1} + \delta\hat{p}_{j+1}, \tag{3.6}$$

where the rounding error vectors $\delta\hat{q}_{j+1}$ and $\delta\hat{p}_{j+1}$ are bounded by

$$|\delta\hat{q}_{j+1}| \le \varepsilon_M |\hat{u}_j / \hat{\beta}_{j+1}| + O(\varepsilon_M^2),$$
$$|\delta\hat{p}_{j+1}| \le \varepsilon_M |\hat{v}_j / \hat{\gamma}_{j+1}| + O(\varepsilon_M^2).$$

From (3.5) and (3.2), we know that the computed $\hat{\alpha}_j$, $\hat{\beta}_{j+1}$ and $\hat{q}_{j+1}$ satisfy

$$\hat{\beta}_{j+1}\hat{q}_{j+1} = A\hat{q}_j - \hat{\alpha}_j\hat{q}_j - \hat{\gamma}_j\hat{q}_{j-1} + f_j, \tag{3.7}$$

where f_j is the sum of roundoff errors in computing the intermediate vector $\hat{u}_j$ and the Lanczos vector $\hat{q}_{j+1}$:

$$f_j = \delta\hat{u}_j + \hat{\beta}_{j+1}\delta\hat{q}_{j+1}.$$

By using bounds for the rounding errors $\delta\hat{u}_j$ and $\delta\hat{q}_{j+1}$, we have

$$\begin{aligned} |f_j| &\le |\delta\hat{u}_j| + |\hat{\beta}_{j+1}\delta\hat{q}_{j+1}| \\ &\le (2+m)\varepsilon_M|A|\,|\hat{q}_j| + 3\varepsilon_M|\hat{\alpha}_j|\,|\hat{q}_j| + 2\varepsilon_M|\hat{\gamma}_j|\,|\hat{q}_{j-1}| + \varepsilon_M|\hat{u}_j| + O(\varepsilon_M^2) \\ &\le (3+m)\varepsilon_M|A|\,|\hat{q}_j| + 4\varepsilon_M|\hat{\alpha}_j|\,|\hat{q}_j| + 3\varepsilon_M|\hat{\gamma}_j|\,|\hat{q}_{j-1}| + O(\varepsilon_M^2). \end{aligned} \tag{3.8}$$

A similar derivation for the computed scalar $\hat{\gamma}_{j+1}$ and the Lanczos vector $\hat{p}_{j+1}$ yields

$$\hat{\gamma}_{j+1}\hat{p}_{j+1} = A^T\hat{p}_j - \hat{\alpha}_j\hat{p}_j - \hat{\beta}_j\hat{p}_{j-1} + g_j, \tag{3.9}$$

where the error vector g_j is bounded by

$$|g_j| \le (3+m)\varepsilon_M|A|\,|\hat{p}_j| + 4\varepsilon_M|\hat{\alpha}_j|\,|\hat{p}_j| + 3\varepsilon_M|\hat{\gamma}_i|\,|\hat{p}_{j-1}| + O(\varepsilon_M^2). \tag{3.10}$$

Summarizing the above discussion and the results of (3.7), (3.8), (3.9), and (3.10), we have the following theorem, which governs all computed quantities.

Theorem 3.1. *Let A be an $n \times n$ real nonsymmetric matrix with at most m nonzero entries in any row or column. Suppose the Lanczos algorithm with starting vectors q_1 and p_1, implemented in floating-point arithmetic with machine precision ε_M, reaches the jth step without breakdown. Let the computed $\hat{\alpha}_i$, $\hat{\beta}_{i+1}$ and $\hat{\gamma}_{i+1}$, $\hat{q}_{i+1}$, $\hat{p}_{i+1}$ for $i = 1, \ldots, j$ satisfy*

$$A\hat{Q}_j = \hat{Q}_j\hat{T}_j + \hat{\beta}_{j+1}\hat{q}_{j+1}e_j^T + F_j, \tag{3.11}$$

$$A^T\hat{P}_j = \hat{P}_j\hat{T}_j^T + \hat{\gamma}_{j+1}\hat{p}_{j+1}e_j^T + G_j, \tag{3.12}$$

where $e_j = (0, 0, \ldots, 0, 1)^T \in \mathbb{R}^j$,

$$\hat{Q}_j = (\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_j), \qquad \hat{P}_j = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_j),$$

$$\hat{T}_j = \begin{pmatrix} \hat{\alpha}_1 & \hat{\gamma}_2 & & \\ \hat{\beta}_2 & \hat{\alpha}_2 & \ddots & \\ & \ddots & \ddots & \hat{\gamma}_j \\ & & \hat{\beta}_j & \hat{\alpha}_j \end{pmatrix}, \qquad \hat{\beta}_i = \pm\hat{\gamma}_i.$$

Then

$$\begin{aligned} |F_j| &\le (3+m)\varepsilon_M|A|\,|\hat{Q}_j| + 4\varepsilon_M|\hat{Q}_j|\,|\hat{T}_j| + O(\varepsilon_M^2), \\ |G_j| &\le (3+m)\varepsilon_M|A|^T|\hat{P}_j| + 4\varepsilon_M|\hat{P}_j|\,|\hat{T}_j|^T + O(\varepsilon_M^2). \end{aligned}$$

In finite-precision arithmetic, we also lose the biorthogonality among the computed Lanczos vectors $\hat{q}_i$ and $\hat{p}_i$. As in the symmetric Lanczos procedure [21, 30], the error, once introduced in some computed Lanczos vectors, is propagated to future steps. Such error propagation can be analyzed by the following corollary, which shows the interesting phenomenon of the loss of biorthonormality among the computed Lanczos vectors.

