APPROXIMATING THE LEADING SINGULAR TRIPLETS OF A LARGE MATRIX FUNCTION*

SARAH W. GAAF† AND VALERIA SIMONCINI‡

Abstract. Given a large square matrix A and a sufficiently regular function f so that $f(A)$ is well defined, we are interested in the approximation of the leading singular values and corresponding singular vectors of $f(A)$, and in particular of $\|f(A)\|$, where $\|\cdot\|$ is the matrix norm induced by the Euclidean vector norm. Since neither $f(A)$ nor $f(A)v$ can be computed exactly, we introduce a new *inexact* Golub-Kahan-Lanczos bidiagonalization procedure, where the inexactness is related to the inaccuracy of the operations $f(A)v$, $f(A)^*v$. Particular outer and inner stopping criteria are devised so as to cope with the lack of a true residual. Numerical experiments with the new algorithm on typical application problems are reported.

1. Introduction. Given a large $n\times n$ complex matrix A and a sufficiently regular function f so that $f(A)$ is well defined, we are interested in approximating the largest singular values and corresponding singular vectors of the matrix function $f(A)$. This computation will also give an approximation to its 2-norm, namely, $\|f(A)\|$, where $\|\cdot\|$ is the matrix norm induced by the Euclidean vector norm, and it is defined as

$$\|f(A)\| = \max_{0\neq\mathbf{x}\in\mathbb{C}^n} \frac{\|f(A)\mathbf{x}\|}{\|\mathbf{x}\|}. \tag{1.1}$$

In our presentation we will chiefly discuss this norm approximation because of its interest in applications. However, we shall keep in mind that the considered procedure allows us to also determine both associated left and right singular vectors, and that a group of singular triplets can be determined simultaneously.

The problem of approximating the norm of a matrix function arises in the solution of stiff linear initial value problems [13],[30], in the evaluation of derivatives and perturbations of matrix functions, which arise for instance in electronic structure theory [14],[29],[24], and in monitoring the magnitude of the inverse of distance matrices [1]. In numerical linear algebra the norm of matrix polynomials may be used in the analysis of iterative procedures, and the norm of rational matrix functions, and in particular of the transfer function may give information on the sensitivity of the matrix itself to perturbations; see, e.g., [34],[5] and their references.

If A were normal, then the approximation could be stated in terms of an eigenvalue problem in A. Indeed, if $A = Q\Lambda Q^*$ is the eigendecomposition of A with Q unitary and Λ diagonal, then $f(A) = Qf(\Lambda)Q^*$ [22], so that the leading singular values of $f(A)$ could be determined by a procedure that approximates the eigenvalues of A.

The problem is significantly more challenging if A is large and non-normal, since there is no relation between eigenvalues and singular values that can be readily exploited during the computation. Moreover, although A may be sparse, in general $f(A)$ will be dense, and it cannot be computed explicitly. We are thus left with procedures that use f and A by means of the action of $f(A)$ to a vector $\mathbf{v}$. The Lanczos bidiagonalization is among the most used strategies for approximating selected singular

*Version of September 19, 2015.

†Department of Mathematics and Computer Science, TU Eindhoven, PO Box 513, 5600 MB, The Netherlands, (s.w.gaaf@tue.nl). This author is supported by a Vidi research grant from the Netherlands Organisation for Scientific Research (NWO).

‡Dipartimento di Matematica, Università di Bologna, Piazza di Porta S. Donato, 5, I-40127 Bologna, Italy, and IMATI-CNR, Pavia (valeria.simoncini@unibo.it).

triplets of a given matrix. Given a matrix F, this procedure generates a sequence of orthonormal vectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots\}$ and $\{\mathbf{u}_1, \mathbf{u}_2, \ldots\}$ by alternating products of $F\mathbf{v}$ and $F^*\mathbf{u}$. In our case, $F = f(A)$ with A of large dimensions, therefore these matrix vector products cannot be computed exactly and the standard Lanczos process fails.

We introduce a novel *inexact* implementation of the Lanczos bidiagonalization process, where at each iteration the action of $f(A)\mathbf{v}$ and $f(A)^*\mathbf{u}$ is approximated with some loose tolerance by means of a projection method. The problem of approximating $f(A)\mathbf{v}$ has seen a great interest growth in the past fifteen years, due to the emerging occurrence of this computation in many scientific and engineering applications; see, e.g., [2],[9],[12],[17],[19],[21],[20], and their references. For our purposes we shall use Krylov subspace methods for approximating $f(A)^*\mathbf{u}$ and $f(A)\mathbf{v}$ at each iteration, equipped with a cheap stopping criterion that may also be adapted to the outer current accuracy. We shall show that the inexactness in the Lanczos bidiagonalization causes the loss of the known symmetry structure of the process. Nonetheless, as is the case in finite precision analysis [26], orthogonality of the basis can be preserved, so that the recurrence maintains its effectiveness.

If a rough approximation to $\|f(A)\|$ is the only quantity of interest, instead of a group of singular triplets, then other approaches could be considered. For instance, f could be approximated by some other more convenient functions, and then the resulting matrix function norm could be more easily estimated. As an alternative, equivalent definitions of $f(A)$ could be used, from which the norm could also be estimated; or, the relation of $\|f(A)\|$ with other norms or with some other spectral tool could be used; some of these approaches are briefly recalled in section 2. Methods in the mentioned classes, however, usually at most provide the order of magnitude of the actual norm and are thus inappropriate if more correct digits are needed.

This paper is organized as follows. Section 2 reviews some methods available for the approximation of $\|f(A)\|$. In section 3 the standard Lanczos bidiagonalization is recalled and the general notation used in this paper is introduced. Section 4 presents the inexact Lanczos bidiagonalization procedure, including the details on the stopping criteria in section 4.1. Section 5 discusses the approximation of the matrix function multiplication, and a stopping criterion for its accuracy, while in section 6 a stopping criterion in the case of an inner flexible strategy is analyzed. In section 6.1 we show how specific spectral properties allow us to make a variable accuracy for the inner iteration feasible, which is finalized in section 6.2. Section 7 focuses on the practical implementation and the numerical results are presented in section 8. We will conclude with some discussion in section 9.

The following notation will be used throughout. The vector $\mathbf{e}_i$ indicates the ith column of the identity matrix of a given dimension. The conjugate transpose of a matrix A will be denoted by A^*. We will use the Matlab-like notation $[\mathbf{x}; \mathbf{y}]$ to denote the column vector

$$\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}, \quad \mathbf{x} \in \mathbb{C}^{n_x}, \mathbf{y} \in \mathbb{C}^{n_y}.$$

The Euclidean vector norm for vectors will be used, namely $\|\mathbf{x}\| = (\sum_{i=1}^{n} |x_i|^2)^{\frac{1}{2}}$, for $\mathbf{x} \in \mathbb{C}^n$. Unless explicitly stated, the induced matrix norm (1.1) will be used for matrices. For $A \in \mathbb{C}^{n\times n}$, spec$(A)$ denotes the set of its eigenvalues, and $W(A) = \{z \in \mathbb{C} : z = (\mathbf{x}^*A\mathbf{x})/(\mathbf{x}^*\mathbf{x}), \mathbf{x} \in \mathbb{C}^n \backslash \{\mathbf{0}\}\}$ is its field of values.

2. Available techniques for estimating the norm. While the Lanczos bidiagonalization is widely recognized as the method of choice for approximating selected

singular triplets of a large matrix, if one is only interested in estimates of $\|f(A)\|_2$ with A non-Hermitian, then rather different procedures could also be used. A simple approach consists of roughly estimating $\|\cdot\|_2$ by using some other matrix norm. For instance,

$$\|f(A)\|_2 \le \sqrt{n}\|f(A)\|_p, \quad p = 1, \infty, \qquad \text{or} \quad \|f(A)\|_2 \le \sqrt{\|f(A)\|_1 \, \|f(A)\|_\infty},$$

where $\|f(A)\|_p$ is once again an induced norm [23, p. 365]. This bound is usually pessimistic, and it is clearly unsatisfactory for n large. The fact that for A large the entries of $f(A)$ are not all readily available provides an additional challenge.

In the following we describe a few approaches available in the literature that are tailored to the matrix function case. Some of them first determine an explicit upper bound for the norm, which only depends on scalar quantities. The core computation will then be to determine a good approximation to the obtained upper bound. The quality of the final estimate of $\|f(A)\|_2$ will thus depend both on the sharpness of the initial upper bound and on the accuracy of the computation. For general non-normal matrices the initial bound is often not very sharp, limiting the quality of the overall estimation. Finally, a computation-oriented estimation is the power method, which directly approximates $\|f(A)\|_2$ as the square root of the largest eigenvalue of $f(A)^* f(A)$. A more detailed list follows.

1. Let $r(A)$ be the numerical radius of A, that is $r(A) = \max\{|z| : z \in W(A)\}$, where $W(A)$ is the field of values of A. Since (see, e.g., [16, Theorem 1.3-1])

   $$r(A) \le \|A\|_2 \le 2\, r(A),$$

   by applying the bounds to $f(A)$ instead of A, it is possible to estimate $\|f(A)\|_2$ by means of $r(f(A))$; see, e.g., [36],[28] for numerical methods to compute the numerical radius of a given matrix. A related special case is given by the exponential function, for which the bound

   $$\|\exp(A)\| \le \exp(\alpha) \tag{2.1}$$

   holds, where α is the largest eigenvalue of the Hermitian part of A, that is of $\frac{1}{2}(A + A^*)$ [19, section 10.1].
2. If it is possible to find $K > 0$ and $\Omega \subset \mathbb{C}$ such that

   $$\|f(A)\|_2 \le K \|f\|_\Omega,$$

   then it is sufficient to estimate $\|f\|_\Omega$; here $\|f\|_\Omega$ is the L_∞-norm of f on Ω. This can be done for instance when f is a polynomial, [6], for which K is known to be less than 11.08 and conjectured to be equal to 2, and Ω coincides with the field of values of A. We refer to the Ph.D. thesis of D. Choi [5], for a discussion on the use if this bound when A is normal, or when A is a contraction; see also [34] for a detailed analysis of this bound when using pseudospectral information. The computationally intensive task is given by the determination of Ω. If Ω coincides with $W(A)$, then the cost of accurately approximating Ω may be higher than that of approximating the single quantity $\|f(A)\|_2$.
3. This approach is in the same spirit as the one above. For $\varepsilon > 0$, let $\sigma_\varepsilon(A) = \{z \in \mathrm{spec}(A + E) : \|E\| < \varepsilon\}$ and assume that f is analytic in $\sigma_\varepsilon(A)$. If L_ε denotes the length of the boundary $\partial\sigma_\varepsilon(A) = \{z \in \mathbb{C} : \|(zI - A)^{-1}\| = \varepsilon^{-1}\}$,

then by using the Cauchy integral expression for $f(A)$ we obtain (see, e.g., [34])

$$\|f(A)\| \leq \frac{L_\varepsilon}{2\pi\varepsilon}\|f\|_{\partial\sigma_\varepsilon}.$$

Although the involved quantities may be easier to compute than in the previous case, the dependence on $\varepsilon > 0$ remains not fully controllable.

4. Using the relation $\|f(A)\|_2^2 = \lambda_{\max}(f(A)^* f(A))$ a run of a few iterations of the power method can give an estimate to $\lambda_{\max}(f(A)^* f(A))$; see, e.g., [19, Algorithm 3.19] for an algorithm specifically designed for the largest singular triplet.

The power method is probably the most appealing approach among the ones listed above. If a rough approximation is required, typically to determine the order of magnitude, then the power method provides a satisfactory answer in few iterations. However, if more than one digit of accuracy is required, then the process may need many iterations to converge. As for with A, the stability in the computation may be highly influenced by the squaring; we refer to section 8 for an example of this well known phenomenon.

