VARIABLE ACCURACY OF MATRIX-VECTOR PRODUCTS
IN PROJECTION METHODS FOR EIGENCOMPUTATION

V. SIMONCINI

Abstract. We analyze the behavior of projection-type schemes, such as the Arnoldi and Lanczos
methods, for the approximation of a few eigenvalues and eigenvectors of a matrix $A$, when $A$ cannot
be applied exactly but only with a possibly large perturbation. This occurs for instance in shift-
and-invert procedures or when dealing with large generalized eigenvalue problems. We theoretically
show that the accuracy with which $A$ is applied at each iteration can be relaxed, as convergence to
specific eigenpairs takes place. We show that the size of the perturbation is allowed to be inversely
proportional to the current residual norm, in a way that also depends on the sensitivity of the matrix
$A$. This result provides a complete understanding of reported experimental evidence in the recent
literature. Moreover, we adapt our theoretical results to devise a practical relaxation criterion to
achieve convergence of the inexact procedure. Numerical experiments validate our analysis.

1. Introduction. We are interested in the behavior of projection-type proce-
dures, such as the Arnoldi or Lanczos methods (see e.g. [18, 2]), for the approximation
of a few of the eigenvalues and corresponding eigenvectors in the eigenvalue problem

$$Ax = \lambda x, \quad \|x\| = 1,$$

where $A$ is an $n \times n$ non-Hermitian matrix and $\| \cdot \|$ is the Euclidean norm. We focus
on the case in which $A$ cannot be applied exactly, but only with a perturbation, that
is, at each iteration the operation $y = Av$ is replaced by

$$y = Av + f,$$

where $f$ can change at each iteration and $\| f \|$ can be monitored. In general, we
expect $\| f \|$ to be much larger than machine precision, so that the standard techniques
of round-off error analysis are not appropriate. This is indeed the case when, for
instance, shift-and-invert procedures are used to find interior eigenvalues of the given
matrix; or when a generalized eigenvalue problem is considered. On large problems,
both these procedures require the (approximate) solution of a linear system to apply
the matrix $A$. Finally, inaccurate products occur when the matrix itself is an operator
that needs to be estimated each time it is applied. As an alternative to a fixed
perturbation tolerance, methods based on shift-and-invert power iterations have for
a long time focused on increasing the accuracy as convergence was taking place, see
e.g. [9, 26], and [2, sec. 11.2] and references therein; see also [27] for a recent analysis
of perturbed power iterations. On the other hand, in the context of projection-
type methods, more recently the case of a decreasing accuracy has been considered
[5, 21, 7, 15]. We notice that an analogous problem has received considerable attention
in the linear system setting, see e.g. [4, 6, 30, 22, 31].

An approximate matrix-vector multiplication significantly perturbs the method,
but in a way that is apparently far less dramatic than the norm of the perturbation
would suggest. A large number of experiments in [5] showed that in many cases con-
vergence towards eigenpairs of $A$ can be achieved despite the fact that $\| f \|$ is allowed
to grow as the iteration progresses. In other words, it was shown in [5] that the ac-
curacy with which $A$ is applied at each iteration can be relaxed, as convergence takes

---


1Dipartimento di Matematica, Università di Bologna, Piazza di Porta S. Donato, 5, I-40127
Bologna, Italy; IMATI-CNR, Pavia and CIRSA, Ravenna, Italy (valeria@dm.unibo.it).
place. It was argued that the size of the perturbation can be related to the inverse of the residual norm of the current eigenvalue approximation. However, these arguments were not supported by theoretical justifications in [5]; the not always consistent behavior of the methods under the analyzed perturbations did not allow the authors to make conclusive statements as of the reliability of using variable accuracy. We aim to fill this gap in this paper.

The aim of this paper is twofold. First, we provide a theoretical understanding of the experimental evidence reported in [5], together with an analysis of the spectral properties that may influence the inexact process. Then, we adapt our theoretical results to devise a practical relaxation criterion to achieve convergence of the inexact procedure. Assuming that the unperturbed process converges, given an approximate eigenpair \((\theta, z)\) obtained by the perturbed process, we will show that the deviation of the computable residual from the true (unobservable) residual \(Az - \theta z\), can be kept below a fixed small tolerance. This result will imply that the final attainable true residual will also fall below the required tolerance.

Our analysis of variable accuracy in the matrix-vector products includes both Ritz and Harmonic Ritz approximations obtained with the Arnoldi method, as well as Ritz spectral information computed by the Lanczos procedure. Although these methods usually provide different approximation quantities and they satisfy different optimality properties (see e.g. [2]), the analysis of their performance under matrix-vector perturbations can be unified within the presented framework.

In this paper we restrict our analysis to the case when the considered method is not restarted. In practice, projection-type schemes require possibly sophisticated restarting procedures such as implicit restarting (see [10]), to limit memory requirements while improving the current available approximation. The problem of handling inexactness when some form or restarting is included will be the topic of future research.

The key idea beyond the success of inexact processes is related to an intrinsic property of Krylov subspace methods. Approximations are generated as \(V_m u\) where the \(m\) columns of the matrix \(V_m\) span the Krylov subspace of dimension \(m\), and \(u\) is the approximate solution of the projected problem. The perturbations \(f\)'s affecting the matrix-vector operation in (1.2) at each iteration, may be collected as subsequent columns of a “perturbation matrix” \(F_m\). It turns out that the fundamental relation associated with the inexact process formally differs from the unperturbed one by an additional term \(F_m u\). Clearly, if one is able to show that the components of \(u\) have a decreasing pattern going towards zero, then the norm of the corresponding columns of \(F_m\) is allowed to grow, that is, larger perturbations are allowed in later iterations, while still yielding a small perturbation term \(F_m u\), in norm. As a result, the perturbed fundamental relation remains significantly close to the unperturbed one, and the approximation process does not appear to be influenced. Therefore, given a problem to be solved by means of projection onto a Krylov subspace, the inexact matrix-vector product in (1.2) can be conveniently exploited if one can show that the approximation vector \(u\) has a decreasing pattern. Intuitively, a decreasing pattern is associated with a “marginal approximation” property, as the Krylov subspace grows. However, a rigorous treatment of this step is far from obvious. As one may suspect, both the difficulty in proving the existence of this pattern, as well as the constraints under which this pattern does settle down, depend on the possible complexity and nonlinearity of the problem. In the linear system setting, the decreasing pattern of \(u\) (in this case the projected system approximate solution) was proved in [22, 30].
In this paper, we solve this problem within eigenvalue computation. We will show that the matrix sensitivity and the nonlinearity of the problem play a crucial role: the conditions under which the pattern arises are different and significantly more stringent than in the linear system setting. Moreover, the analysis becomes even more complex when more than one eigenpair, or more generally an invariant subspace, is sought. Finally, deep results from matrix perturbation theory need to be employed, making the approach and the conclusions of this paper significantly different with respect to what has been done in [22, 30] for linear systems.

