SIAM J. Matrix Anal. Appl.
Vol. **, No. *, pp. 0000–0000

ACCURACY OF THE LANCZOS PROCESS FOR THE EIGENPROBLEM AND SOLUTION OF EQUATIONS[*]

CHRISTOPHER C. PAIGE[†]

Abstract. In [SIAM J. Matrix Anal. Appl., 31 (2010), pp. 2347–2359] it was shown that k steps of the finite precision Lanczos process for tridiagonalizing an $n \times n$ Hermitian matrix A could be viewed as an exact Lanczos process for a $(k+n) \times (k+n)$ augmented Hermitian matrix, producing exactly orthogonal vectors. Here we use this and related results to prove the highly accurate behavior of the finite precision Lanczos process when used for finding the eigensystem of A, or for solving linear systems $Ax = b$. It turns out that the finite precision process mimics the exact process in iterative rather than n-step ways, and makes available backward stable results. These results are also complete, such as making available the complete eigensystem of an A with discrete eigenvalues.

Key words. Lanczos process, orthogonality, rounding error analysis, large sparse matrices, eigenproblem, systems of equations, Conjugate Gradients

AMS subject classifications. 65F10, 65F15, 65F25, 65F50, 65G50, 15A18, 15A57

DOI. **.****/*********

1. Introduction.

Some notation used in this paper is described at the start of section 3. Given $A = A^H \in \mathbb{C}^{n\times n}$ and a vector $v_1 \in \mathbb{C}^n$ of unit-length, *i.e.*, $v_1^H v_1 = 1$, one "good" implementation (see the Appendix for a more precise description of "good") of the Hermitian matrix tridiagonalization process of Cornelius Lanczos, see [16], [23, (2.1)–(2.8)], and, *e.g.*, [8, §§10.1–10.3], uses the following two 2-term recurrences (see also [13]). Compute $u_1 := Av_1$, then for $k = 1, 2, \ldots$ (symbols u and w in (1.1) are used only in this description and the Appendix, nowhere else)

$$\begin{aligned} &\alpha_k := v_k^H u_k, \quad w_k := u_k - v_k\alpha_k, \quad \beta_{k+1} := +(w_k^H w_k)^{1/2}, \\ &\text{stop if } \beta_{k+1} \text{ is small enough, else} \\ &v_{k+1} := w_k/\beta_{k+1}, \quad u_{k+1} := Av_{k+1} - v_k\beta_{k+1}. \end{aligned} \tag{1.1}$$

If we define $V_k \triangleq [v_1, \ldots, v_k] \in \mathbb{C}^{n\times k}$ then in theory this gives after k steps

$$AV_k = V_kT_k + v_{k+1}\beta_{k+1}e_k^T = V_{k+1}T_{k+1,k}, \qquad V_k^H V_k = I_k, \tag{1.2}$$

where the real symmetric tridiagonal matrix T_k has diagonal elements $\alpha_1, \ldots, \alpha_k$ and positive next-to-diagonal elements $\beta_2, \ldots, \beta_k$, and, again in theory, the process necessarily stops in $\ell \leq n$ steps with V_ℓ and T_ℓ, while $\beta_{\ell+1} = 0$.

Lanczos originally presented his tridiagonalization process in [16] for solving the eigenproblem of A, for if T_ℓ has eigendecomposition $T_\ell Y = Y\Lambda$, $Y^T = Y^{-1}$, then $A(V_\ell Y) = (V_\ell Y)\Lambda$. He also mentioned it would be useful for solving linear systems $Ax = b$, and in [17] Lanczos adapted such a solution to the case of general square A, see [17, §3], and then mostly treated the symmetric positive definite subcase. This was equivalent to taking $\beta_1 = \|b\|_2$, $v_1 = b/\beta_1$, and at the k-th step of the Lanczos

[*] Version of March 16, 2019. Received by the editors June 9, 2017; accepted for publication (in revised form) by xxxxxxxxx xxxx *, 2019; published electronically xxxx **, 2019.
http://www.siam.org/journals/simax/**-*/*****.html

[†] ORCID iD 0000-0002-0844-0276. School of Computer Science, McGill University, Montreal, Quebec, Canada, H3A 0E9. (paige@cs.mcgill.ca). The author's work was supported by NSERC of Canada grant OGP0009236. Similar material to that in section 13 here appeared in [26].

process (1.2) computing the approximation $x_k = V_k z_k$ where $T_k z_k = e_1\beta_1$. In theory this gives the solution x at the ℓ-th step, and is mathematically equivalent to Hestenes and Stiefel's method of conjugate gradients (CG) in [15], and, *e.g.*, [8, §11.3].

1.1. Finite precision. With finite precision computation the Lanczos process produces a sequence of n-vectors v_j, each with a Euclidean norm that is 1 to almost machine precision, but with a possible severe loss of orthogonality. In fact V_k can become very rank deficient. Because of this the process can continue indefinitely with β_{k+1} never negligible, so that the resulting algorithms for finding eigenvalues or solving equations behave quite differently from the exact cases.

To simplify the presentation we use the word "essentially" (without quotes) in the sense illustrated by: "essentially equal to" (also "$\approx$") meaning "equal to within $O(\epsilon)\|A\|_2$", and "$\underset{\sim}{\in}$" similarly, where if $\|y\|_2 = 1$, then "$y \underset{\sim}{\in} \text{Range}(P_3)$" means "$y + O(\epsilon)\|A\|_2 \in \text{Range}(P_3)$". Here, together with the computer floating-point precision ϵ, $O(\epsilon)$ may be polynomially dependent on the number of steps k, the dimension n of A, and the maximum number of nonzeros in a row of A, see [25, §3.2]. The bound on the accuracy of computed eigenvalues can grow as k^2, but in [21, §8.7] it was stated: "In practice well separated eigenvalues of A (this includes multiple eigenvalues too) have been found to have an error proportional to k, and since if the maximum possible error is proportional to k^2 the expected error would be proportional to k for stochastic errors, the above bound is probably a very good one."

Definition 1.1. *We say that a possible solution to a given problem involving T_k or A is "backward stable" if it is the exact solution to that problem with a perturbed matrix $A+\delta A$ or T_k+E where the norm of the backward error δA or E is bounded by $O(\epsilon)\|A\|$ in the above sense, even if that solution does not arise from a numerical computation. Examples are $\{\widetilde{X}, \tilde{\Lambda}\}$ in (13.8), $\tilde{x}_k$ in (14.4), and $\tilde{z}_k$ in (14.5).*

In [24] it was shown how a special $(k+n)\times(k+n)$ unitary matrix can be defined from any sequence of k unit-length vectors in $\mathbb{C}^n$. This was used in [25] to show that k steps of a good implementation of the finite precision Lanczos process such as (1.1) produce a tridiagonal matrix T_k that satisfies an exact Lanczos process for a $(k+n)\times(k+n)$ augmented matrix $\mathcal{A}_k$, see Corollary 7.1, where the vectors v_k are the computed vectors normalized to have exact length 1.

Here we use the results in [24, 25, 34] to prove that the Lanczos process eventually makes available at least one of every eigenvalue of A, or the solution of $Ax = b$, in a backward stable manner, but this can take $k \gg n$ steps. The terminology "makes available" is used instead of "produces", because the Lanczos process alone does not produce, for example, a backward stable solution of $Ax = b$. Further computations are needed, and the analyses of these could be combined with the analyses here.

A key part of the analysis here is the unitary matrix $Q^{(k)}$ in (4.4),

$$(1.3)\quad Q^{(k)} \equiv \left[\begin{array}{c|c} Q_{11}^{(k)} & Q_{12}^{(k)} \\ \hline Q_{21}^{(k)} & Q_{22}^{(k)} \end{array}\right] \triangleq \left[\begin{array}{c|c} S_k & (I_k - S_k)V_k^H \\ \hline V_k(I_k - S_k) & I_n - V_k(I_k - S_k)V_k^H \end{array}\right] \in \mathbb{U}^{(k+n)\times(k+n)}.$$

This, along with S_k, is introduced in Theorem 4.1. In the arithmetically exact case of the Lanczos process $S_k = 0$ and $Q_{22}^{(k)} = I_n - V_k V_k^H$, $V_k^H V_k = I_k$, so that if A has no multiple eigenvalues and the process goes to completion, $Q_{22}^{(n)} = 0$. The finite precision implementation mimics this case to the extent that for A with all distinct eigenvalues, $\|Q_{22}^{(k)}\|_F \searrow 0$, and if $Q_{22}^{(k)} = 0$, then n of the eigenvalues of T_k are essentially the n eigenvalues of A, while all converged eigenvalues of T_k are also essentially eigenvalues of A, *i.e.*, T_k can have repeats of eigenvalues of A.

Our purpose is to apply this analysis to obtain an increased understanding of how the Lanczos process performs for large sparse Hermitian matrix problems such as the eigenproblem and solution of linear systems, see, for example, [16, 17, 15, 30, 4]. Because the Golub-Kahan bidiagonalization of a general possibly non-square matrix [7] can be formulated as an Hermitian Lanczos process, the results here can be extended to understanding the use of this bidiagonalization in solving least squares problems, singular value computations, and related problems; see, for example, [7, 31, 32, 6, 33, 12]. This analysis can be useful for more adventurous algorithms, see for example Carson and Demmel [2], and perhaps even for unsymmetric Lanczos and CG-like methods, see for example [28, 37]. Similar ideas could help simplify and improve earlier analyses such as that in [29].

The rest of the paper is organized as follows. In the next section we give a brief and incomplete history, followed by the notation used here with some helpful background. In section 4 we summarize the crucial theorem on obtaining the unitary matrix $Q^{(k)}$ in (1.3) from k unit-length n-vectors, the columns of V_k, while section 5 applies this to show that the Lanczos process is always on a useful path, and section 6 derives some properties of that unitary matrix $Q^{(k)}$ that we need. Section 7 summarizes the result of the rounding error analysis in [25]. This shows that the finite precision Lanczos process behaves as a higher dimensional exact Lanczos process for a slightly perturbed $(k+n) \times (k+n)$ matrix $\mathcal{A}_k$. Section 8 introduces the Singular Value Decomposition (SVD) of S_k in (1.3), and how it defines important subspaces related to V_k. Sections 9–12 are devoted to convergence and rate of convergence of the process, showing how the Lanczos process makes available backward stable eigenpairs of A for those eigenvectors that are represented in the initial vector v_1. When A has no multiple eigenvalues so that eventually $Q_{22}^{(k)} = 0$ in (1.3), sections 13 and 14 show how the Lanczos process makes available backward stable solutions for the eigenproblem and solution of equations. Section 15 gives an example of the Lanczos-CG process solving $Ax = b$. Section 16 discusses how all these results might be extended to the full analyses of various practical methods. Finally there are a few additional comments and a summary in section 17, while the Appendix provides extra explanatory material.

2. A brief history.

The early development of our understanding of the finite precision tridiagonalization process of a symmetric matrix A proposed by Cornelius Lanczos in [16] has been discussed by Parlett [36] and by Meurant [18], see also Meurant and Strakǒs [19]. The work here was initiated with [21, 22, 23], where several of those results were clarified and simplified by Panayotov [35], see also [27], but these and other works seemed incomplete. A breakthrough arose with the realization in [24] that an early idea on loss of orthogonality in modified Gram-Schmidt (MGS) outlined by Björck and Paige in [1] could be extended to apply to *any* sequence of unit-length vectors v_j. This approach was applied in [25] to give an augmented backward stability result for the Hermitian matrix Lanczos process [16]. But using this to prove the convergence and accuracy of methods based on the Lanczos process has not been easy, so to provide tools for this study, many relevant results were derived by Paige and Wülling in [34]. In particular they derived the SVD of S_k in (1.3), and that is very effective in the analysis.

One of the guiding lights in this area has been Beresford Parlett, who, with several students and colleagues greatly improved the use and understanding of the Lanczos process. See, for example, [36] for explanations and clarifications of many of the important ideas and relations. In particular Anne Greenbaum, once a student of Parlett, and (initially quite independently) Zdeněk Strakoš, developed our understanding of the practical behavior of the Lanczos process and its use for both the eigenproblem and CG; see, for example, [9, 10, 39, 11, 40]. To put the present study in context, an augmented result on the stability of the Lanczos process was initiated by Greenbaum in [10]. Following this, Strakoš and coworkers developed illuminating results on the practical behaviors of the Lanczos process and CG via an analysis based on the fundamental relationship with the theory of orthogonal polynomials and Gauss quadrature of the Riemann–Stieltjes integral; see the survey paper [19] for a nice description, and [20] for further developments and an extensive literature survey.

The augmented matrix approach here is based solely on ideas from matrix theory and an extension of the groundbreaking concept of backward stability for numerical algorithms developed largely and very effectively by Wilkinson whose work in, *e.g.*, [41, 42], strongly motivated the work leading to this paper. See the note by Hammarling and Higham [14] for valuable history on Wilkinson and backward error analysis.

3. Additional notation and remarks. We use "$\triangleq$" for "is defined to be", and "$\equiv$" for "is equivalent to". Let I_n denote the $n \times n$ unit matrix, with j-th column e_j. We say $Q_1 \in \mathbb{C}^{n\times k}$ has orthonormal columns if $Q_1^H Q_1 = I_k$ and write $Q_1 \in \mathbb{U}^{n\times k}$. We denote the Frobenius norm by $\|B\|_F$, the Euclidean norm by $\|v\|_2 \triangleq \sqrt{v^H v}$ and the spectral norm by $\|B\|_2 \triangleq \sigma_{\max}(B)$, the maximum singular value of B. We write $V_j \triangleq [v_1, v_2, \ldots, v_j]$ and use Range(B) to denote the range of B.

We often index matrices by dimensional subscripts as in V_k when the $(k+1)$-st matrix can be obtained from the k-th by adding a column, or a column and a row. This holds for $V_k \in \mathbb{C}^{n\times k}$ and $S_k \in \mathbb{C}^{k\times k}$. Otherwise we usually use superscripts, as in $Q^{(k)}$, and then subscripts denote partitioning, as in $Q^{(k)} \equiv [Q_1^{(k)} | Q_2^{(k)}]$. We often omit the particular superscript $\cdot^{(k)}$ when the meaning is clear (but do not omit any other superscripts, *e.g.*, we do not omit $\cdot^{(k+1)}$). The integer ℓ is described just after equation (1.2), it denotes the concluding step of the exact Lanczos process.

Stating definitions and results from [23] will simplify the presentation a little. Assume that for the computed T_k in (1.1)–(1.2), where (with $Y \equiv Y^{(k)}$ and $M \equiv M^{(k)}$)

$$(3.1) \qquad T_k Y = YM, \quad Y^T Y = I_k, \quad M \triangleq \operatorname{diag}(\mu_1, \ldots, \mu_k), \quad Y \triangleq [y_1, y_2, \ldots, y_k].$$

Remark 3.1. If T_k is the leading $k \times k$ block of a real symmetric tridiagonal matrix T_m of the form in (1.2) and $(T_k + \widehat{E}_k)\tilde{y} = \tilde{y}\tilde{\mu}$, $\tilde{y}^H\tilde{y} = 1$, then for all $m > k$

$$\Big[T_m + \widehat{E}_m\Big]\begin{bmatrix}\tilde{y}\\0\end{bmatrix} = \begin{bmatrix}\tilde{y}\\0\end{bmatrix}\tilde{\mu}, \quad \widehat{E}_m \triangleq \begin{bmatrix}\widehat{E}_k & 0\\0 & 0\end{bmatrix} - e_{k+1}(\beta_{k+1}e_k^T\tilde{y})\begin{bmatrix}\tilde{y}\\0\end{bmatrix}^H,$$

and there is an eigenvalue of T_m within $\|\widehat{E}_m\|_2 \le \beta_{k+1}|e_k^T\tilde{y}| + \|\widehat{E}_k\|_2$ of $\tilde{\mu}$, [42, p.87]. We say that $\tilde{\mu}$ (as an approximation to an eigenvalue of any T_m, $m \ge k$) has "converged to within $\|\widehat{E}_m\|_2$". If $\|\widehat{E}_m\|_2 \approx 0$ we say that "$\tilde{\mu}$ has converged", so that $\{\tilde{\mu}, \left[\begin{smallmatrix}\tilde{y}\\0\end{smallmatrix}\right]\}$ is a backward stable eigenpair of T_m in the sense of Definition 1.1. It is largely the nonzero $\beta_{k+1}e_k^T\tilde{y} \approx 0$ that forces us to use many expressions involving "$\approx$". □

Remark 3.2. Orthogonality of v_{k+1} can only be lost in the direction of those $V_k y_j^{(k)}$ for which $\mu_j^{(k)}$ has converged, see [23, (3.18)]. This follows because it was shown in [23] that $(\beta_{k+1}e_k^T y_j^{(k)})v_{k+1}^H V_k y_j^{(k)} \approx 0$, $j = 1 : k$, see Remark 3.1 with $\widehat{E}_k = 0$. □

Remark 3.3. If an eigenvalue μ_j of T_k from the finite precision Lanczos process on $A = A^H$ has converged, then it is essentially an eigenvalue of A, see [23, Theorem 3.1], and in fact a backward stable eigenpair $\{\mu_r, V_k y_r\}$ for A, where $\mu_r \approx \mu_j$, $\|V_k y_r\|_2 \approx 1$, is then available from the Lanczos process, see [23, (3.3) & Corollary 3.1]. □

Remark 3.4. It follows from [23, (3.18)] and Remarks 3.1–3.3 above, see also [21, (8.32)], that orthogonality of the v_j in (1.1)–(1.2) is not lost until the first eigenpair $\{\mu_j^{(k)}, y_j^{(k)}\}$ of T_k, and hence the approximate eigenpair $\{\mu_j^{(k)}, V_k y_j^{(k)}\}$ of A, have essentially converged. □

Remark 3.5. It was proven in [21, §8.6.1, pp.122–126], with a summary in [23, (3.28)], that if $\{\mu_j^{(k)}, y_j^{(k)}\}, \dots, \{\mu_{j+s}^{(k)}, y_{j+s}^{(k)}\}$ are $s+1$ eigenpairs of T_k that are close to each other but sufficiently separated from the rest, then

$$\sum_{i=j}^{j+s} \|V_k y_i^{(k)}\|_2^2 \approx s+1. \tag{3.2}$$

□

Remark 3.6. If an eigenpair $\{\mu_j^{(k)}, y_j^{(k)}\}$ of T_k has converged in the sense of Remark 3.1 so that $\beta_{k+1}|e_k^T y_j^{(k)}| \approx 0$, then $\mu_j^{(k)}$ has essentially converged to an eigenvalue λ_i of A, see Remark 3.3. Also orthogonality of v_{k+1} is not significantly lost in the direction $V_k y_j^{(k)}$ until $\mu_j^{(k)}$ has converged, see Remark 3.2. But such loss of orthogonality allows the same eigenvalue of A to be approximated again later. Therefore we mainly consider the first time any eigenvalue $\mu_j^{(k)}$ of T_k converges to an eigenvalue of A, and call $\{\mu_j^{(k)}, y_j^{(k)}\}$ a "first converged" eigenpair of T_k (with respect to A), and $\{\mu_j^{(k)}, V_k y_j^{(k)}\}$ a "first converged" eigenpair of A. In this case we know from Remark 3.5 that $\|V_k y_j^{(k)}\|_2 \approx 1$ if $\mu_j^{(k)}$ is a well-separated eigenvalue of T_k. □

Remark 3.7. If $AZ = Z\Lambda + R$ with Λ diagonal, $Z^H Z \approx I$, then $(A-E)Z = Z\Lambda$ with $E \triangleq R(Z^H Z)^{-1} Z^H$, and $\|E\|_2 \approx \|R\|_2$. So $\{\Lambda, Z\}$ are backward stable for A if $R \approx 0$. □

4. Obtaining a unitary matrix from unit-length n-vectors.

The next theorem was given in full with proofs in [24]. It allows us to develop an $(k+n)\times(k+n)$ unitary matrix $Q^{(k)}$ from any $n \times k$ matrix V_k with unit-length columns.