3. Lanczos bidiagonalization. We start by recalling the Golub-Kahan bidiagonalization process in our context, in terms of the matrix function $f(A)$; then we will discuss how to actually obtain $f(A)$ times a vector. Let $\mathbf{u}_0 = 0$ and $\beta_1 = 0$, and given the vector $\mathbf{v}_1$ of unit norm, then for $j = 1, \ldots, m$ the following recurrence relations define the Lanczos algorithm

$$\begin{aligned} \beta_{2j}\mathbf{u}_j &= f(A)\mathbf{v}_j - \beta_{2j-1}\mathbf{u}_{j-1} \\ \beta_{2j+1}\mathbf{v}_{j+1} &= f(A)^*\mathbf{u}_j - \beta_{2j}\mathbf{v}_j. \end{aligned} \tag{3.1}$$

The coefficients β_{2j} and β_{2j+1} are computed so that the corresponding vectors $\mathbf{u}_j$ and $\mathbf{v}_{j+1}$ have unit norm. By collecting the two sets of vectors as $U_m = [\mathbf{u}_1, \ldots, \mathbf{u}_m]$, $V_m = [\mathbf{v}_1, \ldots, \mathbf{v}_m]$, we observe that $U_m^* U_m = I$, $V_m^* V_m = I$, $V_m^* \mathbf{v}_{m+1} = 0$. Moreover, the two recurrences can be compactly written as

$$\begin{aligned} f(A)V_m &= U_m B_m \\ f(A)^* U_m &= V_m B_m^* + \beta_{2m+1}\mathbf{v}_{m+1}\mathbf{e}_m^*, \end{aligned} \tag{3.2}$$

where B_m is the following bidiagonal matrix

$$B_m = \begin{bmatrix} \beta_2 & \beta_3 & & \\ & \beta_4 & \beta_5 & \\ & & \ddots & \ddots \end{bmatrix} \in \mathbb{R}^{m\times m}.$$

It can be shown that the columns of V_m span the Krylov subspace $\mathcal{K}_m(f(A)^* f(A), \mathbf{v}_1)$ and the columns of U_m span the Krylov subspace $\mathcal{K}_m(f(A)f(A)^*, f(A)\mathbf{v}_1)$. Define

$$\mathcal{B}_{2m} = \begin{bmatrix} 0 & B_m \\ B_m^* & 0 \end{bmatrix}, \qquad \text{and} \qquad \mathcal{W}_{2m} = \begin{bmatrix} U_m & 0 \\ 0 & V_m \end{bmatrix},$$

and

$$\mathcal{F} = \begin{bmatrix} 0 & f(A) \\ f(A)^* & 0 \end{bmatrix}.$$

Then the recursion (3.2) can be rewritten in the more compact matrix notation ([7, p. 178–186], [15, p. 448–449, p. 495])

$$\mathcal{F}\mathcal{W}_{2m} = \mathcal{W}_{2m}\mathcal{B}_{2m} + \beta_{2m+1}\begin{bmatrix} 0 \\ \mathbf{v}_{m+1} \end{bmatrix}\mathbf{e}_m^*, \quad \mathbf{e}_m \in \mathbb{R}^{2m}. \tag{3.3}$$

For both even and odd j, the eigenvalues of $\mathcal{B}_j$ occur in $\pm$ pairs, with the exception of an extraneous zero eigenvalue in the odd case. Within this setting, is it thus possible to approximate the singular values of $f(A)$ by the positive eigenvalues of $\mathcal{B}_{2m}$, or, equivalently, by the singular values of B_m. In particular, for the largest singular value it holds that (see [22, Corollary 3.1.3, Lemma 3.3.1.]):

$$\sigma_1(B_{j-1}) \leq \sigma_1(B_j) \leq \sigma_1(f(A)), \qquad 2 \leq j \leq m.$$

There are several advantages of the Golub-Kahan bidiagonalization over the simpler power method applied to $f(A)^* f(A)$, which are mainly related to the fact that the eigenvalue squaring in this latter problem may lead to severe loss of information in the case very small or very large singular values arise. In the inexact case the bidiagonal formulation also allows us to better trace the inexactness during the whole approximation process; this is discussed in the next section.

4. Inexact Lanczos bidiagonalization. When neither the explicit computation of the matrix $f(A)$ nor the accurate operation $f(A)\mathbf{v}$ (or $f(A)^*\mathbf{v}$) are feasible, then approximate computations must be performed, resulting in an *inexact* Lanczos bidiagonalization procedure. As a consequence, the recurrence (3.2) needs to be significantly revised so as to acknowledge for the quantities that are actually computed.

For a given $\mathbf{v}$, the exact matrix-vector multiplication $f(A)\mathbf{v}$ has to be replaced by an inner procedure that approximates the resulting vector up to a certain accuracy. The same holds for the operation $f(A)^*\mathbf{u}$ for a given vector $\mathbf{u}$. For the sake of the analysis, at each iteration j we shall formalize this difference by writing, for some matrices C_j and D_j,

$$\begin{aligned} \beta_{2j}\mathbf{u}_j &= (f(A)\mathbf{v}_j + C_j\mathbf{v}_j) - \beta_{2j-1}\mathbf{u}_{j-1} \\ \beta_{2j+1}\mathbf{v}_{j+1} &= (f(A)^*\mathbf{u}_j + D_j\mathbf{u}_j) - \beta_{2j}\mathbf{v}_j, \end{aligned}$$

where C_j, D_j implicitly represent the perturbation induced by the approximate computations. Since in general $f(A)^* + D_j$ is no longer the conjugate transpose of $f(A) + C_j$, orthogonality of a new vector $\mathbf{v}_{m+1}$ has to be enforced by explicit orthogonalization against all previous vectors $\mathbf{v}_j$, $1 \leq j \leq m$. The same holds for the vectors $\mathbf{u}_j$, $j = 1, \ldots, m$. Therefore, instead of one bidiagonal matrix B_m in the exact relation, we now obtain an upper triangular matrix M_m and an upper Hessenberg matrix T_m. This leads to the following relations for the inexact (perturbed) Lanczos bidiagonalization:

$$\begin{aligned} (f(A) + \mathfrak{C}_m)V_m &= U_m M_m \\ (f(A)^* + \mathfrak{D}_m)U_m &= V_m T_m + t_{m+1,m}\mathbf{v}_{m+1}\mathbf{e}_m^*, \end{aligned}$$

where $\mathfrak{C}_m = \sum_{j=1}^m C_j\mathbf{v}_j\mathbf{v}_j^*$ and $\mathfrak{D}_m = \sum_{j=1}^m D_j\mathbf{u}_j\mathbf{u}_j^*$. The matrices V_m and U_m are different from the matrices in the exact relation, but they still have orthonormal columns.

The inexact Lanczos bidiagonalization can also be described using the notation of (3.3). Define

$$\widetilde{\mathcal{B}}_{2m} = \begin{bmatrix} 0 & M_m \\ T_m & 0 \end{bmatrix}, \qquad \mathcal{W}_{2m} = \begin{bmatrix} U_m & 0 \\ 0 & V_m \end{bmatrix},$$

and the perturbation matrix

$$\mathcal{G}_{2m} = \begin{bmatrix} 0 & \mathfrak{C}_m \\ \mathfrak{D}_m & 0 \end{bmatrix} \mathcal{W}_{2m} =: \mathcal{E}_m \mathcal{W}_{2m}.$$

The perturbed relation thus becomes

$$\mathcal{F}\mathcal{W}_{2m} + \mathcal{G}_{2m} = \mathcal{W}_{2m}\widetilde{\mathcal{B}}_{2m} + t_{m+1,m} \begin{bmatrix} 0 \\ \mathbf{v}_{m+1} \end{bmatrix} \mathbf{e}_m^*, \quad \mathbf{e}_m \in \mathbb{R}^{2m}, \tag{4.1}$$

where

$$\mathcal{F}\mathcal{W}_{2m} + \mathcal{G}_{2m} = (\mathcal{F} + \mathcal{E}_m)\mathcal{W}_{2m} = \begin{bmatrix} 0 & f(A) + \mathfrak{C}_m \\ f(A)^* + \mathfrak{D}_m & 0 \end{bmatrix} \mathcal{W}_{2m} =: \widetilde{\mathcal{F}}_{2m}\mathcal{W}_{2m}.$$

In contrast to the exact case, the space spanned by the columns of $\mathcal{W}_{2m}$ is not a Krylov subspace. However, when $t_{m+1,m}$ is small, this new space is close to an invariant subspace of the perturbed matrix $\widetilde{\mathcal{F}}_{2m}$, because then $\widetilde{\mathcal{F}}_{2m}\mathcal{W}_{2m} \approx \mathcal{W}_{2m}\widetilde{\mathcal{B}}_{2m}$. Notice the similarity of (4.1) with equation (3.1) in [31], which shows that with this formulation, the inexact projection problem amounts to solving a structured eigenvalue problem, where the original Hermitian matrix $\mathcal{F}$ has been perturbed by a structured non-Hermitian perturbation $\mathcal{E}_m$. The theory in [31] can then be used to analyze and monitor the inexact computations, although the general results in [31] should be carefully adapted to the new problem structure.

If $\mathcal{E}_m$ is small in norm, the eigenvalues of the *non-Hermitian* matrix $\widetilde{\mathcal{F}}_{2m}$ are small perturbations of the eigenvalues of the *Hermitian* matrix $\mathcal{F}$. Indeed, the eigenvalues of the perturbed matrix $\widetilde{\mathcal{F}}_{2m}$ lie in discs with radius $\|\mathcal{E}_m\|$ and center the (real) eigenvalues of $\mathcal{F}$ (see, e.g., [33, section IV, Theorem 5.1]). Therefore, for small perturbations in the computations, the eigenvalues of the symmetric matrix $\mathcal{F}$ will be perturbed accordingly. On the other hand, in the following we shall consider the case when $\|\mathcal{E}_m\|$ is larger than usually allowed by a perturbation analysis argument, therefore different strategies need to be devised to ensure good approximations to the wanted eigenvalues of $\mathcal{F}$.

Following the standard procedure of the exact case, we should consider the matrix $\widetilde{\mathcal{B}}_{2m}$ to approximate the largest eigenpairs of $\widetilde{\mathcal{F}}_{2m}$, and according to the discussion above, of $\mathcal{F}$. Due to the non-Hermitian structure of $\widetilde{\mathcal{B}}_{2m}$, however, there are different matrices that can provide the sought after singular value information, namely the matrix $\widetilde{\mathcal{B}}_{2m}$ itself, and the two distinct matrices T_m or M_m. The last two matrices yield approximations to the corresponding triplets of $f(A)+\mathfrak{C}_m$ and $f(A)^*+\mathfrak{D}_m$. The following bound between the largest eigenvalue of $\widetilde{\mathcal{B}}_{2m}$ and the largest singular values of T_m and M_m shows that all these quantities can be easily related. Let $\mathbf{q} = [\mathbf{x}; \mathbf{y}]$

and let θ be an eigenvalue of $\widetilde{\mathcal{B}}_{2m}$. Using $\|\mathbf{x}\|\|\mathbf{y}\| \leq \frac{1}{2}\left(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2\right)$ we obtain[1]

$$|\theta| \leq \max_{\mathbf{q}\neq 0} \left| \frac{\mathbf{q}^* \widetilde{\mathcal{B}}_{2m} \mathbf{q}}{\mathbf{q}^*\mathbf{q}} \right| = \max_{[\mathbf{x};\mathbf{y}]\neq 0} \left| \frac{\mathbf{x}^* M_m \mathbf{y} + \mathbf{y}^* T_m \mathbf{x}}{\mathbf{x}^*\mathbf{x} + \mathbf{y}^*\mathbf{y}} \right|$$
$$\leq \max_{[\mathbf{x};\mathbf{y}]\neq 0} \frac{\|\mathbf{x}\|\|M_m\mathbf{y}\|}{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2} + \max_{[\mathbf{x};\mathbf{y}]\neq 0} \frac{\|\mathbf{y}\|\|T_m\mathbf{x}\|}{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2} \leq \frac{1}{2}(\sigma_1(M_m) + \sigma_1(T_m)).$$

If the inexactness of the bidiagonalization is very large, M_m and T_m^* are very different from each other. In this case, the leading singular values of these two matrices - and thus their mean - may be significantly larger than the biggest (in modulo) eigenvalue of $\widetilde{\mathcal{B}}_{2m}$, since they are related to the numerical radius of $\widetilde{\mathcal{B}}_{2m}$, rather than to its spectrum. This motivated us to use the eigenvalues of $\widetilde{\mathcal{B}}_{2m}$ in the approximation, rather than the singular values of its blocks. Moreover, working with $\widetilde{\mathcal{B}}_{2m}$ made the analysis of the relaxed strategy particularly convenient, since known results on relaxed eigenvalue computation could be exploited.