Corollary 3.1. *Assume that the starting vectors* $\hat{q}_1$ *and* $\hat{p}_1$ *satisfy* $\hat{p}_1^T\hat{q}_1 = 1$. *Then the elements* h_{ik} *of the* $j \times j$ *matrix* $H_j = \hat{P}_j^T\hat{Q}_j = (\hat{p}_i^T\hat{q}_k)$ *satisfy the following equalities. For* $i = 1, 2, \ldots, j$:

$$h_{i+1,i+1} = 1 + \delta h_{i+1,i+1}, \tag{3.13}$$

where

$$|\delta h_{i+1,i+1}| \le (n+4)\varepsilon_M \frac{|\hat{u}_i|^T|\hat{v}_i|}{|\hat{u}_i^T\hat{v}_i|} + O(\varepsilon_M^2),$$

and for $i \neq k$:

$$\hat{\beta}_{k+1}h_{i,k+1} - \hat{\gamma}_{i+1}h_{i+1,k} = (\hat{\alpha}_i - \hat{\alpha}_k)h_{ik} - \hat{\gamma}_k h_{i,k-1} + \hat{\beta}_i h_{i-1,k} + \hat{p}_i^T f_k - g_i^T\hat{q}_k, \tag{3.14}$$

where $h_{0,k} = h_{k,0} = 0$.

Proof. Writing (3.5) and (3.6) for i, we have

$$\begin{aligned} h_{i+1,i+1} = \hat{p}_{i+1}^T\hat{q}_{i+1} &= \left(\frac{\hat{v}_i^T}{\hat{\gamma}_{i+1}} + \delta\hat{p}_{i+1}^T\right)\left(\frac{\hat{u}_i}{\hat{\beta}_{i+1}} + \delta\hat{q}_{i+1}\right) \\ &= \frac{\hat{v}_i^T\hat{u}_i + \hat{\beta}_{i+1}\hat{v}_i^T\delta\hat{q}_{i+1} + \hat{\gamma}_{i+1}\delta\hat{p}_{i+1}^T\hat{u}_i}{\hat{\gamma}_{i+1}\hat{\beta}_{i+1}} + O(\varepsilon_M^2) \\ &= \frac{\hat{v}_i^T\hat{u}_i + \zeta_1}{\hat{\gamma}_{i+1}\hat{\beta}_{i+1}} + O(\varepsilon_M^2), \end{aligned}$$

where

$$|\zeta_1| \le |\hat{\beta}_{i+1}\hat{v}_i^T\delta\hat{q}_{i+1}| + |\hat{\gamma}_{i+1}\delta\hat{p}_{i+1}^T\hat{u}_i| \le 2\varepsilon_M|\hat{v}_i|^T|\hat{u}_i| + O(\varepsilon_M^2).$$

From (3.3) and (3.4), we know that

$$\begin{aligned} \hat{\beta}_{i+1}\hat{\gamma}_{i+1} &= \operatorname{sign}(\hat{\omega}_i)\hat{\beta}_{i+1}^2 = \operatorname{sign}(\hat{\omega}_i)(\sqrt{|\hat{\omega}_i|} + \delta\hat{\beta}_{i+1})^2 \\ &= \hat{v}_i^T\hat{u}_i + \delta\hat{\omega}_i + 2\operatorname{sign}(\hat{\omega}_i)\sqrt{|\hat{\omega}_i|}\delta\hat{\beta}_{i+1} + O(\varepsilon_M^2) \\ &= \hat{v}_i^T\hat{u}_i + \zeta_2 + O(\varepsilon_M^2), \end{aligned}$$

where

$$|\zeta_2| \le |\delta\hat{\omega}_i| + 2\sqrt{|\hat{\omega}_i|}|\delta\hat{\beta}_{i+1}| \le (n+2)\varepsilon_M|\hat{v}_i|^T|\hat{u}_i| + O(\varepsilon_M^2).$$

Hence, the quantity $h_{i+1,i+1}$ can be written

$$h_{i+1,i+1} = \frac{\hat{v}_i^T\hat{u}_i + \zeta_1}{\hat{v}_i^T\hat{u}_i + \zeta_2} + O(\varepsilon_M^2) = 1 + \delta h_{i+1,i+1},$$

where by the bounds of ζ_1 and ζ_2,

$$\begin{aligned} |\delta h_{i+1,i+1}| &\le \frac{|\zeta_1| + |\zeta_2|}{|\hat{v}_i^T\hat{u}_i|} + O(\varepsilon_M^2) \\ &\le (n+4)\varepsilon_M\frac{|\hat{v}_i|^T|\hat{u}_i|}{|\hat{v}_i^T\hat{u}_i|} + O(\varepsilon_M^2). \end{aligned}$$

This gives (3.13).

In order to prove (3.14), writing (3.7) and (3.9) for k and i, we have

$$\hat{\beta}_{k+1}\hat{q}_{k+1} = A\hat{q}_k - \hat{\alpha}_k\hat{q}_k - \hat{\gamma}_k\hat{q}_{k+1} + f_k, \tag{3.15}$$

$$\hat{\gamma}_{i+1}\hat{p}_{i+1} = A^T\hat{p}_i - \hat{\alpha}_i\hat{p}_i - \hat{\beta}_i\hat{p}_{i-1} + g_i. \tag{3.16}$$

The result of (3.14) now comes about from $\hat{p}_i^T \times (3.15) - (3.16)^T \times \hat{q}_k$. □

4. Convergence versus loss of biorthogonality

The effects of finite-precision arithmetic and the loss of orthogonality in the symmetric Lanczos procedure have been studied by many people; see, for example, [23, 30, 15]. Paige was the first to provide an understanding of the effects of the loss of orthogonality among the Lanczos vectors. In [24, 30], it is stated that the loss of orthogonality implies convergence of a Ritz pair to an eigenpair. In this section, we shall discuss the effects of rounding errors on the nonsymmetric Lanczos procedure. We shall show that a conclusion similar to Paige's theory still holds, subject to a certain condition.