In section 2 we show that some of the eigenvectors of the representation matrix of $A$ in the approximation subspace have a decreasing pattern. This key result will be used in section 3 to show that in the inexact Arnoldi method, the matrix-vector multiplication can be perturbed in a way that is inversely proportional to this pattern. A practical relaxation strategy for the Arnoldi method is proposed in section 3.1 and numerically tested in section 4. Our theoretical results are subsequently applied to related methods that are currently used as alternatives to general Arnoldi and that are based on the same key relations [2]. In particular, we will discuss the inexact Harmonic Ritz approximations in section 5, and the inexact Lanczos method in section 6. Finally, section 7 summarizes our results and discusses some related issues.

The following notation will be used throughout. For a vector $u$, $\bar{u}$ denotes its conjugate, $u^*$ its conjugate transpose, and $\|u\|$ its 2-norm. For a given $k \times k$ matrix $T$, an eigenpair $(\lambda, u)$ consists of $\lambda \in \mathbb{C}$ and $u \in \mathbb{C}^k$ such that $Tu = \lambda u$, $\|u\| = 1$. An eigentriple $(\lambda, u, v)$ of $T$ is such that $Tu = \lambda u$ and $v^*T = \lambda v^*$ such that $\|u\| = 1$ and $v^*u = 1$. Here $u$ indicates a right eigenvector and $v$ a left eigenvector. Moreover, $I_k$ denotes the identity matrix of size $k$ (the subscript is omitted if clear from the context), while $e_j$ is the $j$th column of the identity matrix of given dimension not smaller than $j$. Finally, $\text{Range}(X)$ is the space generated by the columns of $X$, while $\lambda(H)$ is the set of eigenvalues of a square matrix $H$. Exact precision arithmetic is assumed throughout the paper, and the term inexact refers to an inaccurate computation, whose error is significantly larger than the finite precision arithmetic unit. All experiments were run using Matlab [11].

2. Bounds for the eigenvector components of the Arnoldi matrix.

2.1. Notation. Starting with a unit norm vector $v_1$, the Arnoldi method builds a basis $V_m$ for the Krylov subspace $K_m(A, v_1) = \text{span}\{v_1, Av_1, \ldots, A^{m-1}v_1\}$ satisfying the following Arnoldi relation,

$$AV_m = V_mH_m + v_{m+1}h_{m+1,m}e_m^* = V_{m+1} \begin{bmatrix} H_m \\ h_{m+1,m}e_m^* \end{bmatrix}, \quad V_{m+1}^*V_{m+1} = I.$$  

(2.1)

Matrix $H_m$ is an $m \times m$ upper Hessenberg matrix, and it is the orthogonal projection and restriction of the matrix $A$ onto the Krylov subspace. An eigenpair $(\theta, u)$ of $H_m$ defines the Ritz value, $\theta$, and Ritz vector, $V_m u$, which may be used to approximate some of the eigenpairs of $A$ [2, 18]. The accuracy in the approximation is usually monitored by means of some relative quantity involving the residual $r_m = AV_m u - \theta V_m u$; we refer to [2] and its references for a detailed discussion on stopping criteria for eigenvalue solvers. It is also important to recall that for ill conditioned problems, small residuals do not necessarily imply small errors in the approximate eigenpair [8]; additional care should be taken in this case.

Given the eigenpair $(\theta, u)$ and using (2.1), the residual $r_m$ and its norm can be
cheaply computed as

\begin{equation}
  r_m = v_{m+1} h_{m+1,m} e_m^* u, \quad \|r_m\| = |h_{m+1,m}| |e_m^* u|.
\end{equation}

This relation emphasizes that for the residual to be small, at least one of $h_{m+1,m}$ or $|e_m^* u|$ have to be small. In the former case, the Krylov subspace is close to an invariant subspace. On the other hand, a small $|e_m^* u|$ indicates that the $m$th component of the eigenvector $u$ of $H_m$ is small. No other knowledge of the eigenvector components is commonly employed in the convergence test, although it can be experimentally observed that the absolute values of the components of $u$ tend to exhibit a decreasing pattern if $(\theta, V_m u)$ is a good approximation to an eigenpair of $A$; see e.g. [17, 19].

In Proposition 2.2 we will show that there exists a strong relation between the magnitude of the $k + 1$st component of $u$ and the residual of some Ritz pair after $k$ Arnoldi iterations with $k < m$. This relation can be derived as a consequence of a general approximation theorem for non-normal matrices, cf. e.g. [28]. Below we report the result with our notation. To this end, we introduce some definitions. Given an orthogonal basis $U$ for an invariant subspace of a matrix $H$, and given $Y$ so that $[U, Y]$ is unitary, Range$(U)$ is called a simple invariant subspace of $H$ if the spectra of $U^* H U$ and of $Y^* H Y$ do not intersect; cf. [28, Definition V.1.2]. Moreover, for two square matrices $L_1, L_2$ with disjoint spectra, the function $\text{sep}(L_1, L_2)$ is defined as (see e.g. [28, p. 231])

\[ \text{sep}(L_1, L_2) := \inf_{\|P\|=1} \| PL_1 - L_2 P \|. \]

The definition holds for $\| \cdot \|$ being any consistent family of norms; in the following we shall use the 2-norm for vectors and the induced 2-norm for matrices. If $L_1$ is a scalar, then $\text{sep}(L_1, L_2) = \sigma_{\text{min}}(L_2 - L_1 I)$. An analogous relation holds whenever $L_2$ is a scalar.

**Theorem 2.1.** [28, Theorem V.2.1, p. 230] Let $X = [U, Y]$ be a unitary matrix and let

\[ X^* H_m X = \begin{bmatrix} L_1 & K \\ G & L_2 \end{bmatrix}. \]

Set $\gamma = \|G\|, \eta = \|K\|$. Assume that $L_1$ and $L_2$ have distinct spectra, so that $\delta := \text{sep}(L_1, L_2) > 0$. Then if $\gamma \eta / \delta^2 < \frac{1}{4}$, there is a unique matrix $P$ satisfying $\|P\| < 2 \gamma^2$ such that the columns of $\tilde{U} = (U + Y P)(I + P^* P)^{-\frac{1}{2}}$ form an orthonormal basis of a simple right invariant subspace of $H_m$. The representation of the matrix $H_m$ with respect to $\tilde{U}$ is given by $\tilde{L}_1 = (I + P^* P)^{-\frac{1}{2}} (L_1 + K P)(I + P^* P)^{-\frac{1}{2}}$.