THEOREM 4.1 ([24, Theorem 2.1]). *For integers $n \geq 1$ and $k \geq 1$, and $V_j \triangleq [v_1, \dots, v_j] \in \mathbb{C}^{n\times j}$ with $\|v_j\|_2 = 1$, $j = 1, \dots, k+1$, define the strictly upper triangular matrix S_k, where U_k is the strictly upper triangular part of $V_k^H V_k = I + U_k + U_k^H$,*

$$S_k \triangleq (I_k + U_k)^{-1} U_k \equiv U_k (I_k + U_k)^{-1} \in \mathbb{C}^{k\times k} \tag{4.1}$$

where $I_k \pm S_k$ and $I_k \pm U_k$ are clearly always nonsingular. Then

$$U_k S_k = S_k U_k, \quad U_k = (I_k - S_k)^{-1} S_k \equiv S_k (I_k - S_k)^{-1}, \quad (I_k - S_k)^{-1} = I_k + U_k, \tag{4.2}$$

$$\|S_k\|_2 \leq 1; \quad V_k^H V_k = I \Leftrightarrow \|S_k\|_2 = 0; \quad V_k^H V_k \ \ \textit{singular} \Leftrightarrow \|S_k\|_2 = 1. \tag{4.3}$$

Most importantly, S_k is the unique *strictly upper triangular $k \times k$ matrix such that*

$$Q^{(k)} \equiv \left[\begin{array}{c|c} Q_{11}^{(k)} & Q_{12}^{(k)} \\ \hline Q_{21}^{(k)} & Q_{22}^{(k)} \end{array}\right] \triangleq \left[\begin{array}{c|c} S_k & (I_k - S_k) V_k^H \\ \hline V_k (I_k - S_k) & I_n - V_k (I_k - S_k) V_k^H \end{array}\right] \in \mathbb{U}^{(k+n)\times(k+n)}. \tag{4.4}$$

We also write $Q^{(k)} \equiv \Big[\underset{k}{Q_1^{(k)}} \,\Big|\, \underset{n}{Q_2^{(k)}}\Big]$. *Define* $\begin{bmatrix} s_k \\ 0 \end{bmatrix} \triangleq S_k e_k$, *then with* (4.2) *we have*

$$S_k e_k = (I_k - S_k) U_k e_k = \begin{bmatrix} (I_{k-1} - S_{k-1}) V_{k-1}^H v_k \\ 0 \end{bmatrix}, \quad s_{k+1} = (I_k - S_k) V_k^H v_{k+1}, \tag{4.5}$$

$$Q_1^{(k+1)} = \left[\begin{array}{c} S_{k+1} \\ \hline V_{k+1}(I_{k+1} - S_{k+1}) \end{array}\right] = \left[\begin{array}{c|c} S_k & s_{k+1} \\ 0 & 0 \\ \hline V_k(I_k - S_k) & v_{k+1} - V_k s_{k+1} \end{array}\right]. \quad \square \tag{4.6}$$

Perhaps the simplest proof of Theorem 4.1 so far is that given in [26]. In [24] the above construction was called an *orthonormal augmentation of a sequence of unit-length vectors*, and $Q^{(k)}$ an *augmented unitary matrix*.

5. Applying $Q^{(k)}$ in Theorem 4.1 to the Lanczos process.

The tridiagonal matrix T_k arising from the finite precision version of the Lanczos process in (1.1)–(1.2) can be viewed as the result of a unitary similarity transformation applied to a strange, slightly perturbed, higher dimensional matrix, as we now illustrate.

Theorem 5.1 ([25, Theorem 3.1]). *After k finite precision steps of a good implementation of the Lanczos algorithm with $A = A^H$ and v_1 leading to the computed β_{k+1} and tridiagonal matrix T_k, see* (1.2)*, let $V_{k+1} = [v_1, v_2, \ldots, v_{k+1}]$ be the matrix of computed Lanczos vectors normalized to have unit length. Then if $Q^{(k)} \in \mathbb{U}^{(k+n)\times(k+n)}$ is as in* (4.4) *in Theorem 4.1, and $A_k \triangleq A - v_{k+1}\beta_{k+1}v_k^H - v_k\beta_{k+1}v_{k+1}^H = A_k^H$, we have*

$$Q^{(k)H}\mathcal{A}_k Q^{(k)} = \mathcal{T}_k \triangleq \left[\begin{array}{c|c} T_k & e_k\beta_{k+1}v_{k+1}^H \\ \hline v_{k+1}\beta_{k+1}e_k^T & A_k \end{array}\right], \quad \mathcal{A}_k \triangleq \begin{bmatrix} T_k & 0 \\ 0 & A \end{bmatrix} + H^{(k)}, \tag{5.1}$$

$$T_{k+1,k} \triangleq \begin{bmatrix} T_k \\ \beta_{k+1}e_k^T \end{bmatrix}, \quad H^{(k)} = H^{(k)H} \equiv \begin{bmatrix} H_{11}^{(k)} & H_{12}^{(k)} \\ H_{21}^{(k)} & H_{22}^{(k)} \end{bmatrix}, \quad \|H^{(k)}\|_2 \le O(\epsilon)\|A\|_2. \ \square$$

More precise bounds for $H^{(k)}$ are suggested in [25, §3], with a correction suggested by Carson and Demmel in [2, §5]. The corresponding Lanczos process in Corollary 7.1 might facilitate an understanding of Theorem 5.1.

5.1. The Lanczos process always behaves well.

No matter how large k is in Theorem 5.1, we can in theory apply at most $n-1$ exact unitary similarity transformations to $\mathcal{T}_k$ in (5.1) to complete its tridiagonalization, giving $\widetilde{Q}^{(k)H}\mathcal{A}_k\widetilde{Q}^{(k)} = \widetilde{T}_{k+n}$, so that the eigenvalues of $\widetilde{T}_{k+n}$ are exactly the eigenvalues of $\mathcal{A}_k$, that is, essentially all of the eigenvalues of A together with essentially all of the eigenvalues of T_k.

Not all eigenvalues of T_k will have converged, and so not all eigenvalues of $\widetilde{T}_{k+n}$ will essentially be eigenvalues of A. But every converged eigenvalue of T_k is essentially an eigenvalue of A, see Remark 3.3, and eigenvalues of the developing T_k never lose their level of convergence, see Remark 3.1. Because this is true for all k, it shows that the Lanczos process is always on track for the eigenproblem, the accuracy of approximation to eigenvalues of A is only limited by the slowly growing size of the backward error $H^{(k)}$. It follows from the above that the eigenvalues of T_k essentially lie between the smallest and largest eigenvalues of A. A more precise result is given in Theorem 17.1 in the Appendix.

6. Some properties of $Q^{(k)}$ in (4.4). Our analysis of the computational Lanczos process uses properties of the sub-blocks of $Q^{(k)}$ in (4.4)–(4.6). From (4.5) $s_{k+1} = (I_k - S_k)V_k^H v_{k+1} = Q_{12}^{(k)} v_{k+1}$, so together with (4.4) and (4.6)

$$Q_{22}^{(k)} v_{k+1} = [I_n - V_k(I_k - S_k)V_k^H]v_{k+1} = v_{k+1} - V_k s_{k+1} = Q_{21}^{(k+1)} e_{k+1}, \tag{6.1}$$

$$q^{(k+1)} \triangleq \begin{bmatrix} s_{k+1} \\ v_{k+1} - V_k s_{k+1} \end{bmatrix} = \begin{bmatrix} Q_{12}^{(k)} \\ Q_{22}^{(k)} \end{bmatrix} v_{k+1} = Q_2^{(k)} v_{k+1}. \tag{6.2}$$

For $j = 1 : k+1$ define the orthogonal projectors $\mathcal{P}_j \triangleq I_n - v_j v_j^H$. Because S_k is strictly upper triangular we see $S_1 = 0$, so from (4.4) we have $Q_{22}^{(1)} = \mathcal{P}_1$, and we use

$$\begin{aligned} &Q_{21}^{(k)} = V_k(I_k - S_k), \quad Q_{12}^{(k)} = (I_k - S_k)V_k^H, \quad Q_{22}^{(k)} = I_n - Q_{21}^{(k)} V_k^H, \\ &V_{k+1} = [V_k, v_{k+1}], \qquad I_{k+1} - S_{k+1} = \begin{bmatrix} I_k - S_k & -s_{k+1} \\ 0 & 1 \end{bmatrix}, \end{aligned}$$

to prove several things with (6.2), in particular that $Q_{22}^{(k)} = \mathcal{P}_1 \cdots \mathcal{P}_k$:

$$Q_{21}^{(k+1)} = V_{k+1}(I_{k+1} - S_{k+1}) = [V_k(I_k - S_k), v_{k+1} - V_k s_{k+1}] = [Q_{21}^{(k)}, Q_{22}^{(k)} v_{k+1}], \tag{6.3}$$

$$Q_{12}^{(k+1)} = (I_{k+1} - S_{k+1})V_{k+1}^H = \begin{bmatrix} (I_k - S_k)V_k^H - s_{k+1} v_{k+1}^H \\ v_{k+1}^H \end{bmatrix} = \begin{bmatrix} Q_{12}^{(k)} \mathcal{P}_{k+1} \\ v_{k+1}^H \end{bmatrix}, \tag{6.4}$$

$$Q_{22}^{(k+1)} = I_n - Q_{21}^{(k+1)} V_{k+1}^H = I_n - Q_{21}^{(k)} V_k^H - Q_{22}^{(k)} v_{k+1} v_{k+1}^H = Q_{22}^{(k)}(I_n - v_{k+1} v_{k+1}^H). \tag{6.5}$$

The decrease in $\|Q_{22}^{(k)}\|_F$ is crucial. First $\|Q_{22}^{(k+1)}\|_2 \leq \|Q_{22}^{(k)}\|_2$ because

$$Q_{22}^{(k+1)} Q_{22}^{(k+1)H} = Q_{22}^{(k)}(I_n - v_{k+1} v_{k+1}^H) Q_{22}^{(k)H} = Q_{22}^{(k)} Q_{22}^{(k)H} - Q_{22}^{(k)} v_{k+1} v_{k+1}^H Q_{22}^{(k)H}. \tag{6.6}$$

This and (6.3) with $Q_{22}^{(0)} \triangleq I_n$ show how $\|Q_{22}^{(k)}\|_F$ decreases and $\|Q_{21}^{(k)}\|_F$ increases:

$$\|Q_{22}^{(k+1)}\|_F^2 = \text{trace}[Q_{22}^{(k+1)} Q_{22}^{(k+1)H}] = \|Q_{22}^{(k)}\|_F^2 - \|Q_{22}^{(k)} v_{k+1}\|_2^2, \tag{6.7}$$

$$\|Q_{21}^{(k+1)}\|_F^2 = \|Q_{21}^{(k)}\|_F^2 + \|Q_{22}^{(k)} v_{k+1}\|_2^2 = \textstyle\sum_{j=0}^{k} \|Q_{22}^{(j)} v_{j+1}\|_2^2. \tag{6.8}$$

Ideally $S_{k+1} = 0$, so in (6.1) $Q_{22}^{(k)} v_{k+1} = v_{k+1}$, and $\|Q_{22}^{(k)}\|_F^2 = n - k$ decreases by 1 each step. Computationally, see Remark 3.4, until the first eigenpair converges there is negligible loss of orthogonality, and then $\|Q_{22}^{(k)} v_{k+1}\|_2 \approx \|v_{k+1}\|_2 = 1$ also, but once orthogonality is lost convergence can become very slow.

To facilitate an understanding of the subsequent theory we now give an example indicating how $\|Q_{22}^{(k)} v_{k+1}\|_2 = \|v_{k+1} - V_k s_{k+1}\|_2$ and $\|Q_{22}^{(k)}\|_F$ can behave in practice.

Remark 6.1. All computations are carried out in MATLAB™ using IEEE double precision floating-point arithmetic (unit roundoff $u = 2^{-53} \approx 10^{-16}$). The computed results involving the theoretical $Q^{(k)}$ have rounding errors, and therefore are approximations. To limit cancellation errors we always compute $Q_{22}^{(k)} v_{k+1}$ instead of $v_{k+1} - V_k s_{k+1}$, and compute $Q_{22}^{(k+1)}$ via $Q_{22}^{(k)}(I_n - v_{k+1} v_{k+1}^H)$, see (6.1) and (6.5). For ease of reference we sometimes call $\|Q_{22}^{(k)} v_{k+1}\|_2$ the "Q-change", see (6.7), (6.8). □

EXAMPLE 6.1. *The Lanczos process has difficulty with close eigenvalues, especially when there are one or more very well separated eigenvalues, so to illustrate how*

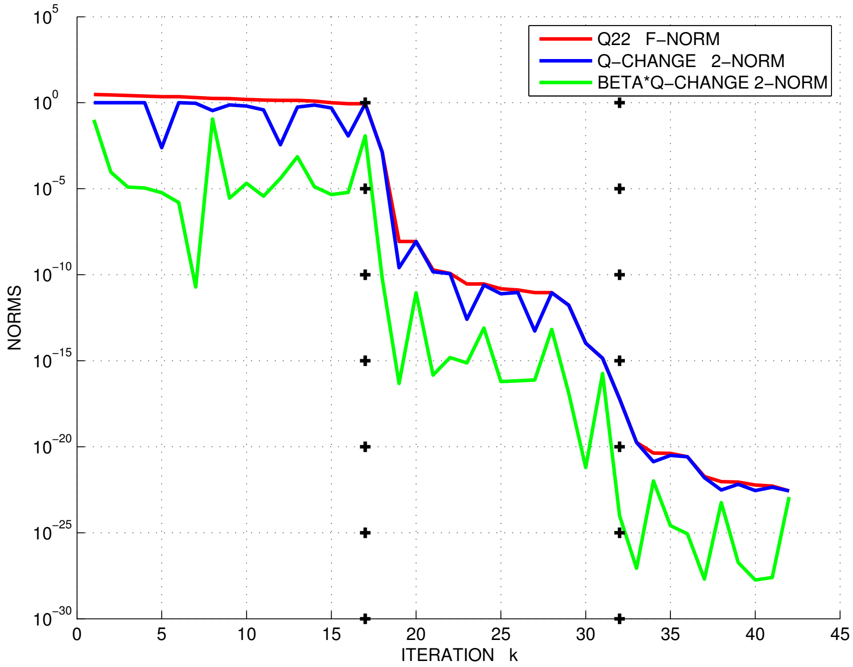

FIG. 6.1. *Plots of* $\|Q_{22}^{(k)}\|_F$, $\|Q_{22}^{(k)}v_{k+1}\|_2$, *and* $\beta_{k+1}\|Q_{22}^{(k)}v_{k+1}\|_2$ *obtained from the Lanczos process with* v_1 *in a 6 dimensional eigensubspace of* $A = A^T \in \mathbb{R}^{10\times 10}$.