From the analysis of §3, we know that at the end of the jth step of the nonsymmetric Lanczos procedure, the computed quantities obey the following three important equalities:

$$A\widehat{Q}_j = \widehat{Q}_j\widehat{T}_j + \hat{\beta}_{j+1}\hat{q}_{j+1}e_j^T + F_j\,, \tag{4.1}$$

$$A^T\widehat{P}_j = \widehat{P}_j\widehat{T}_j^T + \hat{\gamma}_{j+1}\hat{p}_{j+1}e_j^T + G_j\,, \tag{4.2}$$

$$\widehat{P}_j^T\widehat{Q}_j - I_j = C_j + \Delta_j + D_j\,, \tag{4.3}$$

where the rounding error matrices F_j and G_j are bounded as in Theorem 3.1, C_j is a strictly lower triangular matrix, Δ_j a diagonal matrix and D_j a strictly upper triangular matrix.

To simplify our discussion, we make two assumptions, which are also used in the symmetric Lanczos procedure [25, p. 265]. The first assumption is the so-called *local biorthogonality*. It says that the computed Lanczos vectors are biorthogonal to their "neighboring" Lanczos vectors, that is

$$\hat{p}_i^T\hat{q}_{i-1} = 0\,, \quad \hat{p}_{i-1}^T\hat{q}_i = 0 \quad \text{for } i = 2, \ldots, j\,. \tag{4.4}$$

In the matrix notation, local biorthogonality means that the second subdiagonal elements of the strictly lower triangular matrix C_j are zero, and the superdiagonal elements of the strictly upper triangular matrix D_j are also zero.

The second assumption is that the eigenvalue problem for the $j \times j$ tridiagonal matrix $\widehat{T}_j$ is solved exactly, that is,

$$\widehat{T}_j z_i = z_i\theta_i\,, \quad w_i^H\widehat{T}_j = \theta_i w_i^H\,, \quad i = 1, \ldots, j\,. \tag{4.5}$$

With these assumptions, we are now ready to present the next theorem concerning the effects of the loss of biorthogonality. It explains the implication of the failure of the equalities (2.8) and (2.9).

Theorem 4.1. *Assume that the Lanczos algorithm in finite-precision arithmetic satisfies* (4.1) *through* (4.5). *Let*

$$\Delta_j\widehat{T}_j - \widehat{T}_j\Delta_j = K_j - L_j\,,$$
$$\widehat{P}_j^T F_j - G_j^T\widehat{Q}_j = N_j - M_j\,,$$

where K_j and N_j are strictly lower triangular matrices, and L_j and M_j strictly upper triangular matrices. Then the computed Ritz vectors $\hat{x}_i$ $(= \widehat{Q}_j z_i)$ and $\hat{y}_i$

$(= \hat{P}_j w_i)$, *for* $i = 1, \ldots, j$, *satisfy*

$$\hat{p}_{j+1}^T \hat{x}_i = \frac{\phi_{ii}^{(j)}}{\hat{\gamma}_{ji}}, \tag{4.6}$$

$$\hat{y}_i^H \hat{q}_{j+1} = \frac{\psi_{ii}^{(j)}}{\hat{\beta}_{ji}}, \tag{4.7}$$

where

$$\phi_{ii}^{(j)} = w_i^H (K_j + N_j) z_i, \qquad \psi_{ii}^{(j)} = w_i^H (L_j + M_j) z_i,$$
$$\hat{\gamma}_{ji} = \hat{\gamma}_{j+1}(w_i^H e_j), \qquad \hat{\beta}_{ji} = \hat{\beta}_{j+1}(e_j^T z_i).$$

Proof. From $\hat{P}_j^T \times (4.1)$, we have

$$\hat{P}_j^T A \hat{Q}_j = \hat{P}_j^T \hat{Q}_j \hat{T}_j + \hat{\beta}_{j+1} \hat{P}_j^T \hat{q}_{j+1} e_j^T + \hat{P}_j^T F_j. \tag{4.8}$$

On the other hand, by taking the transpose of $\hat{Q}_j^T \times (4.2)$, we have

$$\hat{P}_j^T A \hat{Q}_j = \hat{T}_j \hat{P}_j^T \hat{Q}_j + \hat{\gamma}_{j+1} e_j \hat{p}_{j+1}^T \hat{Q}_j + G_j^T \hat{Q}_j. \tag{4.9}$$

Subtracting (4.9) and (4.8), we get

$$0 = \hat{P}_j^T \hat{Q}_j \hat{T}_j - \hat{T}_j \hat{P}_j^T \hat{Q}_j + \hat{\beta}_{j+1} \hat{P}_j^T \hat{q}_{j+1} e_j^T - \hat{\gamma}_{j+1} e_j \hat{p}_{j+1}^T \hat{Q}_j + \hat{P}_j^T F_j - G_j^T \hat{Q}_j,$$

that is,

$$\begin{aligned} &\hat{\gamma}_{j+1} e_j \hat{p}_{j+1}^T \hat{Q}_j - \hat{\beta}_{j+1} \hat{P}_j^T \hat{q}_{j+1} e_j^T \\ &\quad = (I_j + C_j + \Delta_j + D_j)\hat{T}_j - \hat{T}_j (I_j + C_j + \Delta_j + D_j) + \hat{P}_j^T F_j - G_j^T \hat{Q}_j \\ &\quad = C_j \hat{T}_j - \hat{T}_j C_j + \Delta_j \hat{T}_j - \hat{T}_j \Delta_j + D_j \hat{T}_j - \hat{T}_j D_j + \hat{P}_j^T F_j - G_j^T \hat{Q}_j. \end{aligned} \tag{4.10}$$