4.1. A computable stopping criterion. In this section we analyze a strategy for monitoring the convergence of the inexact bidiagonal iteration. Some stopping criterion, for instance based on the problem residual, needs to be introduced to exit the process. As it is common to other inexact processes, the true problem residual is inaccessible as soon as inexactness takes place, so one could for example use an approximation to the true residual. However, it is unclear whether the computed approximations are still meaningful for the original problem, since they were computed with significantly modified data.

Let $(\theta, \mathbf{q})$ be an eigenpair of $\widetilde{\mathcal{B}}_{2m}$, where $\mathbf{q}$ is a unit vector. As the iterations proceed, $(\theta, \mathcal{W}_{2m}\mathbf{q})$ tends to approximate an eigenpair of $\widetilde{\mathcal{F}}_{2m}$. We would like to ensure that $(\theta, \mathcal{W}_{2m}\mathbf{q})$ also tends to an eigenpair of $\mathcal{F}$. To monitor the convergence of θ and to define a stopping criterion for the outer iteration, the residual is used. We call $\mathcal{F}\mathcal{W}_{2m}\mathbf{q} - \theta\mathcal{W}_{2m}\mathbf{q}$ the *true residual*, which is not available, since $\mathcal{F}$ cannot be applied exactly. We thus introduce the *computed residual*, which is the residual of the actually computed quantities, namely (see (4.1))

$$\mathbf{r}_{2m} := \widetilde{\mathcal{F}}_{2m}\mathcal{W}_{2m}\mathbf{q} - \theta\mathcal{W}_{2m}\mathbf{q} = t_{m+1,m} \begin{bmatrix} 0 \\ \mathbf{v}_{m+1} \end{bmatrix} \mathbf{e}_m^*\mathbf{q}, \quad (\mathbf{e}_m \in \mathbb{R}^{2m}).$$

In the sequel, we shall use the following obvious inequality to estimate the true residual norm:

$$\|\mathcal{F}\mathcal{W}_{2m}\mathbf{q} - \theta\mathcal{W}_{2m}\mathbf{q}\| \leq \|\mathbf{r}_{2m}\| + \|(\mathcal{F}\mathcal{W}_{2m}\mathbf{q} - \theta\mathcal{W}_{2m}\mathbf{q}) - \mathbf{r}_{2m}\|, \tag{4.2}$$

where $\|(\mathcal{F}\mathcal{W}_{2m}\mathbf{q} - \theta\mathcal{W}_{2m}\mathbf{q}) - \mathbf{r}_{2m}\|$ is the gap between the computed and the true residuals, in short the "residual gap". If this gap can be imposed to be small, then the computed residual will give an estimate for the true residual. In this case, convergence can be monitored by only using the (available) computed residual, and the following relative stopping criterion can be used:

$$\text{if} \quad \frac{|t_{m+1,m}\mathbf{e}_m^*\mathbf{q}|}{|\theta|} < \varepsilon_{out} \quad \text{then} \quad \text{stop} \tag{4.3}$$

[1] Another bound can be obtained for the geometric mean, that is $|\theta| \leq \sqrt{\sigma_1(M_m)\sigma_1(T_m)}$.

for some outer tolerance ε_{out}, where θ is the largest (in modulo) eigenvalue. Finally, as the computed residual norm goes to zero, the quantity $\|(\mathcal{F}\mathcal{W}_{2m}\mathbf{q}-\theta\mathcal{W}_{2m}\mathbf{q})-\mathbf{r}_{2m}\|$ will tend to dominate again, playing the role of the final attainable accuracy level.

To see how we can impose the residual gap to be small, and recalling the definition of $\mathcal{G}_{2m}$, we first consider a more convenient expression for $\mathcal{G}_{2m}\mathbf{q}$, with $\mathbf{q}=[\mathbf{x};\mathbf{y}]$, that is

$$\mathcal{G}_{2m}\mathbf{q} = \begin{bmatrix} \mathfrak{C}_m V_m \mathbf{y} \\ \mathfrak{D}_m U_m \mathbf{x} \end{bmatrix} =: \begin{bmatrix} G_m^{(1)}\mathbf{y} \\ G_m^{(2)}\mathbf{x} \end{bmatrix}.$$

Let $G_m^{(\ell)} = [\mathbf{g}_1^{(\ell)}, \dots, \mathbf{g}_m^{(\ell)}]$, for $\ell = 1, 2$. Then

$$\begin{aligned}\|(\mathcal{F}\mathcal{W}_{2m}\mathbf{q} - \theta\mathcal{W}_{2m}\mathbf{q}) - \mathbf{r}_{2m}\| = \|\mathcal{G}_{2m}\mathbf{q}\| &= \left\| \begin{bmatrix} G_m^{(1)}\mathbf{y} \\ G_m^{(2)}\mathbf{x} \end{bmatrix} \right\| = \left\| \sum_{j=1}^{m} \begin{bmatrix} \mathbf{g}_j^{(1)}\mathbf{e}_j^*\mathbf{y} \\ \mathbf{g}_j^{(2)}\mathbf{e}_j^*\mathbf{x} \end{bmatrix} \right\| \\ \leq \sum_{j=1}^{m} \left\| \begin{bmatrix} \mathbf{g}_j^{(1)}\mathbf{e}_j^*\mathbf{y} \\ \mathbf{g}_j^{(2)}\mathbf{e}_j^*\mathbf{x} \end{bmatrix} \right\| &= \sum_{j=1}^{m} (\|\mathbf{g}_j^{(1)}\|^2 \, |\mathbf{e}_j^*\mathbf{y}|^2 + \|\mathbf{g}_j^{(2)}\|^2 \, |\mathbf{e}_j^*\mathbf{x}|^2)^{\frac{1}{2}}. \end{aligned} \tag{4.4}$$

The vectors $\mathbf{g}_j^{(\ell)}$, $\ell = 1, 2$, implicitly carry the error caused by the inexact computation of $f(A)\mathbf{v}$ and $f(A)^*\mathbf{u}$, respectively, in the inner iteration. If every term of this sum is small, the computed residual will be close to the true residual. The following Lemma states how the inaccuracy in the matrix-vector products relates to the residual gap; its proof is based on the corresponding result in [31], however the structure is exploited so as to have a dependence with respect to m instead of $2m$, the size of $\widetilde{\mathcal{B}}_{2m}$.

LEMMA 4.1. *Assume that m iterations of the inexact Lanczos bidiagonalization process have been taken.*

If $\|\mathbf{g}_j^{(1)}\|, \|\mathbf{g}_j^{(2)}\| < \frac{1}{m}\varepsilon$ for $1 \leq j \leq m$, then $\|(\mathcal{F}\mathcal{W}_{2m}\mathbf{q} - \theta\mathcal{W}_{2m}\mathbf{q}) - \mathbf{r}_{2m}\| < \varepsilon$.

Proof. From $\|\mathbf{q}\| = 1$ with $\mathbf{q} = [\mathbf{x};\mathbf{y}]$ it follows that $\|[\mathbf{e}_j^*\mathbf{x}; \mathbf{e}_j^*\mathbf{y}]\| \leq 1$. From (4.4) we obtain

$$\begin{aligned}\|(\mathcal{F}\mathcal{W}_{2m}\mathbf{q} - \theta\mathcal{W}_{2m}\mathbf{q}) - \mathbf{r}_{2m}\| &\leq \sum_{j=1}^{m} (\|\mathbf{g}_j^{(1)}\|^2 \, |\mathbf{e}_j^*\mathbf{y}|^2 + \|\mathbf{g}_j^{(2)}\|^2 \, |\mathbf{e}_j^*\mathbf{x}|^2)^{\frac{1}{2}} \\ &< \sum_{j=1}^{m} \frac{1}{m}\varepsilon (|\mathbf{e}_j^*\mathbf{y}|^2 + |\mathbf{e}_j^*\mathbf{x}|^2)^{\frac{1}{2}} \leq \frac{1}{m}\varepsilon \sum_{j=1}^{m} 1 = \varepsilon. \end{aligned}$$

∎

This result shows that if ε is sufficiently small, then the residual gap will stay below the computed residual norm until convergence. In our experiments, m will play the role of the maximum number of Lanczos bidiagonalization iterations, which is usually set to a number between 50 and 500.

5. Approximation of $f(A)\mathbf{v}$ and $f(A)^*\mathbf{u}$ and a computable inner stopping criterion. The performance of the inexact Lanczos bidiagonalization process depends on the approximation accuracy of the matrix-vector products $f(A)\mathbf{v}$ and $f(A)^*\mathbf{u}$. Due to the size of A, we consider approximating these quantities by means of a projection-type iterative method as follows; we limit our discussion to $f(A)\mathbf{v}$, and a corresponding procedure can be used for $f(A)^*\mathbf{u}$. We also notice that in general, $f(A)\mathbf{v}$ and $f(A)^*\mathbf{u}$ require distinct approximations, also for functions satisfying $f(A^*) = f(A)^*$. Starting with the unit vector $\mathbf{v}$ and the matrix A, we construct a sequence of approximation

subspaces $\mathcal{K}_i$ of $\mathbb{R}^n$, $i = 1, 2, \ldots$, and define the matrix $P_i = [\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \ldots, \mathbf{p}_i] \in \mathbb{C}^{n\times i}$, whose orthonormal columns span the subspace, and $\mathbf{v} = \mathbf{p}_1 = P_i\mathbf{e}_1$, in a way so that the spaces are nested, that is $\mathcal{K}_i \subseteq \mathcal{K}_{i+1}$. Typical such choices are Krylov and rational Krylov subspaces [19],[17]. The desired approximation is then obtained as

$$f(A)\mathbf{v} \approx P_i f(H_i)\mathbf{e}_1, \qquad H_i = P_i^* A P_i.$$

For small i, the reduced non-Hermitian matrix H_i has small size, so that $f(H_i)$ can be computed efficiently by various strategies such as decomposition-type methods [19].

Our stopping criterion of this approximation process is based on an estimation of the error norm, and it uses an approach previously introduced in [25, Proposition 2.2]; see also [35] for an earlier application to the exponential.

PROPOSITION 5.1. [25, Proposition 2.2] *Assume that $i+d$ inner iterations have been executed. Let $\mathbf{z}_{i+d} = P_{i+d}f(H_{i+d})\mathbf{e}_1$ be an approximation to $f(A)\mathbf{v}$ and define $\omega_{i+d} = \|\mathbf{z}_{i+d} - \mathbf{z}_i\|/\|\mathbf{z}_i\|$. If $\|f(A)\mathbf{v} - \mathbf{z}_{i+d}\| \ll \|f(A)\mathbf{v} - \mathbf{z}_i\|$ and $\|f(A)\mathbf{v}\| \approx \|\mathbf{z}_i\|$, then*

$$\|f(A)\mathbf{v} - \mathbf{z}_i\| \approx \frac{\omega_{i+d}}{1-\omega_{i+d}}\,\|\mathbf{z}_i\|. \tag{5.1}$$

The result in (5.1) shows that after $i + d$ iterations it is possible to provide an estimate of the error norm at iteration i. Therefore, we introduce the following stopping criterion for the approximation of $f(A)\mathbf{v}$:

$$\text{if} \quad \frac{\omega_{i+d}}{1-\omega_{i+d}} \le \varepsilon_{in} \quad \text{then} \quad \text{stop}$$

for some inner tolerance ε_{in}. In the numerical experiments presented in section 8 we have used $d = 4$. The accuracy of the inner iteration will influence the final accuracy of the inexact Lanczos bidiagonalization. In the notation of the previous section, if after i inner iterations the stopping criterion is satisfied, we have thus derived the following estimate for the perturbation occurring at the jth Lanczos step,

$$\|f(A)\mathbf{v}_j - \mathbf{z}_i\| = \|C_j\mathbf{v}_j\| \approx \varepsilon_{in}\|\mathbf{z}_i\|;$$

we recall that the matrix C_j is not explictly determined, and it is used to express the inexactness at iteration j in terms of a matrix-vector product with $\mathbf{v}_j$. Note that here $\|C_j\mathbf{v}_j\| = \|\mathbf{g}_j^{(1)}\|$, with the notation in (4.4). An analogous relation holds with respect to $f(A)^*\mathbf{u}_j$ and thus $\|\mathbf{g}_j^{(2)}\|$. We stress here that, since the approximation process changes at each iteration, the threshold for the quantity $\|C_j\mathbf{v}_j\|$ may vary as the Lanczos bidiagonalization proceeds, so that $\varepsilon_{in} = \varepsilon_{in}^{(j)}$. As experienced with other eigenvalue and linear system problems, $\varepsilon_{in}^{(j)}$ may even be allowed to grow during the iteration, without significantly affecting the overall process. This is discussed in the next section.