$\|Q_{22}^{(k)}\|_F^2$ *decays even with this very poor behavior, the process was applied to a random symmetric matrix* $A \in \mathbb{R}^{10\times 10}$ *with eigenvalues* $\lambda_i = i * 0.00001$, $i = 1{:}9$, *and* $\lambda_{10} = 1$. *The initial vector* v_1 *was a random combination of the eigenvectors of* A *corresponding to eigenvalues* $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_{10}$, *to show how the others will also be found.*

The results are plotted in Figure 6.1. Not until $k = 12$ *was* V_k *recognized as having rank 10. It can be seen from a standard plot, but not this semi-log plot, that* $\|Q_{22}^{(k)}\|_F^2$ *follows the correct path, decreasing from 10 by 1 each step until step 5 when the first eigenvalue of* A *has converged to* $O(\epsilon)\|A\|$, *at which point the rate of decrease of* $\|Q_{22}^{(k)}\|_F^2$ *slows. Unlike the exact process that would give* $\beta_7 = 0$, *the computational process only gave* $\beta_7 = 0.000001595485258$, *and so did not stop.*

The line at the top, the red line, represents values of $\|Q_{22}^{(k)}\|_F$. *The second line from the top, the blue line, represents values of* $\|Q_{22}^{(k)}v_{k+1}\|_2$, *the Q-change. Although we know* $\|Q_{22}^{(k)}v_{k+1}\|_2 \le \|Q_{22}^{(k)}\|_2$, *it is remarkable how close* $\|Q_{22}^{(k)}v_{k+1}\|_2$ *was to* $\|Q_{22}^{(k)}\|_2$, *rarely departing too far from it for more than a step at a time. This, with (6.7), is one reason that* $\|Q_{22}^{(k)}\|_F$ *decreases so rapidly here, even after orthogonality and linear independence have been lost. But this does not always happen, see section 15. The rate of convergence is discussed in section 12.1.*

At $k = 19$, $\beta_{k+1} = 0.000000185699582$ *and* $\|Q_{22}^{(k)}v_{k+1}\|_2 = 0.000000000258131$, *so the process could be stopped with* $\beta_{k+1}\|Q_{22}^{(k)}v_{k+1}\|_2 < 10^{-16}$, *see the green, and lowest, line, even though* $\|Q_{22}^{(k)}v_{k+1}\|_2$ *and* $\|Q_{22}^{(k)}\|_F = 0.000000008678201$ *are still a long* *way from zero. Nevertheless, at* $k = 19$*, to* 10^{-15} *all the eigenvalues of* A *have been given accurately, where* $\lambda_{10} = 1$ *appears 5 times, each of* $\lambda_1, \lambda_2, \lambda_3, \lambda_4$*, and* λ_5 *appear twice, but the second value of* $\lambda_3 = 0.000030000000035$ *has not yet converged. Remember* $\|S_k\|_2 \leq 1$ *in* (4.3)*. For later interest the computed version of* S_{19} *here has 9 essentially unit singular values indicating the amount of loss of linear independence of the columns of the ideal* $V_{19} \in \mathbb{R}^{10\times 19}$*, 6 essentially zero singular values, with the others being* 0.000000008678156*,* 0.000000000028064*,* 0.000000000001571*, and* 0.000000000000007*, all less than* $\|Q_{22}^{(k)}\|_F = 0.000000008678201$*.*

The red line shows how $\|Q_{22}^{(k)}\|_F$ *decreases quite rapidly. However* $\|Q_{22}^{(k)}\|_F$ *and* $\|Q_{22}^{(k)} v_{k+1}\|_2$ *are not zero to 15 decimal places until step* $k = 32$*, well after all eigenvalues had been found to full precision at step* $k = 19$*. At step* $k = 32$*,* S_{32} *has 10 essentially zero and 22 essentially unit singular values, to 15 decimal places. It turns out that* $Q_{22}^{(k)} = 0$ *implies that the SVD of* S_k *(see Definition 8.1) is* $S_k = W_1 P_1^H$*, where* $W_1, P_1 \in \mathbb{U}^{k\times(k-n)}$*, and that explains this.*

This example emphasizes that while the *theory* here concentrates on the effect of $\|Q_{22}^{(k)}\|_F$ decreasing until it stabilizes (at zero in cases like this where A has no multiple eigenvalues) we expect the *practical* Lanczos process to produce accurate results well before $\|Q_{22}^{(k)}\|_F$ stabilizes.

7. The "Exact" Finite Precision Lanczos process.

A simple rounding error analysis of a good finite precision implementation of (1.2) gives, see, *e.g.*, [22],

$$(7.1)\quad AV_k = V_k T_k + v_{k+1}\beta_{k+1}e_k^T + E_k = V_{k+1}T_{k+1,k} + E_k, \quad \|E_k\|_{2,F} \leq O(\epsilon)\|A\|_{2,F},$$

but this has limited applicability. We base our analysis on a theorem from [25].

Corollary 7.1 ([25, Corollary 3.2]). *With the assumptions and notation of Theorem 5.1, there is an exact Lanczos process for the Hermitian matrix* $\mathcal{A}_k$ *in* (5.1)*,*

$$(7.2)\quad \left(\begin{bmatrix} T_k & 0 \\ 0 & A \end{bmatrix} + H^{(k)}\right)\begin{bmatrix} S_k \\ V_k(I - S_k) \end{bmatrix} = \begin{bmatrix} S_k \\ V_k(I - S_k) \end{bmatrix} T_k + \begin{bmatrix} s_{k+1} \\ v_{k+1} - V_k s_{k+1} \end{bmatrix} \beta_{k+1} e_k^T,$$

$$(7.3)\quad \left[Q_1^{(k)} \;\middle|\; q^{(k+1)} \right] \triangleq \left[\begin{matrix} S_k \\ V_k(I - S_k) \end{matrix} \;\middle|\; \begin{matrix} s_{k+1} \\ v_{k+1} - V_k s_{k+1} \end{matrix} \right] \in \mathbb{U}^{(k+n)\times(k+1)}, \text{ see } (4.6), (6.2).$$

This follows by multiplying (5.1) *on the left by* $Q^{(k)}$*, and taking the first* k *columns. Here* $q^{(k+1)}$ *is the last column of* $Q_1^{(k+1)}$ *with its zero* $(k+1)$*-st element removed.* □

We call (7.2)–(7.3) the "exact" finite precision Lanczos process because the computed $T_{k+1,k}$ is seen to be the exact result of k steps of an exact Lanczos process with exact orthogonality arising from the strange Hermitian matrix $\mathcal{A}_k$ with its $O(\epsilon)\|A\|_2$ Hermitian backward error $H^{(k)}$, the only rounding error component. To help understanding, if $H^{(k)} = 0$ then S_k and s_{k+1} will be zero, the top block-row of (7.2) will be zero, while the bottom block-row will correspond to the ideal Lanczos process.

Even in practice the first column of S_k is zero, so the first column of $Q_1^{(k)}$ is $(0^T, v_1^T)^T$. But a nonzero rounding error term $H^{(k)}$ leads to some nonzero elements in each s_{k+1}, and so in the top of $q^{(k+1)}$. For the ideal V_{k+1} with unit length columns and the resulting ideal s_{k+1} it is even possible to have some $\|s_{k+1}\|_2 = 1$, so that $v_{k+1} = V_k s_{k+1}$, increasing the rank deficiency by one in going from V_k to V_{k+1}. Nevertheless the augmented system (7.2) is still an exact Lanczos process.

The k-th step of the Lanczos process produces α_k and β_{k+1}, so it seems strange that the full T_k is part of $\mathcal{A}_k$, because T_k then seems to appear fully on both sides of equation (7.2). But this is just an artifact that is necessary to make $\mathcal{A}_k$ Hermitian. Because $e_k^T S_k = 0$, we see from (7.3) that the k-th row of $Q_1^{(k)}$ is zero, and so the k-th column of T_k is not used in $\mathcal{A}_k Q_1^{(k)}$ in (7.2). Thus (7.2) shows how $T_{k+1,k}$ is developed from $T_{k,k-1}$. We use Corollary 7.1 to show just what is happening in the finite precision Lanczos process.

Remark 7.1. Because T_k is tridiagonal the columns of $Q_1^{(k)}$ form an orthonormal basis for the Krylov subspace $\mathcal{K}^k(\mathcal{A}_k, Q_1^{(k)} e_1)$. This means that $B = (Q_1^{(k)})^H \mathcal{A}_k Q_1^{(k)} = T_k$ gives $\beta_{k+1} = \min_B \|\mathcal{A}_k Q_1^{(k)} - Q_1^{(k)} B\|_F$, *i.e.*, T_k minimizes the residual for $Q_1^{(k)}$ as an approximate invariant subspace for the matrix $\mathcal{A}_k$, see Parlett [36, §12.7,13.1]. □

8. The singular value decomposition (SVD) of S_k.

We derive the SVD $S_k = W^{(k)} \Sigma^{(k)} P^{(k)H}$ when S_k in (4.1) arises from any matrix V_k with unit-length columns. We remind the reader that we often omit the superscript $\cdot^{(k)}$ for readability, and write, *e.g.*, $S_k = W \Sigma P^H$, but never omit other superscripts such as $\cdot^{(k+1)}$.

From (4.3) we know that $\sigma_{\max}(S_k) \le 1$, and we show that unit singular values are crucial in the analysis. Also if $V_k^H V_k = I$ then $S_k = 0$ in (4.1), and it helps to label the singular vectors of S_k according to its zero, unit, and in between singular values. Briefly, zero singular values correspond to no loss of orthogonality, unit singular values to loss of linear independence, and intermediate singular values to loss of orthogonality but not loss of linear independence. The rest of this section comes from [34].

Definition 8.1 (Partitioned SVD of S_k, [34, §4]). *Let the $k \times k$ matrix S_k in Theorem 4.1 have m_k unit and n_k zero singular values with SVD*

$$S_k = W \Sigma P^H \equiv W_1 P_1^H + W_2 \Sigma_2 P_2^H, \quad I - S_k S_k^H = W \Gamma^2 W^H \equiv W_2 \Gamma_2^2 W_2^H + W_3 W_3^H, \tag{8.1}$$

$$W \equiv W^{(k)} \equiv [w_1, \ldots, w_k] \equiv [W_1, W_2, W_3] \in \mathbb{U}^{k \times k}, \; W_1 \in \mathbb{U}^{k \times m_k}, \; W_3 \in \mathbb{U}^{k \times n_k},$$

$$P \equiv P^{(k)} \equiv [p_1, ., p_k] \equiv [P_1, P_2, P_3] \in \mathbb{U}^{k \times k}, \; P_1 \in \mathbb{U}^{k \times m_k}, \; P_2 \in \mathbb{U}^{k \times \ell_k}, \; P_3 \in \mathbb{U}^{k \times n_k}, \tag{8.2}$$

$$\Sigma \equiv \Sigma^{(k)} \equiv \mathrm{diag}(\sigma_1, \ldots, \sigma_k) \equiv \mathrm{diag}(I_{m_k}, \Sigma_2, O_{n_k}), \quad \Sigma_2 \in \mathbb{R}^{\ell_k \times \ell_k}, \quad k = \ell_k + m_k + n_k,$$

$$\Gamma^2 \triangleq I_k - \Sigma^2, \quad \Gamma \equiv \Gamma^{(k)} \equiv \mathrm{diag}(\gamma_1, \ldots, \gamma_k) \equiv \mathrm{diag}(O_{m_k}, \Gamma_2, I_{n_k}), \; \Gamma_2 \textit{ positive definite},$$

where the singular values σ_j, $1 \le j \le k$, of S_k in $\Sigma \equiv \Sigma^{(k)}$ are arranged as follows,

$$1 = \sigma_1 = \cdots = \sigma_{m_k} > \sigma_{m_k+1} \ge \cdots \ge \sigma_{m_k+\ell_k} > \sigma_{m_k+\ell_k+1} = \cdots = \sigma_k = 0. \tag{8.3}$$

These singular vectors of S_k combine with (4.4) to reveal key properties of V_k:

$$Q_1^{(k)} P = \left[\begin{array}{c} S_k P \\ \hline V_k(I_k - S_k) P \end{array}\right] = \left[\begin{array}{c|c|c} W_1 & W_2 \Sigma_2 & 0 \\ \hline V_k(P_1 - W_1) & V_k(P_2 - W_2 \Sigma_2) & V_k P_3 \end{array}\right] = \left[\begin{array}{c|c|c} W_1 & W_2 \Sigma_2 & 0 \\ \hline 0 & \widetilde{V}_2 \Gamma_2 & \widetilde{V}_3 \end{array}\right], \tag{8.4}$$

$$Q^{(k)H} \begin{bmatrix} W \\ 0 \end{bmatrix} = \left[\begin{array}{c} S_k^H W \\ \hline V_k(I_k - S_k)^H W \end{array}\right] = \left[\begin{array}{c|c|c} P_1 & P_2 \Sigma_2 & 0 \\ \hline V_k(W_1 - P_1) & V_k(W_2 - P_2 \Sigma_2) & V_k W_3 \end{array}\right] = \left[\begin{array}{c|c|c} P_1 & P_2 \Sigma_2 & 0 \\ \hline 0 & \widehat{V}_2 \Gamma_2 & \widehat{V}_3 \end{array}\right], \tag{8.5}$$

where $[\widetilde{V}_2, \widetilde{V}_3]$ and $[\widehat{V}_2, \widehat{V}_3]$ are defined in the following theorem. The first equality in each of (8.4) and (8.5) follows from the structure of $Q^{(k)}$, and the second by applying (8.1). But the columns in each expression are orthonormal, giving the structure in the fourth expressions. Because $\Gamma_2 > 0$, each of $[\widetilde{V}_2, \widetilde{V}_3]$, $[\widehat{V}_2, \widehat{V}_3]$ has orthonormal columns that span Range(V_k). This structure is used to prove the following theorem.

Theorem 8.2 (Range & null space of V_k, [34, Theorem 4.2]). *With the notation in Theorem 4.1 and Definition 8.1, define* $\widetilde{V}_2 \triangleq V_k(P_2 - W_2\Sigma_2)\Gamma_2^{-1}$, $\widetilde{V}_3 \triangleq V_k P_3$, $\widehat{V}_2 \triangleq V_k(W_2 - P_2\Sigma_2)\Gamma_2^{-1}$ *and* $\widehat{V}_3 \triangleq V_k W_3$*. Let the columns of* $\widehat{V}_0$ *comprise an orthonormal basis of* $\mathrm{Range}(V_k)^{\perp}$*. Then defining* $\widetilde{V}^{(k)} \triangleq [\widehat{V}_0, \widetilde{V}_2, \widetilde{V}_3]$ *and* $\widehat{V}^{(k)} \triangleq [\widehat{V}_0, \widehat{V}_2, \widehat{V}_3]$,

$$\mathrm{Range}(V_k) = \mathrm{Range}([\widetilde{V}_2, \widetilde{V}_3]) = \mathrm{Range}([\widehat{V}_2, \widehat{V}_3]) \perp \mathrm{Range}(\widehat{V}_0), \ \ \mathrm{rank}(V_k) = k - m_k, \tag{8.6}$$

$$\mathcal{N}(V_k) = \mathrm{Range}(P_1 - W_1), \quad P_1 - W_1 \in \mathbb{C}^{k\times m_k}, \quad \mathrm{rank}(P_1 - W_1) = m_k, \tag{8.7}$$

$$\widetilde{V}^{(k)} \equiv \widetilde{V} \triangleq [\widehat{V}_0, \widetilde{V}_2, \widetilde{V}_3] \in \mathbb{U}^{n\times n}, \quad \widehat{V}^{(k)} \equiv \widehat{V} \triangleq [\widehat{V}_0, \widehat{V}_2, \widehat{V}_3] \in \mathbb{U}^{n\times n}, \tag{8.8}$$

$$Q_{22}^{(k)} = [\widehat{V}_0,\ \widetilde{V}_2]\,\mathrm{diag}(I_{n-(k-m_k)}, -\Sigma_2)[\widehat{V}_0,\ \widehat{V}_2]^H = \widehat{V}_0\widehat{V}_0^H - \widetilde{V}_2\Sigma_2\widehat{V}_2^H, \tag{8.9}$$

where this last can be seen by substituting the SVDs (8.1) *and* (8.10) *in* $Q^{(k)}$*, and using the CS-Decomposition* (CSD, see [5, 38], or for example [8, §2.5.4]) *of* $Q^{(k)}$*.* □

$\mathrm{Range}(\widetilde{V}_3^{(k)})$ in (8.4) and (8.8) is crucial for the analysis: we will later show that if an eigenvector of A lies in $\mathrm{Range}(\widetilde{V}_3^{(k)})$ then it is available at step k of the process.

Remark 8.1. In Definition 8.1 it can be seen that W_1 and P_1 are arbitrary up to a right orthogonal transformation $W_1P_1^H = (W_1Z)(P_1Z)^H$, $Z \in \mathbb{U}^{m_k\times m_k}$, while P_3 and W_3 are each arbitrary up to individual right orthogonal transformations. □

Theorem 8.2, with (8.4) and (8.5), gives expansions for $Q_{21}^{(k)}$ and $Q_{12}^{(k)H}$

$$Q_{21}^{(k)} = V_k(I - S_k) = \widetilde{V}_2\Gamma_2P_2^H + \widetilde{V}_3P_3^H, \quad Q_{12}^{(k)H} = V_k(I - S_k)^H = \widehat{V}_2\Gamma_2W_2^H + \widehat{V}_3W_3^H, \tag{8.10}$$

while (4.5) can be expanded using (8.5) to give a new expression for s_{k+1}

$$\begin{aligned} s_{k+1} &= W^{(k)}W^{(k)H}(I - S_k)V_k^H v_{k+1} = W^{(k)}[0,\ \widehat{V}_2^{(k)}\Gamma_2^{(k)},\ \widehat{V}_3^{(k)}]^H v_{k+1} \\ &= W_2^{(k)}\Gamma_2^{(k)}\widehat{V}_2^{(k)H}v_{k+1} + W_3^{(k)}\widehat{V}_3^{(k)H}v_{k+1}, \quad W_1^{(k)H}s_{k+1} = 0. \end{aligned} \tag{8.11}$$

Section 9 discusses non-generic cases of the Lanczos process, while sections 10, 11, and 12 prove convergence and consider rate of convergence. Because sections 10, 11, and 12 are lengthy and difficult, some readers might want to skip from here to section 13 on a first reading to see the important results that follow when $Q_{22}^{(k)} = 0$, *i.e.*, that the Lanczos process makes available backward stable solutions.

9. Early termination of the exact Lanczos process. For $k < \ell$ in section 1, the exact Lanczos process (1.1)–(1.2) gives for $Q^{(k)}$ and $Q^{(k+1)}$ in (4.4)

$$\begin{aligned} Q^{(k)} &\equiv \left[\begin{array}{c|c} Q_{11}^{(k)} & Q_{12}^{(k)} \\ \hline Q_{21}^{(k)} & Q_{22}^{(k)} \end{array}\right] = \left[\begin{array}{c|c} 0 & V_k^H \\ \hline V_k & I_n - V_kV_k^H \end{array}\right], \\ Q_{22}^{(k+1)} &= Q_{22}^{(k)} - v_{k+1}v_{k+1}^H, \quad Q_{21}^{(k+1)} = [Q_{21}^{(k)}, v_{k+1}] = Q_{12}^{(k+1)H}, \end{aligned} \tag{9.1}$$

so that $v_{k+1}v_{k+1}^H$ is taken from $Q_{22}^{(k)}$ while v_{k+1} is added to $Q_{21}^{(k)}$ and $Q_{12}^{(k)H}$.

