Accuracy of the s-step Lanczos method for the symmetric eigenproblem

Erin Carson
James Demmel

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2014-165
http://www.eecs.berkeley.edu/Pubs/TechRpts/2014/EECS-2014-165.html

September 17, 2014

Acknowledgement

This material is based upon work supported by the U.S. Department of Energy Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program under Award Numbers DE-SC0004938, DE-SC0003959, and DE-SC0010200; by the U.S. Department of Energy Office of Science, Office of Advanced Scientific Computing Research, X-Stack program under Award Numbers DE-SC0005136, DE-SC0008699, DE-SC0008700, and AC02-05CH11231; by DARPA Award Number HR0011-12-2-0016, as well as contributions from Intel, Oracle, and MathWorks.

ACCURACY OF THE S-STEP LANCZOS METHOD FOR THE SYMMETRIC EIGENPROBLEM

ERIN CARSON AND JAMES DEMMEL

Abstract. The s-step Lanczos method is an attractive alternative to the classical Lanczos method as it enables an $O(s)$ reduction in data movement over a fixed number of iterations. This can significantly improve performance on modern computers. In order for s-step methods to be widely adopted, it is important to better understand their error properties. Although the s-step Lanczos method is equivalent to the classical Lanczos method in exact arithmetic, empirical observations demonstrate that it can behave quite differently in finite precision.

In this paper, we demonstrate that bounds on accuracy for the finite precision Lanczos method given by Paige [*Lin. Alg. Appl.*, 34:235–258, 1980] can be extended to the s-step Lanczos case assuming a bound on the condition numbers of the computed s-step bases. Our results confirm theoretically what is well-known empirically: the conditioning of the Krylov bases plays a large role in determining finite precision behavior. In particular, if one can guarantee that the basis condition number is not too large throughout the iterations, the accuracy and convergence of eigenvalues in the s-step Lanczos method should be similar to those of classical Lanczos. This indicates that, under certain restrictions, the s-step Lanczos method can be made suitable for use in many practical cases.

Key words. Krylov subspace methods, error analysis, finite precision, roundoff, Lanczos, avoiding communication, orthogonal bases

AMS subject classifications. 65G50, 65F10, 65F15, 65N15, 65N12

1. Introduction.

Given an n-by-n symmetric matrix A and a starting vector v_1 with unit 2-norm, m steps of the Lanczos method [23] theoretically produces the orthonormal matrix $V_m = [v_1, \ldots, v_m]$ and the m-by-m symmetric tridiagonal matrix T_m such that

$$AV_m = V_m T_m + \beta_{m+1} v_{m+1} e_m^T. \tag{1.1}$$

When $m = n$, the eigenvalues of T_n are the eigenvalues of A. In practice, the eigenvalues of T_m are still good approximations to the eigenvalues of A when $m \ll n$, which makes the Lanczos method attractive as an iterative procedure. Many Krylov subspace methods (KSMs), including those for solving linear systems and least squares problems, are based on the Lanczos method. In turn, these various Lanczos-based methods are the core components in numerous scientific applications.

Classical implementations of Krylov methods, the Lanczos method included, require one or more sparse matrix-vector multiplications (SpMVs) and one or more inner product operations in each iteration. These computational kernels are both communication-bound on modern computer architectures. To perform an SpMV, each processor must communicate entries of the source vector it owns to other processors in the parallel algorithm, and in the sequential algorithm the matrix A must be read from slow memory (when it is too large to fit in cache, the most interesting case). Inner products involve a global reduction in the parallel algorithm, and a number of reads and writes to slow memory in the sequential algorithm (depending on the size of the vectors and the size of the fast memory).

Thus, many efforts have focused on communication-avoiding Krylov subspace methods (CA-KSMs), or s-step Krylov methods, which can perform s iterations with $O(s)$ less communication than classical KSMs; see, e.g., [6, 7, 9, 11, 12, 18, 19, 10, 37, 39]. In practice, this can translate into significant speedups for many problems [26, 42].

Equally important to the performance of each iteration is the convergence rate of the method, i.e., the total number of iterations required until the desired convergence criterion is met. Although theoretically the Lanczos process described by (1.1) produces an orthogonal basis and a tridiagonal matrix similar to A after n steps, these properties need not hold in finite precision. The effects of roundoff error on the ideal Lanczos process were known to Lanczos when he published his algorithm in 1950. Since then, much research has been devoted to better understanding this behavior, and to devise more robust and stable algorithms.

Although s-step Krylov methods are mathematically equivalent to their classical counterparts in exact arithmetic, it perhaps comes as no surprise that their finite precision behavior may differ significantly, and that the theories developed for classical methods in finite precision do not hold for the s-step case. It has been empirically observed that the behavior of s-step Krylov methods deviates further from that of the classical method as s increases, and that the severity of this deviation is heavily influenced by the polynomials used for the s-step Krylov bases (see, e.g., [1, 6, 19, 20]).

Arguably the most revolutionary work in the finite precision analysis of classical Lanczos was a series of papers published by Paige [27, 28, 29, 30]. Paige's analysis succinctly describes how rounding errors propagate through the algorithm to impede orthogonality. These results were developed to give theorems which link the loss of orthogonality to convergence of the computed eigenvalues [30]. No analogous theory currently exists for the s-step Lanczos method.

In this paper, we use the complete rounding error analysis of the s-step Lanczos method presented in [4] to extend the analysis of Paige for classical Lanczos to the s-step Lanczos method. Our analysis here of s-step Lanczos very closely follows Paige's rounding error analysis for orthogonality in classical Lanczos [29], and the proofs for accuracy and convergence of eigenvalues presented here follow the approach of [30]. The derived bounds in [4] are very similar to those of Paige for classical Lanczos, but with the addition of an amplification term which depends on the condition number of the Krylov bases computed every s steps. We show here that, based on restrictions on the size of this condition number, the same theorems of Paige apply to the s-step case.

Our results confirm theoretically what is well-known empirically: the conditioning of the Krylov bases plays a large role in determining finite precision behavior. In particular, if one can guarantee that the basis condition number is not too large throughout the iteration, the accuracy and convergence of eigenvalues in the s-step Lanczos method should be similar to those produced by classical Lanczos. This indicates that, under certain restrictions, the s-step Lanczos method is suitable for use in practice.

The remainder of this paper is outlined as follows. In Section 2, we present related work in s-step Krylov methods and the analysis of finite precision Lanczos. In Section 3, we review a variant of the Lanczos method and derive the corresponding s-step Lanczos method. Section 4 summarizes rounding error results for the s-step Lanczos method from [4]. Sections 5 and 6 use the results of Paige [30] to prove results on the accuracy of the computed eigenvalues and rate of convergence of the computed eigenvalues, respectively. Section 7 concludes with a discussion of future work.

2. Related work.

We briefly review related work in s-step Krylov methods as well as work related to the analysis of classical Krylov methods in finite precision.

2.1. s-step Krylov subspace methods.

The term 's-step Krylov method', first used by Chronopoulos and Gear [8], describes variants of Krylov methods where the iteration loop is split into blocks of s iterations. Since the Krylov subspaces required to perform s iterations of updates are known, bases for these subspaces can be computed upfront, inner products between basis vectors can be computed with one block inner product, and then s iterations are performed by updating the coordinates in the generated Krylov bases (see Section 3 for details). Many formulations and variations have been derived over the past few decades with various motivations, namely increasing parallelism (e.g., [8, 39, 40]) and avoiding data movement, both between levels of the memory hierarchy in sequential methods and between processors in parallel methods. A thorough treatment of related work can be found in [19].

Many empirical studies of s-step Krylov methods found that convergence often deteriorated using $s > 5$ due to the inherent instability of the monomial basis. This motivated research into the use of better-conditioned bases (e.g., Newton or Chebyshev polynomials) for the Krylov subspace, which enabled convergence for higher s values (see, e.g., [1, 18, 20, 33]). Hoemmen has used a novel matrix equilibration and balancing approach to achieve similar effects [19].

The term ‘communication-avoiding Krylov methods’ refers to s-step Krylov methods and implementations which aim to improve performance by asymptotically decreasing communication costs, possibly both in computing inner products and computing the s-step bases, for both sequential and parallel algorithms; see [11, 19]. Hoemmen et al. [19, 26] derived communication-avoiding variants of Lanczos, Arnoldi, Conjugate Gradient (CG) and the Generalized Minimum Residual (GMRES) method. Details of nonsymmetric Lanczos-based CA-KSMs, including communication-avoiding versions of Biconjugate Gradient (BICG) and Stabilized Biconjugate Gradient (BICGSTAB) can be found in [6]. Although potential performance improvement is our primary motivation for studying these methods, we use the general term ‘s-step methods’ here as our error analysis is independent of runtime.

Many efforts have been devoted specifically to the s-step Lanczos method. The first s-step Lanczos methods known in the literature are due to Kim and Chronopoulos, who derived a three-term symmetric s-step Lanczos method [21] as well as a three-term nonsymmetric s-step Lanczos method [22]. Hoemmen derived a three-term communication-avoiding Lanczos method, CA-Lanczos [19]. Although the three-term variants require less memory, their numerical accuracy can be worse than implementations which use two coupled two-term recurrences [17]. A two-term communication-avoiding nonsymmetric Lanczos method (called CA-BIOC, based on the ‘BIOC’ version of nonsymmetric Lanczos of Gutknecht [16]) can be found in [2]. This work includes the derivation of a new version of the s-step Lanczos method, equivalent in exact arithmetic to the variant used by Paige [29]. It uses a two-term recurrence like BIOC, but is restricted to the symmetric case and uses a different starting vector.

For s-step KSMs that solve linear systems, increased roundoff error in finite precision can decrease the maximum attainable accuracy of the solution, resulting in a less accurate solution than found by the classical method. A quantitative analysis of roundoff error in CA-CG and CA-BICG can be found in [5]. Based on the work of [38] for conventional KSMs, we have also explored residual replacement strategies for CA-CG and CA-BICG as a method to limit the deviation of true and computed residuals when high accuracy is required [5].

2.2. Error analysis of the Lanczos method.

Lanczos and others recognized early on that rounding errors could cause the Lanczos method to deviate from its ideal theoretical behavior. Since then, various efforts have been devoted to analyzing and improving the finite precision Lanczos method.

Widely considered to be the most significant development was the series of papers by Paige discussed in Section 1. A detailed rounding error analysis for the s-step

Lanczos case, based on Paige's results [29], can be found in [4] and is summarized in Section 4. Another important development was due to Greenbaum and Strakoš, who performed a backward-like error analysis which showed that finite precision Lanczos and CG behave very similarly to the exact algorithms applied to any of a certain class of larger matrices [14]. Paige has recently shown a similar type of augmented stability for the Lanczos process [31]. There are many other analyses of the behavior of various KSMs in finite precision, including some more recent results due to Wülling [43] and Zemke [44]; for a thorough overview of the literature, see [24, 25].

A number of strategies for maintaining the orthogonality among the Lanczos vectors were inspired by the analysis of Paige, such as selective reorthogonalization [32] and partial reorthogonalization [35]. Recently, Gustafsson et al. have extended such reorthogonalization strategies for classical Lanczos to the s-step case [15].

3. The s-step Lanczos method.

The classical Lanczos method is shown in Algorithm 1. We use the same variant used by Paige in his error analysis for the classical Lanczos method [29] to allow easy comparison of results. Note that for simplicity, we assume no breakdown occurs, i.e., $\beta_{m+1} \neq 0$ for $m < n$, and thus breakdown conditions are not discussed here. We now give a derivation of s-step Lanczos, obtained from classical Lanczos in Algorithm 1. The same derivation appears in [4], except in the present version the iterations are 1-indexed rather than 0-indexed to match the notation of Paige.

Algorithm 1 The classical Lanczos method

```
Require: $n$-by-$n$ real symmetric matrix $A$ and length-$n$ vector $v_1$ such that $\|v_1\|_2 = 1$
1: $u_1 = Av_1$
2: for $m = 1, 2, \ldots$ until convergence do
3:     $\alpha_m = v_m^T u_m$
4:     $w_m = u_m - \alpha_m v_m$
5:     $\beta_{m+1} = \|w_m\|_2$
6:     $v_{m+1} = w_m / \beta_{m+1}$
7:     $u_{m+1} = Av_{m+1} - \beta_{m+1} v_m$
8: end for
```

Suppose we are beginning iteration $m = sk + 1$ where $k \in \mathbb{N}$ and $0 < s \in \mathbb{N}$. By induction on lines 6 and 7 of Algorithm 1, it is true that

$$\begin{aligned} v_{sk+j} &\in \mathcal{K}_s(A, v_{sk+1}) + \mathcal{K}_s(A, u_{sk+1}) \quad \text{and} \\ u_{sk+j} &\in \mathcal{K}_{s+1}(A, v_{sk+1}) + \mathcal{K}_{s+1}(A, u_{sk+1}) \end{aligned} \tag{3.1}$$

for $j \in \{1, \ldots, s+1\}$, where $\mathcal{K}_i(A, x) = \text{span}\{x, Ax, \ldots, A^{i-1}x\}$ denotes the Krylov subspace of dimension i of matrix A with respect to vector x. Since $\mathcal{K}_j(A, x) \subseteq \mathcal{K}_i(A, x)$ for $j \leq i$, we can write

$$v_{sk+j}, u_{sk+j} \in \mathcal{K}_{s+1}(A, v_{sk+1}) + \mathcal{K}_{s+1}(A, u_{sk+1})$$

for $j \in \{1, \ldots, s+1\}$. Note that since $u_1 = Av_1$, if $k = 0$ we have

$$v_j, u_j \in \mathcal{K}_{s+2}(A, v_1).$$

for $j \in \{1, \ldots, s+1\}$.