6. Relaxing the inner solution accuracy. The bound in (4.4) on the residual gap suggests that the accuracy on the inner solution approximation can be relaxed as convergence takes place. Indeed, following similar strategies in [32],[31],[4], we observe that it is the product $\|\mathbf{g}_j^{(1)}\|\,|\mathbf{e}_j^*\mathbf{y}|$ in (4.4) that needs to be small, and not each factor, to ensure a small gap; the same for $\|\mathbf{g}_j^{(2)}\|\,|\mathbf{e}_j^*\mathbf{x}|$. Therefore, if $|\mathbf{e}_j^*\mathbf{y}|$ is sufficiently small, indicating that the $(m + j)$th component of the eigenvector $\mathbf{q}$ is small, $\|\mathbf{g}_j^{(1)}\|$

is allowed to be larger, and the required accuracy of ε_{out} can still be achieved. This induces a variable (possibly growing) accuracy in the inner iteration, which drives the size of $\|\mathbf{g}_j^{(1)}\|$. In the following we shall first show that the quantities $|\mathbf{e}_j^*\mathbf{y}|$ and $|\mathbf{e}_j^*\mathbf{x}|$ do tend to decrease as the approximation improves. We then derive a computable expression for the variable stopping tolerance in the approximation of $f(A)\mathbf{v}$ and $f(A)^*\mathbf{u}$ at each iteration of the resulting "relaxed" Lanczos bidiagonalization process. This strategy may be convenient in case the cost approximating $f(A)\mathbf{v}$ and $f(A)^*\mathbf{u}$ is very high, as is the case for instance if an accurate approximation to the leading singular triplets is requested.

6.1. Spectral properties of the approximate singular triplets. To ensure that the magnitude of $\|\mathbf{g}_j^{(\ell)}\|$, $\ell = 1, 2$, can be relaxed in the bound (4.4), we need to verify that $|\mathbf{e}_j^*\mathbf{x}|$ and $|\mathbf{e}_j^*\mathbf{y}|$ become small as convergence takes place. This fact has been verified in the eigenvalue setting in [31], however the peculiar structure of the Lanczos bidiagonal recurrence requires the ad-hoc modifications of the results in [31]. To this end, we first define the submatrix of $\widetilde{\mathcal{B}}_{2m}$ of size $2k$ as

$$\widetilde{\mathcal{B}}_{2k} = \begin{bmatrix} 0 & M_k \\ T_k & 0 \end{bmatrix},$$

where M_k, T_k are the leading portions of the corresponding $m \times m$ matrices. Let $(\theta^{(2k)}, \mathbf{q}^{(2k)})$ be an eigenpair of $\widetilde{\mathcal{B}}_{2k}$, where $\mathbf{q}^{(2k)} = [\mathbf{x}; \mathbf{y}]$ has unit norm, and $\mathbf{x}, \mathbf{y} \in \mathbb{C}^k$. Further, let

$$\widetilde{\mathbf{q}} = \begin{bmatrix} \mathbf{x} \\ 0 \\ \mathbf{y} \\ 0 \end{bmatrix}, \tag{6.1}$$

where the 0-vectors have length $m-k$, and define $\mathcal{X} = [\widetilde{\mathbf{q}}, Y]$, where Y is chosen such that $\mathcal{X}$ is unitary. Define $\underline{\widetilde{\mathcal{B}}}_{2m} = Y^*\widetilde{\mathcal{B}}_{2m}Y \in \mathbb{C}^{(2m-1)\times(2m-1)}$. The following proposition shows that under certain hypotheses some of the components of the approximate eigenvectors do tend to zero as convergence takes place. Its proof is technical and it is postponed to the appendix.

PROPOSITION 6.1. *Let $(\theta^{(2k)}, \mathbf{q}^{(2k)})$ be an eigenpair of $\widetilde{\mathcal{B}}_{2k}$, and $\widetilde{\mathbf{q}}$ be as defined in (6.1). Let $\mathbf{s}_{2m}^* = \widetilde{\mathbf{q}}^*\widetilde{\mathcal{B}}_{2m} - \theta^{(2k)}\widetilde{\mathbf{q}}^*$, $\delta_{2m,2k} = \sigma_{min}(\underline{\widetilde{\mathcal{B}}}_{2m} - \theta^{(2k)}I) > 0$, and $\mathbf{r}_{2k} = t_{k+1,k}\begin{bmatrix} 0 \\ \mathbf{v}_{k+1} \end{bmatrix}\mathbf{e}_k^*\mathbf{q}^{(2k)}$. If*

$$\|\mathbf{r}_{2k}\| < \frac{\delta_{2m,2k}^2}{4\,\|\mathbf{s}_{2m}\|}, \tag{6.2}$$

then there exists a unit norm eigenvector $\mathbf{q} = [\mathbf{x}_1; \mathbf{x}_2; \mathbf{y}_1; \mathbf{y}_2]$ of $\widetilde{\mathcal{B}}_{2m}$ with $\mathbf{x}_1, \mathbf{y}_1 \in \mathbb{C}^k$, $\mathbf{x}_2, \mathbf{y}_2 \in \mathbb{C}^{m-k}$, such that

$$\left\| \begin{bmatrix} \mathbf{x}_2 \\ \mathbf{y}_2 \end{bmatrix} \right\| \leq \frac{\tau}{\sqrt{1+\tau^2}},$$

with $\tau \in \mathbb{R}$, $0 \leq \tau < 2\frac{\|\mathbf{r}_{2k}\|}{\delta_{2m,2k}}$. Moreover, if θ is the eigenvalue associated with $\mathbf{q}$, we have

$$|\theta - \theta^{(2k)}| \leq \|\mathbf{s}_{2m}\|\tau. \tag{6.3}$$

This proposition states that if after $k \leq m$ iterations of Lanczos bidiagonalization the computed residual $\|\mathbf{r}_{2k}\|$ is sufficiently small, then there exists an eigenvector of $\widetilde{\mathcal{B}}_{2m}$ such that some of its components are bounded correspondingly. These are precisely the components that allow us to relax the accuracy in the inner iteration. Note that $\delta_{2m,2k}$ gives an indication of the distance between the spectrum of $\widetilde{\mathcal{B}}_{2m}$ and $\theta^{(2k)}$. It should be kept in mind that for non-normal matrices, the value of $\delta_{2m,2k}$ may be much smaller [33, Example 2.4, p. 234]. On the other hand, since $\widetilde{\mathcal{B}}_{2m}$ is a perturbation to a Hermitian matrix, the quantity $\mathbf{s}_{2m}$ is an approximate residual for $(\theta^{(2k)}, \mathbf{q}^{(2k)})$ as an eigenpair of $\widetilde{\mathcal{B}}_{2m}$, and thus it will be small as m grows. As a consequence, condition (6.2) in the theorem is likely to be satisfied, and the eigenvalue error (6.3) may be much smaller than τ. For this reason, in our pratical implementation we assumed that condition (6.2) is satisfied after the first two iterations.

6.2. Variable accuracy in the inner approximation. In this section we show that relaxation in the inner accuracy at step $k \leq m$ is possible if there exists an eigenpair $(\theta^{2(k-1)}, \mathbf{q}^{2(k-1)})$ of $\widetilde{\mathcal{B}}_{2(k-1)}$ such that

$$\|\mathbf{r}_{2(k-1)}\| < \frac{\delta^2_{2m,2(k-1)}}{4\|\mathbf{s}_{2m}\|}, \tag{6.4}$$

$$\forall \theta_j \in \Lambda(\widetilde{\mathcal{B}}_{2m}), \qquad \theta_j \neq \theta, \qquad |\theta_j - \theta^{2(k-1)}| > 2\frac{\|\mathbf{s}_{2m}\|\|\mathbf{r}_{2(k-1)}\|}{\delta_{2m,2(k-1)}}. \tag{6.5}$$

The first condition (6.4) ensures that there exists an eigenvector $\mathbf{q}$ of $\widetilde{\mathcal{B}}_{2m}$ whose specified components are small, according to Proposition 6.1. Let θ be the eigenvalue associated with this $\mathbf{q}$. The second condition, (6.5), guarantees that the eigenvalue $\theta^{2(k-1)}$ of $\widetilde{\mathcal{B}}_{2(k-1)}$ is a perturbation of the eigenvalue θ of $\widetilde{\mathcal{B}}_{2m}$, which is the final approximation to the original problem. In particular, the two conditions ensure that $\theta^{2(k-1)}$ is closer to θ than to all other eigenvalues θ_j of $\widetilde{\mathcal{B}}_{2m}$. It is also interesting to observe that if (6.4) holds, then (6.5) can be replaced by the stricter but possibly more insightful condition $|\theta_j - \theta^{2(k-1)}| > \delta_{2m,2(k-1)}/2$.

The following theorem states how the use of a variable accuracy will still guarantee a small residual gap, and hence yields a true residual with an accuracy which is bounded by the accuracy of the gap, in agreement with (4.2).

THEOREM 6.2. *Assume m inexact Lanczos bidiagonalization iterations are carried out. Let $(\theta, \mathbf{q})$ be an eigenpair of $\widetilde{\mathcal{B}}_{2m}$, where θ is simple and $\|\mathbf{q}\| = 1$. Given $0 < \varepsilon_{out} \in \mathbb{R}$, with the notation of (4.4) assume that for $k = 1, \ldots, m$*

$$\|\mathbf{g}_k^{(1)}\|, \|\mathbf{g}_k^{(2)}\| \leq \begin{cases} \frac{\delta_{2m,2(k-1)}}{2m\|\mathbf{r}_{2(k-1)}\|}\varepsilon_{out} & \text{if } k > 1, \text{ and there exists } (\mathbf{q}^{2(k-1)}, \theta^{2(k-1)}) \\ & \text{of } \widetilde{\mathcal{B}}_{2(k-1)} \text{ satisfying (6.4) and (6.5),} \\ \frac{1}{m}\varepsilon_{out} & \text{otherwise.} \end{cases} \tag{6.6}$$

Then $\|(\mathcal{F}\mathcal{W}_{2m}\mathbf{q} - \theta\mathcal{W}_{2m}\mathbf{q}) - \mathbf{r}_{2m}\| \leq \varepsilon_{out}$.