The vectors $v_1, \ldots, v_k$, $k \le \ell$, of the process on A span the Krylov subspace

$$\mathcal{K}_k(A, v_1) \triangleq \mathrm{span}\{v_1, Av_1, ..., A^{k-1}v_1\} = \mathrm{Range}(V_k), \tag{9.2}$$

where the Lanczos process can stop with $\beta_{\ell+1} = 0$, $\ell < n$. If A has the eigensystem

$$AX = X\Lambda, \quad X^HX = I_n, \quad X \equiv [x_1, \ldots, x_n], \quad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n), \tag{9.3}$$

and $v_1^H x_i = 0$ then $x_i \perp \mathcal{K}_k(A, v_1)$ for all k, and cannot be found by the process, giving

$$x_i^H Q_{21}^{(k)} = 0, \quad x_i^H Q_{22}^{(k)} = x_i^H, \quad x_i^H Q_{22}^{(k)} v_{k+1} = x_i^H v_{k+1} = 0, \quad \text{for all } k, \tag{9.4}$$

so that $Q_{22}^{(k)}$ can never be zero. These $v_1^H x_i = 0$ can occur for two distinct reasons:

Item 1. The original v_1 might be such that $v_1^H x_i = 0$ when λ_i is a singleton.

Item 2. If A has multiple eigenvalues the ideal Lanczos process will stop with ℓ no greater than the number of distinct eigenvalues of A. Suppose $\lambda_1 = \lambda_2 = \lambda_3$ with $X_1 \triangleq [x_1, x_2, x_3]$. Then X_1 is arbitrary up to multiplication on the right by a unitary matrix, and could theoretically be altered to be $\widehat{X}_1 \triangleq [\hat{x}_1, \hat{x}_2, \hat{x}_3]$ where $v_1^H \widehat{X}_1 = [\xi_1, 0, 0]$, so that $v_1^H \hat{x}_i = 0$ for $i = 2, 3$.

These possibilities are pertinent to our analysis. Example 6.1 tested Item 1, and found that the eigenvectors orthogonal to v_1 were quite quickly introduced by rounding errors. However we cannot analyse this, or assume this will always happen, so we need $|v_1^H x_i| > 0$ to prove convergence of distinct eigenvalues λ_i of A.

Remark 9.1. For Item 2, computed $\|Q_{22}^{(k)}\|_F$ tends to stabilize at a significant nonzero value if A has multiple eigenvalues. □

In the Lanczos process with $v_1 \triangleq b/\beta_1$, $\beta_1 \triangleq \|b\|_2$, neither case of $\ell < n$ limits solving $Ax = b$, nonsingular $A = A^H$. Take $x = V_\ell z$ where $AV_\ell = V_\ell T_\ell$ and $T_\ell z = e_1 \beta_1$.

10. Preliminaries for convergence theory.

Our proof of convergence in section 12 for an eigenvector x_i of A is based on showing that $\widetilde{V}_3^{(k)}$ develops so that $x_i \underset{\sim}{\in} \text{Range}(\widetilde{V}_3^{(k)})$, or equivalently, see (10.7) below, that $x_i^H Q_{22}^{(k)} \approx 0$.

The Lanczos process (7.2) gives, see (4.4), (6.2), and Theorem 5.1 for $H_{ij} \equiv H_{ij}^{(k)}$,

$$(T_k + H_{11}) S_k + H_{12} Q_{21} = S_k T_k + s_{k+1} \beta_{k+1} e_k^T, \tag{10.1}$$

$$(A + H_{22}) Q_{21} + H_{21} S_k = Q_{21} T_k + Q_{22} v_{k+1} \beta_{k+1} e_k^T, \tag{10.2}$$

where $Q_{22}^{(k)} v_{k+1}$ is then appended to $Q_{21}^{(k)}$ to give $Q_{21}^{(k+1)} = [Q_{21}^{(k)}, Q_{22}^{(k)} v_{k+1}]$, see (6.3), while $Q_{22}^{(k)} v_{k+1} v_{k+1}^H$ is subtracted from $Q_{22}^{(k)}$ to give $Q_{22}^{(k+1)} = Q_{22}^{(k)} \mathcal{P}_{k+1}$, see (6.5).

Now $\|Q_{22}^{(k)}\|_F^2$ decreases by $\|Q_{22}^{(k)} v_{k+1}\|_2^2$ each step, see (6.7), where with (9.3)

$$\|Q_{22}^{(k)} v_{k+1}\|_2^2 = \|X^H Q_{22}^{(k)} v_{k+1}\|_2^2 = \textstyle\sum_{i=1}^n |x_i^H Q_{22}^{(k)} v_{k+1}|^2, \tag{10.3}$$

and we now show that $|x_i^H Q_{22}^{(k)} v_{k+1}|^2$ is also the amount that $\|x_i^H Q_{21}^{(k)}\|_2^2$ increases and $\|x_i^H Q_{22}^{(k)}\|_2^2$ decreases each step. Using (6.3) and (6.6),

$$\|x_i^H Q_{21}^{(k+1)}\|_2^2 = \|x_i^H [Q_{21}^{(k)}, Q_{22}^{(k)} v_{k+1}]\|_2^2 = \|x_i^H Q_{21}^{(k)}\|_2^2 + |x_i^H Q_{22}^{(k)} v_{k+1}|^2, \tag{10.4}$$

$$\|x_i^H Q_{22}^{(k+1)}\|_2^2 = \|x_i^H Q_{22}^{(k)}\|_2^2 - |x_i^H Q_{22}^{(k)} v_{k+1}|^2, \tag{10.5}$$

so that $|x_i^H Q_{22}^{(k)} v_{k+1}|^2$ contributes to the decrease in both $\|x_i^H Q_{22}^{(k)}\|_2^2$ and $\|Q_{22}^{(k)}\|_F^2$. The proof of convergence in section 12 shows that $\|x_i^H Q_{22}^{(k)}\|_2 \searrow 0$ based on (10.5). In (12.10) we derive a lower bound on $|x_i^H Q_{22}^{(k)} v_{k+1}|$ to assess the rate of decrease.

If eventually $\|x_i^H Q_{22}^{(k)}\|_2^2 = 0$, then from (4.4), Theorem 8.2, and (8.10),

$$1 = \|x_i^H [Q_{21}^{(k)}, Q_{22}^{(k)}]\|_2^2 = \|x_i^H Q_{21}^{(k)}\|_2^2 + \|x_i^H Q_{22}^{(k)}\|_2^2, \tag{10.6}$$

$$x_i^H Q_{22}^{(k)} = 0 \Leftrightarrow \|x_i^H Q_{21}^{(k)}\|_2^2 = 1 \Leftrightarrow x_i \in \text{Range}(\widetilde{V}_3^{(k)}) \Leftrightarrow Q_{21}^{(k)H} x_i \in \text{Range}(P_3^{(k)}), \tag{10.7}$$

where these results hold when "=" is replaced by "≈" throughout this paragraph.

Fortunately, the desirable subspace $\mathrm{Range}(\widetilde{V}_3^{(k)})$ never decreases.

Lemma 10.1. *For $\widetilde{V}_3^{(k)}$ in Theorem 8.2* $\mathrm{Range}(\widetilde{V}_3^{(k)}) \subseteq \mathrm{Range}(\widetilde{V}_3^{(k+1)})$.

Proof. From (8.8)–(8.9) we see that $\mathrm{Range}(\widetilde{V}_3^{(k)}) = \mathcal{N}(Q_{22}^{(k)H})$, see also [34, (6.4)]. But $Q_{22}^{(k+1)} = Q_{22}^{(k)}(I_k - v_{k+1}v_{k+1}^H)$ in (6.5), so that $\mathrm{Range}(\widetilde{V}_3^{(k)}) \subseteq \mathrm{Range}(\widetilde{V}_3^{(k+1)})$. □

If $x_i \mathrel{\underset{\sim}{\in}} \mathrm{Range}(\widetilde{V}_3^{(m)})$ we now show that the eigenpair $\{\lambda_i, x_i\}$ of A is essentially available from all $k \geq m$ steps of the Lanczos process. We give the proof assuming $x_i \mathrel{\underset{\sim}{\in}} \mathrm{Range}(\widetilde{V}_3^{(m)})$. A proof with the more precise $x_i \in \mathrm{Range}(\widetilde{V}_3^{(m)})$ follows trivially.

Theorem 10.2. *For the Lanczos process (1.1) applied to A resulting in (7.2) with Theorem 4.1 and Definition 8.1, consider the eigensystems (3.1) and (9.3). Let $X_j \in \mathbb{U}^{n\times j}$ such that $AX_j = X_j\Lambda_j$ denote j columns of X where $X_j \mathrel{\underset{\sim}{\in}} \mathrm{Range}(\widetilde{V}_3^{(m)})$, see Theorem 8.2. For $k \geq m$ define $\widetilde{Y}_j \equiv \widetilde{Y}_j^{(k)} \triangleq Q_{21}^{(k)H}X_j$, $\widetilde{V}_3 \equiv \widetilde{V}_3^{(k)} \triangleq V_k P_3^{(k)}$, then*

$$\widetilde{Y}_j \triangleq Q_{21}^{(k)H}X_j \approx P_3^{(k)}\widetilde{V}_3^{(k)H}X_j \mathrel{\underset{\sim}{\in}} \mathrm{Range}(P_3^{(k)}), \quad \widetilde{Y}_j \mathrel{\underset{\sim}{\in}} \mathbb{U}^{k\times j}, \quad S_k\widetilde{Y}_j \approx 0, \tag{10.8}$$

$$X_j \approx V_k\widetilde{Y}_j \approx Q_{21}^{(k)}\widetilde{Y}_j \approx \widetilde{V}_3 P_3^{(k)H}\widetilde{Y}_j, \quad Q_{22}^{(k)H}X_j \approx 0, \quad \widetilde{Y}_j^H V_k^H V_k\widetilde{Y}_j \approx I_j, \tag{10.9}$$

$$[T_k - (Q_{21}^{(k)H}H_{22}^{(k)} + S_k^H H_{12}^{(k)})\widetilde{V}_3 P_3^H]\widetilde{Y}_j \approx \widetilde{Y}_j\Lambda_j, \quad \widetilde{Y}_j^H\widetilde{Y}_j \approx I_j, \quad \beta_{k+1}e_k^T\widetilde{Y}_j \approx 0, \tag{10.10}$$

so for $k \geq m$, $\{\Lambda_j, \widetilde{Y}_j^{(k)}\}$ are j converged backward stable eigenpairs of T_k.

It follows from this that Λ_j and $X_j \approx V_k\widetilde{Y}_j^{(k)}$ are essentially available from the Lanczos process, where $\{\Lambda_j, V_k\widetilde{Y}_j^{(k)}\}$ are j backward stable eigenpairs of A.

Proof. From Lemma 10.1 $X_j \mathrel{\underset{\sim}{\in}} \mathrm{Range}(\widetilde{V}_3^{(m)}) \Rightarrow X_j \mathrel{\underset{\sim}{\in}} \mathrm{Range}(\widetilde{V}_3^{(k)})$ for all $k \geq m$.

Now $\widetilde{Y}_j \triangleq Q_{21}^H X_j \approx P_3\widetilde{V}_3^H X_j$ from (8.10), so $\widetilde{Y}_j \mathrel{\underset{\sim}{\in}} \mathrm{Range}(P_3)$ and $S_k\widetilde{Y}_j \approx 0$ from (8.1). But then $\widetilde{V}_3^H X_j \approx P_3^H\widetilde{Y}_j$, and $X_j \approx \widetilde{V}_3\widetilde{V}_3^H X_j \approx \widetilde{V}_3 P_3^H\widetilde{Y}_j \approx V_k P_3 P_3^H\widetilde{Y}_j \approx V_k\widetilde{Y}_j$. Also $Q_{21}\widetilde{Y}_j = V_k(I - S_k)\widetilde{Y}_j \approx V_k\widetilde{Y}_j \approx X_j$, while $Q_{22}^H X_j \approx 0$ from (8.8)–(8.9), where $V_k\widetilde{Y}_j \approx X_j \in \mathbb{U}^{n\times j}$ and $\widetilde{Y}_j^H\widetilde{Y}_j \approx X_j^H\widetilde{V}_3\widetilde{V}_3^H X_j \approx I_j$ completes (10.8) and (10.9). Next applying X_j^H to the left of (10.2) and replacing $X_j^H Q_{21}$ by $\widetilde{Y}_j^H$ gives with (10.9)

$$\Lambda_j\widetilde{Y}_j^H + X_j^H(H_{22}Q_{21} + H_{21}S_k) \approx \widetilde{Y}_j^H T_k, \quad \widetilde{E}^{(k)} \triangleq (Q_{21}^{(k)H}H_{22}^{(k)} + S_k^H H_{12}^{(k)}),$$

$$\widetilde{Y}_j\Lambda_j \approx T_k\widetilde{Y}_j - \widetilde{E}^{(k)}X_j \approx [T_k - \widetilde{E}^{(k)}\widetilde{V}_3 P_3^H]\widetilde{Y}_j. \tag{10.11}$$

Therefore from Remark 3.7 $\{\Lambda_j, \widetilde{Y}_j\}$ are backward stable eigenpairs of T_k. Finally, multiplying (7.2) on the right by $\widetilde{Y}_j$ gives with (7.3), (10.8), (10.9), and (10.11),

$$\left(\begin{bmatrix} T_k & 0 \\ 0 & A \end{bmatrix} + H^{(k)}\right)\begin{bmatrix} 0 \\ X_j \end{bmatrix} \approx \begin{bmatrix} 0 \\ X_j \end{bmatrix}\Lambda_j + Q_1^{(k)}\widetilde{E}^{(k)}X_j + q^{(k+1)}\beta_{k+1}e_k^T\widetilde{Y}_j,$$

so with $AX_j = X_j\Lambda_j$, $\widetilde{E}^{(k)} \approx 0$, $Q_1^{(k)H}Q_1^{(k)} = I_k$, and $\|q^{(k+1)}\|_2 = 1$, this shows that $\beta_{k+1}e_k^T\widetilde{Y}_j \approx 0$, completing (10.10) and showing that $\{\Lambda_j, \widetilde{Y}_j\}$ are *converged* backward stable eigenpairs of T_k. Then $AV_j\widetilde{Y}_j \approx AX_j = X_j\Lambda_j \approx V_k\widetilde{Y}_j\Lambda_j$, which with $\widetilde{Y}_j^H V_k^H V_k\widetilde{Y}_j \approx I_j$ and Remark 3.7 proves backward stability of $\{\Lambda_j, V_k\widetilde{Y}_j^{(k)}\}$ for A. □

If an eigenpair $\{\mu_j^{(k)}, y_j^{(k)}\}$ of T_k has converged then in the exact case of (1.2) $AV_k y_j^{(k)} \approx V_k y_j^{(k)}\mu^{(k)}$, and $\{\mu_j^{(k)}, V_k y_j^{(k)}\}$ is a backward stable eigenpair for A. The computational Lanczos process modelled by (7.2) parallels this very nicely:

Corollary 10.3. *With $\mathcal{A}_k$ in* (5.1)*, if $\widetilde{V}_3^{(k)}$ in Theorem 8.2 has developed so that $x_j \underset{\approx}{\in} \mathrm{Range}(\widetilde{V}_3^{(k)})$ for some eigenvector x_j of A, $Ax_j = x_j\lambda_j$, then $\tilde{y}_j^{(k)} \triangleq Q_{21}^{(k)H} x_j$ satisfies $T_k\tilde{y}_j^{(k)} \approx \tilde{y}_j^{(k)}\lambda_j$ with $\beta_{k+1}e_k^T\tilde{y}_j^{(k)} \approx 0$, $\|\tilde{y}_j^{(k)}\|_2 \approx 1$, and $S_k\tilde{y}_j^{(k)} \approx 0$. Thus multiplying* (7.2) *on the right by $\tilde{y}_j^{(k)}$ gives*

$$\mathcal{A}_kQ_1^{(k)}\tilde{y}_j^{(k)} \approx Q_1^{(k)}\tilde{y}_j^{(k)}\lambda_j, \quad Q_1^{(k)}\tilde{y}_j^{(k)} = \begin{bmatrix} S_k \\ V_k(I-S_k)\end{bmatrix}\tilde{y}_j^{(k)} \approx \begin{bmatrix} 0 \\ V_k\tilde{y}_j^{(k)}\end{bmatrix}, \quad \|Q_1^{(k)}\tilde{y}_j^{(k)}\|_2 \approx 1.$$

Proof. Because $x_j \underset{\approx}{\in} \mathrm{Range}(\widetilde{V}_3^{(k)})$, all these results follow immediately from Theorem 10.2. Note how this parallels the ideal case, with $\mathcal{A}$ instead of A. □

11. Converged eigenpairs. The next definition arises from Theorem 10.2.

Definition 11.1. *For the Lanczos process described in Corollary 7.1 the eigenpair $\{\lambda_i, x_i\}$ of A has converged (really: "been converged to") if $x_i \underset{\approx}{\in} \mathrm{Range}(\widetilde{V}_3^{(k)})$.*

We now examine what converged eigenpairs of T_k mean for eigenpairs of A. For converged $\{\mu_i^{(k)}, y_i^{(k)}\}$ of T_k, multiplying (10.1) and (10.2) on the right by $y_i^{(k)}$:

$$(11.1) \quad T_ky_i^{(k)} = y_i^{(k)}\mu_i^{(k)}, \quad \|y_i^{(k)}\|_2 = 1, \quad T_kS_ky_i^{(k)} \approx S_ky_i^{(k)}\mu_i^{(k)}, \quad \beta_{k+1}e_k^Ty_i^{(k)} \approx 0,$$

$$(11.2) \quad AQ_{21}^{(k)}y_i^{(k)} = AV_k(I-S_k)y_i^{(k)} \approx V_k(I-S_k)y_i^{(k)}\mu_i^{(k)} = Q_{21}^{(k)}y_i^{(k)}\mu_i^{(k)}.$$

Definition 11.2. *A group of essentially equal eigenvalues of $M = M^H$ is "sufficiently separated" if they are separated by δ from their neighbours where δ is large enough so that if $My \approx y\mu$, $y^Hy = 1$, with μ essentially in this group, then we must have $y \underset{\approx}{\in}$ "the invariant subspace for this group". This requires $O(\epsilon)\|M\|/\delta \approx 0$, see, e.g.,* [8, §8.1.3]*, so this is a strong restriction on δ, but for T_k we remove it later.*

In the next theorem it should not be confusing if we use the same notation Y_t for the original Y_t and Y_t transformed from the right by some unitary transformation.