By the local biorthogonality assumption (4.4), it is easy to see that $C_j \hat{T}_j - \hat{T}_j C_j$ is a strictly lower triangular matrix, and $D_j \hat{T}_j - \hat{T}_j D_j$ is a strictly upper triangular matrix. Since the diagonal elements of $\Delta_j \hat{T}_j - \hat{T}_j \Delta_j$ are zero, we can write

$$\Delta_j \hat{T}_j - \hat{T}_j \Delta_j = K_j - L_j,$$

where K_j is the strictly lower triangular part of $\Delta_j \hat{T}_j - \hat{T}_j \Delta_j$ and $-L_j$ the strictly upper triangular part of it. Note that the rank-one matrix $e_j \hat{p}_{j+1}^T \hat{Q}_j$ has nonzero entries only from $(j, 1)$ through $(j, j-1)$ in the last row, and $\hat{P}_j^T \hat{q}_{j+1} e_j^T$ has nonzero entries only from $(1, j)$ through $(j-1, j)$ in the last column. From these observations and the equality (4.10), we know that the diagonal elements of $\hat{P}_j^T F_j - G_j^T \hat{Q}_j$ must also be zero. Therefore, we can write

$$\hat{P}_j^T F_j - G_j^T \hat{Q}_j = N_j - M_j,$$

where N_j is the strictly lower triangular part of $\hat{P}_j^T F_j - G_j^T \hat{Q}_j$ and $-M_j$ the strictly upper triangular part. By writing down the strictly lower triangular part and the strictly upper triangular part of (4.10), respectively, we have the following important equalities:

$$\hat{\gamma}_{j+1} e_j \hat{p}_{j+1}^T \hat{Q}_j = C_j \hat{T}_j - \hat{T}_j C_j + K_j + N_j, \tag{4.11}$$

$$-\hat{\beta}_{j+1} \hat{P}_j^T \hat{q}_{j+1} e_j^T = D_j \hat{T}_j - \hat{T}_j D_j - L_j - M_j. \tag{4.12}$$

From $w_i^H \times (4.11) \times z_i$, and the assumption (4.5), we have

$$\begin{aligned}\hat{\gamma}_{j+1}(w_i^H e_j)\hat{p}_{j+1}^T \widehat{Q}_j z_i &= w_i^H C_j \widehat{T}_j z_i - w_i^H \widehat{T}_j C_j z_i + w_i^H (K_j + N_j) z_i \\ &= \theta_i w_i^H C_j z_i - \theta_i w_i^H C_j z_i + w_i^H (K_j + N_j) z_i .\end{aligned}$$

Hence, this gives (4.6). Similarly, by $w_i^H \times (4.12) \times z_i$, we have

$$\begin{aligned}-\hat{\beta}_{j+1} w_i^H \widehat{P}_j^T \hat{q}_{j+1}(e_j^T z_i) &= w_i^H D_j \widehat{T}_j z_i - w_i^H \widehat{T}_j D_j z_i - w_i^H (L_j + M_j) z_i \\ &= \theta_i w_i^H D_j z_i - \theta_i w_i^H D_j z_i - w_i^H (L_j + M_j) z_i .\end{aligned}$$

This gives (4.7), and the theorem is proved. □

Equations (4.6) and (4.7) describe the way in which the biorthogonality is lost. Recall that the scalars $\hat{\beta}_{ji}$ and $\hat{\gamma}_{ji}$ are the essential quantities used as the backward error criteria for the computed Ritz triplet $(\theta_i, \hat{x}_i, \hat{y}_i) = (\theta_i, \widehat{Q}_j z_i, \widehat{P}_j w_i)$. Hence, if the quantities $|\phi_{ii}^{(j)}|$ and $|\psi_{ii}^{(j)}|$ are bounded and bounded away from zero, then (4.6) and (4.7) exactly reflect the reciprocal relation between the convergence of the Lanczos procedure (i.e., tiny $\hat{\beta}_{ji}$ and $\hat{\gamma}_{ji}$) and the loss of biorthogonality (i.e., large $\hat{p}_{j+1}^T \hat{x}_i = \hat{p}_{j+1}^T \widehat{Q}_j z_i$ and $\hat{y}_i^H \hat{q}_{j+1} = w_i^H \widehat{P}_i^T \hat{q}_{j+1}$).

In order to estimate $\phi_{ii}^{(j)}$ and $\psi_{ii}^{(j)}$, let us assume $\Delta_j = 0$, i.e., $\hat{p}_i^T \hat{q}_i = 1$, which simplifies the technical details of the analysis and appears to be the case in practice, up to the order of machine precision. Under this assumption, we have $K_j = L_j = 0$ in Theorem 4.1, and moreover, we have

$$\begin{aligned}\phi_{ii}^{(j)} &= w_i^H N_j z_i = w_i^H \times (\text{strictly lower triangular part of } \widehat{P}_j^T F_j - G_j^T \widehat{Q}_j) \times z_i , \\ \psi_{ii}^{(j)} &= w_i^H M_j z_i = w_i^H \times (\text{strictly upper triangular part of } \widehat{P}_j^T F_j - G_j^T \widehat{Q}_j) \times z_i .\end{aligned}$$

By taking the absolute value on both sides of the above two equations, and using the standard consistency conditions for vector and matrix norms, we have

$$|\phi_{ii}^{(j)}| \le (\|\widehat{P}_j^T\|_F \|F_j\|_F + \|G_j^T\|_F \|\widehat{Q}_j\|_F) \|z_i\|_2 \|w_i\|_2$$

and

$$|\psi_{ii}^{(j)}| \le (\|\widehat{P}_j^T\|_F \|F_j\|_F + \|G_j^T\|_F \|\widehat{Q}_j\|_F) \|z_i\|_2 \|w_i\|_2 .$$

By estimating $\|F_j\|_F$ and $\|G_j^T\|_F$ from Theorem 3.1, we have the following corollary, which gives upper bounds for the quantities $\phi_{ii}^{(j)}$ and $\psi_{ii}^{(j)}$.