Proof. Although the strategy used for the proof is similar to that of Theorem 3.1 in [31], the block structure of our problem requires specialized technical details. Suppose that at the $(k-1)$th iteration there exists an eigenpair $(\theta^{2(k-1)}, \mathbf{q}^{2(k-1)})$ of $\widetilde{\mathcal{B}}_{2(k-1)}$ satisfying the conditions (6.4) and (6.5). This implies that $\theta^{2(k-1)}$ is a perturbation of the considered eigenvalue θ of $\widetilde{\mathcal{B}}_{2m}$, since θ is the only eigenvalue of $\widetilde{\mathcal{B}}_{2m}$

such that

$$|\theta - \theta^{2(k-1)}| \leq 2\frac{\|\mathbf{s}_{2m}\|\|\mathbf{r}_{2(k-1)}\|}{\delta_{2m,2(k-1)}}.$$

Let $\mathcal{K} \subset \{1,\ldots,m\}$ be defined such that for each $k \in \mathcal{K}$ there exists a eigenpair $(\mathbf{q}^{2(k-1)}, \theta^{2(k-1)})$ of $\widetilde{\mathcal{B}}_{2(k-1)}$ satisfying the conditions (6.4) and (6.5). Then, similar to the reasoning in the proof of Lemma 4.1 and using (4.4),

$$\begin{aligned}
\|(\mathcal{F}\mathcal{W}_{2m}\mathbf{q} - \theta\mathcal{W}_{2m}\mathbf{q}) - \mathbf{r}_{2m}\| = \|\mathcal{G}_{2m}\mathbf{q}\| &\leq \sum_{k=1}^{m}(\|\mathbf{g}_k^{(1)}\|^2\,|\mathbf{e}_k^*\mathbf{y}|^2 + \|\mathbf{g}_k^{(2)}\|^2\,|\mathbf{e}_k^*\mathbf{x}|^2)^{\frac{1}{2}} \\
&\leq \sum_{k\in\mathcal{K}}(\|\mathbf{g}_k^{(1)}\|^2\,|\mathbf{e}_k^*\mathbf{y}|^2 + \|\mathbf{g}_k^{(2)}\|^2|\mathbf{e}_k^*\mathbf{x}|^2)^{\frac{1}{2}} + \sum_{\substack{k\notin\mathcal{K},\\ k\leq m}}(\|\mathbf{g}_k^{(1)}\|^2\,|\mathbf{e}_k^*\mathbf{y}|^2 + \|\mathbf{g}_k^{(2)}\|^2|\mathbf{e}_k^*\mathbf{x}|^2)^{\frac{1}{2}} \\
&\leq \sum_{k\in\mathcal{K}}\frac{\delta_{2m,2(k-1)}\varepsilon_{out}}{2m\|\mathbf{r}_{2(k-1)}\|}(|\mathbf{e}_k^*\mathbf{y}|^2 + |\mathbf{e}_k^*\mathbf{x}|^2)^{\frac{1}{2}} + \sum_{\substack{k\notin\mathcal{K},\\ k\leq m}}\frac{\varepsilon_{out}}{m}(|\mathbf{e}_k^*\mathbf{y}|^2 + |\mathbf{e}_k^*\mathbf{x}|^2)^{\frac{1}{2}} \\
&\leq \sum_{k\in\mathcal{K}}\frac{\delta_{2m,2(k-1)}\varepsilon_{out}}{2m\|\mathbf{r}_{2(k-1)}\|}2\frac{\|\mathbf{r}_{2(k-1)}\|}{\delta_{2m,2(k-1)}} + \sum_{\substack{k\notin\mathcal{K},\\ k\leq m}}\frac{\varepsilon_{out}}{m} \\
&= \frac{|\mathcal{K}|}{m}\,\varepsilon_{out} + \frac{m-|\mathcal{K}|}{m}\,\varepsilon_{out} = \varepsilon_{out} \qquad \square
\end{aligned}$$

Algorithm 1: Inexact Lanczos bidiagonalization

Input: $A \in \mathbb{C}^{n\times n}$ non-Hermitian, a function f, a maximum number of (outer) iterations m, an (outer) tolerance ε_{out}.

Output: An approximation to the leading singular triplet

1: Choose $\mathbf{v}_1$ with $\|\mathbf{v}_1\| = 1$, and set $V = [\mathbf{v}_1]$, $U = \emptyset$, $M = \emptyset$, $T = \emptyset$.
2: **for** $j = 1,\ldots,m$
3: $\quad\mathbf{z} \approx f(A)\mathbf{v}_j$
4: $\quad\mathbf{u}_j, \mathbf{m}_j \longleftarrow$ `rgs`$(\mathbf{z}, U)$
5: $\quad$Expand basis: $U = [U, \mathbf{u}_j]$
6: $\quad$Expand matrix: $M = [M, \mathbf{m}_j]$ (the old M is first padded with a zero row)
7: $\quad\mathbf{z} \approx f(A)^*\mathbf{u}_j$
8: $\quad\mathbf{v}_{j+1}, \mathbf{t}_j \longleftarrow$ `rgs`$(\mathbf{z}, V)$
9: $\quad$Expand basis: $V = [V, \mathbf{v}_{j+1}]$
10: $\quad$Expand matrix: $T = [T, \mathbf{t}_j]$
11: $\quad K =$ `[zeros(j,j), M ; T , zeros(j,j)]`
12: $\quad[Q, D] \longleftarrow$ `eig`(K)
13: $\quad(\theta, \mathbf{q}) \longleftarrow$ with $\theta = \max_i |D_{ii}|$ (extract $\mathbf{x}, \mathbf{y}$ from $\mathbf{q} = [\mathbf{x}; \mathbf{y}]$ with $\|\mathbf{x}\| = 1, \|\mathbf{y}\| = 1$)
14: $\quad$Convergence check: if $|T(j+1, j)\mathbf{q}(j)|/\theta < \varepsilon_{out}$ then return $(\theta, \mathbf{x}, \mathbf{y})$ and stop
15: $\quad$If required: compute variable tolerance to be used in the next iteration
16: **end**

7. Practical implementation. Algorithm 1 implements the inexact Lanczos bidiagonalization to approximate $\|f(A)\|_2$ and the associated singular vectors. The function rgs$(\mathbf{z}, Z)$ double orthogonalizes the vector $\mathbf{z}$ with respect to the orthogonal columns of Z, and returns the orthogonalization coefficients. The same algorithm can be used to approximate more singular triplets.

At every iteration of the Lanczos bidiagonalization, two inner iterations (line 4 and line 7) approximate the corresponding matrix-vector multiplication $f(A)\mathbf{v}$ and

$f(A)^*\mathbf{u}$, respectively. The inner iteration uses one of the algorithms for approximating the action of a matrix function to a vector, as discussed in section 5. In theory, any such algorithm could be used; in our experiments we employed both the standard and extended Krylov subspace methods.

If the variant with variable inner tolerance is employed, the next inner tolerance is computed at the end of every Lanczos bidiagonalization iteration. To be conservative, during the first two iterations the inner tolerance is $\frac{1}{m}\varepsilon_{out}$, so that $\|\mathbf{g}_k\| \leq \frac{1}{m}\varepsilon_{out}$. Then, in subsequent iterations we assume that (6.4) and (6.5) are always satisfied, and thus we require that the inner stopping criterion is such that

$$\max\{\|\mathbf{g}_k^{(1)}\|, \|\mathbf{g}_k^{(2)}\|\} \leq \frac{\delta_{2m,2(k-1)}}{2m\|\mathbf{r}_{2(k-1)}\|}\varepsilon_{out}.$$

Note that a relative criterion is always used, that is, in practice the quantity to be checked is divided by the current approximation $\theta^{2(k-1)}$. This corresponds to using $\varepsilon_{out}^{(k)} = \theta^{2(k-1)}\varepsilon_{out}$ for some fixed value ε_{out}. Since $\delta_{2m,2(k-1)}$ is not available at iteration k, we consider the following approximation:

$$\delta^{2(k-1)} := \min_{\theta_j \in \Lambda(\widetilde{\mathcal{B}}_{2(k-1)}) \setminus \{\theta^{2(k-1)}\}} |\theta^{2(k-1)} - \theta_j|.$$

In fact, $\delta_{2m,2(k-1)}$ can be much smaller than the computed $\delta^{2(k-1)}$. However, it will not be overrated much when $\theta^{2(k-1)}$ is converging to the corresponding eigenvalue θ of $\widetilde{\mathcal{B}}_{2m}$, since it is related to the sensitivity of $\widetilde{\mathcal{B}}_{2m}$ and not of the matrix $\mathcal{F}$. If the $\delta^{2(k-1)}$ is very small, it constrains the inner accuracy to be very small too. This occurs when the largest eigenvalues of $\widetilde{\mathcal{B}}_{2m}$ are clustered. We refer to section 8.4 for a numerical illustration. We also remark that the computation of $\delta^{2(k-1)}$ does not significantly increase the computational costs, as all the eigenvalues of $\widetilde{\mathcal{B}}_{2(k-1)}$ are already required to obtain the current approximation.

8. Numerical experiments. In this section we report on our numerical experiments to evaluate the performance of the inexact Lanczos bidiagonalization for different combinations of matrices and functions. All experiments were performed with Matlab Version 7.13.0.564 (R2011b) on a Dell Latitude laptop running Ubuntu 14.04 with 4 CPUs at 2.10GHz. We are mainly interested in the first singular triplet of $f(A)$, so as to obtain $\|f(A)\|$. We considered five different matrices, summarized in Table 8.1, all of dimension $n = 10,000$ except A_4. The spectrum of a sample of these matrices of smaller size, $n = 1000$, is reported in Figure 8.1. For A_4, the matrix originating from a cavity driven problem was shifted by $10I$ so that all functions could be treated with all methods; we refer to the Matrix Market site for more information on this problem [27]. For A_5 a 5-point stencil finite difference approximation was used, together with homogeneous Dirichlet boundary conditions. We considered the following functions,

$$\exp(x), \quad \exp(-x), \quad \sqrt{x}, \quad \frac{1}{\sqrt{x}}, \quad \frac{\exp(-\sqrt{x})-1}{x}.$$

We note that all these functions allow for an efficient computation when applied to small scale matrices, by means of specifically designed Matlab functions; see [19]. The performance also critically depends on the choice of the inner method for approximating $f(A)\mathbf{v}$ and $f(A)^*\mathbf{u}$ at each iteration. We shall report our experience

Matrix	Structure	Description
A_1	tridiag$(0, \underline{\lambda_i}, 0.3)$	$\lambda_i = (1 + \rho_i^{(1)}) + i(\rho_i^{(2)} - 0.5)$
A_2	tridiag$(1.5, \underline{2}, -1)$	
A_3	Toeplitz	i-th row: $[4, 0, 0, 0, 0, -2, 0, \underline{10}, 0, 0, 0, 6]$
A_4	shifted E20R1000	Driven cavity problem (Matrix Market) shifted as: $A := A_{e20r1000} + 10I$
A_5	Sparse	Centered Finite Difference discretization of $\mathcal{L}(u) = -\nabla^2 u - 100u_x - 100u_y$

Table 8.1: Description of the selected matrices, all of size $n = 10,000$, except A_4, of size 4241. $\rho_i^{(j)}$ is a random entry taken from a uniform distribution in $(0, 1)$.

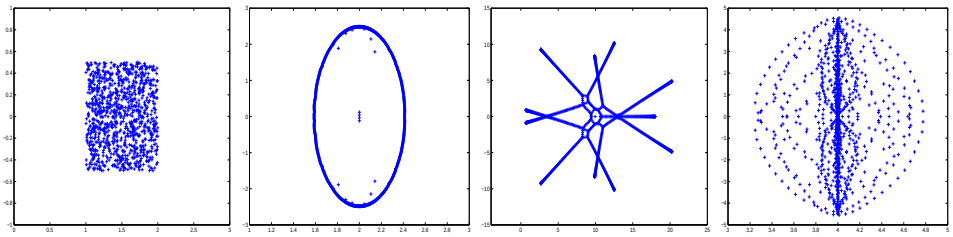

Fig. 8.1: Spectrum of matrices A_1, A_2, A_3, A_5 (from left to right) in Table 8.1, for a smaller size, $n = 1000$.

with the standard and extended Krylov methods. The Rational Krylov method could also be employed for this approximation.

A random vector is used to start the inexact Lanczos process. Convergence is monitored by checking the computed residual norm with respect to the first singular triplet, and the inexact Lanczos bidiagonalization terminates as soon as (4.3) is satisfied; different values for ε_{out} will be considered. In case more than one triplet is desired, then all corresponding residuals should be monitored. In section 8.1 we explore the fixed inner tolerance method, and the dependence of its performance on all the other parameters, including the outer accuracy. Indeed, if only a rough approximation of $\|f(A)\|$ is required, the computational efforts should be proportionally low. In section 8.4 the influence of the variable (relaxed) inner tolerance described in section 6.2 is analyzed, thus a more stringent final accuracy is considered so as to exercise the variable inner threshold.