Theorem 11.3. *With* (3.1) *and the background and results of Corollary 7.1, suppose that $\mu_1^{(k)} \approx \cdots \approx \mu_t^{(k)}$ are sufficiently separated from the other eigenvalues of T_k and are converged so that* (11.1) *and* (11.2) *hold for $i = 1:t$. Then there exists a right unitary transformation of $Y_t \triangleq [y_1^{(k)}, \ldots, y_t^{(k)}]$ and a $(t-1)\times(t-1)$ upper triangular $R \equiv \{\rho_{ij}\} \equiv [r_1, \ldots, r_{t-1}]$ with nonnegative diagonal elements such that*

$$(11.3) \quad S_kY_t \approx Y_t\begin{bmatrix} 0 & R \\ 0 & 0\end{bmatrix}, \quad Y_t^HY_t = I_t, \quad S_ky_1 \approx 0, \quad S_ky_2 \approx y_1\rho_{11}, \quad 0 \le \rho_{11} \le 1,$$

$$(11.4) \quad AV_ky_1 \approx V_ky_1\mu_1^{(k)}, \quad y_1 \approx P_3^{(k)}P_3^{(k)H}y_1, \quad V_ky_1 \approx \widetilde{V}_3^{(k)}P_3^{(k)H}y_1, \quad \|V_ky_1\|_2 \approx 1,$$

where $\{\mu_1^{(k)}, V_ky_1\}$ is a backward stable eigenpair of A. Next with $Y_{2:t} \triangleq [y_2, y_3, \ldots, y_t]$

$$(11.5) \quad \mathcal{A}_kQ_1^{(k)}Y_t \approx Q_1^{(k)}Y_tM_t, \quad M_t \triangleq \mathrm{diag}(\mu_1^{(k)}, \ldots, \mu_t^{(k)}), \quad Y_t^HQ_1^{(k)H}Q_1^{(k)}Y_t = I_t,$$

$$(11.6) \quad \mathcal{A}_k \approx \begin{bmatrix} T_k & 0 \\ 0 & A\end{bmatrix}, \quad Q_1^{(k)}Y_t = \begin{bmatrix} S_k \\ V_k(I-S_k)\end{bmatrix}Y_t \approx \left[\begin{array}{c|c} 0 & Y_{t-1}R \\ V_ky_1 & V_k(Y_{2:t}-Y_{t-1}R)\end{array}\right],$$

$$(11.7) \quad (V_ky_1)^HV_ky_{i+1} \approx (V_ky_1)^HV_kY_{t-1}r_i, \quad 1 \approx \|r_i\|_2^2 + \|V_k(y_{i+1}-Y_{t-1}r_i)\|_2^2,$$

for $i = 1\!:\!t-1$. In particular (11.4) *with* (11.7) *for $i = 1$ gives*

$$\|V_k y_1\|_2 \approx 1,\ (V_k y_1)^H V_k y_2 \approx \rho_{11},\ 1 \approx \rho_{11}^2 + \|V_k y_2\|_2^2 + \rho_{11}^2 - 2\rho_{11}^2,\ \|V_k y_2\|_2 \approx 1. \tag{11.8}$$

Proof. Because T_k essentially has a t-dimensional invariant subspace corresponding to $\mu_1^{(k)} \approx \cdots \approx \mu_t^{(k)}$, from (11.1) we must have $S_k Y_t \underset{\sim}{\in} \text{Range}(Y_t)$. Therefore for any right unitary transformation of Y_t there exists $B = \{\beta_{ij}\} \in \mathbb{C}^{t\times t}$ such that $S_k Y_t \approx Y_t B$. We can take B to be upper triangular via the Schur decomposition, giving $S_k y_1 \approx y_1 \beta_{11}$ so that $\beta_{11} \approx 0$ because S_k has all eigenvalues 0. Next B can be unitarily transformed from the left with $t-1$ rotations to strictly upper triangular form with real nonnegative next to diagonal elements ρ_{ii}, giving upper triangular $R \equiv \{\rho_{ij}\} \equiv [r_1, \ldots, r_{t-1}]$ in (11.3). This with (11.2), Definition 8.1, and Theorem 8.2 gives (11.4), where from Remark 3.7 $\{\mu_1^{(k)}, V_k y_1\}$ is a backward stable eigenpair of A.

Combining (11.3) with Corollary 7.1 gives (11.5) and (11.6), while the orthonormality of the columns of $Q_1^{(k)} Y_t$ in (11.6) gives (11.7), and (11.8) follows. □

If one or more essentially equal eigenvalues of T_k have converged, then (11.4) shows that there is at least one backward stable eigenpair $\{\mu_1^{(k)}, V_k y_1\}$ of A, with $V_k y_1 \underset{\sim}{\in} \text{Range}(\widetilde{V}_3^{(k)})$, the requirement for convergence of an eigenvector of A in Theorem 10.2. But we have not found a proof of convergence based on such converged eigenpairs of T_k. However there are fascinating results for repeated eigenvalues of T_k.

Corollary 11.4 (Repeated eigenvalues). *With the background and results of Theorem 11.3 if $\mu_2^{(k)} \approx \cdots \approx \mu_t^{(k)}$ are repeats of $\mu_1^{(k)}$, so there is only one eigenvector of A corresponding to these t converged eigenvalues of T_k, then in* (11.6)

$$Q_1^{(k)} Y_t = \begin{bmatrix} S_k \\ V_k(I - S_k) \end{bmatrix} Y_t \approx \left[\begin{array}{c|c} 0 & Y_{t-1} \\ V_k y_1 & 0 \end{array}\right],\quad S_k Y_t \approx Y_t J_t,\quad J_t \triangleq \begin{bmatrix} 0 & I_{t-1} \\ 0 & 0 \end{bmatrix}, \tag{11.9}$$

$$S_k y_1 \approx 0;\quad S_k y_j \approx y_{j-1},\quad V_k y_j \approx V_k y_1,\quad j = 2\!:\!t;\quad Y_{2:t} \triangleq [y_2, \ldots, y_t], \tag{11.10}$$

$$y_1 \underset{\sim}{\in} \text{Range}(P_3^{(k)}),\ V_k y_1 \underset{\sim}{\in} \text{Range}(\widetilde{V}_3^{(k)}),\ \|V_k y_1^{(k)}\|_2 \approx 1,\ Y_{2:t} \underset{\sim}{\in} \text{Range}(P_1^{(k)}), \tag{11.11}$$

and the $\{\mu_j^{(k)}, V_k y_j^{(k)}\}$ are essentially identical backward stable eigenpairs for the one eigenpair of A.

Proof. Because there is only one eigenvector of A for these t eigenvectors of T_k, and $\|V_k y_1\|_2 \approx 1$, there exist scalars ζ_j, $j = 1\!:\!t-1$ in (11.6) such that $Q_1^{(k)} y_{j+1}$ gives

$$V_k y_1 \zeta_j \approx V_k y_{j+1} - V_k Y_{t-1} r_j,\quad j = 1\!:\!t-1;\quad V_k y_1(\zeta_1 + \rho_{11}) \approx V_k y_2. \tag{11.12}$$

With (11.8) this last gives $\zeta_1 \approx 0$, $\rho_{11} \approx 1$, $r_1 \approx e_1$, $V_k y_1 \approx V_k y_2$. Since it is true for $i = 1$, suppose that $\zeta_i \approx 0$ and $r_i \approx e_i$ in (11.12) for $i = 1\!:\!j-1 < t-1$, then in (11.6)

$$V_k y_1 \approx V_k y_2 \approx \cdots \approx V_k y_j,\quad Q_1^{(k)} Y_{j+1} \approx \left[\begin{array}{c|c|c} 0 & Y_{j-1} & Y_{t-1} r_j \\ V_k y_1 & 0 & V_k(y_{j+1} - Y_{t-1} r_j) \end{array}\right], \tag{11.13}$$

where orthonormality gives $r_j \approx e_j \rho_{jj}$ and so (11.12) gives $V_k y_1(\zeta_j + \rho_{jj}) \approx V_k y_{j+1}$. But then from (11.7), (11.8), (11.13) $(V_k y_1)^H V_k y_{j+1} \approx (V_k y_1)^H V_k y_j \rho_{jj} \approx \rho_{jj} \geq 0$ and

$$\zeta_j \approx 0,\quad 1 - \rho_{jj}^2 \approx \|V_k(y_{j+1} - y_1 \rho_{jj})\|_2^2 \approx \|V_k y_{j+1}\|_2^2 - \rho_{jj}^2,\quad \|V_k y_{j+1}\|_2 \approx 1,\quad \rho_{jj} \approx 1,$$

so that $V_k y_{j+1} \approx V_k y_1$. Therefore $\zeta_j \approx 0$, $r_j \approx e_j$, and $V_k y_{j+1} \approx V_k y_1$ for $j = 1\!:\!t-1$, proving (11.9) and (11.10). From (11.10) $\|S_k y_j\|_2 \approx \|y_{j-1}\|_2 = 1$, so (11.11) follows from Definition 8.1 and Theorem 8.2. The backward stability follows from (11.4), (11.13), and Remark 3.7. The results hold for $t=1$, except then there is no $Y_{2:t}$. $\square$

Remark 11.1. When there are repeats, (11.10) shows that $y_1,\ldots,y_t$ essentially form the start of a Jordan chain of principal vectors of S_k, see, *e.g.*, [42, §39, pp.42–3]. Therefore if there is a mix of different repeats and non-repeats in a converged group of close eigenvalues $\mu_1^{(k)}\approx\cdots\approx\mu_t^{(k)}$ of T_k, in theory there is a unitary transformation of Y_t that will group each chain in its correct order, leading to Jordan blocks of the form shown in (11.9), so that we do not require "sufficient separation" to split all the converged eigenvectors into their respective blocks. Each block starts with a $y_j \underset{\approx}{\in} \mathrm{Range}(P_3^{(k)})$ followed by its repeats, if any, in $\mathrm{Range}(P_1^{(k)})$, see (11.11). It follows that a converged eigenpair $\{\mu_j^{(k)}, y_j^{(k)}\}$ of T_k in (7.2) has $\mu_j^{(k)}$ essentially an eigenvalue of A, and it is either first converged, see Remark 3.6, with $y_j^{(k)} \underset{\approx}{\in} \mathrm{Range}(P_3^{(k)})$, or it is a repeat with $y_j^{(k)} \underset{\approx}{\in} \mathrm{Range}(P_1^{(k)})$, and in each case $\{\mu_j^{(k)}, V_k y_j^{(k)}\}$ is a backward stable eigenpair for A with $\|V_k y_j^{(k)}\|_2\approx 1$. $\square$

DEFINITION 11.5. *As soon as an eigenvalue $\mu_j^{(k)}$ in $\{\mu_j^{(k)}, y_j^{(k)}\}$ of T_k has first converged to an eigenvalue λ_i in $\{\lambda_i, x_i\}$ of A, we define $\{\mu_j^{(k)}, V_k y_j^{(k)}\}$ with $\|V_k y_j^{(k)}\|_2\approx 1$ as our approximation to $\{\lambda_i, x_i\}$, so that this common practice is ideal.*

To complete this section we show that if there are no repeats in $\mu_1\approx\cdots\approx\mu_t$ in Theorem 11.3 then there is a Y_t giving $V_kY_t \underset{\approx}{\in} \mathbb{U}^{n\times t}$.

COROLLARY 11.6 (Unrelated eigenvalues). *With the background and results of Theorem 11.3 if none of $\mu_2^{(k)}\approx\cdots\approx\mu_t^{(k)}$ are repeats, so there are exactly t converged eigenvectors of A corresponding to these t eigenvalues of T_k, then in (11.6) $R\approx 0$ and*

$$(11.14)\qquad Y_t^HY_t=I_t,\quad Y_t \underset{\approx}{\in} \mathrm{Range}(P_3^{(k)}),\quad V_kY_t \underset{\approx}{\in} \mathrm{Range}(\widetilde{V}_3^{(k)}),\quad V_kY_t \underset{\approx}{\in} \mathbb{U}^{n\times t},$$

$$(11.15)\qquad S_kY_t\approx 0,\quad Q_1^{(k)}Y_t=\begin{bmatrix} S_k \\ V_k(I-S_k)\end{bmatrix}Y_t\approx\begin{bmatrix} 0 & \ldots & 0 \\ V_ky_1 & \ldots & V_ky_t\end{bmatrix},$$

where each $\{\mu_j^{(k)}, V_k y_j^{(k)}\}$ is a backward stable eigenpair of A.

Proof. There are no repeats, so $\{\mu_j^{(k)}, y_j^{(k)}\}$, $j=1:t$, are all first converged, and from Remark 11.1 $Y_t \underset{\approx}{\in} \mathrm{Range}(P_3^{(k)})$, so $Y_t\approx P_3^{(k)}Z_t$ for some $Z_t\in\mathbb{U}^{n_k\times t}$. This gives $V_kY_t\approx V_kP_3^{(k)}Z_t=\widetilde{V}_3^{(k)}Z_t\in\mathbb{U}^{n\times t}$, completing (11.14). But $Y_t \underset{\approx}{\in} \mathrm{Range}(P_3^{(k)}) \Rightarrow S_kY_t\approx 0$, see Definition 8.1, proving (11.15). Multiplying (10.2) on the right by Y_t and using (11.14)–(11.15) with Remark 3.7 proves the backward stability. $\square$

The fact that the $\|V_ky_j\|_2\approx 1$ in Corollaries 11.4, 11.6, improves on Remark 3.5.

12. Convergence of the finite precision Lanczos process. Here we examine convergence and rate of convergence of the Lanczos process in the sense of Definition 11.1. We do this for each eigenvalue λ_i of A with x_i represented in v_1.

LEMMA 12.1. *For the Lanczos process (1.1) applied to A with initial unit-length vector v_1 resulting in (7.2) with (4.4), consider the eigensystem (9.3). If $|x_i^Hv_1|>0$ then $\|x_i^HQ_{21}^{(k)}\|_2\geq\|x_i^HQ_{21}^{(1)}\|_2=|x_i^Hv_1|>0$ and $\|x_i^HQ_{21}^{(k)}\|_2^2$ will increase, and $\|x_i^HQ_{22}^{(k)}\|_2^2$ decrease, strictly monotonically by $|x_i^HQ_{22}^{(k)}v_{k+1}|^2$ each step unless $x_i^HQ_{22}^{(k)}v_{k+1}=0$.*

Proof. Because $Q_{21}^{(1)}e_1=v_1$, if $|x_i^Hv_1|>0$, we see that $\|x_i^HQ_{21}^{(1)}\|_2=|x_i^Hv_1|>0$. Then (10.4)–(10.5) show that $\|x_i^HQ_{21}^{(k)}\|_2^2$ will increase and $\|x_i^HQ_{22}^{(k)}\|_2^2$ decrease by $|x_i^HQ_{22}^{(k)}v_{k+1}|^2$ each step unless $x_i^HQ_{22}^{(k)}v_{k+1}=0$. $\square$

We want to prove that $\|x_i^H Q_{22}^{(k)}\|_2^2 \searrow 0$ for each relevant eigenvector x_i of A, so from Lemma 12.1 we need to understand what $x_i^H Q_{22}^{(k)} v_{k+1} = 0$ implies. If $AX_i = X_i\lambda_i$, maximizing $\|\widetilde{V}_3^{(k)H} x_i\|_2 \le 1$, the length of the projection of x_i on $\mathrm{Range}(\widetilde{V}_3^{(k)})$ over $x_i \in \mathrm{Range}(X_i)$, $\|x_i\|_2 = 1$, maximizes $\|x_i^H Q_{21}^{(k)}\|_2^2$ and minimizes $\|x_i^H Q_{22}^{(k)}\|_2^2$.

THEOREM 12.2. *Assume the conditions in Lemma 12.1, where for each essentially multiple eigenvalue λ_i of A with $AX_i \approx X_i\lambda_i$ we take a single eigenvector $x_i \in \mathrm{Range}(X_i)$, $x_i^H x_i = 1$, that maximizes $\|\widetilde{V}_3^{(k)H} x_i\|_2 \le 1$. If $|x_i^H v_1| > 0$ then*

$$(12.1)\quad x_i^H Q_{22}^{(k)} v_{k+1} \approx 0 \;\Rightarrow\; T_k(Q_{21}^{(k)H} x_i) \approx (Q_{21}^{(k)H} x_i)\lambda_i; \quad \|Q_{21}^{(k)H} x_i\|_2 \ge |x_i^H v_1| > 0.$$

This shows that λ_i is essentially an eigenvalue of T_k, but does not prove that it is a converged *eigenvalue of T_k. If $x_i^H Q_{22}^{(j)} v_{j+1} \approx 0$ for $j = k, k+1$ then*

$$(12.2)\quad T_k(Q_{21}^{(k)H} x_i) \approx (Q_{21}^{(k)H} x_i)\lambda_i, \qquad \beta_{k+1} e_k^T (Q_{21}^{(k)H} x_i) \approx 0,$$

$$(12.3)\quad Q_{21}^{(k)H} x_i \underset{\approx}{\in} \mathrm{Range}(P_3^{(k)}), \quad \|Q_{21}^{(k)H} x_i\|_2 \approx 1, \quad x_i \underset{\approx}{\in} \mathrm{Range}(\widetilde{V}_3^{(k)}), \quad x_i^H Q_{22}^{(k)} \approx 0,$$

so that $\{\lambda_i, Q_{21}^{(k)H} x_i\}$ is a converged backward stable eigenpair of T_k. Then x_i has been converged to, and λ_i and x_i are essentially available from the Lanczos process.