For $k > 0$, we then define the 'basis matrix' $\mathcal{Y}_k = [\mathcal{V}_k, \mathcal{U}_k]$, where $\mathcal{V}_k$ and $\mathcal{U}_k$ are size n-by-$(s+1)$ matrices whose columns form bases for $\mathcal{K}_{s+1}(A, v_{sk+1})$ and

$\mathcal{K}_{s+1}(A, u_{sk+1})$, respectively. For $k = 0$, we define $\mathcal{Y}_0$ to be a size n-by-$(s+2)$ matrix whose columns form a basis for $\mathcal{K}_{s+2}(A, v_1)$. Then by (3.1), we can represent v_{sk+j} and u_{sk+j}, for $j \in \{1, \ldots, s+1\}$, by their coordinates (denoted with primes) in $\mathcal{Y}_k$, i.e.,

$$v_{sk+j} = \mathcal{Y}_k v'_{k,j}, \qquad u_{sk+j} = \mathcal{Y}_k u'_{k,j}. \tag{3.2}$$

Note that for $k = 0$, the coordinate vectors are length-$(s+2)$ and for $k > 0$, the coordinate vectors are length-$(2s+2)$. We can write a similar equation for auxiliary vector w_{sk+j}, i.e., $w_{sk+j} = \mathcal{Y}_k w'_{k,j}$ for $j \in \{1, \ldots, s\}$. We define also the Gram matrix $G_k = \mathcal{Y}_k^T \mathcal{Y}_k$, which is size $(s+2)$-by-$(s+2)$ for $k = 0$ and $(2s+2)$-by-$(2s+2)$ for $k > 0$. Using this matrix, the inner products in lines 3 and 5 can be written

$$\alpha_{sk+j} = v_{sk+j}^T u_{sk+j} = v'^T_{k,j} \mathcal{Y}_k^T \mathcal{Y}_k u'_{k,j} = v'^T_{k,j} G_k u'_{k,j} \quad \text{and} \tag{3.3}$$

$$\beta_{sk+j+1} = (w_{sk+j}^T w_{sk+j})^{1/2} = (w'^T_{k,j} \mathcal{Y}_k^T \mathcal{Y}_k w'_{k,j})^{1/2} = (w'^T_{k,j} G_k w'_{k,j})^{1/2}. \tag{3.4}$$

We assume that the bases are generated via polynomial recurrences represented by matrix $\mathcal{B}_k$, which is in general upper Hessenberg but often tridiagonal in practice. The recurrence can thus be written in matrix form as

$$A\underline{\mathcal{Y}}_k = \mathcal{Y}_k \mathcal{B}_k \tag{3.5}$$

where, for $k = 0$, $\mathcal{B}_k$ is size $(s+2)$-by-$(s+2)$ and $\underline{\mathcal{Y}}_0 = \big[\mathcal{Y}_0[I_{s+1}, 0_{s+1,1}]^T, 0_{n,1}\big]$, and for $k > 0$, $\mathcal{B}_k$ is size $(2s+2)$-by-$(2s+2)$ and $\underline{\mathcal{Y}}_k = \big[\mathcal{V}_k[I_s, 0_{s,1}]^T, 0_{n,1}, \mathcal{U}_k[I_s, 0_{s,1}]^T, 0_{n,1}\big]$. Note that then above $\mathcal{B}_k$ has zeros in column $s+2$ when $k = 0$ and zeros in columns $s+1$ and $2s+1$ for $k > 0$. Therefore, using (3.5), for $j \in \{1, \ldots, s\}$,

$$Av_{sk+j+1} = A\mathcal{Y}_k v'_{k,j+1} = A\underline{\mathcal{Y}}_k v'_{k,j+1} = \mathcal{Y}_k \mathcal{B}_k v'_{k,j+1}. \tag{3.6}$$

Thus, to compute iterations $sk+2$ through $sk+s+1$ in s-step Lanczos, we first generate basis matrix $\mathcal{Y}_k$ such that (3.6) holds, and we compute the Gram matrix G_k from $\mathcal{Y}_k$. Then updates to the length-n vectors can be performed by updating instead the length-$(2s+2)$ coordinates for those vectors in $\mathcal{Y}_k$. Inner products and multiplications with A become smaller operations which can be performed locally, as in (3.3), (3.4), and (3.6). The complete s-step Lanczos algorithm is presented in Algorithm 2. Note that in Algorithm 2 we show the length-n vector updates in each inner iteration (lines 16 and 18) for clarity, although these vectors play no part in the inner loop iteration updates. In practice, the basis change operation (3.2) can be performed on a block of coordinate vectors at the end of each outer loop to recover v_{sk+i} and u_{sk+i}, for $i \in \{2, \ldots, s+1\}$.

4. Rounding errors in the s-step Lanczos method. The analysis in [4] used a standard model of floating point arithmetic where it is assumed that the computations are carried out on a machine with relative precision ϵ (see, e.g., [13]). Terms with ϵ of order > 1, which have negligible effect on the results, are ignored. We also ignore underflow and overflow. Quantities computed in finite precision arithmetic are decorated with hats, e.g., if we are to compute the expression $\alpha = v^T u$ in finite precision, we get $\hat{\alpha} = fl(v^T u)$. Throughout this analysis, e_i denotes the i^{th} column of an identity matrix of appropriate size and $\|\cdot\|$ denotes the 2-norm, unless otherwise specified.

It is also assumed throughout that the generated bases $\hat{\mathcal{U}}_k$ and $\hat{\mathcal{V}}_k$ are numerically full rank. That is, all singular values of $\hat{\mathcal{U}}_k$ and $\hat{\mathcal{V}}_k$ are greater than $\epsilon n \cdot 2^{\lfloor \log_2 \theta_1 \rfloor}$

Algorithm 2 The s-step Lanczos method

```
Require: $n$-by-$n$ real symmetric matrix $A$ and length-$n$ vector $v_1$ such that $\|v_1\|_2 = 1$
$u_1 = Av_1$
for $k = 0, 1, \dots$ until convergence do
    Compute $\mathcal{Y}_k$ with change of basis matrix $\mathcal{B}_k$
    Compute $G_k = \mathcal{Y}_k^T \mathcal{Y}_k$
    $v'_{k,1} = e_1$
    if $k = 0$ then
        $u'_{0,1} = \mathcal{B}_k e_1$
    else
        $u'_{k,1} = e_{s+2}$
    end if
    for $j = 1, 2, \dots, s$ do
        $\alpha_{sk+j} = v'^T_{k,j} G_k u'_{k,j}$
        $w'_{k,j} = u'_{k,j} - \alpha_{sk+j} v'_{k,j}$
        $\beta_{sk+j+1} = (w'^T_{k,j} G_k w'_{k,j})^{1/2}$
        $v'_{k,j+1} = w'_{k,j} / \beta_{sk+j+1}$
        $v_{sk+j+1} = \mathcal{Y}_k v'_{k,j+1}$
        $u'_{k,j+1} = \mathcal{B}_k v'_{k,j+1} - \beta_{sk+j+1} v'_{k,j}$
        $u_{sk+j+1} = \mathcal{Y}_k u'_{k,j+1}$
    end for
end for
```

where θ_1 is the largest singular value of $\hat{\mathcal{U}}_k$ or $\hat{\mathcal{V}}_k$, respectively. The results of [4] are summarized in the following theorem.

THEOREM 4.1. *Assume that Algorithm 2 is implemented in floating point arithmetic with relative precision ϵ and applied for $m = sk+j$ steps to the n-by-n real symmetric matrix A, starting with vector v_1 with $\|v_1\|_2 = 1$. Let $\sigma = \|A\|_2$, $\theta\sigma = \||A|\|_2$ and $\tau_k \sigma = \||\mathcal{B}_k|\|_2$, where $\mathcal{B}_k$ is defined in* (3.6), *and let*

$$\bar{\Gamma}_k = \max_{i \in \{0,\dots,k\}} \|\hat{\mathcal{Y}}_i^+\|_2 \| |\hat{\mathcal{Y}}_i| \|_2 \geq 1 \quad \textit{and} \quad \bar{\tau}_k = \max_{i \in \{0,\dots,k\}} \tau_i, \tag{4.1}$$

where above we use the superscript '+' to denote the Moore-Penrose pseudoinverse, i.e., $\hat{\mathcal{Y}}_i^+ = (\hat{\mathcal{Y}}_i^T \hat{\mathcal{Y}}_i)^{-1} \hat{\mathcal{Y}}_i^T$. Then $\hat{\alpha}_i$, $\hat{\beta}_{i+1}$, and $\hat{v}_{i+1}$ will be computed, for $i \in \{1, \dots, m\}$, such that

$$A\hat{V}_m = \hat{V}_m \hat{T}_m + \hat{\beta}_{m+1} \hat{v}_{m+1} e_m^T + \delta \hat{V}_m, \tag{4.2}$$

with

$$\begin{aligned} \hat{V}_m &= [\hat{v}_1, \hat{v}_2, \dots, \hat{v}_m], \\ \delta\hat{V}_m &= [\delta\hat{v}_1, \delta\hat{v}_2, \dots, \delta\hat{v}_m], \quad \textit{and} \\ \hat{T}_m &= \begin{bmatrix} \hat{\alpha}_1 & \hat{\beta}_2 & & \\ \hat{\beta}_2 & \ddots & \ddots & \\ & \ddots & \ddots & \hat{\beta}_m \\ & & \hat{\beta}_m & \hat{\alpha}_m \end{bmatrix}, \end{aligned}$$

and

$$\|\delta\hat{v}_i\|_2 \leq \epsilon_1\sigma, \tag{4.3}$$

$$\begin{aligned} \hat{\beta}_{i+1}|\hat{v}_i^T\hat{v}_{i+1}| &\leq 2\epsilon_0\sigma, \\ |\hat{v}_{i+1}^T\hat{v}_{i+1} - 1| &\leq \epsilon_0/2, \quad \textit{and} \\ \left|\hat{\beta}_{i+1}^2 + \hat{\alpha}_i^2 + \hat{\beta}_i^2 - \|A\hat{v}_i\|_2^2\right| &\leq 4i(3\epsilon_0 + \epsilon_1)\sigma^2, \end{aligned} \tag{4.4}$$

where we have used the notation

$$\epsilon_0 \equiv 2\epsilon(n{+}11s{+}15)\bar{\Gamma}_k^2 \quad \textit{and} \quad \epsilon_1 \equiv 2\epsilon\big((n{+}2s{+}5)\theta + (4s{+}9)\bar{\tau}_k + 10s{+}16\big)\bar{\Gamma}_k^2. \tag{4.5}$$

Furthermore, if R_m is the strictly upper triangular matrix such that

$$\hat{V}_m^T\hat{V}_m = R_m^T + diag(\hat{V}_m^T\hat{V}_m) + R_m, \tag{4.6}$$

then

$$\hat{T}_m R_m - R_m\hat{T}_m = \hat{\beta}_{m+1}\hat{V}_m^T\hat{v}_{m+1}e_m^T + \delta R_m, \tag{4.7}$$

where δR_m is upper triangular with elements ρ such that

$$\begin{aligned} |\rho_{1,1}| &\leq 2\epsilon_0\sigma, \quad \textit{and for } i \in \{2,\ldots,m\}, \\ |\rho_{i,i}| &\leq 4\epsilon_0\sigma, \\ |\rho_{i-1,i}| &\leq 2(\epsilon_0+\epsilon_1)\sigma, \quad \textit{and} \\ |\rho_{\ell,i}| &\leq 2\epsilon_1\sigma, \quad \textit{where } \ell \in \{1,\ldots,i{-}2\}. \end{aligned} \tag{4.8}$$

Note that the above theorem is obtained by substituting the notation in (4.5) into the bounds of Theorem 4.1 in [4]. This sacrifices some tightness in the bounds in favor of simplifying the notation. These bounds are also a factor of 2 larger than those that appear in [4], to match the notation of Paige. Also note that the value of $\bar{\Gamma}_k$ in (4.1) is likely a large overestimate. This causes our bounds for the $s = 1$ case to be larger tham those of Paige for classical Lanczos. To obtain tighter bounds, in iteration $m = sk + j$, one can instead use, e.g.,

$$\bar{\Gamma}_k \equiv \max\left\{\frac{\||\hat{\mathcal{Y}}_\ell||\mathcal{B}_\ell||\hat{v}'_{\ell,i}|\|_2}{\||\mathcal{B}_\ell|\|_2\|\hat{\mathcal{Y}}_\ell\hat{v}'_{\ell,i}\|_2}, \max_{x\in\{\hat{w}'_{\ell,i},\hat{u}'_{\ell,i},\hat{v}_{\ell,i},\hat{v}_{\ell,i+1}\}} \frac{\||\hat{\mathcal{Y}}_\ell||x|\|_2}{\|\hat{\mathcal{Y}}_\ell x\|_2}\right\},$$

where the maximum is over $\ell \leq k$, and $i \leq j$ if $\ell = k$, $i \leq s$ if $\ell < k$ (see [4, Section 5]).