**2.2. Spectral properties of the Arnoldi matrix.** We first use the result of Theorem 2.1 in the case of the approximation of one eigenpair, and then generalize it to the approximation of an invariant subspace.

Consider the principal submatrix of $H_m$ of size $k$, $H_k$, i.e.

\begin{equation}
  H_m = \begin{bmatrix} H_k \\ h_{k+1,k} e_1 e_k^* \end{bmatrix}, \quad H_k \in \mathbb{C}^{k \times k},
\end{equation}

and let $u^{(k)}$ be an eigenvector of $H_k$. Let $Y$ be a matrix such that the matrix

\[ X = \begin{bmatrix} u^{(k)} \\ 0 \end{bmatrix}, Y \in \mathbb{C}^{m \times m} \]
is unitary, where here and in the following, \( \mathbf{0} \) pads with zeros the bottom part of a vector with a total of \( m \) components. Then

\[
(2.4) \quad H_m := Y^* H_m Y \in \mathbb{C}^{(m-1) \times (m-1)},
\]

is the orthogonal projection and restriction of \( H_m \) onto the range of \( Y \), the space orthogonal to the space spanned by \( u^{(k)} \).

**Proposition 2.2.** Let \((\theta^{(k)}, u^{(k)})\) be an eigenpair of \( H_k \), \( r_k = e_{k+1} h_{k+1,k} e_k^* u^{(k)} \), \( \delta_{m,k} = \sigma_{\min}(H_m - \theta^{(k)} \mathbf{I}) > 0 \) with \( H_m \) defined in (2.4), and \( s^* = [(u^{(k)})^*, \Omega^*] H_m - \theta^{(k)} [(u^{(k)})^*, \Omega^*] \). If

\[
(2.5) \quad \|r_k\| < \frac{\delta_{m,k}^2}{4\|s_m\|},
\]

then there exists a unit norm eigenvector \( u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \) of \( H_m \) with \( u_1 \in \mathbb{C}^k \), such that

\[
(2.6) \quad \|u_2\| \leq \frac{\tau}{\sqrt{1 + \tau^2}}, \quad \text{with} \quad \tau \in \mathbb{R}, \quad 0 \leq \tau < 2\|r_k\| \frac{\delta_{m,k}}{\delta_{m,k}}.
\]

Moreover, if \( \theta \) is the eigenvalue associated with \( u \), we have

\[
(2.7) \quad |\theta - \theta^{(k)}| \leq \|s_m\| \tau.
\]

**Proof.** Let \( Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \) be such that the matrix \( X = \left[ \begin{bmatrix} u^{(k)} \\ \mathbf{0} \end{bmatrix}, Y \right] \in \mathbb{C}^{m \times m} \) is unitary. Note that this implies \( Y_1^* u^{(k)} = 0 \) and \( Y_2^* Y_2 = \mathbf{I} \). Using (2.3) gives

\[
X^* H_m X = \begin{bmatrix} \theta^{(k)} & K \\ G & H_m \end{bmatrix}, \quad G = Y_2^* h_{k+1,k} e_1 e_k^* u^{(k)}, \quad K = (u^{(k)})^* [H_k, H_s] Y.
\]

Since \( Y_2 \) has orthonormal rows,

\[
(2.8) \quad \gamma := \|G\| = \|Y_2^* h_{k+1,k} e_1 e_k^* u^{(k)}\| = |h_{k+1,k} e_k^* u^{(k)}| = \|r_k\|.
\]

Moreover, since \( [(u^{(k)})^*, \Omega^*] Y = 0 \), \( \|K\| = \|[(u^{(k)})^*, \Omega^*] H_m Y\| = \|s_m^* Y\| \leq \|s_m\| \). Using Theorem 2.1, if \( \frac{4\|s_m\|}{\delta_{m,k}} < \frac{1}{4} \), i.e. if (2.5) holds, then there exists a vector \( p \in \mathbb{C}^{m-1} \) satisfying

\[
\tau := \|p\| < \frac{\gamma}{2\|s_m\|}
\]

such that the unit norm vector

\[
u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \frac{1}{\sqrt{1 + \|p\|^2}} \left( \begin{bmatrix} u^{(k)} \\ \mathbf{0} \end{bmatrix} \mathbf{I} + \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} p \right)
\]

is an eigenvector of \( H_m \). To complete the proof, we notice that

\[
\|u_2\| = \frac{1}{\sqrt{1 + \|p\|^2}} \|Y_2 p\| \leq \frac{1}{\sqrt{1 + \|p\|^2}} \|p\|.
\]
The bound for \(|\theta - \theta^{(k)}|\) follows from the representation of \(H_m\) with respect to \(u\) in Theorem 2.1, that is \(|\theta - \theta^{(k)}| = \|Kp\| \leq \|K\| \|p\| \leq \|s_m\| \tau\). Proposition 2.2 shows that if the residual \(r_k\) is sufficiently small, then there exists an eigenvector \(u\) of \(H_m\) whose last \(m - k\) components can be bounded by a quantity involving \(\|r_k\|\) and thus they are also small. Each of the last \(m - k\) components of \(u\) is bounded by \(\tau/\sqrt{1 + \tau^2}\), hence

\[
(2.9) \quad |e_j^u| \leq \frac{\tau}{\sqrt{1 + \tau^2}} \leq \tau \leq 2\|r_k\|/\delta_{m,k}, \quad j = k + 1, \ldots, m.
\]

The bound (2.9) is most interesting for \(j = k + 1\). Indeed, Proposition 2.2 can be applied to other principal submatrices of \(H_m\), of size \(k_1\) larger than \(k\). In this case, and if a more accurate Ritz pair is computed with \(H_{k_1}\), the last \(m - k_1\) components of the corresponding eigenvector are smaller than those for \(H_k\). In practice, this implies that for \(k\) sufficiently large, so that \(\|r_k\|\) satisfies (2.5), the last \(m - k\) components of \(u\) have an almost monotonically decreasing pattern.