Proof. Lemma 12.1 and multiplying (10.2) on the left by x_i^H gives (12.1), since

$$(12.4)\qquad \lambda_i x_i^H Q_{21}^{(k)} + x_i^H (H_{22} Q_{21}^{(k)} + H_{21} S_k) = x_i^H Q_{21}^{(k)} T_k + x_i^H Q_{22}^{(k)} v_{k+1} \beta_{k+1} e_k^T.$$

Now suppose that $x_i^H Q_{22}^{(j)} v_{j+1} \approx 0$ for $j = k, k+1$. Then from (6.3) and (12.4)

$$\begin{aligned} &x_i^H Q_{21}^{(k+1)} = [x_i^H Q_{21}^{(k)}, x_i^H Q_{22}^{(k)} v_{k+1}] \approx [x_i^H Q_{21}^{(k)}, 0], \\ &[\lambda_i x_i^H Q_{21}^{(k+1)} + x_i^H (H_{22}^{(k+1)} Q_{21}^{(k+1)} + H_{21}^{(k+1)} S_{k+1})] e_{k+1} \approx x_i^H Q_{21}^{(k+1)} T_{k+1} e_{k+1} \\ \approx\ & x_i^H (H_{22}^{(k+1)} Q_{21}^{(k+1)} + H_{21}^{(k+1)} S_{k+1}) e_{k+1} \approx x_i^H Q_{21}^{(k)} e_k \beta_{k+1} \approx 0, \end{aligned}$$

which with (12.1) proves (12.2). Now $Q_{21}^H x_i = (P_2 \Gamma_2 \widetilde{V}_2^H + P_3 \widetilde{V}_3^H) x_i \perp \mathrm{Range}(P_1)$ from (8.2) and (8.10), so from Remark 11.1 the converged eigenvector $Q_{21}^H x_i$ of T_k in (12.2) cannot be a repeat and must be first converged satisfying $Q_{21}^H x_i \underset{\approx}{\in} \mathrm{Range}(P_3^{(k)})$ as desired in (12.3), see also (11.4). Therefore we have from (8.10) with (12.1)

$$(12.5)\qquad Q_{21}^{(k)H} x_i \approx P_3^{(k)} \widetilde{V}_3^{(k)H} x_i, \quad Q_{21}^{(k)} Q_{21}^{(k)H} x_i \approx \widetilde{V}_3^{(k)} \widetilde{V}_3^{(k)H} x_i, \quad \|\widetilde{V}_3^{(k)H} x_i\|_2 > 0.$$

Multiplying (10.2) on the right by $Q_{21}^{(k)H} x_i$ and using (12.2) and (12.5) gives

$$(12.6)\qquad A \widetilde{V}_3^{(k)} \widetilde{V}_3^{(k)H} x_i \approx \widetilde{V}_3^{(k)} \widetilde{V}_3^{(k)H} x_i \lambda_i, \qquad \|\widetilde{V}_3^{(k)} \widetilde{V}_3^{(k)H} x_i\|_2 > 0.$$

Suppose there are $t \ge 1$ eigenvalues of A essentially equal to λ_i that are sufficiently separated from the rest, see Definition 11.2, so that $AX_i \approx X_i\lambda_i$, $X_i \in \mathbb{U}^{n\times t}$. It then follows from (12.6) that $\exists\, z \in \mathbb{C}^t$ such that $X_i z \approx \widetilde{V}_3^{(k)} \widetilde{V}_3^{(k)H} x_i \in \mathrm{Range}(\widetilde{V}_3^{(k)})$. But by definition x_i, $\|x_i\|_2 = 1$, maximizes $\|\widetilde{V}_3^{(k)H} x_i\|_2 \le 1$ over $x_i \in \mathrm{Range}(X_i)$. Therefore $x_i \underset{\approx}{\in} \mathrm{Range}(\widetilde{V}_3^{(k)})$, so that (12.3) follows from (8.8)–(8.9). Since $x_i \underset{\approx}{\in} \mathrm{Range}(\widetilde{V}_3^{(k)})$ it follows from Theorem 10.2 that x_i has been converged to, and λ_i and x_i are essentially available from the Lanczos process.

This is what we want, because we need only prove that at least one eigenvector $x_i \in \mathrm{Range}(X_i)$ with $|x_i^H v_1| > 0$ will converge, see Item 2 in section 9. ◻

The next theorem proves that the Lanczos process obtains first convergences.

Theorem 12.3. *Assume the conditions, and the choice of eigenvector for any essentially multiple eigenvalue, in Theorem 12.2. Then the computational Lanczos process modelled in (7.2)–(7.3) eventually makes available backward stable approximations to every such eigenpair $\{\lambda_i, x_i\}$ of A for which $|x_i^H v_1| > 0$. If A has distinct eigenvalues and $|x_i^H v_1| > 0$, $i = 1:n$, then $\|Q_{22}^{(k)}\|_F^2$ decreases monotonically until $Q_{22}^{(k)} \approx 0$, when all the eigenvalues of A will have been satisfactorily approximated by n eigenvalues of T_k, ensuring $\widetilde{V}_3^{(k)} \in \mathbb{U}^{n\times n}$ and with the notation in Definition 8.1 and Theorem 8.2, $\widehat{V}_0^{(k)}$, $\widetilde{V}_2^{(k)}$, $\widehat{V}_2^{(k)}$, $P_2^{(k)}$, and $W_2^{(k)}$ are nonexistent, while*

$$\widetilde{V}_3^{(k)} \in \mathbb{U}^{n\times n}, \quad Q_{22}^{(k)} = 0, \quad P^{(k)} = [P_1^{(k)}, P_3^{(k)}], \quad Q_{21}^{(k)} = \widetilde{V}_3^{(k)} P_3^{(k)H}. \tag{12.7}$$

Proof. Theorem 12.2 shows that for any such eigenvector x_i of A with $|x_i^H v_1| > 0$, two consecutive steps with $x_i^H Q_{22}^{(k)} v_{k+1} \approx 0$ imply that $x_i^H Q_{22}^{(k)} \approx 0$, see (12.3). Therefore from Lemma 12.1 we see that $\|x_i^H Q_{22}^{(k)}\|_2$ must decrease at least every second step until it is essentially zero, giving $x_i \underset{\approx}{\in} \mathrm{Range}(\widetilde{V}_3^{(k)})$, see (12.3). Then from Theorem 10.2 backward stable approximations to λ_i and x_i are available.

If A has n distinct eigenvalues each with $|x_i^H v_1| > 0$, then it necessarily follows that $\|X^H Q_{22}^{(k)}\|_F = \|Q_{22}^{(k)}\|_F$ decreases until every eigenvector x_i of A has been found and satisfies $x_i \underset{\approx}{\in} \mathrm{Range}(\widetilde{V}_3^{(k)})$. But this implies that $X \underset{\approx}{\in} \mathrm{Range}(\widetilde{V}_3^{(k)})$ in (9.3), so that $\widetilde{V}_3^{(k)} \in \mathbb{U}^{n\times n}$. This with Theorem 8.2 and Definition 8.1 shows that $n_k = n$, $\widehat{V}_0^{(k)}$, $\widetilde{V}_2^{(k)}$, $\widehat{V}_2^{(k)}$, $P_2^{(k)}$, and $W_2^{(k)}$ do not exist, so that $Q_{22}^{(k)} = 0$ and $P^{(k)} = [P_1^{(k)}, P_3^{(k)}]$. Then (8.10) completes (12.7). ◻

12.1. Rate of convergence. We assume $\beta_{k+1} > 0$ and use two measures of eigenvalue separation. Define $\delta_{i,j}^{(k)}$ for $i = 1:n$ and $\epsilon_{j,i}^{(k)}$ for $j = 1:k$ via (3.1) and (9.3):

$$\delta_{i,j}^{(k)} \triangleq |\lambda_i - \mu_j^{(k)}| \triangleq \min_{m=1:k} |\lambda_i - \mu_m^{(k)}|, \qquad \epsilon_{j,i}^{(k)} \triangleq |\lambda_i - \mu_j^{(k)}| \triangleq \min_{m=1:n} |\lambda_m - \mu_j^{(k)}|. \tag{12.8}$$

To depict how $\{\lambda_i, x_i\}$ is converged to, multiply (12.4) on the right by Y in (3.1):

$$x_i^H Q_{21} Y (\lambda_i I_k - M) = \left[x_i^H Q_{22} v_{k+1} \beta_{k+1} e_k^T - x_i^H (H_{21} S_k + H_{22} Q_{21})\right] Y, \quad i = 1:n. \tag{12.9}$$

Either $\delta_{i,j}^{(k)} = 0$ here, or $(\lambda_i I_k - M)$ is nonsingular, in which case

$$\begin{aligned} x_i^H Q_{21} Y &= \left[x_i^H Q_{22} v_{k+1} \beta_{k+1} e_k^T - x_i^H (H_{21} S_k + H_{22} Q_{21})\right] Y (\lambda_i I_k - M)^{-1}, \\ \|x_i^H Q_{21} Y\|_2 &\le \left\| \left[x_i^H Q_{22} v_{k+1} \beta_{k+1} e_k^T - x_i^H (H_{21} S_k + H_{22} Q_{21})\right] Y \right\|_2 / \delta_{i,j}^{(k)}. \end{aligned}$$

Either way we get a lower bound on the residual $|x_i^H Q_{22}^{(k)} v_{k+1}| \beta_{k+1}$ in (12.4)

$$|x_i^H Q_{22}^{(k)} v_{k+1}| \beta_{k+1} \ge \|x_i^H Q_{21}^{(k)}\|_2 \delta_{i,j}^{(k)} - \|x_i^H (H_{21}^{(k)} S_k + H_{22}^{(k)} Q_{21}^{(k)})\|_2. \tag{12.10}$$

Similarly, multiplying (10.2) on the right by $y_j^{(k)}$ in $T_k y_j^{(k)} = y_j^{(k)} \mu_j^{(k)}$ gives (12.11). Then multiplying this on the left by X^H gives a useful lower bound. For $j = 1 : k$,

$$
\begin{aligned}
&(12.11) \quad (A + H_{22}) Q_{21} y_j^{(k)} + H_{21} S_k y_j^{(k)} = Q_{21} y_j^{(k)} \mu_j^{(k)} + Q_{22} v_{k+1} \beta_{k+1} e_k^T y_j^{(k)}, \\
&\qquad\qquad (\Lambda - \mu_j^{(k)} I_n) X^H Q_{21} y_j^{(k)} = X^H \left[Q_{22} v_{k+1} \beta_{k+1} e_k^T y_j^{(k)} - (H_{21} S_k + H_{22} Q_{21}) y_j^{(k)} \right], \\
&(12.12) \quad \| Q_{22}^{(k)} v_{k+1} \|_2 \beta_{k+1} | e_k^T y_j^{(k)} | \geq \| Q_{21}^{(k)} y_j^{(k)} \|_2 \epsilon_{j,i}^{(k)} - \| (H_{21}^{(k)} S_k + H_{22}^{(k)} Q_{21}^{(k)}) y_j^{(k)} \|_2.
\end{aligned}
$$

We saw in Theorem 10.2 that if $\widetilde{V}_3^{(k)}$ has developed so that $x_i \underset{\sim}{\in} \text{Range}(\widetilde{V}_3^{(k)})$, then x_i has been accurately approximated by the Lanczos process, where from (8.8)–(8.10) $\|x_i^H Q_{21}^{(k)}\|_2 \approx 1$ and $x_i^H Q_{22}^{(k)} \approx 0$. So although initially $\|x_i^H Q_{21}^{(k)}\|_2 \geq |x_i^H v_1|$ in (12.10) could be small, it will essentially equal one by the time $x_i \underset{\sim}{\in} \text{Range}(\widetilde{V}_3^{(k)})$.

Until the first eigenpair converges we will have $S_k \approx 0$, and since from (4.4) $\|S_k y_j^{(k)}\|_2^2 + \|Q_{21}^{(k)} y_j^{(k)}\|_2^2 = 1$, this gives $\|Q_{21}^{(k)} y_j^{(k)}\|_2^2 \approx 1$ in (12.12). Note with (8.10) that $Q_{21}^{(k)} y_j^{(k)} = (\widetilde{V}_2 \Gamma_2 P_2^H + \widetilde{V}_3 P_3^H) y_j^{(k)}$, where all first converged $y_j^{(k)} \underset{\sim}{\in} \text{Range}(P_3^{(k)})$, see Remark 11.1, giving $\|Q_{21}^{(k)} y_j^{(k)}\|_2 \approx 1$ for these in (12.12), while $y_j^{(k)} \underset{\sim}{\in} \text{Range}(P_1^{(k)})$ for all repeats, giving $\|Q_{21}^{(k)} y_j^{(k)}\|_2 \approx 0$ for such repeats.

Initially $\|Q_{22}^{(k)}\|_F^2$ decreases and $\|Q_{21}^{(k)}\|_F^2$ increases by about 1 per step until the first eigenpair of A converges, see Remark 3.4, but after that (12.10) and (12.12) can give insight on such changes. Until then, $Q_{21}^{(k)H} Q_{21}^{(k)} \approx V_k^H V_k \approx I_k$.

After that, we need bounds on how $\|x_i^H Q_{22}^{(k)}\|_2 \searrow 0$. First, (12.10) shows that

$$
|x_i^H Q_{22}^{(k)} v_{k+1}| \geq (\|x_i^H Q_{21}^{(k)}\|_2 \delta_{i,j}^{(k)} - \|x_i^H (H_{21}^{(k)} S_k + H_{22}^{(k)} Q_{21}^{(k)})\|_2) / \beta_{k+1},
$$

where $0 < \beta_{k+1} \leq \|A\|_2$. Thus because $0 < |x_i^H v_1| \leq \|x_i^H Q_{21}^{(k)}\|_2 \nearrow 1$, see Lemma 12.1 and (10.4), if there are no $\mu_j^{(k)}$ close to λ_i, see (12.8), $|x_i^H Q_{22}^{(k)} v_{k+1}|$ will be significant, and cause a significant decrease in $\|x_i^H Q_{22}^{(k)}\|_2$, see (10.5).

Alternatively, it can be seen from (12.4) that $|x_i^H Q_{22}^{(k)} v_{k+1}| \beta_{k+1}$ is essentially the norm of the residual when taking $\{\lambda_i, Q_{21}^H x_i\}$ as an approximate eigenpair of T_k, and so the larger this residual, the larger the decrease $|x_i^H Q_{22}^{(k)} v_{k+1}|^2$ in $\|x_i^H Q_{22}^{(k)}\|_2^2$ will tend to be. So usually convergence will be good, however not always, as we now argue.

Remark 12.1. The bound (12.10) gives a possible explanation for the slowness seen in Example 6.1. When λ_i is one of a group of very close eigenvalues of A, $\delta_{i,j}^{(k)}$ can be small, not because of the closeness of the $\mu_j^{(k)}$ of T_k that will eventually converge to λ_i, but because of the closeness of λ_i to other eigenvalues of T_k that have already converged to close neighbours of λ_i. Small $\delta_{i,j}^{(k)}$ might allow $|x_i^H Q_{22}^{(k)} v_{k+1}|$ to be unusually small, slowing the convergence to λ_i. This dependence on closeness of eigenvalues would make it difficult, or impossible, to predict the rate of convergence in general, so the best we can do here is to prove the convergence as in Theorem 12.3, and indicate the possible rates of convergence as in (12.10) and (12.12). □

Another approach is to notice from (12.11) that $Q_{22}^{(k)} v_{k+1} \beta_{k+1} e_k^T y_j^{(k)}$ is essentially the residual when taking $\{\mu_j^{(k)}, Q_{21}^{(k)} y_j^{(k)}\}$ as an approximate eigenpair of A. This is bounded in (12.12), where if $\mu_j^{(k)}$ has first converged to some eigenvalue of A, then $Q_{21}^{(k)} y_j^{(k)} \approx V_k y_j^{(k)}$ and $\|Q_{21}^{(k)} y_j^{(k)}\|_2 \approx 1$, see Corollary 10.3. We see from (12.12) that if $\mu_j^{(k)}$ is not close to any eigenvalue of A, then the decrease $\|Q_{22}^{(k)} v_{k+1}\|_2^2$ in $\|Q_{22}^{(k)}\|_F^2$, see (6.7), will be significant. But as in Remark 12.1 the rate of decrease could be slow if $\mu_j^{(k)}$ was converging to an eigenvalue in a group of close eigenvalues of A.

13. Accuracy of the Lanczos process for the eigenproblem when $Q_{22}^{(k)}\!=\!0$. Theorem 12.3 showed that when A has distinct eigenvalues and $|x_i^H v_1| > 0$, $i\!=\!1\!:\!n$, then $\|Q_{22}^{(k)}\|_F^2$ decreases until $Q_{22}^{(k)} = 0$. Here we give some new results and repeat some of those in sections 10 to 12, because $Q_{22}^{(k)} = 0$ quickly leads to very clean results.