Corollary 4.1. *Assume that* $\Delta_j = 0$ *in Theorem* 4.1. *Then* $\phi_{ii}^{(j)}$ *and* $\psi_{ii}^{(j)}$ *satisfy*

$$|\phi_{ii}^{(j)}| \le \varepsilon_M \operatorname{cond}(\theta_i)(2(3+m)\|A\|_F + 8\|\widehat{T}_j\|_F) + O(\varepsilon_M^2), \tag{4.13}$$

$$|\psi_{ii}^{(j)}| \le \varepsilon_M \operatorname{cond}(\theta_i)(2(3+m)\|A\|_F + 8\|\widehat{T}_j\|_F) + O(\varepsilon_M^2), \tag{4.14}$$

where

$$\operatorname{cond}(\theta_i) = \|\widehat{Q}_j\|_F \|\widehat{P}_j\|_F \|z_i\|_2 \|w_i\|_2 .$$

The quantity $\operatorname{cond}(\theta_i)$ *is the condition number of the computed Ritz value* θ.

Observe that in the symmetric Lanczos procedure, $\|\widehat{Q}_j\|_F = \|\widehat{P}_j\|_F$ is bounded by the constant $\sqrt{j}$, and $\|z_i\|_2 = \|w_i\|_2 = 1$, i.e., $\operatorname{cond}(\theta_i) = j$,

and $\|\widehat{T}_j\|_F$ is also bounded; hence $|\phi_{ii}^{(j)}| = |\psi_{ii}^{(j)}| = O(jn\varepsilon_M\|A\|)$, which is just the result obtained by Paige [20, 21] and a key fact to explain Paige's theory [25, 30]. Unfortunately, for the nonsymmetric Lanczos procedure, because of possibly small ω_j (i.e., near breakdown), the Lanczos vectors $\|\hat{q}_i\|_2$ and $\|\hat{p}_i\|_2$ could grow unboundedly. It is suggested to accumulate the quantities $\|\widehat{Q}_j\|_F^2 = \sum_{i=1}^{j} \|\hat{q}_i\|_2^2$ and $\|\widehat{P}_j\|_F^2 = \sum_{i=1}^{j} \|\hat{p}_i\|_2^2$, which only costs about $4jn$ flops. We can thereby obtain a computable bound for $\text{cond}(\theta_i)$ in practice. Theorem 4.1 and Corollary 4.1 say that if the orthogonality between $\hat{p}_{j+1}$ and $\hat{x}_i$ (respectively $\hat{q}_{j+1}$ and $\hat{y}_i$) is lost, then the value $|\hat{\gamma}_{ji}|$ is proportional to $|\phi_{ii}^{(j)}|$ (resp. $|\hat{\beta}_{ji}|$ is proportional to $|\psi_{ii}^{(j)}|$). Given the upper bounds (4.13) and (4.14), and supposing that $\text{cond}(\theta_i)$ is reasonably bounded, the loss of biorthogonality implies that $|\hat{\gamma}_{ji}|$ and $|\hat{\beta}_{ji}|$ are small. Therefore, in the best case we can state that if the effects of finite-precision arithmetic, F_j and G_j in (3.11) and (3.12), are small, then small residuals tell us that the computed eigenvalues are eigenvalues of matrices close to the given matrix. In the next section, we shall verify this claim by numerical examples.

To end this section, we recall that in the nonsymmetric Lanczos algorithm, even without breakdown (i.e., $\omega_i \neq 0$), the procedure is still susceptible to potential instabilities (near breakdown), i.e., at least some ω_i is tiny. Consequently, huge intermediate quantities $\|\hat{q}_i\|_2$ and $\|\hat{p}_i\|_2$ could appear. If this happens, we will have a huge condition number $\text{cond}(\theta_i)$, and the implication of the loss of biorthogonality to the small residuals may no longer hold. The look-ahead Lanczos strategies proposed by Parlett, Taylor, and Liu [26] and Freund, Gutknecht, and Nachtigal [9] provide ways to control the occurrence of potentially huge intermediate quantities by skipping over steps in which a breakdown or instabilities would occur in the standard procedure. An error analysis of these look-ahead Lanczos algorithms has not been given. Further investigations of these schemes is definitely needed.

5. Numerical examples

In this section, we present three numerical examples to see the practical numerical behavior of the convergence of a Ritz value versus the loss of biorthogonality among the Lanczos vectors in the nonsymmetric Lanczos algorithm as discussed in the previous section.

A set of experimental Fortran 77 subroutines have been developed, which return the desired intermediate quantities to allow us to observe the details of numerical behavior of the nonsymmetric Lanczos algorithm in practice. The eigenvalue problem of the resulting nonsymmetric tridiagonal matrix $\widehat{T}_j$ in the Lanczos algorithm is solved by the subroutine DGEEVX, an expert driver routine in LAPACK [1], which allows us not only to compute the eigenvalues, right and left eigenvectors, but also to compute the condition numbers of the eigenvalues and eigenvectors. There is literature [22, 12, 5] on the solution of the eigenvalue problem of a nonsymmetric tridiagonal matrix which takes advantage of the tridiagonal structure.

All numerical experiments are carried out on a HP Apollo 400 workstation with machine accuracy $\varepsilon_M \approx 1.11 \times 10^{-16}$, with underflow and overflow threshold 2.23×10^{-308} and $1.80 \times 10^{+308}$, respectively.