It is remarkable that in our setting, \(\|r_k\|\) has a double role. On the one hand, \(\|r_k\| = \|G\|\) (cf. (2.8)), which measures the accuracy of

\[
(2.10) \quad \begin{bmatrix} \theta^{(k)} \\ V^{(k)} \\ 0 \end{bmatrix}
\]

as an approximate eigenpair of \(H_m\) in connection with the general approximation theorem. On the other hand,

\[
r_k = v_{k+1}h_{k+1,k}e_k^u = AV_k u^{(k)} - \theta^{(k)}V_k u^{(k)}
\]

is the residual of \((\theta^{(k)}, V_k u^{(k)})\) as approximate eigenpair of \(A\). This double role is what makes our eigenvector component analysis possible.

In Proposition 2.2, \(s_m\) is the left residual vector of the approximate pair (2.10) of \(H_m\). For non-normal problems, \(\|s_m\|\) is bound not to be small, and in general the estimate \(\|s_m\| \approx \|A\|\) can be used.

Remark 2.3. The function \(\delta_{m,k}\) provides a condition number of the eigenvector problem [28, p. 241] and it reflects the proximity of \(\theta^{(k)}\) to \(H_m\), although it may be much smaller than the distance between the spectrum of \(H_m\) and \(\theta^{(k)}\) for non-normal matrices [28, Example 2.4, p. 234]. We notice that if (2.10) is a good approximation to the eigenpair \((\theta, u)\) of \(H_m\), then \(\delta_{m,k}\) is close to the norm of the reduced resolvent of \(H_m\). These comments will be used when we will derive a practical relaxation criterion for the matrix-vector product with \(A\).

Remark 2.4. Proposition 2.2 states the existence of an eigenvector \(u\) of \(H_m\) satisfying (2.6). On the other hand, we are interested in the characterization of the components of a specific eigenvector \(u\) of \(H_m\). To be able to correctly identify the eigenvector determined by Proposition 2.2 as the analyzed vector \(u\), we will need some further hypotheses. In particular, we will require that after \(k\) iterations, \(\theta^{(k)}\) and \(u^{(k)}\) provide sufficiently good approximations to an eigenpair of \(H_m\) and of \(A\). Unfortunately this apparently restrictive condition is observed to be required in practice. In particular, eigenvector components do not exhibit a decreasing pattern until the Ritz pair reaches its final, possibly superlinear, asymptotic convergence rate; see [3] for a discussion on the occurrence of superlinear convergence rate.

We next generalize our result to the case of an invariant subspace.
**Proposition 2.5.** Let the columns of $U^{(k)}$ be an orthonormal basis for a simple invariant subspace of $H_k$ of dimension $\ell$, with representation matrix $L^{(k)} = (U^{(k)})^* H_k U^{(k)}$. Let $\delta_{m,k} = \text{sep}(L^{(k)}, L_2) > 0$, $R_k = v_{k+1} h_{k+1,k} e_k U^{(k)}$ and $S_m = [(U^{(k)})^*, O^*] H_m - L^{(k)} [(U^{(k)})^*, O^*]$, where $O$ is the $(m-k) \times \ell$ zero matrix. If 
\[
\|R_k\| < \frac{\delta_{m,k}}{4\|S_m\|},
\]
then there exists a matrix $U = \begin{bmatrix} U_1 & U_2 \end{bmatrix}$, $U^* U = I$, with $U_1 \in \mathbb{C}^{k \times \ell}$ and whose columns span a simple invariant subspace of $H_m$, such that
\[
\|U_2\| \leq \frac{\tau}{\sqrt{1 + \tau^2}}, \quad \text{with} \quad \tau \in \mathbb{R}, \quad 0 \leq \tau < 2\frac{\|R_k\|}{\delta_{k,m}}. \tag{2.11}
\]

**Proof.** The proof follows the lines of that of Proposition 2.2. \qed

The setting of Proposition 2.5 is most appropriate whenever the sought after eigenvalues form a cluster separated from the rest of the spectrum. Indeed, $\|R_k\|$ is required to be less than $\delta_{m,k}/\|S_m\|$, and $\delta_{m,k}$ may be very small if the whole cluster is not sufficiently well captured by $\text{Range}(U^{(k)})$. The result of Proposition 2.5 is particularly helpful when dealing with close to defective eigenvalues, whose eigenvector bases can be very ill conditioned. The use of invariant subspaces considerably simplifies the analysis [28].

We would also like to comment on the case when $H_m$, and thus $H_k$, is tridiagonal. Whenever $H_m$ is tridiagonal, e.g. for Hermitian problems and in the non-Hermitian Lanczos process, the relation between the $k+1$st component of an eigenvector of $H_m$ and the residual at step $k$ can be derived directly [20]. However, in the context of inexact matrix-vector multiplication, the resulting matrix $H_m$ is not tridiagonal but upper Hessenberg, even in the inexact Lanczos recurrence (cf. section 6), therefore the more general result is required.

As an alternative to Proposition 2.2, Jasper van den Eshof [29] noticed that a related result could be obtained as follows. Let $(\theta^{(k)}, u^{(k)})$ be as before, with associated residual $r_k$, and let $\tilde{u}^{(k)} = [u^{(k)}; 0]$. If $H_m = QTQ^*$ is the Schur decomposition of $H_m$, with $Q = [u, Q_2]$ unitary and
\[
T = \begin{bmatrix} \theta & a^* \\ 0 & T_{22} \end{bmatrix},
\]
then we have (cf. [2, formula (7.107), p.230])
\[
\sin \theta(\tilde{u}^{(k)}, u) \leq \frac{\|r_k\|}{\sigma_{\min}(T_{22} - \theta^{(k)} I)}. \tag{2.12}
\]
Therefore, using $u = [u_1; u_2]$ we can write $u = \tilde{u}^{(k)} (\tilde{u}^{(k)})^* u + (I - \tilde{u}^{(k)} (\tilde{u}^{(k)})^*) u$ so that $u_2 = [0, I] (I - \tilde{u}^{(k)} (\tilde{u}^{(k)})^*) u$ from which we obtain
\[
\|u_2\| \leq \| (I - \tilde{u}^{(k)} (\tilde{u}^{(k)})^*) u \| = \sin \theta(\tilde{u}^{(k)}, u) \leq \frac{\|r_k\|}{\sigma_{\min}(T_{22} - \theta^{(k)} I)}. \tag{2.13}
\]
The relation is very similar to that in (2.6), however, $\sigma_{\min}(T_{22} - \theta^{(k)} I)$ is not the same quantity as $\delta_{m,k}$, as in the former the matrix $T_{22}$ is completely specified by the spectral properties of the target matrix $H_m$, and does not take into account the approximate eigenvector. Nonetheless, it is not clear whether the bound in (2.6), which also includes a condition on the residual, is sharper than the one above. Other characterizations using perturbation theorems can be found e.g. in [28, sec. V.2.2].