Using (4.8), it can be shown that

$$\|\delta R_m\|_F^2 \leq 2\sigma^2\Big(2\big(5m-4\big)\epsilon_0^2 + 4(m{-}1)\epsilon_0\epsilon_1 + m(m{-}1)\epsilon_1^2\Big) \tag{4.9}$$

where subscript F denotes the Frobenius norm. If we define

$$\epsilon_2 \equiv \sqrt{2}\max(6\epsilon_0, \epsilon_1), \tag{4.10}$$

then (4.9) gives

$$\|\delta R_m\|_F \leq m\sigma\epsilon_2. \tag{4.11}$$

4.1. Assumptions. In order to make use of Paige's analysis [30], we must make the similar assumptions that

$$\hat{\beta}_{i+1} \neq 0 \text{ for } i \in \{1, \ldots, m\}, \quad m(3\epsilon_0 + \epsilon_1) \leq 1, \text{ and } \epsilon_0 < \frac{1}{12}. \tag{4.12}$$

These assumptions are used throughout the analysis. Note that (4.12) means that in order to guarantee the applicability of Paige's results for classical Lanczos to the s-step Lanczos case, we must have

$$\bar{\Gamma}_k^2 < \Big(24\epsilon(n+11s+15)\Big)^{-1} = O\big(1/(n\epsilon)\big). \tag{4.13}$$

Since the bounds that will be presented, as well as the bounds in the theorem above, are not tight, this condition on $\bar{\Gamma}_k^2$ could be overly restrictive in practice. In paragraphs labeled 'Comments' in the subsequent section, we comment on what happens to the bounds and analysis in the case that $\bar{\Gamma}_k^2$ exceeds this value, i.e., at least one computed s-step basis is ill-conditioned. As stated previously, we also assume the no underflow or overflow occurs, and that all computed s-step Krylov bases are numerically full rank.

5. Accuracy of eigenvalues.

Theorem 4.1 is in the same form as Paige's equivalent theorem for classical Lanczos (see [30]), except our definitions of ϵ_0 and ϵ_1 are about a factor $\bar{\Gamma}_k^2$ larger (assuming $s \ll n$). This additional amplification term, which can be bounded in terms of the maximum condition number of the computed s-step Krylov bases, has significant consequences for the algorithm as we will see in the next two sections. The equivalent forms of our theorem and Paige's theorem allow us to immediately apply his results from [30], in which bounds are given in terms of ϵ_2, to the s-step case; the only thing that changes in the s-step case is the value of ϵ_0 and ϵ_1, and thus ϵ_2.

In this and the subsequent section, we reproduce the theorems and proofs of Paige, and discuss their application to the s-step Lanczos method. *Note that the present authors claim no contribution to the analysis techniques used here.* In fact, much of the text in the following sections is taken verbatim from Paige [30], with only the notation changed to match the algorithms in Section 3.

Our contribution is showing that the theorems of Paige also apply to the s-step Lanczos method under the assumption that (4.12) (and thus also (4.13)) holds. To aid the reader, we have also added a few more intermediate steps in the analysis that were omitted from [30]. Also note that the text which discusses the meaning of the results for the s-step case is our own.

Let the eigendecomposition of $\hat{T}_m$ be

$$\hat{T}_m Q^{(m)} = Q^{(m)} \operatorname{diag}\big(\mu_i^{(m)}\big), \tag{5.1}$$

for $i \in \{1, \ldots, m\}$, where the orthonormal matrix $Q^{(m)}$ has i^{th} column $q_i^{(m)}$ and (ℓ, i) element $\eta_{\ell,i}^{(m)}$, and the eigenvalues are ordered

$$\mu_1^{(m)} > \mu_2^{(m)} > \cdots > \mu_m^{(m)}.$$

Note that it is assumed that the decomposition (5.1) is computed exactly. If $\mu_i^{(m)}$ is an approximation to an eigenvalue λ_i of A, then the corresponding approximate eigenvector is $z_i^{(m)}$, the ith column of

$$Z^{(m)} \equiv \hat{V}_m Q^{(m)}. \tag{5.2}$$

We now review some properties of $\hat{T}_m$. Let $\nu_i^{(m)}$, for $i \in \{1, \dots, m-1\}$, be the eigenvalues of the matrix obtained by removing the $(t+1)$st row and column of $\hat{T}_m$, ordered so that

$$\mu_1^{(m)} \geq \nu_1^{(m)} \geq \mu_2^{(m)} \geq \cdots \geq \nu_{m-1}^{(m)} \geq \mu_m^{(m)}. \tag{5.3}$$

It was shown in [36] that

$$\left(\eta_{t+1,i}^{(m)}\right)^2 = \prod_{\ell=1,\ell\neq i}^{m} \delta_\ell(t+1,i,m) \tag{5.4}$$

$$\delta_\ell(t+1,i,m) \equiv \begin{cases} \frac{\mu_i^{(m)}-\nu_\ell^{(m)}}{\mu_i^{(m)}-\mu_\ell^{(m)}} & \ell = 1,2,\dots,i-1 \\ \frac{\mu_i^{(m)}-\nu_{\ell-1}^{(m)}}{\mu_i^{(m)}-\mu_\ell^{(m)}} & \ell = i+1,\dots,m \end{cases}$$

$$0 \leq \delta_\ell(t+1,i,m) \leq 1, \ell = 1,\dots,i-1,i+1,\dots,m. \tag{5.5}$$

If we apply $\hat{T}_m$ to the r^{th} eigenvector of $\hat{T}_t$, where $1 \leq r \leq t < m$,

$$\hat{T}_m \begin{bmatrix} q_r^{(t)} \\ 0_{m-t,1} \end{bmatrix} = \begin{bmatrix} \mu_r^{(t)} q_r^{(t)} \\ \hat{\beta}_{t+1}\eta_{t,r}^{(t)} e_1 \end{bmatrix} \tag{5.6}$$

and from [41],

$$\delta_{t,r} \equiv \hat{\beta}_{t+1}|\eta_{t,r}^{(t)}| \geq \min_i |\mu_i^{(m)} - \mu_r^{(t)}|. \tag{5.7}$$

Definition 5.1. *[30, Definition 1] We say that an eigenvalue $\mu_r^{(t)}$ of $\hat{T}_t$* has stabilized to within $\delta_{t,r}$ *if, for every $m > t$, we know there is an eigenvalue of $\hat{T}_m$ within $\delta_{t,r}$ of $\mu_r^{(t)}$. We will say $\mu_r^{(t)}$ has* stabilized *when we know it has stabilized to within $\gamma(m+1)^\omega \sigma\epsilon_2$ where γ and ω are small positive constants.*

From (5.7), we can see that after t steps $\mu_r^{(t)}$ has necessarily stabilized to within $\delta_{t,r}$. Multiplying (5.6) by $q_i^{(m)T}$, $i \in \{1,\dots,m\}$, gives

$$\left(\mu_i^{(m)} - \mu_r^{(t)}\right) q_i^{(m)T} \begin{bmatrix} q_r^{(t)} \\ 0_{m-t,1} \end{bmatrix} = \hat{\beta}_{t+1}\eta_{t+1,i}^{(m)}\eta_{t,r}^{(t)}. \tag{5.8}$$

Another result is obtained by applying eigenvectors of $\hat{T}_m$ to each side of (4.7); i.e., multiplying on the left by $q_\ell^{(m)T}$ and multiplying on the right by $q_i^{(m)}$ for some $i, \ell \in \{1,\dots,m\}$:

$$\begin{aligned} q_\ell^{(m)T}\left(\hat{T}_m R_m - R_m \hat{T}_m\right) q_i^{(m)} &= q_\ell^{(m)T}\left(\hat{\beta}_{m+1}\hat{V}_m^T \hat{v}_{m+1} e_m^T + \delta R_m\right) q_i^{(m)} \\ \left(q_\ell^{(m)T}\hat{T}_m^T\right) R_m q_i^{(m)} - q_\ell^{(m)T} R_m \left(\hat{T}_m q_i^{(m)}\right) &= \hat{\beta}_{m+1}\left(q_\ell^{(m)T}\hat{V}_m^T\right)\hat{v}_{m+1}\left(e_m^T q_i^{(m)}\right) \\ &\quad + q_\ell^{(m)T}\delta R_m q_i^{(m)}. \end{aligned}$$

Using (5.1) and (5.2), the above becomes

$$\begin{aligned} &(q_\ell^{(m)T}\mu_\ell^{(m)}) R_m q_i^{(m)} - q_\ell^{(m)T} R_m (q_i^{(m)}\mu_i^{(m)}) \\ &\qquad\qquad = \hat{\beta}_{m+1} z_\ell^{(m)T} \hat{v}_{m+1}\eta_{m,i}^{(m)} + q_\ell^{(m)T}\delta R_m q_i^{(m)}, \end{aligned}$$

and rearranging, we obtain

$$\big(\mu_\ell^{(m)} - \mu_i^{(m)}\big) q_\ell^{(m)T} R_m q_i^{(m)} = \hat{\beta}_{m+1} z_\ell^{(m)T} \hat{v}_{m+1} \eta_{m,i}^{(m)} + \epsilon_{\ell,i}^{(m)}, \tag{5.9}$$

where

$$\epsilon_{\ell,i}^{(m)} \equiv q_\ell^{(m)T} \delta R_m q_i^{(m)},$$

and

$$|\epsilon_{\ell,i}^{(m)}| \leq m\sigma\epsilon_2, \tag{5.10}$$

which follows from (4.11). If we take $i = \ell$, the left hand side of (5.9) is zero, and we get

$$z_i^{(m)T} \hat{v}_{m+1} = -\frac{\epsilon_{i,i}^{(m)}}{\hat{\beta}_{m+1} \eta_{m,i}^{(m)}}. \tag{5.11}$$

We have

$$|z_i^{(m)T} \hat{v}_{m+1}| \leq \|z_i^{(m)}\| \cdot \|\hat{v}_{m+1}\| \leq (1 + \epsilon_0/4)\, \|z_i^{(m)}\|,$$

and from (5.7) and (5.10),

$$|z_i^{(m)T} \hat{v}_{m+1}| \leq \frac{m\sigma\epsilon_2}{\min_g |\mu_g^{(m)} - \mu_i^{(m)}|}.$$

Therefore, $z_i^{(m)}$ is almost orthogonal to $\hat{v}_{m+1}$ (i.e., $|z_i^{(m)T} \hat{v}_{m+1}| \approx 0$) if we have not yet obtained a small eigenvalue interval about $\mu_i^{(m)}$, the eigenvector approximation $z_i^{(m)}$ does not have a small norm, and $\bar{\Gamma}_k$, and thus ϵ_0 and ϵ_2, are small.

Definition 5.2. *[30, Definition 2] We will say that an eigenpair* (μ, z) represents an eigenpair of A to within δ *if we know that*

$$\frac{\|Az - \mu z\|}{\|z\|} \leq \delta.$$

Thus if (μ, z) represents an eigenpair of A to within δ, then (μ, z) is an exact eigenpair of A perturbed by a matrix whose 2-norm is no greater than δ, and if μ is the Rayleigh quotient of A with z, then the perturbation will be taken symmetric.

Multiplying (4.2) on the right by $q_i^{(m)}$, we get

$$A\hat{V}_m q_i^{(m)} = \hat{V}_m \hat{T}_m q_i^{(m)} + \hat{\beta}_{m+1} \hat{v}_{m+1} e_m^T q_i^{(m)} + \delta\hat{V}_m q_i^{(m)}.$$

Using (5.1) and (5.2), this can be written

$$\begin{aligned} Az_i^{(m)} &= \hat{V}_m q_i^{(m)} \mu_i^{(m)} + \hat{\beta}_{m+1} \hat{v}_{m+1} \eta_{m,i}^{(m)} + \delta\hat{V}_m q_i^{(m)} \\ &= z_i^{(m)} \mu_i^{(m)} + \hat{\beta}_{m+1} \hat{v}_{m+1} \eta_{m,i}^{(m)} + \delta\hat{V}_m q_i^{(m)}, \end{aligned}$$

and thus we obtain

$$Az_i^{(m)} - \mu_i^{(m)} z_i^{(m)} = \hat{\beta}_{m+1} \eta_{m,i}^{(m)} \hat{v}_{m+1} + \delta\hat{V}_m q_i^{(m)}. \tag{5.12}$$

Now, using the above and (4.4), (4.3), and (5.7), if λ_ℓ are the eigenvalues of A, then

$$\begin{aligned}\min_\ell |\lambda_\ell - \mu_i^{(m)}| &\leq \frac{\|Az_i^{(m)} - \mu_i^{(m)} z_i^{(m)}\|}{\|z_i^{(m)}\|} \\ &\leq \frac{\|\hat{\beta}_{m+1}\eta_{m,i}^{(m)}\hat{v}_{m+1} + \delta\hat{V}_m q_i^{(m)}\|}{\|z_i^{(m)}\|} \\ &\leq \frac{\delta_{m,i}(1+\epsilon_0) + m^{1/2}\sigma\epsilon_1}{\|z_i^{(m)}\|}, \end{aligned} \tag{5.13}$$

and if

$$\|z_i^{(m)}\| \approx 1, \tag{5.14}$$

then $(\mu_i^{(m)}, z_i^{(m)})$ represents an eigenpair of A to within about $\delta_{m,i}$. Unfortunately, one can not expect (5.14) to hold in finite precision.

From (5.2) and (4.6), we see that

$$\begin{aligned}\|z_i^{(m)}\|^2 = z_i^{(m)T} z_i^{(m)} &= q_i^{(m)T}\hat{V}_m^T\hat{V}_m q_i^{(m)} \\ &= q_i^{(m)T}\big(R_m^T + \operatorname{diag}(\hat{V}_m^T\hat{V}_m) + R_m\big)q_i^{(m)} \\ &= 2q_i^{(m)T}R_m q_i^{(m)} + q_i^{(m)T}\operatorname{diag}(\hat{V}_m^T\hat{V}_m)q_i^{(m)},\end{aligned}$$

and subtracting 1 from both sides,

$$\|z_i^{(m)}\|^2 - 1 = 2q_i^{(m)T}R_m q_i^{(m)} + q_i^{(m)T}\operatorname{diag}(\hat{V}_m^T\hat{V}_m - I_m)q_i^{(m)}. \tag{5.15}$$

By (4.4), the last term on the right has magnitude bounded by $\epsilon_0/2$.