8.1. Assessing the effectiveness of the inexact bidiagonalization. We analyze the performance of the inexact method when approximating $\|f(A_i)\|$ together with the associated singular vectors. To this end, we need to monitor the number of iterations of both the outer and the two inner iterations, together with the execution time required to reach the required tolerance. In particular, we show both the total and average number of inner iterations. We also display the distance between the final first approximate singular value, $\tilde{\sigma}_1$ and the second approximate singular value, $\tilde{\sigma}_2$: a small relative distance implies that the method will take more iterations to converge. Moreover, this distance cannot be easily predicted from the matrix A, although it significantly influences the computation. For instance, the largest (in

matr	function f	$\tilde{\sigma}_1$	$\frac{\tilde{\sigma}_1-\tilde{\sigma}_2}{\tilde{\sigma}_1}$	tot # outer	tot # inner	average # inner	exec time
A_1	$\exp(-x)$	0.463506	3.34e-02	14	308	11.0	0.39
	$\sqrt{x}$	1.50143	1.29e-02	15	351	11.7	0.44
	$\frac{\exp(-\sqrt{x})-1}{x}$	0.727351	2.11e-02	14	414	14.8	0.70
	$\exp(x)$	9.19137	1.76e-02	16	352	11.0	0.59
	$1/\sqrt{x}$	1.10350	2.33e-02	13	364	14.0	0.73
A_2	$\exp(-x)$	0.222621	1.78e-02	11	308	14.0	0.24
	$\sqrt{x}$	1.7921	2.05e-02	8	252	15.8	0.28
	$\frac{\exp(-\sqrt{x})-1}{x}$	0.469829	1.81e-02	10	429	21.4	0.68
	$\exp(x)$	12.1783	9.01e-02	5	139	13.9	0.14
	$1/\sqrt{x}$	0.814356	2.26e-02	8	324	20.2	0.44
A_3	$\exp(-x)$	0.508086	1.48e-02	13	709	27.3	0.61
	$\sqrt{x}$	4.56086	1.85e-02	12	1036	43.2	3.11
	$\frac{\exp(-\sqrt{x})-1}{x}$	0.615673	1.84e-02	12	1953	81.4	18.29
	$\exp(x)$	$6.75709 \cdot 10^8$	1.45e-02	13	694	26.7	0.83
	$1/\sqrt{x}$	0.959018	1.70e-02	12	1852	77.2	14.56
A_4	$\exp(-x)$	0.000172183	2.42e-01	6	456	38.0	0.58
	$\sqrt{x}$	6.09177	2.78e-02	14	454	16.2	0.81
	$\frac{\exp(-\sqrt{x})-1}{x}$	0.118301	3.07e-02	10	486	24.3	1.15
	$\exp(x)$	$3.15141 \cdot 10^{10}$	1.35e-01	5	397	39.7	0.49
	$1/\sqrt{x}$	0.354039	5.55e-02	9	394	21.9	0.81
A_5	$\exp(-x)$	0.99709	3.39e-02	7	223	15.9	0.34
	$\sqrt{x}$	2.81987	1.16e-02	16	5145	160.8	193.35
	$\frac{\exp(-\sqrt{x})-1}{x}$	6.93384	2.44e-01	4	1570	196.2	111.67
	$\exp(x)$	2959.17	1.84e-02	14	450	16.1	0.65
	$1/\sqrt{x}$	7.36692	2.31e-01	4	1564	195.5	112.22

Table 8.2: Inexact Lanczos bidiagonalization for approximating the leading singular triplet of $f(A)$; outer tolerance $\varepsilon = 10^{-2}$.

modulo) eigenvalues of $\mathcal{F}$ associated with the matrix function $A_2^{\frac{1}{2}}$ are:

-1.7965100, 1.7965100, 1.7964169, -1.7964169, 1.7962429, -1.7962424 .

Although this fact does not constitute a difficulty if just the order of magnitude of $\|A_2^{\frac{1}{2}}\|$ is sought, it indicates that requiring a more accurate approximation will lead to significantly more expensive computations. This problem can be readily observed by comparing the outer number of iterations in Table 8.2 and Table 8.3, where we report the results of our experiments for $\varepsilon_{out} = 10^{-2}$ and $\varepsilon_{out} = 10^{-4}$, respectively. In both cases, the inner tolerance was set to $\varepsilon_{in} = \varepsilon_{out}/(m_{\max})$, where $m_{\max} = 1000$, so that $\varepsilon_{in} = 10^{-7}$ for the more stringent outer tolerance. For all examples, the first six significant digits of $\tilde{\sigma}_1$ are reported.

Comparing the two tables also shows that the singular values are as accurate as the outer tolerance can predict: for smaller ε_{out} already the third singular value digit changes, that is it still has to reach its final (exact) value. This is obviously also related to the relative distance from the second singular value, which is better captured for a smaller ε_{out}.

We also observe that the choice of f strongly influences the overall performance: the bidiagonalization process may take the same number of (outer) iterations for two different selections of f, and yet the total computational cost may be significantly different (see A_1 and A_3 in Table 8.2). As a consequence, the number of outer iterations is not a realistic measure of the algorithm complexity.

matr	function f	$\tilde{\sigma}_1$	$\frac{\tilde{\sigma}_1-\tilde{\sigma}_2}{\tilde{\sigma}_1}$	tot # outer	tot # inner	average # inner	exec time
A_1	$\exp(-x)$	0.463735	2.04e-02	24	624	13.0	0.87
	$\sqrt{x}$	1.50496	8.76e-04	53	1775	16.7	3.27
	$\frac{\exp(-\sqrt{x})-1}{x}$	0.728200	7.62e-03	29	1160	20.0	2.47
	$\exp(x)$	9.19576	9.22e-03	32	832	13.0	1.21
	$\frac{1}{\sqrt{x}}$	1.10504	5.52e-03	29	1156	19.9	2.21
A_2	$\exp(-x)$	0.223129	4.88e-05	209	7104	17.0	26.72
	$\sqrt{x}$	1.79651	5.18e-05	162	8069	24.9	18.92
	$\frac{\exp(-\sqrt{x})-1}{x}$	0.470776	3.85e-05	193	12320	31.9	42.03
	$\exp(x)$	12.1825	8.39e-04	47	1596	17.0	1.41
	$\frac{1}{\sqrt{x}}$	0.816492	5.90e-05	150	9210	30.7	29.43
A_3	$\exp(-x)$	0.509010	4.43e-05	224	14544	32.5	31.55
	$\sqrt{x}$	4.57175	3.35e-05	250	41844	83.7	533.17
	$\frac{\exp(-\sqrt{x})-1}{x}$	0.616989	1.20e-04	155	39968	128.9	1078.46
	$\exp(x)$	$6.77296\cdot10^{8}$	1.17e-04	183	11660	31.9	19.91
	$\frac{1}{\sqrt{x}}$	0.960790	2.12e-05	312	77958	124.9	1827.25
A_4	$\exp(-x)$	0.000172195	2.32e-01	9	783	43.5	1.26
	$\sqrt{x}$	6.09289	2.38e-02	22	1103	25.1	2.25
	$\frac{\exp(-\sqrt{x})-1}{x}$	0.118347	2.44e-02	16	1109	34.7	3.16
	$\exp(x)$	$3.15148\cdot10^{10}$	1.34e-01	7	626	44.7	0.87
	$\frac{1}{\sqrt{x}}$	0.354473	1.18e-02	19	1227	32.3	3.91
A_5	$\exp(-x)$	0.998062	7.74e-03	24	911	19.0	1.37
	$\sqrt{x}$	2.82811	1.67-04	185	70926	191.7	4059.40
	$\frac{\exp(-\sqrt{x})-1}{x}$	6.93435	2.32-01	7	2814	201.0	186.39
	$\exp(x)$	2975.18	2.91e-03	55	2091	19.0	3.20
	$\frac{1}{\sqrt{x}}$	7.36768	2.17-01	7	2814	201.0	197.62

Table 8.3: Inexact Lanczos bidiagonalization for approximating the leading singular triplet of $f(A)$; outer tolerance $\varepsilon = 10^{-4}$.

On a negative side, we observe that in both tables the method performs poorly on A_5 for $f(x) = \sqrt{x}$. For this particular matrix, the inner method takes very many iterations during the whole Lanczos process, with a number of inner iterations that is close to the average throughout. We anticipate that this is not the case for the power method, where as the outer iterations proceed, drastically fewer iterations are required in the inner approximation. This phenomenon seems to be peculiar to this combination of function and matrix, since in all other cases the performance of the Lanczos and power methods is more similar, and it will be further investigated in a future study.

Finally, for the exponential functions $\exp(x), \exp(-x)$ we computed the upper bound in (2.1) by using the Matlab function `eigs` applied to $\frac{1}{2}(A + A^*)$. In all cases except matrix A_4 the estimate is pretty sharp. On the other hand, for $\exp(A_4)$ the upper bound was $2 \cdot 10^{13}$, which is three orders of magnitude larger than the actual norm; for[2] $\exp(-A_4)$ the upper bound was 0.004737, which is more than one order of magnitude larger than the actual value, 0.000172. This example illustrates that, as discussed in section 2, the accuracy of this type of estimate cannot be easily monitored, especially in the case of non-normal matrices.

[2]This bound is obtained for $\exp(\hat{A})$ with $\hat{A} = -A_4$.

matr	function	tot # outer	tot # inner	$\tilde{\sigma}_1$	residual norm	exec time
A_1	$\exp(-x)$	51	1071	0.46327	9.8648e-03	1.5
	$\sqrt{x}$	93	2010	1.5028	9.8607e-03	2.8
	$\frac{\exp(-\sqrt{x})-1}{x}$	61	1782	0.71879	9.8904e-03	3.5
	$\exp(x)$	65	1409	9.1666	9.9964e-03	2.0
	$1/\sqrt{x}$	69	1942	1.0938	9.8670e-03	3.3
A_2	$\exp(-x)$	36	899	0.22238	9.8636e-03	0.8
	$\sqrt{x}$	37	979	1.7903	9.7995e-03	1.1
	$\frac{\exp(-\sqrt{x})-1}{x}$	36	1216	0.46921	9.7553e-03	2.1
	$\exp(x)$	9	232	12.176	9.4623e-03	0.2
	$1/\sqrt{x}$	36	1215	0.81375	9.8715e-03	1.8
A_3	$\exp(-x)$	38	1605	0.50724	9.9024e-03	1.5
	$\sqrt{x}$	41	1000	4.5564	9.8934e-03	1.3
	$\frac{\exp(-\sqrt{x})-1}{x}$	34	4448	0.61486	9.9901e-03	39.1
	$\exp(x)$	38	1699	$6.7455{\cdot}10^{8}$	9.7718e-03	1.7
	$1/\sqrt{x}$	36	4684	0.95774	9.7264e-03	36.3
A_4	$\exp(-x)$	11	825	0.00017219	5.8988e-03	1.0
	$\sqrt{x}$	56	1710	6.0870	9.9037e-03	3.0
	$\frac{\exp(-\sqrt{x})-1}{x}$	28	1309	0.11823	9.5839e-03	3.3
	$\exp(x)$	10	775	$3.1510{\cdot}10^{10}$	8.8031e-03	1.1
	$1/\sqrt{x}$	33	1361	0.35405	9.9455e-03	2.8
A_5	$\exp(-x)$	15	376	0.99643	9.9168e-03	0.52
	$\sqrt{x}$	52	1479	2.81329	9.9649e-03	18.47
	$\frac{\exp(-\sqrt{x})-1}{x}$	7	2745	6.93355	9.6566e-03	189.19
	$\exp(x)$	55	1363	2956.34	9.8499e-03	1.79
	$1/\sqrt{x}$	8	3137	7.36721	6.9885e-03	202.71

Table 8.4: Power method for approximating the leading singular triplet of $f(A)$; outer tolerance $\varepsilon = 10^{-2}$

8.2. Comparisons with the power method. We wish to compare the performance of the new method with that of the power method, as described in section 2. Since in most cases, the leading singular values are not well isolated, we expect that the power method will be slow if an accurate approximation is required. Therefore, we only report results for $\varepsilon = 10^{-2}$. Moreover, our experience is that since the computation is inexact, the product $f(A)^*(f(A)\mathbf{v})$ may give complex values, since the computed actions of $f(A)$ and $f(A)^*$ are not the conjugate of each other. As a result, the approximate eigenvalue may be complex, though with a small imaginary part, and the quantity that is actually computed is given by

$$\lambda^{(k)} = \left| \frac{(\mathbf{v}^{(k)})^* f(A)^* f(A) \mathbf{v}^{(k)}}{(\mathbf{v}^{(k)})^* \mathbf{v}^{(k)}} \right|,$$

where $\mathbf{v}^{(k)}$ is the power method direction after k iterations. Consequently, at convergence we obtain $\tilde{\sigma}_1 \approx \sqrt{\lambda^{(k)}}$. The stopping criterion is based on the relative eigenvalue residual norm, that is

$$\|\mathbf{y}^{(k)} - \lambda^{(k)} \mathbf{v}^{(k)}\| / \lambda^{(k)} \leq \varepsilon_{out},$$

where $\mathbf{y}^{(k)}$ is the result of the approximation of $f(A)^*(f(A)\mathbf{v}^{(k)})$. Note that we kept the same tolerance as for the Lanczos bidiagonalization, although a more stringent tolerance may be required in practice. Table 8.4 collects the results for all test cases.