We would also like to stress that, as shown in Proposition 2.5, the result we have used can be naturally generalized to the case of a simple invariant subspace.
3. Inexact Arnoldi method. When $A$ is not applied exactly, at each iteration the operation $y = Av$ is replaced by \((1.2)\). Letting $F_m = [f_1, \ldots, f_m]$ be the matrix whose columns collect all perturbations, we obtain the following inexact (perturbed) Arnoldi relation,

\[
AV_m + F_m = V_mH_m + h_{m+1,m}v_{m+1}e^*_m. \tag{3.1}
\]

As in the exact case, matrix $V_m$ has orthonormal columns, however, the space generated by its columns is no longer a Krylov subspace associated with $A$. Moreover, both $V_m$ and the upper Hessenberg matrix $H_m$ are different from those one would obtain with the exact process, and $F_m$ measures the inexactness of the perturbed Arnoldi relation. When looking for specific eigenpairs, large perturbations may be allowed and still maintain the convergence of the eigenvalue residual to a low final accuracy. Let $(\theta, u)$ be an eigenpair of $H_m$. We have

\[
AV_m u - \theta V_m u = V_mH_m u - \theta V_m u + h_{m+1,m}v_{m+1}e^*_m u - F_m u = h_{m+1,m}v_{m+1}e^*_m u - F_m u.
\]

We call the quantity $AV_m u - \theta V_m u$ the \textit{true residual}, whereas we call the vector $r_m = h_{m+1,m}v_{m+1}e^*_m u$ the \textit{computed residual}, which can be monitored during the recurrence. Note that the true residual cannot be computed when $A$ is not applied exactly. In particular, in the inexact case, the true and computed residuals differ by the vector $F_m u$. We have

\[
\|(AV_m u - \theta V_m u) - r_m\| = \|F_m u\| = |||f_1, \ldots, f_m||u||
\]

\[
= \|\sum_{k=1}^m f_k(e^*_k u)\| \leq \sum_{k=1}^m ||f_k|| \|e^*_k u\|.
\]

The distance between the true and the computed residuals is small when each addend $\|f_k\| \|e^*_k u\|$, $k = 1, \ldots, m$ in the last sum is small. This occurs when either of the two terms $\|f_k\|$ and $\|e^*_k u\|$ is small, and not necessarily both, as long as the other term remains bounded by some $O(1)$, say, constant. Therefore, if $\|e^*_k u\|$ is small, $\|f_k\|$ is allowed to be large and still provide a small gap between true and computed residuals.

We next make this statement more precise.

Assume that a maximum of $m$ iterations of inexact Arnoldi can be carried out, and let $(\theta, V_m u)$ be the best Ritz approximation with $H_m$ to the sought after eigenpair of $A$. Relaxation in the matrix-vector product at step $k < m$ is possible if there exists an eigenpair of $H_{k-1}$, denoted by $(\theta^{(k-1)}, u^{(k-1)})$, that is sufficiently close to the eigenpair $(\theta, u)$ of $H_m$ or, to an eigenpair of $A$. More precisely, using the notation introduced in Proposition 2.2, for relaxation to take place at step $k$, it must hold

\[
\|r_{k-1}\| < \frac{\delta_{m,k-1}}{4\|s_m\|}, \tag{3.3}
\]

\[
\forall \theta_j \in \Lambda(H_m), \theta_j \neq \theta, \quad |\theta^{(k-1)} - \theta_j| > \frac{2\|s_m\| \|r_{k-1}\|}{\delta_{m,k-1}}. \tag{3.4}
\]

Condition (3.4) ensures that the eigenpair $(\theta^{(k-1)}, u^{(k-1)})$ in Proposition 2.2, is indeed a perturbation of the analyzed eigenpair $(\theta, u)$ of $H_m$; cf. Remark 2.4. It requires that $\theta^{(k-1)}$ is closer to $\theta$ than to other eigenvalues of $H_m$. This closeness is measured in terms of the quantities $\|r_{k-1}\|$ and $\delta_{m,k-1}$, associated with $\theta^{(k-1)}$. If both (3.3)
and (3.4) hold, then \((\theta, u)\) is the only eigenpair of \(H_m\) such that (2.6) and (2.7) hold at iteration \(k - 1\). When 
\[
\theta^{(k-1)} \approx \theta, \quad \begin{bmatrix} u^{(k-1)} \\ 0 \end{bmatrix} \approx u,
\]
we can write 
\[
\delta_{m,k-1} \leq \min_{\theta_i \in \Lambda(H_m)} |\theta_i - \theta^{(k-1)}| \approx \min_{\theta_j \in \Lambda(H_m) \setminus \{\theta\}} |\theta_j - \theta^{(k-1)}|,
\]
in which case, (3.4) follows from (3.3). For simplicity, here and below we assume that \(\theta\) is a simple eigenvalue of \(H_m\). A generalization of the result of Theorem 3.1 can be obtained by using Proposition 2.5.

**Theorem 3.1.** Assume \(m\) inexact Arnoldi iterations have been carried out and let \((\theta, u)\) be an eigenpair of \(H_m\) with \(\theta\) simple Ritz value and \(\|u\| = 1\). Given any \(\varepsilon \in \mathbb{R}, \varepsilon > 0\), assume that for \(k = 1, \ldots, m\),
\[
\begin{cases}
\frac{\delta_{m,k-1}}{2m\|r_{k-1}\|} \|e^*_k u\|, & \text{if } k > 1 \text{ and there exists } (\theta^{(k-1)}, u^{(k-1)}) \text{ of } H_{k-1} \\
\frac{1}{m}, & \text{otherwise.}
\end{cases}
\]
(3.5) \(\|f_k\| \leq \frac{\delta_{m,k-1}}{2m\|r_{k-1}\|} \|e^*_k u\|, \|e^*_k u\| \leq 1\)

Then 
\[
\| (AV_m u - \theta V_m u) - r_m \| \leq \varepsilon.
\]

*Proof.* If at step \(k - 1\) there exists an eigenpair \((\theta^{(k-1)}, u^{(k-1)})\) of \(H_{k-1}\) satisfying (3.3), (3.4), then \(\theta\) is the only eigenvalue of \(H_m\) such that 
\[
|\theta - \theta^{(k-1)}| \leq \frac{2\|s_m\| \|r_{k-1}\|}{\delta_{m,k-1}}.
\]
Hence, Proposition 2.2 ensures that \(\theta^{(k-1)}\) is a perturbation of the considered eigenvalue \(\theta\) of \(H_m\).