Using (5.2), we can write

$$\hat{V}_t^T = Q^{(t)}Z^{(t)T},$$

and multiplying on the right by $\hat{v}_{t+1}$, we get

$$\hat{V}_t^T\hat{v}_{t+1} = Q^{(t)}b_t \quad \text{where } b_t = Z^{(t)T}\hat{v}_{t+1}. \tag{5.16}$$

Using (5.11), we can write

$$e_r^T b_t = e_r^T Z^{(t)T}\hat{v}_{t+1} = z_r^{(t)T}\hat{v}_{t+1} = -\frac{\epsilon_{r,r}^{(t)}}{\hat{\beta}_{t+1}\eta_{t,r}^{(t)}}.$$

Now, by definition, we have

$$R_m = \sum_{t=1}^{m-1}[\hat{V}_t, 0_{n,m-t}]^T\hat{v}_{t+1}e_{t+1}^T,$$

and substituting in (5.16), we obtain

$$R_m = \sum_{t=1}^{m-1}\begin{bmatrix} Q^{(t)} & 0_{t,m-t} \\ 0_{m-t,t} & 0_{m-t,m-t}\end{bmatrix}\begin{bmatrix} b_t \\ 0_{m-t,1}\end{bmatrix}e_{t+1}^T.$$

Multiplying on the left and right by $q_i^{(m)T}$ and $q_i^{(m)}$, respectively, we obtain

$$\begin{aligned} q_i^{(m)T} R_m q_i^{(m)} &= q_i^{(m)T}\Big(\sum_{t=1}^{m-1} \begin{bmatrix} Q^{(t)} & 0_{t,m-t} \\ 0_{m-t,t} & 0_{m-t,m-t} \end{bmatrix} \begin{bmatrix} b_t \\ 0_{m-t,1} \end{bmatrix} e_{t+1}^T q_i^{(m)} \Big) \\ &= q_i^{(m)T} \sum_{t=1}^{m-1} \eta_{t+1,i}^{(m)} \begin{bmatrix} Q^{(t)} & 0_{t,m-t} \\ 0_{m-t,t} & 0_{m-t,m-t} \end{bmatrix} \begin{bmatrix} b_t \\ 0_{m-t,1} \end{bmatrix} \\ &= q_i^{(m)T} \sum_{t=1}^{m-1} \eta_{t+1,i}^{(m)} \sum_{r=1}^{t} \begin{bmatrix} q_r^{(t)} \\ 0_{m-t,1} \end{bmatrix} e_r^T \begin{bmatrix} b_t \\ 0_{m-t,1} \end{bmatrix} \\ &= \sum_{t=1}^{m-1} \eta_{t+1,i}^{(m)} \sum_{r=1}^{t} [e_r^T, 0_{1,m-t}] \begin{bmatrix} b_t \\ 0_{m-t,1} \end{bmatrix} q_i^{(m)T} \begin{bmatrix} q_r^{(t)} \\ 0_{m-t,1} \end{bmatrix} \\ &= -\sum_{t=1}^{m-1} \eta_{t+1,i}^{(m)} \sum_{r=1}^{t} \frac{\epsilon_{r,r}^{(t)}}{\hat\beta_{t+1}\eta_{t,r}^{(t)}} q_i^{(m)T} \begin{bmatrix} q_r^{(t)} \\ 0_{m-t,1} \end{bmatrix}. \end{aligned} \tag{5.17}$$

By (5.8) we have

$$q_i^{(m)T} \begin{bmatrix} q_r^{(t)} \\ 0_{m-t,1} \end{bmatrix} / (\hat\beta_{t+1}\eta_{t,r}^{(t)}) = \frac{\eta_{t+1,i}^{(m)}}{\mu_i^{(m)} - \mu_r^{(t)}},$$

and substituting this into the right hand side of (5.17), we get

$$q_i^{(m)T} R_m q_i^{(m)} = -\sum_{t=1}^{m-1} \big(\eta_{t+1,i}^{(m)}\big)^2 \sum_{r=1}^{t} \frac{\epsilon_{r,r}^{(t)}}{\mu_i^{(m)} - \mu_r^{(t)}}. \tag{5.18}$$

We substitute the expression in (5.4) on the right hand side of (5.18) to obtain

$$q_i^{(m)T} R_m q_i^{(m)} = -\sum_{t=1}^{m-1} \left(\prod_{\substack{\ell=1 \\ \ell\neq i}}^{m} \delta_\ell(t+1,i,m) \cdot \sum_{r=1}^{t} \frac{\epsilon_{r,r}^{(t)}}{\mu_i^{(m)} - \mu_r^{(t)}} \right).$$

Note that t of the $\nu_\ell^{(m)}$ in (5.3) and (5.4) are the eigenvalues $\mu_r^{(t)}$. For $r \in \{1,\dots,t\}$, we let $c(r)$ denote the index such that the numerator of $\delta_{c(r)}(t+1,i,m)$ cancels with $1/(\mu_i^{(m)} - \mu_r^{(t)})$ in (5.18), i.e., $\nu_{c(r)}^{(m)} = \mu_r^{(t)}$ for $c(r) \in \{1,\dots,i-1\}$ and $\nu_{c(r)-1}^{(m)} = \mu_r^{(t)}$ for $c(r) \in \{i+1,\dots,m\}$. Then the previous equation can be written

$$q_i^{(m)T} R_m q_i^{(m)} = -\sum_{t=1}^{m-1} \sum_{r=1}^{t} \left(\frac{\epsilon_{r,r}^{(t)}}{\mu_i^{(m)} - \mu_{c(r)}^{(m)}} \cdot \prod_{\substack{\ell=1 \\ \ell\neq i \\ \ell\neq c(r)}}^{m} \delta_\ell(t+1,i,m) \right). \tag{5.19}$$

From (5.15), under the assumptions in (4.12), $\|z_i^{(m)}\|$ will be significantly different from unity only if the right hand sides of these last three numbered equations are large. In this case (5.17) shows there must be a small $\delta_{t,r} = \hat\beta_{t+1}|\eta_{t,r}^{(t)}|$, and some $\mu_r^{(t)}$ has therefore stabilized. Equation (5.18) shows that some $\mu_r^{(t)}$ must be close to $\mu_i^{(m)}$, and combining this with (5.17) we will show that at least one such $\mu_r^{(t)}$ has stabilized. Finally from (5.19), we see that there is at least one $\mu_{c(r)}^{(m)}$ close to $\mu_i^{(m)}$, so that $\mu_i^{(m)}$ cannot be a well-separated eigenvalue of $\hat{T}_m$. Conversely, if $\mu_i^{(m)}$ is a well-separated eigenvalue of $\hat{T}_m$, then (5.14) holds, and if $\mu_i^{(m)}$ has stabilized, then it and $z_i^{(m)}$ are a satisfactory approximation to an eigenpair of A.

Note that if the assumptions in (4.12) do not hold, $\|z_i^{(m)}\|$ can be significantly differ from unity if $|q_i^{(m)T} R_m q_i^{(m)}|$ is large and/or if $\epsilon_0/2$ is large (e.g., due to a large $\bar{\Gamma}_k^2$; see (4.5)). If $\|z_i^{(m)}\|$ is significantly different from unity and $\epsilon_0/2$ is large, we can not necessarily draw meaningful conclusions about the eigenvalues of $\hat{T}_m$ via (5.17), (5.18), and (5.19) based on the size of $\|z_i^{(m)}\|$.

We will now quantify these claims. We first seek to obtain an upper bound on $|q_i^{(m)T} R_m q_i^{(m)}|$. We note from (5.10) and (4.11) that

$$\sum_{r=1}^{t} \left(\epsilon_{r,r}^{(t)}\right)^2 \le \sum_{r=1}^{t} \sum_{c=1}^{t} \left(\epsilon_{r,c}^{(t)}\right)^2 = \|\delta R_t\|_F^2 \le t^2 \sigma^2 \epsilon_2^2, \tag{5.20}$$

and using the Cauchy-Schwarz inequality,

$$\left(\sum_{r=1}^{t} |\epsilon_{r,r}^{(t)}|\right)^2 \le \sum_{r=1}^{t} \left(\epsilon_{r,r}^{(t)}\right)^2 \sum_{r=1}^{t} 1 \le t^3 \sigma^2 \epsilon_2^2. \tag{5.21}$$

Now, using (5.19) and the bound in (5.5), we can write

$$\begin{aligned} |q_i^{(m)T} R_m q_i^{(m)}| &\le \sum_{t=1}^{m-1} \sum_{r=1}^{t} \frac{|\epsilon_{r,r}^{(t)}|}{|\mu_i^{(m)} - \mu_{c(r)}^{(m)}|} \\ &\le \frac{1}{\min_{\ell \ne i} |\mu_i^{(m)} - \mu_\ell^{(m)}|} \sum_{t=1}^{m-1} \sum_{r=1}^{t} |\epsilon_{r,r}^{(t)}|, \end{aligned}$$

and then using (5.21),

$$\begin{aligned} |q_i^{(m)T} R_m q_i^{(m)}| &\le \frac{\sigma \epsilon_2}{\min_{\ell \ne i} |\mu_i^{(m)} - \mu_\ell^{(m)}|} \sum_{t=1}^{m-1} t^{3/2} \\ &\le \frac{\sigma \epsilon_2}{\min_{\ell \ne i} |\mu_i^{(m)} - \mu_\ell^{(m)}|} \int_{t=0}^{m} t^{3/2} dt \\ &= \frac{\sigma \epsilon_2}{\min_{\ell \ne i} |\mu_i^{(m)} - \mu_\ell^{(m)}|} \cdot \left(\frac{2}{5} m^{5/2}\right) \\ &= \frac{m^{5/2} \sigma \epsilon_2}{(5/2) \min_{\ell \ne i} |\mu_i^{(m)} - \mu_\ell^{(m)}|}. \end{aligned} \tag{5.22}$$

This bound is weak, but it shows that if

$$\min_{\ell \ne i} |\mu_i^{(m)} - \mu_\ell^{(m)}| \ge m^{5/2} \sigma \epsilon_2, \tag{5.23}$$

then from (5.22), $|q_i^{(m)T} R_m q_i^{(m)}| \le 2/5$, and substituting this into (5.15), we have

$$\big|\|z_i^{(m)}\|^2 - 1\big| \le 2\big|q_i^{(m)T} R_m q_i^{(m)}\big| + \frac{\epsilon_0}{2} \le \frac{4}{5} + \frac{\epsilon_0}{2}.$$

Thus,

$$\sqrt{1-\left(\frac{4}{5}+\frac{\epsilon_0}{2}\right)} \leq \|z_i^{(m)}\| \leq \sqrt{1+\left(\frac{4}{5}+\frac{\epsilon_0}{2}\right)}, \tag{5.24}$$

and with the condition that $\epsilon_0 = 2\epsilon(n{+}11s{+}15)\bar{\Gamma}_k^2 < 1/12$ (see (4.12)), we can then guarantee that

$$0.39 < \|z_i^{(m)}\| < 1.4,^* \tag{5.25}$$

which has implications for (5.13).

Comments. Note that we could slightly loosen the bound (4.12) on ϵ_0 and still carry through much of the preceding analysis, although in (5.24) we have assumed that $\epsilon_0/2 < 1 - 4/5 = 1/5$. If we instead have $\epsilon_0/2 \geq 1/5$, we get the trivial bound $0 \leq \|z_i^{(m)}\|^2$. This bound is not useful because in the worst case, $z_i^{(m)}$ is the 0-vector, which indicates either breakdown of the method or rank-deficiency of some $\hat{\mathcal{Y}}_k$.

Note that from (5.2) and (4.4),

$$\begin{aligned}\left|\sum_{i=1}^{m}\|z_i^{(m)}\|_2^2 - m\right| &= \left|\,\|Z^{(m)}\|_F^2 - m\,\right| \\ &= \left|\,\|\hat{V}_m Q^{(m)}\|_F^2 - m\,\right| \\ &= \left|\,\|\hat{V}_m\|_F^2 - m\,\right| \\ &= \left|\sum_{i=1}^{m}\|\hat{v}_i\|_2^2 - m\right| \\ &\leq \left| m\left(1+\frac{\epsilon_0}{2}\right) - m \right| = \frac{m\epsilon_0}{2}.\end{aligned}$$

It was also proven in [27] that if $\mu_i^{(m)}, \ldots, \mu_{i+c}^{(m)}$ are $c+1$ eigenvalues of $\hat{T}_m$ which are close to each other but separate from the rest, then

$$\sum_{\ell=i}^{i+c}\|z_\ell^{(m)}\|^2 \approx c+1. \tag{5.26}$$

This means that it is possible to have several close eigenvalues of $\hat{T}_m$ corresponding to one simple eigenvalue of A. If this is the case, then the columns of

$$Z_c \equiv [z_i^{(m)}, \ldots, z_{i+c}^{(m)}]$$

will all correspond to one eigenvector z of A having $z^T z = 1$. We now prove another result.