In Theorem 12.3 $\widetilde{V}_3^{(k)} \in \mathbb{U}^{n\times n} \Rightarrow Q_{22}^{(k)} = 0$, $Q_{21}^{(k)} = \widetilde{V}_3^{(k)} P_3^{(k)H}$, $P^{(k)} = [P_1^{(k)}, P_3^{(k)}]$. From Remark 8.1 and (8.10) we can choose $P_3^{(k)} = \widetilde{P}_3^{(k)}$ so that for this $\widetilde{V}_3^{(k)} \in \mathbb{U}^{n\times n}$,

$$(13.1)\qquad \widetilde{V}_3^{(k)} = V_k \widetilde{P}_3^{(k)} = I_n, \qquad Q_{21}^{(k)} \equiv V_k(I - S_k) = \widetilde{V}_3^{(k)} \widetilde{P}_3^{(k)H} = \widetilde{P}_3^{(k)H}.$$

It then follows from (8.1) and (8.11) that $S_k = W_1^{(k)} P_1^{(k)H}$, $s_{k+1} = W_3^{(k)} \widehat{V}_3^{(k)H} v_{k+1}$, while from (6.2) $v_{k+1} - V_k s_{k+1} = Q_{22}^{(k)} v_{k+1} = 0$, so (7.2)–(7.3) give at step k

$$(13.2)\quad \left(\begin{bmatrix} T_k & 0 \\ 0 & A \end{bmatrix} + \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}\right) \begin{bmatrix} W_1 P_1^H \\ \widetilde{P}_3^H \end{bmatrix} = \begin{bmatrix} W_1 P_1^H \\ \widetilde{P}_3^H \end{bmatrix} T_k + \begin{bmatrix} W_3 \widehat{V}_3^H v_{k+1} \\ 0 \end{bmatrix} \beta_{k+1} e_k^T.$$

From the bottom row we use $P = [P_1, \widetilde{P}_3] \in \mathbb{U}^{k\times k}$ to derive the following results:

$$\begin{aligned} &(A\!+\!H_{22})\widetilde{P}_3^H = \widetilde{P}_3^H T_k - H_{21} W_1 P_1^H, && \widetilde{P}_3^H T_k P_1 = H_{21} W_1, \\ &T_k \widetilde{P}_3 = \widetilde{P}_3(A + H_{22}) + P_1 W_1^H H_{12}, && (A\!+\!H_{22}) = (\widetilde{P}_3^H T_k \widetilde{P}_3), \end{aligned}$$

$$(13.3)\quad (T_k - P_1 W_1^H H_{12} \widetilde{P}_3^H)\widetilde{P}_3 = \widetilde{P}_3(A + H_{22}),$$

$$(13.4)\quad P^H T_k P = \left[\begin{array}{c|c} P_1^H T_k P_1 & W_1^H H_{12} \\ H_{21} W_1 & \widetilde{P}_3^H T_k \widetilde{P}_3 \end{array}\right] = \left[\begin{array}{c|c} P_1^H T_k P_1 & W_1^H H_{12} \\ H_{21} W_1 & A\!+\!H_{22} \end{array}\right] \approx \left[\begin{array}{c|c} P_1^H T_k P_1 & 0 \\ \hline 0 & A\!+\!H_{22} \end{array}\right],$$

$$T_k P_1 = P P^H T_k P_1 = P_1(P_1^H T_k P_1) + \widetilde{P}_3 H_{21} W_1,$$

$$(13.5)\quad (T_k - \widetilde{P}_3 H_{21} W_1 P_1^H) P_1 = P_1(P_1^H T_k P_1).$$

Because $H_{21} W_1 \approx 0$ in (13.4), the eigenvalues of T_k can be split into two groups, the n that are essentially the eigenvalues of $\widetilde{P}_3^H T_k \widetilde{P}_3 = A + H_{22}$, and the $k - n$ that are essentially the eigenvalues of $P_1^H T_k P_1$. The eigenvalues of $\widetilde{P}_3^H T_k \widetilde{P}_3$ are exactly all of the eigenvalues of $A\!+\!H_{22}$, but also essentially n of the eigenvalues of T_k.

With (13.1), (6.3) with $Q_{22}^{(k)} = 0$ gives

$$(13.6)\qquad \widetilde{P}_3^{(k+1)H} = Q_{21}^{(k+1)} = \begin{bmatrix} Q_{21}^{(k)} & 0 \end{bmatrix} = \begin{bmatrix} \widetilde{P}_3^{(k)H} & 0 \end{bmatrix},$$

so for $j\!>\!k$ the matrix of eigenvectors $\widetilde{P}_3^{(j)}$ in (13.3) is not meaningfully changed from $\widetilde{P}_3^{(k)}$, while the eigenvalues of $\widetilde{P}_3^{(j)H} T_j \widetilde{P}_3^{(j)} = A\!+\!H_{22}^{(j)}$ are essentially the same.

It seems that the roles of A and T_k have been reversed in (13.3), but on completion of the exact case we have $AV_\ell = V_\ell T_\ell$ for some $\ell \le n$, where this can also be written $T_\ell Z_\ell = Z_\ell A$ with $Z_\ell \triangleq V_\ell^H$, so that this "completed" finite precision case (13.3) in some sense parallels the exact completed case.

The next development might seem strange because we are showing what is available from the Lanczos process, not how to compute it. Consider the eigensystem of $(A+H_{22})$: $(A+H_{22})\widetilde{X} = \widetilde{X}\tilde{\Lambda}$ where $\tilde{\Lambda} \triangleq \mathrm{diag}(\tilde{\lambda}_1,\ldots,\tilde{\lambda}_n)$ and $\widetilde{X} \in \mathbb{U}^{n\times n}$. Define $\widetilde{Y} \triangleq \widetilde{P}_3\widetilde{X} \in \mathbb{U}^{k\times n}$; then from (13.3) with $\widetilde{T}_k \triangleq T_k - P_1W_1^HH_{12}\widetilde{P}_3^H$ and $V_k\widetilde{P}_3 = I_n$,

$$\widetilde{T}_k\widetilde{Y} = \widetilde{T}_k\widetilde{P}_3\widetilde{X} = \widetilde{P}_3(A+H_{22})\widetilde{X} = \widetilde{P}_3\widetilde{X}\tilde{\Lambda} = \widetilde{Y}\tilde{\Lambda}, \tag{13.7}$$

$$(A+H_{22})\widetilde{X} = \widetilde{X}\tilde{\Lambda}, \quad V_k\widetilde{Y} = V_k\widetilde{P}_3\widetilde{X} = \widetilde{X}, \quad (A+H_{22})V_k\widetilde{Y} = V_k\widetilde{Y}\tilde{\Lambda}. \tag{13.8}$$

Therefore, for the computed T_k once $Q_{22}^{(k)} = 0$, n backward stable eigenpairs $\tilde{\Lambda}$ and $\widetilde{Y}$ of T_k lead to a backward stable eigendecomposition of A. This has a parallel format to the computational solution, where $\widetilde{Y}$ and $\tilde{\Lambda}$ would be computed from T_k, and then $V_k\widetilde{Y}$ formed to give the eigenvector matrix $\widetilde{X}$.

So when $Q_{22}^{(k)} = 0$, a complete backward stable eigendecomposition of A is available from the computational Lanczos process.

In (13.3) and (13.5) we essentially have two eigensubspaces of T_k, Range$(P_1^{(k)})$ and Range$(\widetilde{P}_3^{(k)})$ where $P^{(k)} = [P_1^{(k)}, \widetilde{P}_3^{(k)}] \in \mathbb{U}^{k\times k}$. The n first converged eigenvectors of T_k essentially lie in Range$(\widetilde{P}_3^{(k)})$, see Remark 11.1, while all the others including the converged repeats essentially lie in Range$(P_1^{(k)})$.

All converged eigenvalues of T_k are essentially eigenvalues of A, see Remark 3.3. It follows that the eigenvalues of $P_1^HT_kP_1$ that have converged, see (13.4) and (13.5), must essentially be repeats of those of $\widetilde{P}_3^HT_k\widetilde{P}_3$, *i.e.*, superfluous, but not misleading.

14. Accuracy of the Lanczos process for solving systems of equations.

For solving linear systems $Ax=b$ with Hermitian positive definite A, in theory the method of Conjugate Gradients (CG) [15] is equivalent to taking $v_1\beta_1 = b$, $v_1^Hv_1 = 1$, and computing approximations $x_k = V_kz_k$, where $T_kz_k = e_1\beta_1$ with T_k and V_{k+1} coming from the Lanczos process $AV_k = V_kT_k + v_{k+1}\beta_{k+1}e_k^T$. This gives the residual

$$r_k \triangleq b - Ax_k = b - AV_kz_k = b - V_kT_kz_k - v_{k+1}\beta_{k+1}e_k^Tz_k = -v_{k+1}\beta_{k+1}e_k^Tz_k, \tag{14.1}$$

and in theory this is zero no later than the n-th step. In practice we would stop if, *e.g.*, $\|r_k\|_2 = \beta_{k+1}|e_k^Tz_k| \le O(\epsilon)(\|A\|_2\|x_k\|_2 + \|b\|_2)$, or earlier.

The above version is all we need here, but note that instead of solving $T_kz_k = e_1\beta_1$ explicitly, the computational method defines $y_k \triangleq L_k^Tz_k$, thereby allowing it to sequentially factorize T_k, carry out two forward solves, and form x_k in the sequence

$$T_k = L_kL_k^T, \quad L_kC_k^H = V_k^H, \quad L_ky_k = e_1\beta_1, \quad x_k = C_ky_k \ (= V_kT_k^{-1}e_1\beta_1), \tag{14.2}$$

where L_k is lower bidiagonal. We call this the "Lanczos-CG" method. For it to be certain of working in practice we require T_k to be positive definite. Theorem 17.1 in the Appendix shows that for a good finite precision implementation of the Lanczos process such as (1.1) the eigenvalues of T_k essentially lie between the extreme eigenvalues of A, so that T_k will be positive definite if A is sufficiently positive definite.

Assuming $Q_{22}^{(k)} = 0$, we give an analysis of the Lanczos process for use in solving $Ax = b$, $A^H = A$, whether A is positive definite or not. From (13.1), (13.2), and (4.2)

$$\widetilde{P}_3v_1 = (I-S_k)^HV_k^Hv_1 = (I-S_k)^H(I+U_k+U_k^H)e_1 = (I-S_k)^H(I+U_k^H)e_1 = e_1. \tag{14.3}$$

At step k let $\tilde{x}_k$ be the solution of

$$(A+H_{22})\tilde{x}_k = b = v_1\beta_1. \tag{14.4}$$

Multiply (13.3) on the right by $\tilde{x}_k$, define $\tilde{z}_k \triangleq \widetilde{P}_3\tilde{x}_k$, and use $\widetilde{P}_3 v_1 = e_1$ from (14.3) and $\widetilde{V}_3 = V_k\widetilde{P}_3 = I_n$ from (13.1), to give

$$(14.5) \qquad (T_k - P_1W_1^H H_{12}\widetilde{P}_3^H)\tilde{z}_k = e_1\beta_1, \quad \tilde{z}_k \triangleq \widetilde{P}_3\tilde{x}_k, \quad \tilde{x}_k = V_k\widetilde{P}_3\tilde{x}_k = V_k\tilde{z}_k.$$

In the exact Lanczos-CG case (14.1) we took $x_k = V_k z_k$, where $T_k z_k = e_1\beta_1$. Here we also have $\tilde{x}_k = V_k\tilde{z}_k$, where $\tilde{z}_k$ in (14.5) is a backward stable solution to $T_k z_k = e_1\beta_1$. Thus from (14.4) $\tilde{x}_k = V_k\tilde{z}_k$ is seen to be a backward stable solution to $Ax = b$ where $\tilde{z}_k$ is a backward stable solution to $T_k z_k = e_1\beta_1$, and this is as good as can be expected with finite precision.

It is important to realize that this proof assuming $Q_{22}^{(k)} = 0$ did not require A to be positive definite, or even nonsingular. All it required was a solution to (14.4). In fact if $A + H_{22}$ is singular and $\tilde{x}_k$ is the minimum norm solution to (14.4), then $\tilde{z}_k$ would essentially be the minimum norm solution to (14.5). Therefore the Lanczos process makes available backward stable solutions to all compatible systems $Ax = b$ with $A^H = A$ in this case.

Unlike the proofs for eigenvalues in section 12, the proof here for solving $Ax = b$ assumed $Q_{22}^{(k)} = 0$, and so ignored the case of possible multiple eigenvalues. But if a multiple eigenvalue λ_i has an eigensubspace spanned be the columns of X_i, $X_i^H X_i = I$, the Lanczos process need only converge to $X_iX_i^H b$ for this subspace, and this is presumably what happens for each multiple eigenvalue in Theorem 12.2, so this would presumably lead to a proof for convergence where there are multiple eigenvalues.

15. Lanczos-CG for $Ax = b$. Example 15.1 shows how slow $\|Q_{22}^{(k)}\|_F^2$ can be in decreasing to zero, even though convergence to the solution is fast, see Figure 15.1.

EXAMPLE 15.1. *The matrix* $A = \text{gallery}('\text{wathen}', 20, 20)$; *in Matlab is* 1281×1281 *symmetric positive definite with random elements. It has no multiple eigenvalues, but several are equal to the 5th figure. We took random x elements in* $[-1, 1]$ *and*

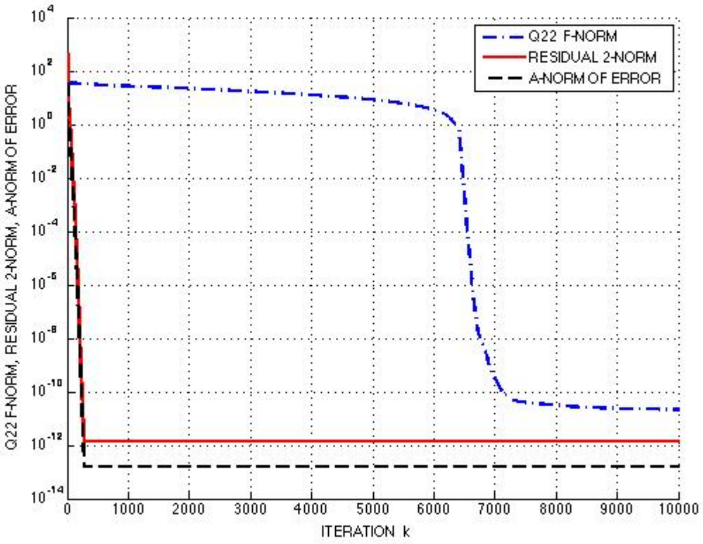

FIG. 15.1. *Plots of* $\|x - x_k\|_A$, $\|b - Ax_k\|_2$, *and* $\|Q_{22}^{(k)}\|_F$, $A = A^T \in \mathbb{R}^{1281\times1281}$.

$b \triangleq A * x$. Matlab's condest(A) *computed a lower bound for the 1-norm condition number of A of about 1200, with $\|x\|_2 = 20.4013$, and $\|b\|_2 = 1768.6$. Lanczos-CG solving $Ax = b$ gave $\|x - x_k\|_A = 1.54*10^{-13}$ and the residual $\|b - Ax_k\|_2 = 1.40*10^{-12}$ in 279 steps, and stayed there, while $\|Q_{22}^{(k)}\|_F^2$ decreased from 1281 to 1037 in 279 steps, where the ideal value would be $1281 - 279 = 1002$. The value of $\|Q_{22}^{(k)}\|_F$ decreased extremely slowly, only reaching 10^{-10} at step 7123. Of course this is only an indication of the theoretical $\|Q_{22}^{(k)}\|_F$ value, because the computed $\|v_j\|_2$ are not precisely 1, and because of the rounding errors in computing $\|Q_{22}^{(k)}\|_2$. This very slow, then late but relatively quick decrease is not at all well understood, and is one of several properties requiring more research. Even so, a satisfying solution x_k was found quite quickly.*

With full reorthogonalization the Lanczos-CG solution stopped improving at $k = 261$ with $\|x - x_k\|_A = 1.64 * 10^{-13}$ and the residual $\|b - Ax_k\|_2 = 1.39 * 10^{-12}$, and these stayed there, while $\|Q_{22}^{(k)}\|_F^2$ decreased from 1281 to the correct 1020 at $k = 261$. Thus Lanczos-CG obtained just as accurate a solution as with full reorthogonalization in only 18 more steps, even though $\|Q_{22}^{(k)}\|_F^2$ would not be zero in the ideal process until $k = 1281$, and not essentially zero until much later in the computational process.

16. Practical computations based on the Lanczos process. In practice we would like to provide some preprocessing of the problem such as preconditioning of A, or choice of v_1, in order to obtain the desired solution in a reasonable number of steps. But as long as Corollary 7.1 still holds, everything in this paper will apply.

The analysis shows that the finite precision Lanczos process does make available solutions that are backward stable. It is then up to the remaining computations in any method to obtain these. For example different solution of equations methods solve something like $T_k z_k = e_1\beta_1$ and compute something like $x_k = V_k z_k$ in different ways, see for example (14.2). Previously the Lanczos process was considered to be the weak part of such methods, and the remaining computations were considered to be faultless in comparison. Now we see that the analyses of the remaining computations should be included to show each overall method is backward stable.

For the eigenproblem of A that seems straightforward. We can find the eigenvalues of T_k in a backward stable manner, and can tell which eigenvectors have converged. So as in Definition 11.5 we would take the k for which $\mu_j^{(k)}$ has first converged, then $\|V_k y_j^{(k)}\|_2 \approx 1$ and $\{\mu_j^{(k)}, V_k y_j^{(k)}\}$ is a backward stable eigenpair for A.

The analysis for solution of equations is less obvious. The analysis (14.3)–(14.5) did not require A or T_k to be positive definite, it only required a solution of (14.4), showing that the Lanczos process can be used to solve any compatible system of equations with $A^H = A$, so that is not a difficulty. If A is not positive definite then T_k could be singular for some k, and other methods than (14.2) are needed, such as SYMMLQ [30] or MINRES-QLP [4]. The analyses of these could be combined with the analysis of the Lanczos process to certify the overall methods. We would also like to show that CG as implemented in [15] is backward stable.

Because of the close relationships between Golub-Kahan bidiagonalization (GKB) [7] for general non-square matrices and Lanczos tridiagonalization for Hermitian matrices, a variant of the analysis can presumably be used to prove that methods based on the GKB are equally successful. Problems with skew symmetric matrices can be handled via both the Lanczos process and GKB, see [12], and can also be analyzed.

17. Comments and summary. The papers [24, 25, 34] come in a sequence, each built on the earlier ones, and all leading to this one, which uses those results to prove the reliability and convergence of the Lanczos process for the eigenproblem and solution of equations. We have shown in Theorem 12.3 that the finite precision Lanczos process on a Hermitian matrix A essentially makes available a backward stable eigenpair of A for every eigenvalue λ_i of A with $|x_i^H v_1|>0$, and in sections 13 and 14 that when A has discrete eigenvalues having $|x_i^H v_1|>0$, $i=1:n$, the process behaves very like the exact process in that eventually T_k makes available a complete set of eigenpairs of A, or the solution of $Ax=b$, in a backward stable manner. But because of the possibility of deriving many repeats of eigenvalues of A, the finite precision Lanczos process can take many more than the ideal number of steps. Many have suspected this for many years, so it is pleasing to see that it is true.