Let \(K\) be the subset of \(\{2, \ldots, m\}\) such that for each \(k \in K\) there exists an eigenpair \((\theta^{(k-1)}, u^{(k-1)})\) of \(H_{k-1}\) satisfying (3.3) and (3.4). Then, using (2.9),
\[
\| (AV_m u - \theta V_m u) - r_m \| = \| F_m u \| \leq \sum_{k=1}^{m} \| f_k \| \| e^*_k u \| \\
\leq \sum_{k \in K} \| f_k \| \| e^*_k u \| + \sum_{k \notin K, k \leq m} \| f_k \| \| e^*_k u \| \\
\leq \sum_{k \in K} \frac{\delta_{m,k-1}}{2m\|r_{k-1}\|} \| e^*_k u \| + \sum_{k \notin K, k \leq m} \frac{1}{m} \| e^*_k u \| \\
\leq \frac{\|s_m\| \|r_{k-1}\|}{\delta_{m,k-1}} + \sum_{k \notin K, k \leq m} \frac{1}{m} \| e^*_k u \| \\
= \frac{|K|}{m} \varepsilon + \frac{m - |K|}{m} \varepsilon = \varepsilon.
\]

If the conditions of Theorem 3.1 hold, then the difference between the two residuals is less than some fixed value \(\varepsilon\). For a sufficiently good starting vector, the norm
of the computed residual tends to zero as \( m \) goes to infinity, therefore, \( \varepsilon \) also provides a bound for the final attainable accuracy of the true residual. For this reason, it is natural to relate the value of \( \varepsilon \) to the threshold in the eigenvalue problem stopping criterion.

Note that \( \delta_{m,k-1} \) may dramatically influence the size of the perturbation. For a sensitive matrix \( H_m \), \( \delta_{m,k-1} \) may be very small and thus force high accuracy in the matrix-vector product to maintain convergence. On the other hand, it is also important to realize that \( \delta_{m,k-1} \) is related to the sensitivity of \( H_m \), and not of the original matrix \( A \). Since \( H_m \) is a projection of \( A \) onto a possibly much smaller space, in general we expect \( H_m \) to be less sensitive to perturbations than \( A \).

### 3.1. A practical relaxation strategy.

The result of Theorem 3.1 suggests that we could derive a practical criterion for relaxing the accuracy with which \( A \) is applied at each iteration. Unfortunately, the criterion in (3.5) requires crucial information that is not available at iteration \( k \), namely \( m(k)_{1} = \min(H_{m} - \theta^{(k-1)} I) \).

This quantity emphasizes the sensitivity of the eigenproblem with \( H_{m} \), and cannot cheaply be replaced. We therefore suggest a relaxation strategy that mimics (3.5), while sacrificing some accuracy in \( m(k)_{1} \).

Assume that a maximum of \( m \) inexact iterations are to be carried out. At the first iteration, we require that \( \|f_{1}\| \) be less than or equal to \( \frac{1}{m} \varepsilon \). At iteration \( k > 1 \) we require that the perturbation satisfies

\[
\|f_{k}\| \leq \frac{\min\{\alpha, \delta_{(k-1)}\}}{2m \|r_{k-1}\|} \varepsilon, \quad \delta_{(k-1)} := \min_{\partial_{j} \in \Lambda(H_{k-1}) \setminus \{\theta^{(k-1)}\}} |\theta^{(k-1)} - \theta_{j}|,
\]

where \( \alpha \) is an estimate of \( \|A\| \), and it is included to make the condition invariant\(^1\) with respect to a scaling of \( A \). The quantity \( \delta_{(k-1)} \) is related to the distance of the Ritz value \( \theta^{(k-1)} \) from the rest of the spectrum of \( H_{k-1} \). Clearly, \( \delta_{(k-1)} \) may in general be very different from \( m(k)_{1} \) (cf. also Remark 2.3). On the other hand, \( m(k)_{1} \) will not be too overestimated when \( \theta^{(k-1)} \) and its nearby eigenvalues have stabilized to the corresponding eigenvalues of a matrix \( H_{m} \) that is not very sensitive to perturbations.

We should add that, in our numerical experiments, we assume that conditions (3.3) and (3.4) are always satisfied. We only require the residual to be less than one, otherwise the unit value is used instead of the residual in the test.

In (3.6), \( \varepsilon \) is some fixed constant, naturally related to the final stopping tolerance for the eigenvalue computation. In the numerical experiments of section 4, we require that the final residual be less than \( \varepsilon \) in norm. In general, \( \varepsilon \) may include some information about the eigenproblem, one could for instance let \( \varepsilon \) depend on the current approximation, that is, \( \varepsilon_{k-1} = |\theta^{(k-1)}| \varepsilon \), with \( \varepsilon \) fixed. This choice of \( \varepsilon_{k-1} \) is associated with the following stopping criterion for the eigenvalue solver,

\[
\frac{\|r_{k-1}\|}{|\theta^{(k-1)}|} \leq \varepsilon,
\]

which is commonly employed in practical implementations; see e.g. Example 4.2.

**Remark 3.2.** The proposed dynamic perturbation criterion is tailored to the sought after invariant subspace of \( A \). Assuming that convergence is achieved in the unperturbed case, our analysis shows that specific Ritz pairs obtained by the inexact procedure can still converge to the wanted eigenpairs of \( A \), up to a certain tolerance.

\(^1\)We thank J. Langou for pointing this out.
Other Ritz pairs may be significantly perturbed, if in the exact scheme they approximate eigenpairs of $A$ with a slower convergence rate. In Example 4.1 we report an experiment where eigenvalues do converge at different rates, and the inexact method delivers significantly different results from the exact process (cf. Table 4.1). Our theory formalizes a similar consideration stated in [22, sec. 11].

4. Numerical experiments. In this section we report on some numerical experiments that support our theoretical results for the inexact Arnoldi method using Ritz pairs.