Lemma 5.1. *(see [30, Lemma 3.1]) Let $\hat{T}_m$ and $\hat{V}_m$ be the result of $m = sk+j$ steps of the s-step Lanczos method with* (4.5) *and* (4.10), *and let R_m be the strictly*

[*] Note that these bounds differ from those given by Paige in [30], which are $0.42 < \|z_i^{(m)}\| < 1.4$; the present authors suspect that the bounds in [30] were obtained using $\epsilon_0 < 1/100$ rather than the specified $\epsilon_0 < 1/12$, the former being the value used by Paige in his earlier work [27]. This slight change in bound carries through the remainder of this paper, resulting in different constants than those in [30]; the fundamental results and conclusions remain unchanged.

upper triangular matrix defined in (4.6)*. Then for each eigenpair* $(\mu_i^{(m)}, q_i^{(m)})$ *of* $\hat{T}_m$*, there exists a pair of integers* (r,t) *with* $0 \leq r \leq t < m$ *such that*

$$\delta_{t,r} \equiv \hat{\beta}_{t+1} |\eta_{t,r}^{(t)}| \leq \psi_{i,m} \quad \textit{and} \quad |\mu_i^{(m)} - \mu_r^{(t)}| \leq \psi_{i,m},$$

where

$$\psi_{i,m} \equiv \frac{m^2 \sigma \epsilon_2}{\left|\sqrt{3}\, q_i^{(m)T} R_m q_i^{(m)}\right|}.$$

Proof. For $r \leq t < m$ we define, using (5.8),

$$\gamma_{r,t} \equiv (\hat{\beta}_{t+1} \eta_{t,r}^{(t)})^{-1} q_i^{(m)T} \begin{bmatrix} q_r^{(t)} \\ 0_{m-t,1} \end{bmatrix} = \frac{\eta_{t+1,i}^{(m)}}{\mu_i^{(m)} - \mu_r^{(t)}}. \tag{5.27}$$

Using this notation and (5.17), we can write

$$\begin{aligned} q_i^{(m)T} R_m q_i^{(m)} &= -\sum_{t=1}^{m-1} \eta_{t+1,i}^{(m)} \sum_{r=1}^{t} \gamma_{r,t} \epsilon_{r,r}^{(t)}, \\ &\equiv -e^T C \bar{q}, \end{aligned}$$

where above, e is the vector with every element unity, C is an $(m-1)$-by-$(m-1)$ upper triangular matrix with (r,t) element $\gamma_{r,t}\epsilon_{r,r}^{(t)}$, and $\bar{q}$ contains the last $m-1$ elements of $q_i^{(m)}$. Letting E be the $(m-1)$-square matrix with (r,t) element $\epsilon_{r,r}^{(t)}$ and combining this with (5.20) gives

$$\begin{aligned} |q_i^{(m)T} R_m q_i^{(m)}| \leq \|e^T C\|_2 \cdot \|\bar{q}\|_2 \leq \|C^T e\|_2 \leq \|C\|_2 \|e\|_2 &\leq m^{1/2} \|C\|_F \\ &\leq m^{1/2} \cdot \max_{r \leq t < m} |\gamma_{r,t}| \cdot \|E\|_F. \end{aligned} \tag{5.28}$$

Using (5.20), we can write

$$\|E\|_F^2 = \sum_{t=1}^{m-1} \sum_{r=1}^{t} \left(\epsilon_{r,r}^{(t)}\right)^2 \leq \sigma^2 \epsilon_2^2 \sum_{t=1}^{m-1} t^2 \leq \frac{\sigma^2 \epsilon_2^2 m^3}{3},$$

and thus we can take the square root above and substitute into (5.28) to get

$$|q_i^{(m)T} R_m q_i^{(m)}| \leq \frac{m^2 \sigma \epsilon_2 |\gamma_{r,t}|}{\sqrt{3}},$$

where we take the r and t giving the maximum value of $|\gamma_{r,t}|$. For this r and t, substituting in the expression for $\gamma_{r,t}$ from (5.27) into the bound above, rearranging, and using $\eta_{t+1,i}^{(m)} \leq 1$ then gives the desired results

$$\delta_{t,r} = \hat{\beta}_{t+1} |\eta_{t,r}^{(t)}| \leq \frac{m^2 \sigma \epsilon_2}{\left|\sqrt{3}\, q_i^{(m)T} R_m q_i^{(m)}\right|} \quad \text{and} \tag{5.29}$$

$$|\mu_i^{(m)} - \mu_r^{(t)}| \leq \frac{m^2 \sigma \epsilon_2}{\left|\sqrt{3}\, q_i^{(m)T} R_m q_i^{(m)}\right|}. \tag{5.30}$$

□

Comments. For classical Lanczos, these bounds show that if $\|z_i^{(m)}\|_2$ is significantly different from unity, then for some $t < m$ there is an eigenvalue of $\hat{T}_t$ which has stabilized and is close to $\mu_i^{(m)}$ [30]. For the s-step Lanczos case, the same holds with the assumptions in (4.12). These assumptions are necessary because otherwise, for s-step Lanczos, $\|z_i^{(m)}\|$ can significantly differ from unity if $|q_i^{(m)T} R_m q_i^{(m)}|$ is large and/or if $\epsilon_0/2$ is large (due to a large $\bar{\Gamma}_k^2$, see (4.5)). If $\|z_i^{(m)}\|$ is much different from unity and $\epsilon_0/2$ is large, we can not necessarily say that there is an eigenvalue of $\hat{T}_t$ which has stabilized to within a meaningful bound regardless of the size of $|q_i^{(m)T} R_m q_i^{(m)}|$.

THEOREM 5.2. *(see [30, Theorem 3.1]) If, with the conditions of Lemma 5.1, an eigenvalue $\mu_i^{(m)}$ of $\hat{T}_m$ produced by s-step Lanczos is stabilized so that*

$$\delta_{m,i} \equiv \hat{\beta}_{m+1}|\eta_{m,i}^{(m)}| \le \sqrt{3}m^2\sigma\epsilon_2, \tag{5.31}$$

and $\epsilon_0 < 1/12$, then for some eigenvalue λ_c of A,

$$|\lambda_c - \mu_i^{(m)}| \le (m+1)^3\sigma\epsilon_2. \tag{5.32}$$

Proof. Suppose (5.31) holds.
(i) If

$$|q_i^{(m)T} R_m q_i^{(m)}| \le \frac{3}{8} - \frac{\epsilon_0}{2}, \tag{5.33}$$

then by (5.15) and (4.4) we have

$$\begin{aligned}
\|z_i^{(m)}\| &\ge \sqrt{1 - \left(2|q_i^{(m)T} R_m q_i^{(m)}| + \frac{\epsilon_0}{2}\right)} \\
&\ge \sqrt{1 - 2\left(\frac{3}{8} - \frac{\epsilon_0}{2}\right) - \frac{\epsilon_0}{2}} \\
&\ge \sqrt{\frac{1}{4} + \frac{\epsilon_0}{2}} \\
&\ge \frac{1}{2}.
\end{aligned} \tag{5.34}$$

Substituting (4.10) and (4.5) into (5.13), we obtain

$$\begin{aligned}
\min_{\ell} |\lambda_\ell - \mu_i^{(m)}| &\le \frac{\delta_{m,i}(1+\epsilon_0) + \sqrt{m}\sigma\epsilon_1}{\|z_i^{(m)}\|} \\
&\le 2 \cdot \left(\sqrt{3}m^2\sigma\epsilon_2 \cdot \frac{13}{12} + \frac{\sqrt{m}\sigma\epsilon_2}{\sqrt{2}}\right) \\
&\le \sigma\epsilon_2\left(\frac{13\sqrt{3}}{6} \cdot m^2 + \sqrt{2m}\right).
\end{aligned}$$

Since $m \ge 1$, $m^3 \ge m^2$ and $m \ge \sqrt{m}$, it follows that

$$(m+1)^3 = m^3 + 3m^2 + 3m + 1 > 4m^2 + 3\sqrt{m} \ge \frac{13\sqrt{3}}{6}m^2 + \sqrt{2m}.$$

From this it follows that (5.32) holds.

In the other case that (5.33) is false, take $\ell = 1$ and write

$$t_1 = m, \quad r_1 = i. \tag{5.35}$$

(ii) In this case we know from (5.29) and (5.30) that there exist positive integers $r_{\ell+1}$ and $t_{\ell+1}$ with

$$r_{\ell+1} \le t_{\ell+1} < t_\ell \tag{5.36}$$

such that

$$\max\left(\delta_{t_{\ell+1},r_{\ell+1}}, |\mu_{r_\ell}^{t_\ell} - \mu_{r_{\ell+1}}^{t_{\ell+1}}|\right) \le \frac{t_\ell^2 \sigma\epsilon_2}{\sqrt{3}\left(\frac{3}{8} - \frac{\epsilon_0}{2}\right)} \le \frac{t_\ell^2 \sigma\epsilon_2}{\sqrt{3}\left(\frac{1}{3}\right)} \le \sqrt{3} t_\ell^2 \sigma\epsilon_2. \tag{5.37}$$

If the equivalent of (5.33), and thus (5.34), holds for $(r_{\ell+1}, t_{\ell+1})$, i.e.,

$$|q_{r_{\ell+1}}^{(t_{\ell+1})T} R_{t_{\ell+1}} q_{r_{\ell+1}}^{(t_{\ell+1})}| \le 3/8 - \epsilon_0/2,$$

then using (5.13), for some eigenvalue λ_c of A,

$$\begin{aligned}
|\lambda_c - \mu_{r_{\ell+1}}^{(t_{\ell+1})}| &\le 2\left(\sqrt{3} t_\ell^2 \sigma\epsilon_2 (1+\epsilon_0) + \sqrt{t_{\ell+1}}\sigma\epsilon_1\right) \\
&\le 2\left(\sqrt{3} t_\ell^2 \sigma\epsilon_2 (1+\epsilon_0) + \frac{\sqrt{t_{\ell+1}}\sigma\epsilon_2}{\sqrt{2}}\right) \\
&= \left(2\sqrt{3} t_\ell^2 (1+\epsilon_0) + \sqrt{2t_{\ell+1}}\right)\sigma\epsilon_2 \\
&\le \left(2\sqrt{3} t_\ell^2 \left(1 + \frac{1}{12}\right) + \sqrt{2t_{\ell+1}}\right)\sigma\epsilon_2 \\
&= \left(\frac{13\sqrt{3} t_\ell^2}{6} + \sqrt{2t_{\ell+1}}\right)\sigma\epsilon_2,
\end{aligned}$$

which gives

$$\begin{aligned}
|\lambda_c - \mu_i^{(m)}| &\le |\lambda_c - \mu_{r_{\ell+1}}^{(t_{\ell+1})}| + \sum_{p=1}^{\ell} |\mu_{r_p}^{(t_p)} - \mu_{r_{p+1}}^{(t_{p+1})}| \\
&\le \left(\frac{13\sqrt{3} t_\ell^2}{6} + \sqrt{2t_{\ell+1}} + \sqrt{3}\sum_{p=1}^{\ell} t_p^2\right)\sigma\epsilon_2 \\
&\le \left(\frac{13\sqrt{3} m^2}{6} + \sqrt{2m} + \frac{\sqrt{3} m(m+1)(2m+1)}{6}\right)\sigma\epsilon_2 \\
&\le (m+1)^3 \sigma\epsilon_2,
\end{aligned}$$

as required by (5.32), where the penultimate inequality follows from (5.35) and (5.36). If the equivalent of (5.33) does not hold, then replace ℓ by $\ell + 1$ and return to (ii).

We see that $\hat{T}_1 = \hat{\alpha}_1$, $q_1^{(1)} = 1$, so that $z_1^{(1)} = v_1$ satisfies (5.34), proving that we must encounter an $(r_{\ell+1}, t_{\ell+1})$ pair satisfying (5.34), which completes the proof. □

Comments. It is clear from the derivation that the bound (5.32) is not tight, and thus should in no way be considered an indication of the maximum attainable accuracy. In practice we can still observe convergence of Ritz values to eigenvalues of A with larger $\bar{\Gamma}_k^2$ than allowed by $\epsilon_0 < 1/12$. The form of the bound in (5.33) makes the assumption that $\epsilon_0 < 3/4$, and the bound $\epsilon_0 < 1/12$ is used in (5.34). As a result of the form (5.33), (5.37) requires that $3/8 - \epsilon_0/2 \geq 1/3$, which gives $\epsilon_0 \leq 1/12$. The present authors believe that the restriction on the size of ϵ_0 could be loosened by a constant factor by changing the form of the right hand side of (5.33) such that meaningful bounds are still obtained. This remains future work.

The following shows we have an eigenvalue with a superior error bound to (5.32) and that we also have a good eigenvector approximation.

Corollary 5.3. *(see [30, Corollary 3.1]) If* (5.31) *holds, then for the final* (r, t) *pair in Theorem 5.2,* $(\mu_r^{(t)}, \hat{V}_t q_r^{(t)})$ *is an exact eigenpair for a matrix within* $6t^2\sigma\epsilon_2$ *of* A.