It would be nice to give a more simple derivation of the results in section 11, and to obtain a greater understanding of how slow the convergence can be, see Remark 12.1. When the relevant analyses have been done it might be useful to show what properties of the exact processes hold for the finite precision processes. For example it follows from [15], see [12, (12.1)–(12.3)], that in theory, *i.e.*, with exact arithmetic,

$$\|x-x_k^{LSQR}\|_2\le\|x-x_k^{LSMR}\|_2,\ \ \|r_k^{LSQR}\|_2\le\|r_k^{LSMR}\|_2,\ \ \|A^T r_k^{LSMR}\|_2\le\|A^T r_k^{LSQR}\|_2$$

for solutions x_k and residuals $r_k \triangleq b-Ax_k$ of $\min_x\|b-Ax\|_2$ where A has full column rank, using LSQR [31] and LSMR [6]. Do these still hold with finite precision? What optimality properties of such methods still hold?

Of course a most useful topic will be to turn the knowledge gained here into practical computational advantage, perhaps by transforming the original problem and/or by developing improved computational algorithms.

17.1. Summary of the finite precision convergence. This is a brief summary of some of the more important theoretical and experimental observations.

1. There may be several eigenvalues of T_k for any one eigenvalue of A, but every converged eigenvalue of T_k is essentially an eigenvalue of A, and eigenvalues of the developing T_k never lose their level of convergence. The Lanczos process is always on track for the eigenproblem, the accuracy of approximation is only limited by the slowly growing size of the backward error $H^{(k)}$ in (5.1).
2. $\|Q_{22}^{(k)}\|_F^2$ will decrease by approximately 1 each step until the first eigenpair $\{\lambda_i, x_i\}$ of A has been found, at which point orthogonality can be lost.
3. Once orthogonality has been lost $\|Q_{22}^{(k)}\|_F^2$ will usually start to decrease at a slower rate, but will continue decreasing until all eigenpairs $\{\lambda_i, x_i\}$ of A corresponding to distinct eigenvalues whose eigenvectors are not orthogonal to v_1, have been found. Rounding errors will usually extend this to $x_i \perp v_1$.
4. If $|x_i^H v_1|>0$ then $\|x_i^H Q_{22}^{(k)}\|_F^2 \searrow 0$ until $x_i^H Q_{22}^{(k)} \approx 0$, at which point backward stable approximations to λ_i and x_i are available, see Theorem 12.3. But once orthogonality is lost, the rate that $\|x_i^H Q_{22}^{(k)}\|_F^2 \searrow 0$ is problem dependent. The decrease can be very slow if λ_i is one of several very close eigenvalues, and finding a general lower bound on the rate of decrease would be a daunting task, if at all possible. One key point is that when it gets there, the Lanczos process never converges to wrong answers, see section 5.1.
5. If A has no multiple eigenvalues then $\|Q_{22}^{(k)}\|_F^2 \searrow 0$, possibly very slowly.
6. If A has r repeated eigenvalues then for $k\ge$ some j, $Q_{22}^{(k)}$ seems to stagnate at $Q_{22}^{(j)} \approx \widetilde{V}_2^{(j)}\Sigma_2^{(j)}\widehat{V}_2^{(j)H}$ with $\Sigma_2^{(j)}\in\mathbb{R}^{r\times r}$. Then while Range$(\widetilde{V}_3^{(j)})$ is the eigensubspace for the distinct eigenvalues of A, Range$(\widetilde{V}_2^{(j)}) \perp$ Range$(\widetilde{V}_3^{(j)})$ appeared to be the eigensubspace for the repeated eigenvalues of A.

7. Lanczos-CG can converge in $\ll$ or $\gg n$ steps for $Ax = b$, see section 15.

Appendix. This clarifies some ideas. First, what is a “good” implementation? The general step for $k>1$ of algorithm (1.1) is mathematically equivalent to

$$
\begin{aligned}
&w_k := Av_k - v_k\alpha_k - v_{k-1}\gamma_k, \quad \alpha_k := v_k^H A v_k, \quad \gamma_k := \beta_k,\\
&\beta_{k+1} := +(w_k^H w_k)^{1/2}, \qquad v_{k+1} := w_k/\beta_{k+1}.
\end{aligned} \tag{17.1}
$$

However to obtain orthogonality, we could instead have taken $\gamma_k \triangleq v_{k-1}^H A v_k$. It was shown in [21, §7.3 & §9.3] that this alternate choice is numerically unreliable. On the other hand, implementations that are essentially equivalent to (17.1) with its choice of coefficients were shown in [21, 22, 23] to have good properties, and we refer to these as “good” implementations. In particular, in the real case the two 2-term recurrences in (1.1) give the best error bounds, see [22, section 2], and apparently best performance. But in the complex case (17.1) might be preferable, for if u, w, B, and C are real with $(B+iC)^H = (B+iC)$, $i \triangleq \sqrt{-1}$, then

$$
\begin{aligned}
&B^T = B, \quad C^T = -C, \quad u^T C u = 0, \quad (u+iw)^H(u+iw) = u^T u + w^T w,\\
&(u+iw)^H(B+iC)(u+iw) = u^T B u + w^T B w + 2w^T C u,
\end{aligned}
$$

so that it is straightforward to compute α_k and β_{k+1} to be real in (17.1).

For the complex case the best way to compute real α_k in (1.1) is not so clear, but some numerical tests in [3] indicated that we can take the real part of the computed α_k, and this can be superior to using (17.1).

Next, section 5.1 mentioned that T_k is positive definite if A is sufficiently so.

THEOREM 17.1. *For Hermitian $H^{(k)}$ in Theorem 5.1 the maximum $\lambda_{\max}(T_k)$ and minimum $\lambda_{\min}(T_k)$ are bounded as follows:*

$$
\lambda_{\min}(A) - \textstyle\sum_{i=1}^k \|H^{(i)}\|_2 \le \lambda_{\min}(T_k) \le \lambda_{\max}(T_k) \le \lambda_{\max}(A) + \textstyle\sum_{i=1}^k \|H^{(i)}\|_2. \tag{17.2}
$$

Proof. In Corollary 7.1 let $\widetilde{Q}_1^{(k)}$ be $Q_1^{(k)}$ less its zero k-th row, and let $\widetilde{H}^{(k)}$ be $H^{(k)}$ without its k-th row and column, then from (7.2)–(7.3)

$$
\widetilde{Q}_1^{(k)H}\widetilde{Q}_1^{(k)} = I_k, \quad T_k = Q_1^{(k)H}\mathcal{A}_k Q_1^{(k)} = \widetilde{Q}_1^{(k)H}[\mathrm{diag}(T_{k-1}, A) + \widetilde{H}^{(k)}]\widetilde{Q}_1^{(k)}, \tag{17.3}
$$

where this is also true for $k = 1$ if we define T_0 to be nonexistent. Now expand $\widetilde{Q}_1^{(k)}$ to a full unitary matrix $\widetilde{Q}^{(k)} = [\widetilde{Q}_1^{(k)}, \widetilde{Q}_2^{(k)}] \in \mathbb{U}^{(k+n-1)\times(k+n-1)}$. Because T_k is the leading principal $k \times k$ submatrix of $\widetilde{Q}^{(k)H}[\mathrm{diag}(T_{k-1}, A) + \widetilde{H}^{(k)}]\widetilde{Q}^{(k)}$, it follows from the separation theorem, see for example [42, Ch.2 §47, p.103], that

$$
\mathrm{W}[T_k] \subseteq \mathrm{W}[\mathrm{diag}(T_{k-1}, A) + \widetilde{H}^{(k)}], \qquad k = 1, 2, 3, \ldots,
$$

where $\mathrm{W}[M] \triangleq \{x^H M x : \|x\|_2 = 1\}$ is the numerical range of M, see *e.g.*, [8, (7.1.4)]. Then from the eigenvalues of the sum of two matrices, see *e.g.*, [42, Ch.2 §44, p.101],

$$
\begin{aligned}
&\lambda_{\max}(T_k) \le \max\{\lambda_{\max}(T_{k-1}), \lambda_{\max}(A)\} + \lambda_{\max}(\widetilde{H}^{(k)}),\\
&\lambda_{\max}(T_1) \le \lambda_{\max}(A) + \lambda_{\max}(\widetilde{H}^{(1)}) \le \lambda_{\max}(A) + \|\widetilde{H}^{(1)}\|_2.
\end{aligned}
$$

This shows that $\max\{\lambda_{\max}(T_1), \lambda_{\max}(A)\} \le \lambda_{\max}(A) + \|\widetilde{H}^{(1)}\|_2$, so

$$
\lambda_{\max}(T_2) \le \max\{\lambda_{\max}(T_1), \lambda_{\max}(A)\} + \|\widetilde{H}^{(2)}\|_2 \le \lambda_{\max}(A) + \|\widetilde{H}^{(1)}\|_2 + \|\widetilde{H}^{(2)}\|_2,
$$

etc. This with $\|\widetilde{H}^{(i)}\|_2 \le \|H^{(i)}\|_2$ leads to the upper bound on $\lambda_{\max}(T_k)$ in (17.2). The lower bound on $\lambda_{\min}(T_k)$ follows similarly using $\lambda_{\min}(\widetilde{H}^{(i)}) \ge -\|\widetilde{H}^{(i)}\|_2$. □

Acknowledgements. Daniel Szyld and Valeria Simoncini were very helpful as editors for this long and complicated paper, while three referees gave exceptional suggestions to improve it. I particularly wish to thank Zdeněk Strakoš for his many contributions to this subject and his renewing my interest in the Lanczos process and related algorithms; Jörg Liesen, Beresford Parlett, and Mike Saunders for encouraging this work; Xiao-Wen Chang and David Titley-Peloquin for discussions and suggestions on this topic and helping to improve this paper; and Wolfgang Wülling, not only for the ideas he contributed, but for his delightful humor that kept me happily trudging on through even the most confusing and difficult parts of this research.

My interest in this area was initiated and strongly motivated by the ground breaking work on backward rounding error analysis of Jim Wilkinson. Above all, I remember Gene Golub very fondly for initially recognizing the worth of this research, and for his friendship and support during my career.

REFERENCES

[1] Å. Björck and C. C. Paige, *Loss and recapture of orthogonality in the modified Gram–Schmidt algorithm*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 176–190. https://doi.org/10.1137/0613015

[2] E. Carson and J. W. Demmel, *Accuracy of the s-step Lanczos method for the symmetric eigenproblem in finite precision*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 793–819. https://doi.org/10.1137/140990735

[3] X.-W. Chang, Personal communication (2018).

[4] S.-C. T. Choi, C. C. Paige, and M. A. Saunders, MINRES-QLP: *A Krylov subspace method for indefinite or singular symmetric systems*, SIAM J. Sci. Comput., 33 (2011), pp. 1810–1836. https://doi.org/10.1137/100787921

[5] C. Davis and W. M. Kahan, *Some new bounds on perturbations of subspaces*, Bull. Amer. Math. Soc., 75 (1969), pp. 863–868.

[6] D. Fong and M. A. Saunders, *An iterative algorithm for sparse least-squares problems*, SIAM J. Sci. Comput., 33:5 (2011) pp. 2950-2971. https://doi.org/10.1137/10079687X

[7] G. H. Golub and W. Kahan, *Calculating the singular values and pseudo-inverse of a matrix*, SIAM J. Numer. Anal., 2 (1965), pp. 205–224. https://doi.org/10.1137/0702016

[8] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed., The Johns Hopkins University Press, Baltimore, 2013. ISBN:9781421407944

[9] A. Greenbaum, *Convergence Properties of the Conjugate Gradient Algorithm in Exact and Finite Precision Arithmetic*, Ph.D. thesis, University of California, Berkeley, 1981.

[10] A. Greenbaum, *Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences*, Linear Algebra Appl., 113(1989), pp. 7–63. https://doi.org/10.1016/0024-3795(89)90285-1

[11] A. Greenbaum and Z. Strakoš, *Predicting the behavior of finite precision Lanczos and conjugate gradient computations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 121–137. https://doi.org/10.1137/0613011

[12] C. Greif, C. C. Paige, D. Titley-Peloquin, and J. M. Varah, *Numerical equivalences among Krylov subspace algorithms for skew-symmetric matrices*, SIAM J. Matrix Anal. Appl., 37 (2016), pp. 1071–1087. https://doi.org/10.1137/15M1030078

[13] M. H. Gutknecht and Z. Strakoš, *Accuracy of two three-term and three two-term recurrences for Krylov space solvers*, SIAM J. Matrix Anal. Appl., 22 (Jan. 2001), pp. 213–229. https://doi.org/10.1137/S0895479897331862

[14] S. Hammarling and N. J. Higham, *Wilkinson and Backward Error Analysis*, Website of the Numerical Linear Algebra Group, School of Mathematics, University of Manchester (2019). https://nla-group.org/2019/02/18/wilkinson-and-backward-error-analysis/

[15] M. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–436. https://doi.org/10.6028/jres.049.044

[16] C. Lanczos, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Research Nat. Bur. Standards, 45 (1950), pp. 255–282. https://doi.org/10.6028/jres.045.026

[17] C. Lanczos, *Solution of systems of linear equations by minimized iterations*, J. Research Nat. Bur. Standards, 49 (1952), pp. 33–53. https://doi.org/10.6028/jres.049.006

[18] G. Meurant, *The Lanczos and Conjugate Gradient Algorithms*, SIAM, Philadelphia, 2006.

https://doi.org/10.1137/1.9780898718140

[19] G. Meurant and Z. Strakoš, *The Lanczos and conjugate gradient algorithms in finite precision arithmetic*, Acta Numerica, Cambridge University Press, 15 (2006), pp. 471–542. https://doi.org/10.1017/S096249290626001X

[20] D. P. O'Leary, Z. Strakoš, and P. Tichý, *On sensitivity of Gauss-Christoffel quadrature*, Numer. Math., 107(1) (2007), pp. 147–174. https://doi.org/10.1007/s00211-007-0078-x

[21] C. C. Paige, *The Computation of Eigenvalues and Eigenvectors of Very Large Sparse Matrices.* PhD thesis, London University, London, England, 1971.

[22] C. C. Paige, *Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix*, J. Inst. Math. Appl., 18 (1976), pp. 341–349. https://doi.org/10.1093/imamat/18.3.341

[23] C. C. Paige, *Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem*, Linear Algebra Appl., 34 (1980), pp. 235–258. https://doi.org/10.1016/0024-3795(80)90167-6

[24] C. C. Paige, *A useful form of unitary matrix obtained from any sequence of unit 2-norm n-vectors*, SIAM J. Matrix Anal. Appl., 31 (2009), pp. 565–583. https://doi.org/10.1137/080725167

[25] C. C. Paige, *An augmented stability result for the Lanczos Hermitian matrix tridiagonalization process*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 2347–2359. https://doi.org/10.1137/090761343

[26] C. C. Paige, *The Effects of Loss of Orthogonality on Large Scale Numerical Computations*, In: O. Gervasi et al. (Eds.): Computational Science and Its Applications, ICCSA (2018), pp. 429–439. Lecture Notes in Computer Science, vol 10962. Springer, Cham. https://doi.org/10.1007/978-3-319-95168-3_29

[27] C. C. Paige and I. Panayotov, *Hessenberg matrix properties and Ritz vectors in the finite-precision Lanczos tridiagonalization process*, SIAM J. Matrix Anal. Appl., 32 (2011), pp. 1079–1094. https://doi.org/10.1137/100796285

[28] C. C. Paige, I. Panayotov, and J.-P. M. Zemke, *An augmented analysis of the perturbed two-sided Lanczos tridiagonalization process*, Linear Algebra Appl., 447 (2014), pp. 119–132. http://doi.org/10.1016/j.laa.2013.05.009

[29] C. C. Paige, M. Rozložník, and Z. Strakoš, *Modified Gram–Schmidt (MGS), least squares, and backward stability of MGS-GMRES*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 264–284. https://doi.org/10.1137/050630416

[30] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629. https://doi.org/10.1137/0712047

[31] C. C. Paige and M. A. Saunders, *LSQR: An algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Software, 8 (1982), pp. 43–71. https://doi.org/10.1145/355984.355989

[32] C. C. Paige and M. A. Saunders, *ALGORITHM* 583, *LSQR: Sparse linear equations and sparse least squares problems*, ACM Trans. Math. Software, 8 (1982), pp. 195–209. https://doi.org/10.1145/355993.356000

[33] C. C. Paige and Z. Strakoš, *Scaled total least squares fundamentals*, Numer. Math., 91 (2002), pp. 117–146. https://doi.org/10.1007/s002110100314

[34] C. C. Paige and W. Wülling, *Properties of a unitary matrix obtained from a sequence of normalized vectors*, SIAM J. Matrix Anal. Appl., 35 (2014), pp. 526–545. https://doi.org/10.1137/120897687

[35] I. Panayotov, *Eigenvalue Estimation with the Rayleigh-Ritz and Lanczos methods*, PhD thesis, McGill University, Montréal, Canada, 2010.

[36] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Classics in Appl. Math. 20, SIAM, Philadelphia, 1998. https://doi.org/10.1137/1.9781611971163

[37] M. A. Saunders, H. D. Simon, and E. L. Yip, *Two Conjugate-Gradient-type methods for unsymmetric linear equations*, SIAM J. Numer. Anal., 25 (1988), pp. 927–940. https://doi.org/10.1137/0725052

[38] G. W. Stewart, *On the perturbation of pseudo-inverses, projections, and linear least squares problems*, SIAM Rev., 19 (1977), pp. 634–662.

[39] Z. Strakoš, *On the real convergence rate of the conjugate gradient method*, Linear Algebra Appl., 154–156 (1991), pp. 535–549. https://doi.org/10.1016/0024-3795(91)90393-B

[40] Z. Strakoš, *Convergence and numerical behavior of the Krylov space methods*, in G. Winter Althaus and E. Spedicato eds., *NATO ASI Institute, Algorithms for Large Sparse Linear Algebraic Systems: The State of the Art and Applications in Science and Engineering*, Kluwer Academic, pp. 175–197, 1998. ISBN: 978-0-7923-4975-4

[41] J. H. Wilkinson, *Rounding Errors in Algebraic Processes*, Notes on Applied Science No. 32, Her Majesty's Stationery Office, London, 1963. (Also published by Prentice-Hall, Engle-

wood Cliffs, NJ, USA, 1964. Reprinted by Dover Publications, New York, 1994).

[42] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, UK, 1965. (1988 Paperback version, ISBN: 9780198534181)