Example 1. This example is from [25], where A is a diagonal matrix

$$A = \text{diag}(0,\ 1 \times 10^{-4},\ 2 \times 10^{-4},\ 3 \times 10^{-4},\ 4 \times 10^{-4},\ 1).$$

The starting vectors are

$$u_1 = (1, 1, 1, 1, 1, 1)^T, \qquad v_1 = (1, 1, 1, 1, 1, -1)^T.$$

The Lanczos procedure generates a sequence of nonsymmetric tridiagonal matrices $\hat{T}_j$ with increasing number of Lanczos steps j. The following table illustrates the convergence of a Ritz value in terms of residuals to the largest eigenvalue $\lambda_{\max} = \lambda_1 = 1.0$ of A, and the loss of biorthogonality among the Lanczos vectors.

j	$\lvert\hat{p}_{j+1}^T \hat{x}_1\rvert$	$\lvert\hat{y}_{j1}\rvert$	$\lvert\hat{y}_1^H \hat{q}_{j+1}\rvert$	$\lvert\hat{\beta}_{j1}\rvert$
2	$0.13 \cdot 10^{-12}$	$0.26 \cdot 10^{-3}$	$0.13 \cdot 10^{-12}$	$0.25 \cdot 10^{-3}$
3	$0.31 \cdot 10^{-7}$	$0.31 \cdot 10^{-7}$	$0.28 \cdot 10^{-8}$	$0.31 \cdot 10^{-7}$
4	$-.24 \cdot 10^{-4}$	$0.31 \cdot 10^{-11}$	$0.24 \cdot 10^{-4}$	$0.31 \cdot 10^{-11}$
5	$0.31 \cdot 10^{0}$	$0.22 \cdot 10^{-15}$	$0.31 \cdot 10^{0}$	$0.22 \cdot 10^{-15}$
6	$0.82 \cdot 10^{0}$	$0.82 \cdot 10^{-16}$	$0.81 \cdot 10^{0}$	$0.82 \cdot 10^{-16}$

We note that in this example the corresponding Ritz value is well conditioned, $\phi_{11}^{(j)} \approx \psi_{11}^{(j)} \approx 10^{-16}$ for all j. As predicted in Theorem 4.1, the loss of biorthogonality accompanies the convergence of a Ritz value to the largest eigenvalue λ_1 in terms of small residuals.

Example 2. The second numerical example is for the Frank matrix:

$$A = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 2 & \cdots & 2 \\ & 2 & 3 & \cdots & 3 \\ & & \ddots & \cdots & \cdot \\ & & & n-1 & n \end{pmatrix}.$$

The Frank matrix has determinant 1. The eigenvalues of the Frank matrix may be obtained in terms of the zeros of Hermite polynomials. They are positive and occur in reciprocal pairs. For more details about the Frank matrix, the reader may refer to [13, 17]. In this experiment, $n = 30$, the largest eigenvalue of A is

$$\lambda_{\max} = \lambda_1 = 0.9620062229328506 \cdot 10^2.$$

We take the starting vectors u_1 and v_1 in the nonsymmetric Lanczos algorithm as random vectors from the normal distribution. The following table illustrates (4.6) and (4.7) in the context of convergence versus loss of biorthogonality between Lanczos vectors.

j	$\lvert\hat{p}_{j+1}^T \hat{x}_1\rvert$	$\lvert\hat{y}_{j1}\rvert$	$\lvert\hat{y}_1^H \hat{q}_{j+1}\rvert$	$\lvert\hat{\beta}_{j1}\rvert$
10	$0.31 \cdot 10^{-10}$	$0.56 \cdot 10^{-2}$	$0.13 \cdot 10^{-10}$	$0.56 \cdot 10^{-2}$
15	$0.15 \cdot 10^{-5}$	$0.16 \cdot 10^{-6}$	$0.51 \cdot 10^{-6}$	$0.16 \cdot 10^{-6}$
20	$0.77 \cdot 10^{0}$	$0.22 \cdot 10^{-12}$	$0.21 \cdot 10^{0}$	$0.22 \cdot 10^{-12}$

At $j = 20$, we have $\|\tilde{Q}_j\|_F \approx 1.81 \times 10^3$, $\|\hat{P}_j\|_F \approx 2.5 \times 10^2$. The observed $\phi_{ii}^{(j)} \approx \psi_{ii}^{(j)} \approx 10^{-12}$. When the Lanczos algorithm is stopped at $j = 20$, the computed largest eigenvalue has the relative accuracy

$$\frac{|\lambda_{\max} - (\text{computed } \lambda_{\max})|}{|\lambda_{\max}|} \approx 4.136 \times 10^{-14}.$$

Example 3. The third example is for a so-called Brusselator matrix, which comes from modeling the concentration waves in reaction and transport interaction of some chemical solutions in a tubular reactor [29]. This test example is also used by Saad in connection with Arnoldi's method [32]. In this model, the concentrations $x(t, z)$ and $y(t, z)$ of two reacting and diffusing components satisfy

$$\frac{\partial x}{\partial t} = \frac{D_x}{L^2}\frac{\partial^2 x}{\partial^2 z} + f(x, y),$$
$$\frac{\partial y}{\partial t} = \frac{D_y}{L^2}\frac{\partial^2 y}{\partial^2 z} + g(x, y),$$

with boundary conditions

$$x(0, z) = x_0(z), \qquad y(0, z) = y_0(z),$$
$$x(0, t) = x(1, t) = x^*, \qquad y(0, t) = y(1, t) = y^*,$$

where $0 \le z \le 1$ is the space coordinate along the tube, t is time, and f and g are chosen as a Brusselator wave model,

$$f(x, y) = \zeta_1 - (\zeta_2 + 1)x + x^2 y, \qquad g(x, y) = \zeta_2 x - x^2 y,$$

with the set of parameters

$$D_x = 0.008, \quad D_y = \tfrac{1}{2}D_x, \quad \zeta_1 = 2, \quad \zeta_2 = 5.45, \quad L = 0.51302.$$