Proof. From Theorem 5.2, if there is an i, $1 \leq i \leq m$ such that (5.31) holds, then there exist r and t, $1 \leq r \leq t \leq m$ such that

$$\delta_{t,r} \leq \sqrt{3}t^2\sigma\epsilon_2 \quad \text{and} \quad \|z_r^{(t)}\| \geq \frac{1}{2},$$

and both $\mu_r^{(t)}$ and $\mu_i^{(m)}$ are close to the same eigenvalue of A. It follows from (5.12) that

$$\begin{aligned}
(A + \delta A_r^{(t)}) z_r^{(t)} &= \mu_r^{(t)} z_r^{(t)}, \text{ with} && (5.38)\\
\delta A_r^{(t)} &\equiv -(\hat{\beta}_{t+1}\eta_{t,r}^{(t)}\hat{v}_{t+1} + \delta V_t q_r^{(t)}) \frac{z_r^{(t)T}}{\|z_r^{(t)}\|^2}, \text{ and} \\
\|\delta A_r^{(t)}\| &\leq \left(|\delta_{t,r}| \cdot \|\hat{v}_{t+1}\| + \|\delta\hat{V}_t\| \cdot \|q_r^{(t)}\| \right) \frac{1}{\|z_r^{(t)}\|} \\
&\leq 2\left(\sqrt{3}t^2\sigma\epsilon_2(1+\epsilon_0) + \sqrt{t}\sigma\epsilon_1 \right) && (5.39)\\
&\leq \left(\frac{13\sqrt{3}t^2\sigma\epsilon_2}{6} + \frac{2\sqrt{t}\sigma\epsilon_2}{\sqrt{2}} \right) \\
&\leq \left(\frac{13\sqrt{3}}{6}t^2 + \sqrt{2t} \right)\sigma\epsilon_2 \\
&\leq \left(\frac{13\sqrt{3}}{6} + \sqrt{2} \right) t^2\sigma\epsilon_2,
\end{aligned}$$

which gives

$$\|\delta A_r^{(t)}\| \leq 6t^2\sigma\epsilon_2, ^{\dagger} \qquad (5.40)$$

where we have used (4.3), (4.4), and (4.10).

[†] Note that in [30], Paige obtains 5 as the leading coefficient in (5.40) rather than 6. The present authors are unable to determine how Paige obtained this result; one possiblity is that the ϵ_0 term in (5.39) was dropped, as it results in an ϵ^2 term on the right hand side.

So, $z_r^{(t)}$, which lies in the range of $\hat{V}_r$, is an exact eigenvector of a matrix close to A, and $\mu_r^{(t)}$ is the corresponding exact eigenvalue. □

As in the classical Lanczos case, the above is also the result we obtain for an eigenvalue of $\hat{T}_m$ produced by s-step Lanczos that is stabilized and well-separated (see definitions in (5.31) and (5.23), respectively).

Paige showed that one can also consider the accuracy of the $\mu_i^{(m)}$ as Rayleigh quotients [30]. With no rounding errors, $\mu_i^{(m)}$ is the Rayleigh quotient of A with $z_i^{(m)}$ and this gives the best bound from (5.12) and (5.13) with $\epsilon = 0$, i.e., in exact arithmetic. Here (5.11) and (5.12) can be combined to give

$$z_i^{(m)T} A z_i^{(m)} - \mu_i^{(m)} z_i^{(m)T} z_i^{(m)} = -\epsilon_{i,i}^{(m)} + z_i^{(m)T} \delta\hat{V}_m q_i^{(m)},$$

so if $\|z_i^{(m)}\| \approx 1$, then $\mu_i^{(m)}$ is close to the Rayleigh quotient

$$\varrho_i^{(m)} = z_i^{(m)T} A z_i^{(m)} / z_i^{(m)T} z_i^{(m)}.$$

If (5.23) holds, then $\|z_i^{(m)}\| > 0.39$, and thus dividing both sides of the above equation by $z_i^{(m)T} z_i^{(m)}$, we can write the bound

$$|\varrho_i^{(m)} - \mu_i^{(m)}| \leq \frac{|\epsilon_{i,i}^{(m)}|}{\|z_i^{(m)}\|^2} + \frac{\|\delta\hat{V}_m\| \cdot \|q_i^{(m)}\|}{\|z_i^{(m)}\|}.$$

Applying (5.10), (4.3), and (4.10) to the right hand side above, we can write the bound

$$|\varrho_i^{(m)} - \mu_i^{(m)}| \leq \left(\frac{1}{0.39^2} + \frac{1}{\sqrt{2} \cdot 0.39} \right) m\sigma\epsilon_2 \leq 9m\sigma\epsilon_2.$$

If $\|z_i^{(m)}\|$ is small, then it is unlikely that $\mu_i^{(m)}$ will be very close to $\varrho_i^{(m)}$, since a small $z_i^{(m)}$ will probably be inaccurate due to rounding errors. The equation (5.26) suggests that at least one of a group of close eigenvalues will have corresponding $\|z_i^{(m)}\| \gtrsim 1$. In fact, using (5.18), (5.21), and an argument similar to that used in Theorem 5.2, it can be shown that every $\mu_i^{(m)}$ lies within $m^{5/2}\sigma\epsilon_2$ of a Rayleigh quotient of A, and so with (4.5) and (4.10), all the $\mu_i^{(m)}$ lie in the interval

$$\lambda_{\min} - m^{5/2}\sigma\epsilon_2 \leq \mu_i^{(m)} \leq \lambda_{\max} + m^{5/2}\sigma\epsilon_2.$$

This differs from the bound on the distance of $\mu_i^{(m)}$ from an eigenvalue of A in (5.32), which requires that $\mu_i^{(m)}$ has stabilized.

We emphasize that whatever the size of $\delta_{m,i}$, the eigenvalue $\mu_i^{(m)}$ of $\hat{T}_m$ with eigenvector $q_i^{(m)}$ has necessarily stabilized to within $\delta_{m,i} \equiv \hat{\beta}_{m+1}|e_m^T q_i^{(m)}|$. If $\mu_i^{(m)}$ is a separated eigenvalue of $\hat{T}_m$ so that (5.23) holds, then it follows from (5.25), (5.12), and (5.13) that

$$\begin{aligned}
\frac{\|Az_i^{(m)} - \mu_i^{(m)} z_i^{(m)}\|}{\|z_i^{(m)}\|} &\leq \frac{\delta_{m,i}(1+\epsilon_0) + \sqrt{m}\sigma\epsilon_1}{\|z_i^{(m)}\|} \\
&\leq \left(\frac{13}{0.39 \cdot 12} \right) \left(\delta_{m,i} + \sqrt{m}\sigma\epsilon_1 \right) \\
&\leq 3(\delta_{m,i} + \sqrt{m}\sigma\epsilon_1),
\end{aligned}$$

which means that the eigenpair $(\mu_i^{(m)}, \hat{V}_m q_i^{(m)})$ represents an eigenpair of A to within

$$3\big(\delta_{m,i} + \sqrt{m}\sigma\epsilon_1\big). \tag{5.41}$$

On the other hand, if $\mu_i^{(m)}$ is one of a close group of eigenvalues of $\hat{T}_m$, so that (5.23) does not hold, then we have found a good approximation to an eigenvalue of A. In this case either (5.33) holds, in which case (5.34), (5.12), and (5.13) show that $(\mu_i^{(m)}, \hat{V}_m q_i^{(m)})$ represents an eigenpair of A to within (5.41), or there exists $1 \le r \le t < m$ such that

$$\max\Big(\delta_{t,r}, |\mu_i^{(m)} - \mu_r^{(t)}|\Big) \le \sqrt{3}m^2\sigma\epsilon_2,$$

from Lemma 5.1. Then, it follows from Theorem 5.2 that $\mu_i^{(m)}$ is within $\Big((m+1)^3 + \sqrt{3}m^2\Big)\sigma\epsilon_2$ of an eigenvalue of A. The $\delta_{m,i}$ and $\mu_i^{(m)}$ can be computed from $\hat{T}_m$ quite quickly, and these results show how we can obtain intervals from them which are known to contain eigenvalues of A, whether $\delta_{m,i}$ is large or small.

6. Convergence of eigenvalues.

Theorem 5.2 showed that, assuming (4.12) holds, if an eigenvalue of $\hat{T}_m$ has stabilized to within $\sqrt{3}m^2\sigma\epsilon_2$, then it is within $(m+1)^3\sigma\epsilon_2$ of an eigenvalue of A, regardless of how many other eigenvalues of $\hat{T}_m$ are close, and Corollary 5.3 showed we had an eigenpair of a matrix within $6m^2\sigma\epsilon_2$ of A. It is now shown that, assuming (4.12), eigenvalues do stabilize to this accuracy using the s-step Lanczos method, and we can give an indication of how quickly this occurs.

It was shown in [27] that at least one eigenvalue of $\hat{T}_m$ must have stabilized by when $m = n$. This is based on (5.11), which indicates that significant loss of orthogonality implies stabilization of at least one eigenvalue. Using (5.16) and (5.11), and the fact that $\|\hat{V}_t^T \hat{v}_{t+1}\|_2^2 = \sum_{i=1}^t |\hat{v}_i^T \hat{v}_{t+1}|^2$, we can write

$$\begin{aligned}
\|R_m\|_F^2 &= \sum_{t=1}^{m-1}\sum_{i=1}^{t} |\hat{v}_i^T \hat{v}_{t+1}|^2 \\
&= \sum_{t=1}^{m-1} \|\hat{V}_t^T \hat{v}_{t+1}\|_2^2 \\
&= \sum_{t=1}^{m-1} \|Q^{(t)} Z^{(t)T} \hat{v}_{t+1}\|_2^2 \\
&= \sum_{t=1}^{m-1} \|Z^{(t)T} \hat{v}_{t+1}\|_2^2 \\
&= \sum_{t=1}^{m-1}\sum_{i=1}^{t} |z_i^{(t)T} \hat{v}_{t+1}|^2 \\
&\le \sum_{t=1}^{m-1}\sum_{i=1}^{t} \left(\frac{|\epsilon_{i,i}^{(t)}|}{|\hat{\beta}_{t+1}\eta_{t,i}^{(t)}|}\right)^2.
\end{aligned}$$

Then, if at step m

$$\delta_{\ell,i} \equiv \hat{\beta}_{\ell+1}|\eta_{\ell,i}^{(\ell)}| \ge \sqrt{3}m^2\sigma\epsilon_2, \quad 1 \le i \le \ell < m, \tag{6.1}$$

we have, with the bound (5.10),

$$\|R_m\|_F^2 \leq \sum_{t=1}^{m-1} t \cdot \frac{t^2\sigma^2\epsilon_2^2}{3m^4\sigma^2\epsilon_2^2} = \frac{1}{3m^4}\sum_{t=1}^{m-1} t^3 = \frac{1}{3m^4}\left(\frac{(m-1)m}{2}\right)^2 \leq \frac{1}{12}.$$

Let $\sigma_1 \geq \cdots \geq \sigma_m$ be the singular values of $\hat{V}_m$. Lemma 2.2 of Rump [34] states that given a matrix $X \in \mathbb{R}^{n\times m}$, if $\|I - X^TX\|_2 \leq \alpha < 1$, then $\sqrt{1-\alpha} \leq \sigma_i(X) \leq \sqrt{1+\alpha}$. By the above, we have

$$\begin{aligned}\|I - \hat{V}_m^T\hat{V}_m\|_2 &\leq 2\|R_m\|_2 + \|\mathrm{diag}(I - \hat{V}_m^T\hat{V}_m)\|_2\\ &\leq 2\|R_m\|_F + \epsilon_0/2\\ &\leq 2/\sqrt{12} + 1/24\\ &< 1.\end{aligned}$$

Then with $\alpha = 2/\sqrt{12} + 1/24$, we apply the result of Rump to obtain the bounds

$$0.61 < \sqrt{1 - \frac{2}{\sqrt{12}} - \frac{1}{24}} \leq \sigma_i(\hat{V}_m) \leq \sqrt{1 + \frac{2}{\sqrt{12}} + \frac{1}{24}} < 1.3,\ ^{\ddagger} \tag{6.2}$$

for $i \in \{1, \ldots, m\}$.

Note that if (6.1) does not hold, then we already have an eigenpair of a matrix close to A. If we now consider the $q_i^{(m)}$ that minimizes $\delta_{m,i}$ for $\hat{T}_m$, we see from (5.16), (5.10), and (5.11) that

$$\begin{aligned}\|\hat{\beta}_{m+1}\eta_{m,i}^{(m)}\hat{V}_m^T\hat{v}_{m+1}\| &\leq \|\hat{\beta}_{m+1}\eta_{m,i}^{(m)}Q^{(m)}Z^{(m)}\hat{v}_{m+1}\|\\ &\leq |\hat{\beta}_{m+1}\eta_{m,i}^{(m)}| \cdot \|Q^{(m)}\| \cdot \|Z^{(m)}\hat{v}_{m+1}\|\\ &\leq |\hat{\beta}_{m+1}\eta_{m,i}^{(m)}| \cdot \sqrt{m} \cdot \frac{m\sigma\epsilon_2}{|\hat{\beta}_{m+1}\eta_{m,i}^{(m)}|}\\ &\leq m^{3/2}\sigma\epsilon_2. \end{aligned}\tag{6.3}$$

Theorem 6.1. *(see [30, Theorem 4.1]) For the s-step Lanczos method, if $n(3\epsilon_0 + \epsilon_1) \leq 1$ and $\epsilon_0 < 1/12$, then at least one eigenvalue of $\hat{T}_n$ must be within $(n+1)^3\sigma\epsilon_2$ of an eigenvalue of the $n \times n$ matrix A, and there exist $r \leq t \leq n$ such that $(\mu_r^{(t)}, z_r^{(t)})$ is an exact eigenpair of a matrix within $6t^2\sigma\epsilon_2$ of A.*

Proof. If (6.1) does not hold for $m = n$, then an eigenvalue has stabilized to that accuracy before $m = n$. Otherwise, (6.1) holds for $m = n$, so from (6.2), $\hat{V}_n$ is nonsingular, and then (6.3) shows that for the smallest $\delta_{n,i}$ of $\hat{T}_n$,

$$\delta_{n,i} \leq \frac{n^{3/2}\sigma\epsilon_2}{0.6} \leq \sqrt{3}n^2\sigma\epsilon_2$$

[‡] Again, note that these bounds differ from those given by Paige in [30], which are $0.41 < \sigma_i(\hat{V}_m) < 1.6$; the present authors suspect that the bounds in [30] were obtained by squaring both sides of the bound and using $\epsilon_0 < 1/100$ rather than the specified $\epsilon_0 < 1/12$, the former being the value used by Paige in his earlier work [27]. This slight change in bound carries through the remainder of this paper, resulting in different constants than those in [30]; the fundamental results and conclusions remain unchanged.

since $\|\hat{V}_m^T \hat{v}_{m+1}\| \geq \sigma_m \|\hat{v}_{m+1}\| > 0.6$ from (6.2) and (4.4). So at least one eigenvalue must have stabilized to within $\sqrt{3}m^2\sigma\epsilon_2$ by $m = n$, and from Theorem 5.2 this eigenvalue must be within $(n+1)^3\sigma\epsilon_2$ of an eigenvalue of A. In fact, Corollary 5.3 shows that there is an exact eigenpair $(\mu_r^{(t)}, z_r^{(t)})$, $r \leq t \leq n$, of a matrix within $6t^2\sigma\epsilon_2$ of A. □

This shows that the s-step Lanczos algorithm gives at least one eigenvalue of A to high accuracy by $m = n$, assuming restrictions on the sizes of ϵ_0 and ϵ_1. We now extend Paige's results on how quickly we can expect to find eigenvalues and eigenvectors of A using the s-step Lanczos method in practice. We first consider the Krylov sequence on which the Lanczos algorithm and several other methods are based. For symmetric A, one way of using m steps of the Krylov sequence is to form an $n \times m$ matrix V whose columns span the range of

$$[v_1, Av_1, \ldots, A^{m-1}v_1] \tag{6.4}$$

and use the eigenvalues of

$$V^T AVq = \mu V^T Vq \tag{6.5}$$

as approximations to some of the eigenvalues of A.