Extensive computational experiments were carried out in [5], where the family of relaxation criteria

$$\|f_k\| \leq \frac{10^{-\alpha_0}}{\|f_{k-1}\|} \varepsilon$$

was introduced, with $\alpha_0 = 0, 1, 2$, while $\varepsilon = \|A\| \varepsilon$, with $\epsilon$ equal to the required final residual accuracy. The authors reported the number of times the inexact procedure successfully achieved the required accuracy in approximating the selected eigenpairs of a wide range of matrices. They showed that in 42% of the tests the criterion (4.1) was fulfilled for $\alpha_0 = 0$, up to 81% for $\alpha_0 = 2$. In the following we shall report experiments for $\alpha_0 = 0$. We remark that our theory provides an understanding of the role of the numerator $10^{-\alpha_0}$, and it explains the reason why for $\alpha_0 = 0$ the effect of the perturbation may be severely underestimated.

![Fig. 4.1. Example 4.1. Left: spectrum of $A$. Right: Convergence curves of the exact method (dash-dotted), inexact Arnoldi with (3.6) (solid) and inexact Arnoldi with (4.1) (dashed). The increasing curves report the values of the perturbation norms, $\|f_k\|$, in (3.6) and (4.1) (labeled BF).](image)

**Example 4.1.** We consider the $900 \times 900$ matrix stemming from the centered finite difference discretization of the operator

$$Lu = -\Delta u + 100((x + y)u)_x + 100((x + y)u)_y,$$

on the unit square, and we seek the eigenvalue with largest real part, $\lambda \approx 7.5127$; cf. Figure 4.1(left). In Figure 4.1(right) we report the convergence history for the exact Arnoldi method for $m = 130$ (blue dash-dotted line), the inexact method with
the flexible accuracy criterion in (3.6) with $\alpha = 10$ (black solid line) and the flexible accuracy criterion adopted by Bouras and Fraysse (dashed line); cf. (4.1). The two increasing curves report the values of $\| f_k \|$ for the two inexact methods, with $\varepsilon = 10^{-8}$. The starting vector for the iterative process was taken to be the normalized vector of all ones. Inexactness of $A$ was simulated by adding a random perturbation vector $f_k$, whose norm was equal to the right-hand side of (3.5) and (4.1) (with $\alpha_0 = 0$), for our criterion and for that of Bouras and Fraysse, respectively.

Both inexact approaches replicate the exact convergence curve until they reach their final attainable accuracy. Note that the original Bouras and Fraysse criterion does not allow the method to fall below the required final residual accuracy, while this is achieved by the criterion in (3.6).

In Table 4.1 we report the last 7 computed Ritz values after $m = 100$ iterations, with the exact Arnoldi method and with the inexact method, when either the new relaxation strategy or the strategy (4.1) are employed. The first column reports the corresponding eigenvalues of $A$. We notice that for the two inexact procedures, most Ritz values differ from those computed by the exact Arnoldi method, and only the first 3-4 digits remain unaltered. This is not the case for the sought after eigenvalue $\lambda \approx 7.5126$, for which the exact and inexact Ritz values coincide with several digits of accuracy. We can thus confirm that the perturbation does affect the convergence of Ritz values that do not converge with the same rate as those that guided the perturbation magnitude; see Remark 3.2.

We next report the results obtained when looking for a group of eigenvalues, namely the three largest eigenvalues of $A$ with both the exact and inexact methods. We considered a starting vector with random entries normally distributed (Matlab function \texttt{randn}, with initial state random number generator), since the previously chosen constant vector had small components onto the second largest eigenvalue. In the table below we display the four largest Ritz values obtained after $m = 150$ iterations of the exact Arnoldi and inexact Arnoldi methods, with $\varepsilon = 10^{-8}$.

<table>
<thead>
<tr>
<th>Method</th>
<th>$\theta_1$</th>
<th>$\theta_3$</th>
<th>$\theta_2$</th>
<th>$\theta_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>exact Arnoldi</td>
<td>6.856543090</td>
<td>7.12220153908</td>
<td>7.18570250215</td>
<td>7.51269629278988</td>
</tr>
<tr>
<td>inexact Arnoldi</td>
<td>6.856516751</td>
<td>7.12220154236</td>
<td>7.18570250124</td>
<td>7.5126962927843</td>
</tr>
</tbody>
</table>

In the inexact process, the perturbation was monitored by using the Frobenius norm of the residual matrix of the three largest Ritz values. The final perturbation norm was equal to $\| f_{150} \| = 2 \cdot 10^{-4}$. We observe in the table that these three Ritz
values deviate from those of the exact process of at most $O(10^{-9})$. On the other hand, the forth Ritz value only matches that of the exact Arnoldi process, to five decimal digits. Therefore, the accuracy in the inexact recurrence is again lost for the approximate eigenpairs that are not involved in the perturbation tolerance. We also note that the largest eigenvalue has more accurate digits with respect to the exact process, than the other eigenvalues. In particular, it has almost full accuracy, in spite of a $2 \cdot 10^{-4}$ perturbation in norm. This phenomenon confirms the fact that for groups of eigenvalues, it is the slowest one converging that drives the allowed perturbation.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{example42.png}
\caption{Example 4.2. Left: spectrum of $A$. Right: convergence curves for inverted Arnoldi (dash-dotted curve) and inexact inverted Arnoldi (solid curve).}
\end{figure}

**Example 4.2.** We next present a typical setting where the flexible accuracy in the matrix-vector product can be fully appreciated. We consider the matrix $\text{SHERMAN5}$ from the Matrix Market repository [12]. This is a non-Hermitian, indefinite real matrix of size $n = 3312$, and was also employed in [14] to analyze the performance of Arnoldi-type methods. The spectrum of the matrix is reported in the left plot of Figure 4.2. We approximate the (real) eigenvalue closest to zero, $\lambda \approx 4.6925 \cdot 10^{-2}$, by means of an inverted Arnoldi process. The generating vector $v_1$ was the normalized vector of all ones. At each inverted Arnoldi iteration, the operation $y = A^{-1}v$ should be carried out. At each (outer) inexact iteration, a system with $A$, namely $Ay = v$, is approximately solved with preconditioned GMRES with zero starting guess. The matlab incomplete LU factorization with $\text{tol} = 10^{-3}$ was used as right preconditioner. The GMRES iteration terminated as soon as the system residual norm reached a certain tolerance $\text{tol}$. The inner stopping criterion we used to approximately solve $Ay = v_k$ at step $k$ is

$$
\|v_k - Ay\| \leq \frac{\min\{\alpha, \delta^{(k-1)}\}}{2m\|r_{k-1}\|/\|\theta^{(k-1)}\|} \varepsilon, \quad \varepsilon = 10^{-10}, \quad \alpha = 400.
$$

In the *exact* case, we assume that we cannot afford to solve with $A$ exactly, therefore we approximately solve the inner system with a fixed tolerance, $\text{tol} = 10^{-10}$. A total of $m = 12$ outer iterations was carried out. The results of our experiment are reported in the right plot of Figure 4.2, where the magnitude of the final inner residual at each iteration is also plotted. The eigenvalue convergence curves cannot
be distinguished until the final accuracy is reached. Note that in the case of variable inner tolerance, preconditioned GMRES took 22 iterations to solve the first inner system at the required accuracy, whereas only 5 iterations were needed at the 8th inverted Arnoldi iteration, to reach a residual of the order of $10^{-2}$.