As expected, the power method is more expensive than the Lanczos procedure, on average four to five times more expensive, in all those cases when the first singular value is not well separated from the second one. Only for the cases of good separation, for instance with A_5 and the functions $(\exp(\sqrt{x})-1)/x$ and $1/\sqrt{x}$, convergence is reached in very few iterations, and the power method is competitive.

We also implemented the power method as described in [19, Algorithm 3.19], using the relative singular value residual as stopping criterion. The performance, both in terms of inner and outer number of iterations, is comparable to that of Table 8.4. Finally, we stress that in both implementations the stopping criterion involves inexact matrix-vector products, therefore the monitored quantity is not the true residual of the corresponding problem.

8.3. Numerical tests with the extended Krylov subspace. If high accuracy is required for the final approximation, so that a more stringent outer tolerance is used, then the inner iteration also requires more computational effort, as its stopping tolerance is also decreased. In this case, it may be appropriate to use more effective methods. One such possibility is the extended Krylov subspace method [8],[25], which may be convenient in case the considered function requires good approximation of the eigenvalues of A closest to the origin. In Table 8.5 we report the runs for $\varepsilon_{out} = 10^{-4}$ when EKSM is used; these numbers should be compared with those in Table 8.3. We notice that EKSM requires the solution of a system with A (or A^*) at each iteration; to limit computational costs, a sparse LU factorization of A was performed and stored once for all at the beginning of the Lanczos bidiagonalization, and used repeatedly in the inner iteration. This represents a tremendous saving with respect to more general rational approximations, where solves with $(A - \tau_j I)$ have to be performed at each inner iteration, with τ_j varying with the inner step.

In Table 8.5 all cases where EKSM provides faster convergence, that is lower execution time, are marked in boldface. It is clear that EKSM is beneficial when good approximations to both ends of the spectrum are required, as is the case for x^α. The lack of improvement in the case of the exponential is not unexpected, as it is known, at least in the Hermitian case, that only one extreme of the spectrum needs to be captured for a fast approximation of $\exp(A)\mathbf{v}$.

We also remark that EKSM could also be employed as inner method in the case of the power iteration used in section 8.2.

8.4. Numerical tests with variable accuracy. In the previous sections, for $\varepsilon_{out} = 10^{-4}$ the inner tolerance was set to the fixed value $\varepsilon_{in} = 10^{-7}$. Here we explore the performance of the inexact computation when the inner tolerance is relaxed.

A relaxed inner accuracy is most convenient when each inner iteration is expensive, so as to profit from a lower number of inner iterations. Therefore, we report on our experience with the extended Krylov subspace as inner method, as the method requires one system solve with the coefficient matrix at each iteration. A more stringent outer tolerance was used, that is $\varepsilon_{out} = 10^{-7}$, than in previous experiments, so as to clearly see the relaxation in the inner tolerance; we also used $m_{\max} = 50$ as maximum number of iterations, so as to balance the much smaller ε_{out} for determining the initial inner tolerance.

Figure 8.2 shows the performance of the relaxation strategy for A_5 and $f(x) = 1/\sqrt{x}$. The plot shows the outer convergence history as the bidiagonalization proceeds, and the corresponding variable inner tolerance. The digits next to each iteration report the actual numbers of inner iterations by means of EKSM to reach the required inner accuracy for approximating $f(A)\mathbf{v}_j$; similar numbers were observed for $f(A)^*\mathbf{u}_j$.

matr	function	$\tilde{\sigma}_1$	$\frac{\tilde{\sigma}_1-\tilde{\sigma}_2}{\tilde{\sigma}_1}$	tot # outer	tot # inner	average # inner	exec time
A_1	$\exp(-x)$	0.463735	2.04e-02	24	480	10.0	3.49
	$\sqrt{x}$	1.50496	8.76e-04	53	954	9.0	8.24
	$\frac{\exp(-\sqrt{x})-1}{x}$	0.728200	7.62e-03	29	522	9.0	4.29
	$\exp(x)$	9.19576	9.22e-03	32	704	11.0	5.82
	$\frac{1}{\sqrt{x}}$	1.10504	5.52e-03	29	522	9.0	4.59
A_2	$\exp(-x)$	0.223129	4.88e-05	209	5434	13.0	36.71
	$\sqrt{x}$	1.79651	5.18e-05	162	3564	11.0	20.03
	$\frac{\exp(-\sqrt{x})-1}{x}$	0.470776	3.85e-05	193	4246	11.0	**29.99**
	$\exp(x)$	12.1825	8.39e-04	47	1408	15.0	5.07
	$\frac{1}{\sqrt{x}}$	0.816492	5.90e-05	150	3300	11.0	**18.70**
A_3	$\exp(-x)$	0.509010	4.43e-05	224	11827	26.4	106.31
	$\sqrt{x}$	4.57175	3.35e-05	250	9402	18.8	**84.26**
	$\frac{\exp(-\sqrt{x})-1}{x}$	0.616989	1.20e-04	155	5578	18.0	**40.02**
	$\exp(x)$	$6.77296 \cdot 10^8$	1.17e-04	183	11169	33	112.76
	$\frac{1}{\sqrt{x}}$	0.960790	2.12e-05	312	11449	18.3	**125.20**
A_4	$\exp(-x)$	0.000172195	2.32e-01	9	376	20.9	4.20
	$\sqrt{x}$	6.09289	2.38e-02	22	483	11.0	5.99
	$\frac{\exp(-\sqrt{x})-1}{x}$	0.118347	2.44e-02	16	318	9.9	4.32
	$\exp(x)$	$3.15148 \cdot 10^{10}$	1.34e-01	7	527	37.6	4.66
	$\frac{1}{\sqrt{x}}$	0.354473	1.18e-02	19	416	10.9	4.88
A_5	$\exp(-x)$	0.998062	7.74e-03	24	887	18.5	11.95
	$\sqrt{x}$	2.82811	1.67e-04	185	8165	22.1	**121.99**
	$\frac{\exp(-\sqrt{x})-1}{x}$	6.93435	2.32e-01	7	294	21.0	**5.01**
	$\exp(x)$	2975.18	2.91e-03	55	2090	19.0	25.19
	$\frac{1}{\sqrt{x}}$	7.36768	2.17e-01	7	294	21.0	**4.32**

Table 8.5: Inexact Lanczos bidiagonalization, outer tolerance $\varepsilon = 10^{-4}$, inner approximation: extended Krylov subspace method.

Table 8.6 reports the values of $\delta_{2m,2(k-1)}$ and $\delta^{2(k-1)}$ during the iterations displayed in Figure 8.2; see the discussion on these parameters at the end of section 7. For this specific example, the values of $\delta^{2(k-1)}$ are a good estimate for the actual $\delta_{2m,2(k-1)}$ even at an early stage of the iteration (we recall here that no relaxed strategy is used in the first two iterations).

We also experimented with the approximation of more than one triplet. We report on our findings for A_1 and again $f(x) = 1/\sqrt{x}$ (similar accuracies were obtained for other functions for the same matrix); to explore the variable inner accuracy we used $\varepsilon_{out} = 10^{-9}$ and $m_{\max} = 100$. Table 8.7 shows the largest ten singular values obtained with a fixed inner tolerance of 10^{-11} ($\tilde{\sigma}_j$, second column), and with a relaxed inner tolerance ($\tilde{\sigma}_j^{(fl)}$, third column), which a-posteriori we observed to go from 10^{-11} up to 10^{-5}. The last column reports the relative error $|\tilde{\sigma}_j - \tilde{\sigma}_j^{(fl)}|/\tilde{\sigma}_j$. In both cases, the inexact Lanczos iteration was stopped as soon as the outer stopping criterion was satisfied for the largest singular value. While in the fixed inner tolerance case the number of iterations varied between 28 and 30, in the flexible case a number of iterations as low as 15 was needed to satisfy the inner criterion at the last stage of the convergence process. After exiting the flexible procedure, however, the first ten approximate singular values are very close to those obtained with the fixed inner tolerance, much closer than warranted by the final inner accuracy of 10^{-5}. This shows in particular that the flexible inner tolerance is conservative, and more accurate

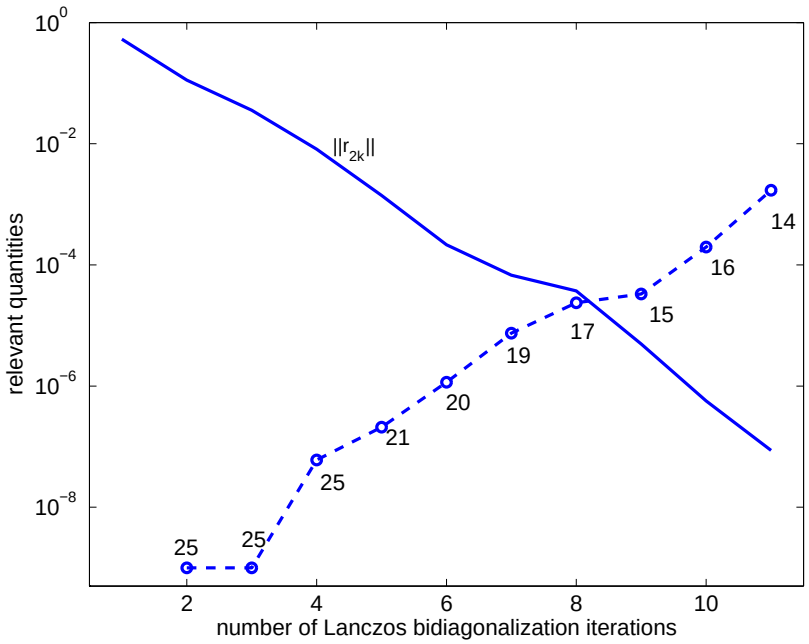

Fig. 8.2: Relaxed inner iteration for variable stopping tolerance, to approximate $\|A_5^{-1/2}\|$, with $\varepsilon_{out} = 10^{-7}$.

$k-1$	$\delta^{2(k-1)}$	$\delta_{2m,2(k-1)}$
1	-	2.3147e-01
2	-	7.7043e-01
3	2.1738e+00	9.5073e-01
4	1.7049e+00	9.6671e-01
5	1.6151e+00	9.6744e-01
6	1.6030e+00	9.6746e-01
7	1.6017e+00	9.6746e-01
8	1.3696e+00	9.6746e-01
9	9.7931e-01	9.6746e-01
10	9.6834e-01	9.6746e-01
11	9.6757e-01	9.6746e-01

Table 8.6: Values of $\delta^{2(k-1)}$ and $\delta_{2m,2(k-1)}$ as the relaxed iteration proceeds, with data as in Figure 8.2.

approximations are usually expected. We refer the reader to [31] for a more detailed analysis of the approximation of more than one eigenpair (and thus more than one triplet in our context).