In the presence of rounding errors, m steps of the s-step Lanczos algorithm with full reorthogonalization form an $m \times m$ matrix $\hat{T}$ and an $n \times m$ matrix $\hat{V}$ such that the columns of $\hat{V}$ span the exact Krylov subspace of $A + \delta A$ starting with v_1.

Theorem 6.2. *(see [30, Theorem 4.2]) For m iterations of the s-step Lanczos method, with* (4.5), (4.10), *and m such that* (6.1) *holds, the m Lanczos vectors (columns of $\hat{V}_m$) span a Krylov subspace of a matrix within* $(3m)^{1/2}\sigma\epsilon_2$ *of A.*

Proof. From (4.2),

$$\begin{aligned} A\hat{V}_m &= \hat{V}_m\hat{T}_m + \hat{\beta}_{m+1}\hat{v}_{m+1}e_m^T + \delta\hat{V}_m \\ &= \hat{V}_{m+1}\hat{T}_{m+1,m} + \delta\hat{V}_m \end{aligned}$$

where $\hat{T}_{m+1,m}$ is the matrix of the first m columns of $\hat{T}_{m+1}$. Then with (6.2), (4.3), and (4.10),

$$\begin{aligned} (A + \delta A_m)\hat{V}_m &= \hat{V}_{m+1}\hat{T}_{m+1,m}, \text{ with} \\ \delta A_m &\equiv -\delta\hat{V}_m(\hat{V}_m^T\hat{V}_m)^{-1}\hat{V}_m^T, \quad \text{and} \\ \|\delta A_m\|_F &= \text{trace}(\delta A_m \delta A_m^T)^{1/2} \\ &= \left(\sum_{i=1}^m |(\delta A_m \delta A_m^T)_{i,i}|\right)^{1/2} \\ &= \left(\sum_{i=1}^m |(\delta\hat{V}_m(\hat{V}_m^T\hat{V}_m)^{-1}\delta\hat{V}_m^T)_{i,i}|\right)^{1/2} \\ &\leq (m \cdot \sigma\epsilon_1 \cdot \frac{1}{0.6^2} \cdot \sigma\epsilon_1)^{1/2} \\ &\leq \sqrt{\frac{m\sigma^2\epsilon_1^2}{0.6^2}} = \frac{\sqrt{m}\sigma\epsilon_1}{0.6} \leq \frac{\sqrt{m}\sigma\epsilon_2}{\sqrt{2}\cdot 0.6} \\ &\leq (3m)^{1/2}\sigma\epsilon_2. \end{aligned}$$

□

Comments. From this we see that for $i \leq m+1$, $\hat{v}_1, \ldots, \hat{v}_i$ computed by the s-step Lanczos method span the same space as the first i Krylov vectors for $A+\delta A_m$ starting with $\hat{v}_1$. This is analogous to the important result of Paige for classical Lanczos: until an eigenvalue of $\hat{T}_{m-1}$ has stabilized, i.e., while (6.1) holds, the vectors $\hat{v}_1, \ldots, \hat{v}_{m+1}$ computed correspond to an exact Krylov sequence for the matrix $A + \delta A_m$. As a result of this and (5.38), assuming that the s-step bases generated in each outer loop are conditioned such that (4.12) holds, the s-step Lanczos algorithm can be thought of as a numerically stable way of computing a Krylov sequence, at least until the corresponding Krylov subspace contains an exact eigenvector of a matrix within $6m^2\sigma\epsilon_2$ of A. Note that a similar result using the technique of writing the finite precision Lanczos recurrence as a recurrence for perturbed A has been used in analysis of the s-step biconjugate gradient method [3].

However, a Krylov subspace can be very sensitive to small perturbations in A. In the case where $\hat{T}_m$ and $\hat{V}_m$ are used to solve the eigenproblem of A, if we follow (6.4) and (6.5), we want the eigenvalues and eigenvectors of $\hat{T}_m$ to be close to those of

$$\hat{V}_m^T A \hat{V}_m q = \mu \hat{V}_m^T \hat{V}_m q, \quad q^T q = 1, \tag{6.6}$$

as would be the case with classical Lanczos with full reorthogonalization (see [27]). If (6.1) holds, then the range of $\hat{V}_m$ is close to what we expect from the s-step Lanczos method with full reorthogonalization, and thus the eigenvalues of (6.6) would be close (how close depends on the value of ϵ_2) to those obtained using full reorthogonalization. (Note that performing full reorthogonalization in the s-step Lanczos method would reintroduce undesirable communication.)

Theorem 6.3. *(see [30, Theorem 4.3]) If $\hat{V}_m$ comes from the s-step Lanczos method with* (4.5) *and* (4.10)*, and* (6.1) *holds, then for any μ and q which satisfy* (6.6)*, $(\mu, \hat{V}_m q)$ is an exact eigenpair for a matrix within $\left(2\delta + 2m^{1/2}\sigma\epsilon_2\right)$ of A, where*

$$\eta \equiv e_m^T q, \quad \delta \equiv \hat{\beta}_{m+1}|\eta|.$$

Proof. Define

$$r \equiv A\hat{V}_m q - \mu \hat{V}_m q. \tag{6.7}$$

Then

$$r = \hat{V}_m(\hat{T}_m - \mu I)q + \hat{\beta}_{m+1}\eta\hat{v}_{m+1} + \delta\hat{V}_m q$$

where we have used (4.2). Since from (6.6), $\hat{V}_m^T r = 0$,

$$(\hat{T}_m - \mu I)q = -(\hat{V}_m^T\hat{V}_m)^{-1}\hat{V}_m^T(\hat{\beta}_{m+1}\eta\hat{v}_{m+1} + \delta\hat{V}_m q), \tag{6.8}$$

$$r = P_m(\hat{\beta}_{m+1}\eta\hat{v}_{m+1} + \delta\hat{V}_m q), \tag{6.9}$$

where $P_m = I - \hat{V}_m(\hat{V}_m^T\hat{V}_m)^{-1}\hat{V}_m^T$ is the projector orthogonal to the range of $\hat{V}_m$. Using (6.1) and (6.3), we can bound

$$\|\hat{V}_m^T\hat{v}_{m+1}\| \leq \frac{m^{3/2}\sigma\epsilon_2}{\sqrt{3}m^2\sigma\epsilon_2} \leq (3m)^{-1/2}. \tag{6.10}$$

We can write

$$\begin{aligned}\|P_m \hat{v}_{m+1}\|^2 &= \hat{v}_{m+1}^T P_m \hat{v}_{m+1} \\ &= \hat{v}_{m+1}^T \hat{v}_{m+1} - \hat{v}_{m+1}^T \hat{V}_m (\hat{V}_m^T \hat{V}_m)^{-1} \hat{V}_m^T \hat{v}_{m+1},\end{aligned}$$

and then using (4.4), (6.2), and (6.10),

$$1 - \frac{\epsilon_0}{2} - \frac{1}{m} \leq 1 - \frac{\epsilon_0}{2} - \left(\frac{1}{\sqrt{3m}} \cdot \frac{1}{0.6^2} \cdot \frac{1}{\sqrt{3m}} \right) \leq \|P_m \hat{v}_{m+1}\|^2 \leq 1 + \frac{\epsilon_0}{2}. \tag{6.11}$$

Using (6.9), (6.11), (4.3), and (4.10), we can write the bound

$$\begin{aligned}\|r\| &\leq \|P_m \hat{v}_{m+1}\| |\hat{\beta}_{m+1} \eta| + \|P_m\| \|\delta \hat{V}_m q\| \\ &\leq (1 + \epsilon_0)\, \delta + \sqrt{m} \sigma \epsilon_1) \\ &\leq (1 + \epsilon_0)\, \delta + \frac{\sqrt{m} \sigma \epsilon_2}{\sqrt{2}}.\end{aligned}$$

Finally, from (6.2) we have

$$0.61 < \|\hat{V}_m q\| < 1.3.$$

Then from (6.7),

$$(A - \delta A) \hat{V}_m q = \mu \hat{V}_m q, \quad \text{where} \quad \delta A \equiv \frac{r q^T \hat{V}_m^T}{\|\hat{V}_m q\|^2},$$

with

$$\begin{aligned}\|\delta A\|_F = \frac{\|r\|}{\|\hat{V}_m q\|} &\leq \frac{1}{0.6} \left((1 + \epsilon_0)\, \delta + \frac{m^{1/2} \sigma \epsilon_2}{\sqrt{2}} \right) \\ &\leq \frac{13}{12 \cdot 0.6}\, \delta + \frac{1}{\sqrt{2} \cdot 0.6}\, m^{1/2} \sigma \epsilon_2 \\ &\leq 2\delta + 2 m^{1/2} \sigma \epsilon_2. \end{aligned} \tag{6.12}$$

□

Ordering the eigenvalues of $\hat{T}_m$ such that

$$\delta_{m,1} \geq \delta_{m,2} \geq \cdots \geq \delta_{m,m},$$

and assuming (6.1) holds for $\ell = m$, then for any eigenpair of (6.6), (6.8) gives, using (6.2), (6.3), (4.3), and (4.10),

$$\begin{aligned}
\|\hat{T}_m q - \mu q\| &\leq \|(\hat{V}_m^T \hat{V}_m)^{-1}\| \left(\|\hat{\beta}_{m+1}\eta \hat{V}_m^T \hat{v}_{m+1}\| + \|\hat{V}_m\| \|\delta \hat{V}_m q\| \right) \\
&= \|(\hat{V}_m^T \hat{V}_m)^{-1}\| \left(\left\| \hat{\beta}_{m+1}\eta \cdot \frac{\hat{\beta}_{m+1}\eta_{m,i}^{(m)}}{\hat{\beta}_{m+1}\eta_{m,i}^{(m)}} \cdot \hat{V}_m^T \hat{v}_{m+1} \right\| + \|\hat{V}_m\| \|\delta \hat{V}_m q\| \right) \\
&= \|(\hat{V}_m^T \hat{V}_m)^{-1}\| \left(\left\| \frac{\hat{\beta}_{m+1}\eta}{\hat{\beta}_{m+1}\eta_{m,i}^{(m)}} \cdot \hat{\beta}_{m+1}\eta_{m,i}^{(m)} \hat{V}_m^T \hat{v}_{m+1} \right\| + \|\hat{V}_m\| \|\delta \hat{V}_m q\| \right) \\
&= \|(\hat{V}_m^T \hat{V}_m)^{-1}\| \left(\frac{\delta}{\delta_{m,i}} \cdot \|\hat{\beta}_{m+1}\eta_{m,i}^{(m)} \hat{V}_m^T \hat{v}_{m+1}\| + \|\hat{V}_m\| \|\delta \hat{V}_m q\| \right) \\
&\leq \|(\hat{V}_m^T \hat{V}_m)^{-1}\| \left(\frac{\delta}{\delta_{m,m}} \cdot \|\hat{\beta}_{m+1}\eta_{m,i}^{(m)} \hat{V}_m^T \hat{v}_{m+1}\| + \|\hat{V}_m\| \|\delta \hat{V}_m q\| \right) \\
&\leq \frac{1}{0.6^2} \cdot \frac{m^{3/2}\sigma\epsilon_2\delta}{\delta_{m,m}} + \frac{1}{0.6} \cdot \frac{\sqrt{m}\sigma\epsilon_2}{\sqrt{2}} \\
&\leq \frac{3m^{3/2}\sigma\epsilon_2\delta}{\delta_{m,m}} + 2m^{1/2}\sigma\epsilon_2 \\
&\leq \left(2 + \frac{3m\delta}{\delta_{m,m}} \right) m^{1/2}\sigma\epsilon_2. \qquad (6.13)
\end{aligned}$$