If we discretize the interval $[0, 1]$ using k interior points and mesh size $h = 1/(k+1)$, then the discrete vector is of the form $(x^T, y^T)^T$, where x and y are k-dimensional vectors. If f_h and g_h denote the corresponding discretized functions f and g, then the Jacobian is a 2×2 block matrix in which the diagonal blocks $(1, 1)$ and $(2, 2)$ are the matrices

$$\frac{1}{h^2}\frac{D_x}{L^2}\,\text{Tridiag}\{1, -2, 1\} + \frac{\partial f_h(x, y)}{\partial x}$$

and

$$\frac{1}{h^2}\frac{D_y}{L^2}\,\text{Tridiag}\{1, -2, 1\} + \frac{\partial g_h(x, y)}{\partial y},$$

respectively, while the blocks $(1, 2)$ and $(2, 1)$ of the Jacobian are

$$\frac{\partial f_h(x, y)}{\partial y} \quad \text{and} \quad \frac{\partial g_h(x, y)}{\partial x},$$

respectively. We denote by A the resulting $2k \times 2k$ Jacobian matrix. The exact eigenvalues are known for this problem, since there exists a quadratic relation between the eigenvalues of the matrix A and those of the classical difference matrix $\text{Tridiag}\{1, -2, 1\}$. The order of the Jacobian in this example is 200. The largest eigenvalue of A is then

$$\lambda_{\max} = \lambda_1 = -0.1235506957879173 \cdot 10^4.$$

We take the starting vectors u_1 and v_1 in the nonsymmetric Lanczos algorithm as random vectors from the normal distribution. The following table presents information analogous to that given before.

j	$\lvert\hat{p}_{j+1}^T\hat{x}_1\rvert$	$\lvert\hat{\gamma}_{j1}\rvert$	$\lvert\hat{y}_1^H\hat{q}_{j+1}\rvert$	$\lvert\hat{\beta}_{j1}\rvert$
50	$0.37\cdot 10^{-9}$	$0.33\cdot 10^{0}$	$0.12\cdot 10^{-9}$	$0.33\cdot 10^{0}$
70	$0.47\cdot 10^{-9}$	$0.26\cdot 10^{0}$	$0.44\cdot 10^{-9}$	$0.22\cdot 10^{0}$
90	$0.54\cdot 10^{-8}$	$0.79\cdot 10^{-1}$	$0.67\cdot 10^{-8}$	$0.79\cdot 10^{-1}$
100	$0.29\cdot 10^{-7}$	$0.64\cdot 10^{-2}$	$0.41\cdot 10^{-7}$	$0.64\cdot 10^{-2}$
105	$0.27\cdot 01^{-3}$	$0.61\cdot 10^{-6}$	$0.11\cdot 10^{-3}$	$0.61\cdot 10^{-6}$
110	$0.23\cdot 10^{-1}$	$0.69\cdot 10^{-9}$	$0.28\cdot 10^{0}$	$0.69\cdot 10^{-9}$

From this table, we see that in the first 90 Lanczos steps, with no sign of convergence of Ritz values, the biorthogonality is well preserved. Once the biorthogonality is gradually lost, the Ritz values start converging. In this example, $\|\widehat{Q}_j\|_F \approx \|\widehat{P}_j\|_F \approx 1.5\times 10^3$ at $j = 110$, and the observed $\phi_{ii}^{(j)} \approx \psi_{ii}^{(j)} \approx 5.3\times 10^{-10}$. At $j = 110$ of the Lanczos procedure, the computed largest Ritz value has a relative accuracy comparable to the largest eigenvalue λ_1 of A,

$$\frac{|\lambda_{\max} - (\text{computed } \lambda_{\max})|}{|\lambda_{\max}|} \approx 3.1010\times 10^{-8}.$$

6. Conclusion and future work

In this paper, an error analysis of the nonsymmetric Lanczos algorithm in finite-precision arithmetic is presented. We have seen that for the nonsymmetric Lanczos algorithm without breakdown, if Ritz values are well conditioned, then the loss of biorthogonality among the computed Lanczos vectors implies the convergence of the Ritz values in terms of small residuals. This observation extends the results obtained by Paige for the Lanczos algorithm for the symmetric eigenvalue problem. In the symmetric case, Ritz values are always well conditioned. The results of our error analysis also provide insight into the need for robustness schemes, such as the look-ahead strategies proposed by Parlett, Taylor, and Liu [26] and Freund, Gutknecht, and Nachtigal [9], to avoid the potential breakdown and instability in the nonsymmetric Lanczos procedure.

This is only a first step in the error analysis of the nonsymmetric Lanczos scheme. In future work, we plan to conduct the error analysis of the variants of the nonsymmetric Lanczos algorithm [26, 5, 9], and study the effects of finite-precision arithmetic on the convergence of Ritz triplets.

Acknowledgments

The author would like to acknowledge Jim Demmel, Anne Greenbaum, Nick Higham, and Zdenek Strakos for fruitful discussions on this work, and Nick and Zdenek for their invaluable comments on the manuscript.