9. Final considerations. We have explored the use of an inexact Lanczos bidiagonalization method for approximating the leading singular triplet of a large matrix function, and in particular its 2-norm. Although several strategies are known

j	$\tilde{\sigma}_j$	$\tilde{\sigma}_j^{(fl)}$	$\lvert\tilde{\sigma}_j - \tilde{\sigma}_j^{(fl)}\rvert/\tilde{\sigma}_j$
1	1.117020718026223	1.117020718026212	9.939144428645173e-15
2	1.107884805324699	1.107884805324724	2.264769787972212e-14
3	1.098394607515649	1.098394607513931	1.564063676705998e-12
4	1.095557563655289	1.095557550135225	1.234080654983002e-08
5	1.087939266226247	1.087939266157844	6.287397018996839e-11
6	1.081739455193175	1.081739454786304	3.761265981356762e-10
7	1.077326677541678	1.077326677826174	2.640758724893079e-10
8	1.070641401649297	1.070641400153385	1.397211109700097e-09
9	1.064637797334345	1.064637795718615	1.517633507177347e-09
10	1.055679471834666	1.055679470916211	8.700132135409333e-10

Table 8.7: First ten approximate singular values of $A_1^{-1/2}$ with fixed tolerance ($\varepsilon_{out} = 10^{-9}$), and relaxed inner tolerance.

to provide rough estimates of a matrix function 2-norm, more accurate approximations require a careful implementation of available approaches, since neither $f(A)$ nor products of the type $f(A)\mathbf{v}$ are available exactly. In particular, we showed that the Lanczos bidiagonalization yields a non-Hermitian perturbation of the original Hermitian matrix, and the recurrence needs to be revisited. Our numerical experiments showed that the computational complexity may vary significantly depending on the requested final accuracy, since the two inner iterations for the approximation of $f(A)\mathbf{v}$ and $f(A)^*\mathbf{u}$ may be very time and memory consuming. We showed that the relaxed strategy alleviates this problem whenever accurate approximations are required. However, for particular selections of matrices and functions, the approximation of $f(A)\mathbf{v}$ can still be very expensive, and some other strategies could be exploited, such as restarting; see, e.g., [10],[11],[17] and references therein. Finally, our approach could be used to estimate the norm of other matrix objects, such as the geometric mean [3], or the *derivatives* of matrix functions, such as the Fréchet derivative of the matrix exponential or of other functions [18].

REFERENCES

[1] B. J. C. Baxter, *Norm estimates for inverses of Toeplitz distance matrices*, J. Approximation theory, 79 (1994), pp. 222–242.

[2] M. Benzi, P. Boito, and N. Razouk, *Decay properties of spectral projectors with applications to electronic structure*, SIAM Review, 55 (2013), pp. 3–64.

[3] R. Bhatia and P. Grover, *Norm inequalities related to the matrix geometric mean*, Lin. Alg. Appl., 437 (2012), pp. 726–733.

[4] A. Bouras and V. Frayssé, *Inexact matrix-vector products in Krylov methods for solving linear systems: A relaxation strategy*, SIAM J. Matrix Analysis and Appl., 26 (2005), pp. 660 – 678.

[5] D. Choi, *Estimating Norms of Matrix Functions using Numerical Ranges*, PhD thesis, University of Washington, 2013.

[6] M. Crouzeix, *Numerical range and functional calculus in Hilbert space*, J. Functional Analysis, 244 (2007), pp. 668–690.

[7] J. K. Cullum and R. A. Willoughby, *Lanczos Algorithms for Large Symmetric Eigenvalue Computations: Vol. 1: Theory*, vol. 41, SIAM, 2002.

[8] V. Druskin and L. Knizhnerman, *Extended Krylov subspaces: approximation of the matrix square root and related functions*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 755–771.

[9] V. Druskin, L. Knizhnerman, and M. Zaslavsky, *Solution of large scale evolutionary prob-*

lems using rational krylov subspaces with optimized shifts, Tech. Rep. 5, 2009.
[10] M. Eiermann and O. Ernst, *A restarted Krylov subspace method for the evaluation of matrix functions*, SIAM J. Numer. Anal., 44 (2006), pp. 2481–2504.
[11] A. Frommer, S. Güttel, and M. Schweitzer, *Convergence of restarted Krylov subspace methods for Stieltjes functions of matrices*, SIAM J. Matrix Anal. Appl, (2014).
[12] A. Frommer and V. Simoncini, *Matrix functions*, tech. rep., Dipartimento di Matematica, Bologna, I, March 2006. To appear on SIAM J. Scient. Computing.
[13] C. W. Gear, *Numerical solution of ordinary differential equations: Is there anything left to do?*, SIAM Rev., 23 (1981), pp. 10–24.
[14] M. Gil', *Perturbation of functions of diagonalizable matrices*, Electronic J. of Linear Algebra, 20 (2010), pp. 303–313.
[15] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The John Hopkins University Press, Baltimore, London, 3rd ed., 1996.
[16] K. Gustafson and D. K. M. Rao, *Numerical Range: The Field of Values of Linear Operators and Matrices*, Universitext, Springer, New York, NY, USA, 1997.
[17] S. Güttel, *Rational Krylov approximation of matrix functions: Numerical methods and optimal pole selection*, GAMM Mitteilungen, 36 (2013), pp. 8–31.
[18] N. Higham and S. Relton, *Estimating the condition number of the frechet derivative of a matrix function*, SIAM J. Sci. Comput., (2014).
[19] N. J. Higham, *Matrix Functions – Theory and Applications*, SIAM, Philadelphia, USA, 2008.
[20] M. Hochbruck and C. Lubich, *Exponential integrators for quantum-classical molecular dynamics*, BIT, Numerical Mathematics, 39 (1999), pp. 620–645.
[21] M. Hochbruck and A. Ostermann, *Exponential integrators*, Acta Numerica, 19 (2010), pp. 209–286.
[22] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, 1991.
[23] ———, *Matrix Analysis*, Cambridge University Press, Cambridge, II ed., 2013.
[24] E. Jarlebring and E. H. Rubensson, *On the condition number and perturbation of matrix functions for hermitian matrices.* arXiv:1206.1762v1, June 2012.
[25] L. Knizhnerman and V. Simoncini, *A new investigation of the extended Krylov subspace method for matrix function evaluations*, Numerical Linear Algebra with Applications, 17 (2010), pp. 615–638.
[26] R. M. Larsen, *Lanczos bidiagonalization with partial reorthogonalization*, Tech. Rep. DAIMI PB-357, Department of Computer Science, Aarhus University, 1998.
[27] Matrix Market, *A visual repository of test data for use in comparative studies of algorithms for numerical linear algebra, Mathematical and Computational Sciences Division, National Institute of Standards and Technology.* Online at `http://math.nist.gov/MatrixMarket`.
[28] E. Mengi and M. Overton, *Algorithms for the computation of the pseudospectral radius and the numerical radius of a matrix*, IMA J. Num. An., 25 (2005), pp. 648–669.
[29] E. H. Rubensson, *Controlling errors in recursive Fermi-Dirac operator expansions with applications in electronic structure theory*, SIAM J. Sci. Comput., 34 (2012), pp. B1–B23.
[30] B. Schmitt, *Norm bounds for rational matrix functions*, Numerische Mathematik, 42 (1983), pp. 379–389.
[31] V. Simoncini, *Variable accuracy of matrix-vector products in projection methods for eigencomputation*, SIAM J. Numer. Anal., 43 (2005), pp. 1155–1174.
[32] V. Simoncini and D. B. Szyld, *Theory of inexact Krylov subspace methods and applications to scientific computing*, SIAM J. Sci. Comput., 25 (2003), pp. 454–477.
[33] G. W. Stewart and J.-G. Sun, *Matrix perturbation theory*, Computer Science and Scientific Computing, Academic Press Inc., Boston, MA, (1990).
[34] L. Trefethen and M. Embree, *Spectra and pseudospectra. The behavior of non-normal matrices and operators*, Princeton University Press, 2005.
[35] J. van den Eshof and M. Hochbruck, *Preconditioning Lanczos approximations to the matrix exponential*, SIAM J. Sci. Comput., 25 (2006), pp. 1438–1457.
[36] G. A. Watson, *Computing the numerical radius*, Lin. Alg. Appl., 234 (1996), pp. 163–172.

Appendix A. Proof of Proposition 6.1. Define the submatrix of $\widetilde{\mathcal{B}}_{2m}$ of size $2k$ as

$$\widetilde{\mathcal{B}}_{2k} = \begin{bmatrix} 0 & M_k \\ T_k & 0 \end{bmatrix}, \qquad \text{i.e.} \qquad \widetilde{\mathcal{B}}_{2m} = \begin{bmatrix} 0 & 0 & M_k & M_\star \\ 0 & 0 & 0 & \star \\ T_k & T_\star & 0 & 0 \\ t_{k+1,k}\mathbf{e}_1\mathbf{e}_k^* & \star & 0 & 0 \end{bmatrix}.$$

Define the vector $\tilde{\mathbf{q}} = \frac{1}{\sqrt{2}}[\mathbf{x}; 0; \mathbf{y}; 0]$, where the 0-vectors have length $m-k$. Let $\mathcal{X} = [\tilde{\mathbf{q}}, Y]$ be such that $\mathcal{X}$ is unitary, where $Y = [Y_1; Y_2; Y_3; Y_4]$. This implies that $\frac{1}{\sqrt{2}}(Y_1^*\mathbf{x} + Y_3^*\mathbf{y}) = 0$, $Y_4Y_4^* = I = Y_2Y_2^*$ and $Y_2Y_4^* = Y_4Y_2^* = 0$. Now, write

$$\mathcal{X}^*\widetilde{\mathcal{B}}_{2m}\mathcal{X} = \begin{bmatrix} \tilde{\mathbf{q}}^*\widetilde{\mathcal{B}}_{2m}\tilde{\mathbf{q}} & \tilde{\mathbf{q}}^*\widetilde{\mathcal{B}}_{2m}Y \\ Y^*\widetilde{\mathcal{B}}_{2m}\tilde{\mathbf{q}} & Y^*\widetilde{\mathcal{B}}_{2m}Y \end{bmatrix} = \begin{bmatrix} \theta^{(2k)} & \mathbf{g}_1^* \\ \mathbf{g}_2 & \underline{\widetilde{\mathcal{B}}_{2m}} \end{bmatrix}.$$

Here

$$\|\mathbf{g}_2\| = \|Y^*\widetilde{\mathcal{B}}_{2m}\tilde{\mathbf{q}}\| = \|\frac{1}{\sqrt{2}}Y^*\begin{bmatrix} M_k\mathbf{y} \\ 0 \\ T_k\mathbf{x} \\ t_{k+1,k}\mathbf{e}_1\mathbf{e}_k^*\mathbf{x} \end{bmatrix}\| = \|\frac{1}{\sqrt{2}}Y_4^*t_{k+1,k}\mathbf{e}_1\mathbf{e}_k^*\mathbf{x}\| = \|\mathbf{r}_{2k}\|.$$

Further, since $\mathbf{s}_{2m}^*Y = \tilde{\mathbf{q}}^*\widetilde{\mathcal{B}}_{2m}Y - \theta^{(2k)}\tilde{\mathbf{q}}^*Y = \tilde{\mathbf{q}}^*\widetilde{\mathcal{B}}_{2m}Y$, we have

$$\|\mathbf{g}_1\| = \|\tilde{\mathbf{q}}^*\widetilde{\mathcal{B}}_{2m}Y\| = \|\mathbf{s}_{2m}^*Y\| \le \|\mathbf{s}_{2m}\|.$$

Now, by [33, Theorem 2.1, p.230],

$$\text{if } \frac{\|\mathbf{r}_{2k}\|\,\|\mathbf{s}_{2m}\|}{\delta_{2m,2k}^2} < \frac{1}{4}, \quad \text{i.e.,} \quad \|\mathbf{r}_{2k}\| < \frac{\delta_{2m,2k}^2}{4\|\mathbf{s}_{2m}\|},$$

then there exists a vector $\mathbf{p} \in \mathcal{C}^{2m-1}$ satisfying $\tau = \|\mathbf{p}\| < 2\frac{\|\mathbf{r}_{2k}\|}{\delta_{2m,2k}}$, such that the unit norm vector

$$\mathbf{q} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \frac{1}{\sqrt{1+\|\mathbf{p}\|^2}}\left(\frac{1}{\sqrt{2}}\begin{bmatrix} \mathbf{x} \\ 0 \\ \mathbf{y} \\ 0 \end{bmatrix} + \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix}\mathbf{p}\right)$$

is an eigenvector of $\widetilde{\mathcal{B}}_{2m}$. Moreover,

$$\left\|\begin{bmatrix} \mathbf{x}_2 \\ \mathbf{y}_2 \end{bmatrix}\right\| = \left\|\frac{1}{\sqrt{1+\tau^2}}\begin{bmatrix} Y_2 \\ Y_4 \end{bmatrix}\mathbf{p}\right\| \le \frac{\tau}{\sqrt{1+\tau^2}}.$$

Further, this same theorem states that $|\theta - \theta^{(2k)}| = \|\mathbf{g}_1^*\mathbf{p}\| \le \|\mathbf{g}_1\|\|\mathbf{p}\| \le \|\mathbf{s}_{2m}\|\tau$. □