From this we can write

$$\begin{aligned}
\left(2 + \frac{3m\delta}{\delta_{m,m}} \right) m^{1/2}\sigma\epsilon_2 &\geq \|\hat{T}_m q - \mu q\|_2 \\
&= \left\| Q^{(m)} \cdot \operatorname{diag}(\mu_i^{(m)}) \cdot Q^{(m)T} q - \mu q \right\|_2 \\
&= \left\| \left(\operatorname{diag}(\mu_i^{(m)}) - \mu I \right) Q^{(m)T} q \right\|_2 \\
&\geq \min_i |\mu_i^{(m)} - \mu| \cdot \|Q^{(m)T} q\| \\
&= \min_i |\mu_i^{(m)} - \mu|.
\end{aligned}$$

Then, from (6.1), $\delta_{m,m} \geq \sqrt{3}m^2\sigma\epsilon_2$, and thus

$$\begin{aligned}
|\mu_x^{(m)} - \mu| \equiv \min_i |\mu_i^{(m)} - \mu| &\leq \left(2 + \frac{3m\delta}{\delta_{m,m}} \right) m^{1/2}\sigma\epsilon_2 \\
&\leq 2m^{1/2}\sigma\epsilon_2 + \frac{3m^{3/2}\sigma\epsilon_2\delta}{\sqrt{3}m^2\sigma\epsilon_2} \\
&\leq 2m^{1/2}\sigma\epsilon_2 + \frac{\sqrt{3}\delta}{\sqrt{m}}. \qquad (6.14)
\end{aligned}$$

Then, for any $t > m$,

$$\hat{T}_t \begin{bmatrix} q \\ 0_{t-m,1} \end{bmatrix} = \begin{bmatrix} \hat{T}_m q \\ \hat{\beta}_{m+1}\eta e_1 \end{bmatrix},$$

and together with (6.13),

$$\begin{aligned}\min_i |\mu_i^{(t)} - \mu| &\leq 2m^{1/2}\sigma\epsilon_2 + \delta\biggl(1 + m^3\left(\frac{3\sigma\epsilon_2}{\delta_{m,m}}\right)^2\biggr)^{1/2} \\ &\leq 2m^{1/2}\sigma\epsilon_2 + \delta\biggl(1 + \left(\frac{3m^3\sigma^2\epsilon_2^2}{m^4\sigma^2\epsilon_2^2}\right)^2\biggr)^{1/2} \\ &\leq 2m^{1/2}\sigma\epsilon_2 + \delta\left(1 + \frac{3}{m}\right)^{1/2}, \qquad (6.15)\end{aligned}$$

where we have again used $\delta_{m,m} \geq \sqrt{3}m^2\sigma\epsilon_2$. Equations (6.14) and (6.15) can then be combined to give

$$\begin{aligned}\min_i |\mu_i^{(t)} - \mu_x^{(m)}| &\leq 2m^{1/2}\sigma\epsilon_2 + \frac{\sqrt{3}\delta}{\sqrt{m}} + 2m^{1/2}\sigma\epsilon_2 + \delta\left(1 + \frac{3}{m}\right)^{1/2} \\ &\leq 4m^{1/2}\sigma\epsilon_2 + \delta\left(\frac{\sqrt{3}}{\sqrt{m}} + \sqrt{1 + 3/m}\right) \\ &\leq 4m^{1/2}\sigma\epsilon_2 + 4\delta.\end{aligned}$$

This means that, assuming ϵ_2 is small enough, an eigenvalue of $\hat{T}_m$ close to μ has stabilized to about 4δ, where from (6.12), μ is within 2δ of an eigenvalue of A.

It can also be shown that for each $\mu_i^{(m)}$ of $\hat{T}_m$,

$$\begin{aligned}\min_{\mu \text{ in } (6.6)} |\mu - \mu_i^{(m)}| &\leq \Bigl(2 + \frac{3m\delta_{m,i}}{\delta_{m,m}}\Bigr)m^{1/2}\sigma\epsilon_2 \\ &\leq 2m^{1/2}\sigma\epsilon_2 + \frac{\sqrt{3}\delta_{m,i}}{\sqrt{m}}.\end{aligned}$$

This means that when $(\mu_i^{(m)}, \hat{V}_m q_i^{(m)})$ represents an eigenpair of A to within about $\delta_{m,i}$, there is a μ of (6.6) within about $\delta_{m,i}$ of $\mu_i^{(m)}$, assuming $m \geq 3$.

Thus, assuming no breakdown, and assuming the restrictions on the size of $\bar{\Gamma}_k$ (see (4.12)), these results say the same thing for the s-step Lanczos case as in the classical Lanczos case: *until an eigenvalue has stabilized, the s-step Lanczos algorithm behaves very much like the error-free process, or the algorithm with reorthogonalization.*

7. Future work. In this paper, we have shown that the results of Paige for classical Lanczos [30] also apply to the s-step Lanczos method as long as the computed s-step bases remain well-conditioned. As in the classical Lanczos case, the upper bounds in this paper and in [4] are likely large overestimates. We stress, as did Paige, that the value of these bounds is in the *insight* they give rather than their tightness. In practice, the present authors have observed that accurate eigenvalue estimates of A can be found with much looser restrictions than indicated by (4.12), and in some cases even in spite of a numerically rank-deficient basis.

Our analysis and extension of Paige's results confirms the empirical observation that the conditioning of the Krylov bases plays a large role in determining finite precision behavior, and also indicates that the s-step method can be made suitable for practical use in many cases, offering both speed and accuracy. The next step is to extend the subsequent analyses of Paige, in which a type of augmented backward stability for the classical Lanczos method is proved [31].

Another area of interest is the development of practical techniques for improving s-step Lanczos based on our results. This could include strategies for reorthogonalizing the Lanczos vectors, (re)orthogonalizing the generated Krylov basis vectors, or controlling the basis conditioning such that (4.12) holds. The bounds could also be used for guiding the use of extended or mixed precision in s-step Krylov methods; that is, rather than control the conditioning of the computed s-step base, the requirements in (4.12) could be met by decreasing the unit roundoff ϵ using techniques either in hardware or software.

Acknowledgements. This material is based upon work supported by the U.S. Department of Energy Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program under Award Numbers DE-SC0004938, DE-SC0003959, and DE-SC0010200; by the U.S. Department of Energy Office of Science, Office of Advanced Scientific Computing Research, X-Stack program under Award Numbers DE-SC0005136, DE-SC0008699, DE-SC0008700, and AC02-05CH11231; by DARPA Award Number HR0011-12-2-0016, as well as contributions from Intel, Oracle, and MathWorks.

We thank our colleague Nicholas Knight for many fruitful discussions related to this paper.

REFERENCES

[1] Z. Bai, D. Hu, and L. Reichel, *A Newton basis GMRES implementation*, IMA J. Numer. Anal., 14 (1994), pp. 563–581.

[2] G. Ballard, E. Carson, J. Demmel, M. Hoemmen, N. Knight, and O. Schwartz, *Communication lower bounds and optimal algorithms for numerical linear algebra*, Acta Numerica, 23 (2014), pp. 1–155.

[3] E. Carson and J. Demmel, *Analysis of the finite precision s-step biconjugate gradient method*, Tech. Report UCB/EECS-2014-18, EECS Dept., U.C. Berkeley, Mar 2014.

[4] ———, *Error analysis of the s-step Lanczos method in finite precision*, Tech. Report UCB/EECS-2014-55, EECS Dept., U.C. Berkeley, May 2014.

[5] ———, *A residual replacement strategy for improving the maximum attainable accuracy of s-step Krylov subspace methods*, SIAM J. Matrix Anal. Appl., 35 (2014), pp. 22–43.

[6] E. Carson, N. Knight, and J. Demmel, *Avoiding communication in nonsymmetric Lanczos-based Krylov subspace methods*, SIAM J. Sci. Comp., 35 (2013).

[7] A. Chronopoulos and C. Gear, *On the efficient implementation of preconditioned s-step conjugate gradient methods on multiprocessors with memory hierarchy*, Parallel Comput., 11 (1989), pp. 37–53.

[8] ———, *s-step iterative methods for symmetric linear systems*, J. Comput. Appl. Math, 25 (1989), pp. 153–168.

[9] A. Chronopoulos and C. Swanson, *Parallel iterative s-step methods for unsymmetric linear systems*, Parallel Comput., 22 (1996), pp. 623–641.

[10] E. de Sturler, *A performance model for Krylov subspace methods on mesh-based parallel computers*, Parallel Comput., 22 (1996), pp. 57–74.

[11] J. Demmel, M. Hoemmen, M. Mohiyuddin, and K. Yelick, *Avoiding communication in computing Krylov subspaces*, Tech. Report UCB/EECS-2007-123, EECS Dept., U.C. Berkeley, Oct 2007.

[12] D. Gannon and J. Van Rosendale, *On the impact of communication complexity on the design of parallel numerical algorithms*, Trans. Comput., 100 (1984), pp. 1180–1194.

[13] G. Golub and C. Van Loan, *Matrix computations*, JHU Press, Baltimore, MD, 3 ed., 1996.

[14] 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.

[15] M. Gustafsson, J. Demmel, and S. Holmgren, *Numerical evaluation of the communication-avoiding Lanczos algorithm*, Tech. Report ISSN 1404-3203/2012-001, Department of Information Technology, Uppsala University, Feb. 2012.

[16] M. Gutknecht, *Lanczos-type solvers for nonsymmetric linear systems of equations*, Acta Numer., 6 (1997), pp. 271–398.

[17] M. Gutknecht and Z. Strakoš, *Accuracy of two three-term and three two-term recurrences for Krylov space solvers*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 213–229.

[18] A. Hindmarsh and H. Walker, *Note on a Householder implementation of the GMRES method*, Tech. Report UCID-20899, Lawrence Livermore National Lab., CA., 1986.

[19] M. Hoemmen, *Communication-avoiding Krylov subspace methods*, PhD thesis, EECS Dept., U.C. Berkeley, 2010.

[20] W. Joubert and G. Carey, *Parallelizable restarted iterative methods for nonsymmetric linear systems. Part I: theory*, Int. J. Comput. Math., 44 (1992), pp. 243–267.

[21] S. Kim and A. Chronopoulos, *A class of Lanczos-like algorithms implemented on parallel computers*, Parallel Comput., 17 (1991), pp. 763–778.

[22] ———, *An efficient nonsymmetric Lanczos method on parallel vector computers*, J. Comput. Appl. Math., 42 (1992), pp. 357–374.

[23] C. Lanczos, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Res. Natn. Bur. Stand., 45 (1950), pp. 255–282.

[24] G. Meurant, *The Lanczos and conjugate gradient algorithms: from theory to finite precision computations*, SIAM, 2006.

[25] G. Meurant and Z. Strakoš, *The Lanczos and conjugate gradient algorithms in finite precision arithmetic*, Acta Numer., 15 (2006), pp. 471–542.

[26] M. Mohiyuddin, M. Hoemmen, J. Demmel, and K. Yelick, *Minimizing communication in sparse matrix solvers*, in Proc. ACM/IEEE Conference on Supercomputing, 2009.

[27] C. Paige, *The computation of eigenvalues and eigenvectors of very large sparse matrices*, PhD thesis, London University, London, UK, 1971.

[28] ———, *Computational variants of the Lanczos method for the eigenproblem*, IMA J. Appl. Math., 10 (1972), pp. 373–381.

[29] ———, *Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix*, IMA J. Appl. Math., 18 (1976), pp. 341–349.

[30] ———, *Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem*, Linear Algebra Appl., 34 (1980), pp. 235–258.

[31] ———, *An augmented stability result for the Lanczos hermitian matrix tridiagonalization process*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 2347–2359.

[32] B. Parlett and D. Scott, *The Lanczos algorithm with selective orthogonalization*, Math. Comput., 33 (1979), pp. 217–238.

[33] B. Philippe and L. Reichel, *On the generation of Krylov subspace bases*, Appl. Numer. Math, 62 (2012), pp. 1171–1186.

[34] S. Rump, *Verified bounds for singular values, in particular for the spectral norm of a matrix and its inverse*, BIT Numer. Math., 51 (2011), pp. 367–384.

[35] H. Simon, *The Lanczos algorithm with partial reorthogonalization*, Math. Comput., 42 (1984), pp. 115–142.

[36] R. Thompson and P. McEnteggert, *Principal submatrices ii: The upper and lower quadratic inequalities*, Lin. Alg. Appl., 1 (1968), pp. 211–243.

[37] S. Toledo, *Quantitative performance modeling of scientific computations and creating locality in numerical algorithms*, PhD thesis, MIT, 1995.

[38] H. Van der Vorst and Q. Ye, *Residual replacement strategies for Krylov subspace iterative methods for the convergence of true residuals*, SIAM J. Sci. Comput., 22 (1999), pp. 835–852.

[39] J. Van Rosendale, *Minimizing inner product data dependencies in conjugate gradient iteration*, Tech. Report 172178, ICASE-NASA, 1983.

[40] H. Walker, *Implementation of the GMRES method using Householder transformations*, SIAM J. Sci. Stat. Comput., 9 (1988), pp. 152–163.

[41] J. Wilkinson, *The algebraic eigenvalue problem*, vol. 87, Oxford Univ. Press, 1965.

[42] S. Williams, M. Lijewski, A. Almgren, B. Van Straalen, E. Carson, N. Knight, and J. Demmel, *s-step Krylov subspace methods as bottom solvers for geometric multigrid*, in International Symposium on Parallel and Distributed Processing, IEEE, 2014.

[43] W. Wülling, *On stabilization and convergence of clustered Ritz values in the Lanczos method*, SIAM J. Matrix Anal. Appl., 27 (2005), pp. 891–908.

[44] J. Zemke, *Krylov subspace methods in finite precision: a unified approach*, PhD thesis, Technische Universität Hamburg-Harburg, 2003.