BIBLIOGRAPHY

1. E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. Mckenney, S. Ostrouchov, and D. Sorensen, *LAPACK user's guide*, SIAM, Philadelphia, PA, 1992.
2. Z. Bai, *A collection of test matrices for the large sparse nonsymmetric eigenvalue problem*, University of Kentucky, Department of Mathematics, RR-93-03, Aug. 1993.
3. D. Boley, S. Elhay, G. H. Golub, and M. H. Gutknecht, *Nonsymmetric Lanczos and finding orthogonal polynomials associated with indefinite weights*, Numerical Analysis Report NA-90-09, Stanford, Aug. 1990.
4. J. Cullum and R. A. Willoughby, *Lanczos algorithms for large symmetric eigenvalue computations*, Vol. 1, *Theory*, Vol. 2, *Programs*; Birkhäuser, Basel, 1985.
5. ——, *A practical procedure for computing eigenvalues of large sparse nonsymmetric matrices*, Large Scale Eigenvalue Problems (J. Cullum and R. A. Willoughby, eds.), North-Holland, Amsterdam, 1986, pp. 193–240.
6. E. R. Davidson, *Super-matrix methods*, Comput. Phys. Comm. **53** (1989), 49–60.
7. I. S. Duff and J. A. Scott, *Computing selected eigenvalues of sparse unsymmetric matrices using subspace iteration*, RAL-91-056, Rutherford Appleton Laboratory, Oxon, England, 1991.
8. T. Ericsson and A. Ruhe, *Lanczos algorithms and field of value rotations for symmetric matrix pencils*, Linear Algebra Appl. **88/89** (1987), pp. 733–746.
9. R. W. Freund, M. H. Gutknecht, and N. M. Nachtigal, *An implementation of the look-ahead Lanczos algorithm for non-Hermitian matrices*, Part I, Tech. Rep. 90.45, RIACS, NASA Ames Research Center, Nov. 1990.
10. W. B. Gragg, *Matrix interpretations and applications of the continued fraction algorithm*, Rocky Mountain J. Math. **5** (1974), 213–225.
11. J. Grcar, *Analyses of the Lanczos algorithm and of the approximation problem in Richardson's method*, Ph.D. Thesis, Univ. of Illinois at Urbana-Champaign, 1981.
12. G. H. Golub and T. N. Robertson, *A generalized Bairstow algorithm*, Comm. ACM **10** (1967), 371–373.
13. G. H. Golub and J. H. Wilkinson, *Ill-conditioned eigensystems and the computation of the Jordan canonical form*, SIAM Rev. **18** (1976), 578–619.
14. G. H. Golub and C. F. Van Loan, *Matrix computations*, 2nd ed., The Johns Hopkins Univ. Press, Baltimore, MD, 1989.
15. A. Greenbaum, *Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences*, Linear Algebra Appl. **113** (1989), 7–63.
16. M. H. Gutknecht, *A completed theory of the nonsymmetric Lanczos process and related algorithms*. Part I, II, IPS Res. Rep. No. 90-10, Zürich, 1990.
17. N. J. Higham, *Algorithm* 694: *A Collection of Test Matrices in MATLAB*, ACM Trans. Math. Software **17** (1991), 289–305.
18. W. Kahan, B. N. Parlett, and E. Jiang, *Residual bounds on approximate eigensystems of nonnormal matrices*, SIAM J. Numer. Anal. **19** (1982), 470–484.
19. C. Lanczos, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Res. Nat. Bur. Standards **45** (1950), 255–282.
20. C. Paige, *Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix*, J. Inst. Math. Appl. **18** (1976), 341–349.
21. ——, *Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem*, Linear Algebra Appl. **34** (1980), 235–258.
22. B. N. Parlett, *Laguerre's method applied to the matrix eigenvalue problem*, Math. Comp. **18** (1964), 464–485.
23. B. N. Parlett and D. S. Scott, *The Lanczos algorithm with selective reorthogonalization*, Math. Comp. **33** (1979), 217–238.

24. B. N. Parlett, *A new look at the Lanczos algorithm for solving symmetric systems of linear equations*, Linear Algebra Appl. **29** (1980), 323–346.
25. ——, *The symmetric eigenvalue problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
26. B. N. Parlett, D. R. Taylor, and Z. Liu, *A look-ahead Lanczos algorithm for unsymmetric matrices*, Math. Comp. **44** (1985), 105–124.
27. B. N. Parlett, *Reduction to tridiagonal form and minimal realizations*, SIAM J. Math. Anal. Appl. **13** (1992), 567–593.
28. A. Ruhe, *Rational Krylov sequence methods for eigenvalue computation*, Linear Algebra Appl. **58** (1984), 391–405.
29. P. Raschman, M. Kubicek, and M. Maros, *Waves in distributed chemical systems: experiments and computations*, New Approaches to Nonlinear Problems in Dynamics—Proc. Asilomar Conf. Ground, Pacific Grove, California, 1979 (P. J. Holmes, ed.). The Engineering Foundation, SIAM, Philadelphia, PA, 1980, pp. 271–288.
30. H. Simon, *Analysis of the symmetric Lanczos algorithm with reorthogonalization methods*, Linear Algebra Appl. **61** (1984), 101–131.
31. Y. Saad, *Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices*, Linear Algebra Appl. **34** (1980), 269–295.
32. ——, *Numerical solution of large nonsymmetric eigenvalue problems*, Comput. Phys. Comm. **53** (1989), 71–90.
33. ——, *Numerical methods for large eigenvalue problems*, Halsted Press, Div. of John Wiley & Sons, Inc., New York, 1992.
34. D. C. Sorensen, *Implicit application of polynomial filters in a k-step Arnoldi method*, SIAM J. Matrix Anal. Appl. **13** (1992), 357–385.
35. G. W. Stewart, *SRRIT—A FORTRAN subroutine to calculate the dominant invariant subspace of a nonsymmetric matrix*, University of Maryland, Department of Computer Science, TR-514, 1978.
36. W. J. Stewart and A. Jennings, *A simultaneous iteration algorithm for real matrices*, ACM Trans. Math. Software **7** (1981), 184–198.
37. Z. Strakos and A. Greenbaum, *Open questions in the convergence analysis of the Lanczos process for the real symmetric eigenvalue problem*, IMA, University of Minnesota, IMA preprint 924, 1992.
38. J. H. Wilkinson, *The algebraic eigenvalue problem*, Oxford University Press, Oxford, 1965.

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF KENTUCKY, LEXINGTON, KENTUCKY 40506
E-mail address: na.bai@na-net.ornl.gov