5. Harmonic Ritz approximation. It has been shown [13, 16, 24] that when looking for interior eigenvalues of Hermitian as well as of non-Hermitian matrices, harmonic Ritz values may be preferred to Ritz values. Harmonic Ritz pairs are pairs $(\theta, V_m u)$ where $\theta$ and $u$ are the eigenvalues and corresponding eigenvectors of the generalized eigenvalue problem

$$(H_m^* H_m + |h_{m+1,m}|^2 e_m e_m^*) u = \theta H_m^* u, \quad \|u\| = 1,$$

or, equivalently, of the standard eigenvalue problem

$$(H_m + |h_{m+1,m}|^2 (H_m^*)^{-1} e_m e_m^*) u = \theta u, \quad \|u\| = 1.$$  

In the following, we let

$$\tilde{H}_m = H_m + |h_{m+1,m}|^2 (H_m^*)^{-1} e_m e_m^*.$$  

It was observed in [14] that a better choice as approximation to eigenpairs of $A$ is the pair $(\rho, V_m u)$, where $\rho$ is the Rayleigh quotient of $u$, that is $\rho = u^* H_m u$. Using (5.1), for the associated residual we have

$$A V_m u - \rho V_m u = V_m H_m u + h_{m+1,m} v_{m+1} e_m^* u - \rho V_m u$$

$$= V_{m+1} \begin{pmatrix} H_m u - \rho u & e_m^* u \\ h_{m+1, m} e_m^* u & \end{pmatrix}.$$  

In the following we will use the computed residual in (5.3), namely we define

$$r_m := V_{m+1} \begin{pmatrix} H_m u - \rho u & e_m^* u \\ h_{m+1, m} e_m^* u & \end{pmatrix},$$

which differs from the true residual $A V_m u - \rho V_m u$ in the inexact case.

A result similar to that of Proposition 2.2 can be derived in terms of the matrix $\tilde{H}_m$ in (5.2).

**Proposition 5.1.** For $k < m$, let $V_k u^{(k)}$ be a harmonic Ritz vector associated with $H_k$, with $\|u^{(k)}\| = 1$, and let $\rho^{(k)} = (u^{(k)})^* H_k u^{(k)}$. Moreover, let $X = \begin{bmatrix} 1 & u^{(k)} \end{bmatrix}$, $Y$ be a unitary matrix and $\tilde{H}_m = Y^* \tilde{H}_m Y$. Let $\delta_{m,k} = \sigma_{\min}(\tilde{H}_m - \rho^{(k)} I)$.

$$r_k = V_{k+1} \begin{pmatrix} H_k u^{(k)} - \rho^{(k)} u^{(k)} \\ h_{k+1,k} e_k^* u^{(k)} \end{pmatrix}, \quad s_m = [(u^{(k)})^*, \Omega^*] \tilde{H}_m - \rho^{(k)} [(u^{(k)})^*, \Omega^*].$$

If $\delta_{m,k} > 0$ and

$$\|r_k\| \leq \frac{\delta_{m,k}^2}{4\|s_m\|},$$

then there exists a unit norm eigenvector $u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ of $\tilde{H}_m$ with $u_1 \in \mathbb{C}^k$, such that

$$\|u_2\| \leq \frac{\tau}{\sqrt{1 + \tau^2}} \leq \tau, \quad \text{where} \quad \tau \in \mathbb{R}, \quad 0 \leq \tau < 2 \frac{\|r_k\|}{\delta_{m,k}}.$$
The following result is similar to Theorem 3.1 and its proof is therefore omitted.

**Theorem 5.2.** Assume m inexact Arnoldi iterations have been carried out and let \( V_{m}u \) be a harmonic Ritz vector associated with \( H_{m} \), with \( \|u\| = 1 \), and let \( \rho \) be the Rayleigh quotient of \( u \), \( \rho = u^{*}H_{m}u \). Given any \( \varepsilon \in \mathbb{R} \), \( \varepsilon > 0 \), assume that for \( k = 1, \ldots, m \),

\[
\|f_{k}\| \leq \begin{cases} 
\frac{d_{m,k-1}}{2^m\|r_{k-1}\|}\varepsilon, & \text{if } k > 1 \text{ and there exists } V_{k-1}u^{(k-1)} \text{ harmonic Ritz vector of } H_{k-1} \text{ satisfying (5.6) and (5.7) with } \\
\rho^{(k-1)} = (u^{(k-1)})^{*}H_{k-1}u^{(k-1)} \text{ } \\
\frac{1}{m}\varepsilon, & \text{otherwise.}
\end{cases}
\]
Then \( \| (AV_m u - \rho V_m u) - r_m \| \leq \varepsilon \).

The deviation of the inexact Harmonic process from its unperturbed counterpart does not significantly differ from that of inexact and exact Arnoldi. As an example, we report the convergence behavior of the exact and inexact Harmonic Ritz approximations for the matrix in Example 4.1. In Figure 5.1 the convergence to the largest eigenvalue in modulo is depicted. The matrix-vector product was perturbed by using the following variant of (3.6),

\[
(5.9) \quad \| f_k \| \leq \frac{\min \{ \alpha, \delta^{(k-1)} \}}{2m| r_{k-1} |} \varepsilon, \quad \delta^{(k-1)} := \min_{\rho_j \neq \rho^{(k-1)}} | \rho^{(k-1)} - \rho_j |,
\]

where each \( \rho_j \) is the Rayleigh quotient associated with the \( j \)th unit norm Harmonic Ritz vector \( u_j^{(j)} \) of \( H_{k-1} \).

We also report the results after \( m = 150 \) iterations with the same matrix, when looking for the first three eigenvalues. Similar strategies as in Example 4.1 were used.

<table>
<thead>
<tr>
<th>method</th>
<th>( \rho_4 )</th>
<th>( \rho_3 )</th>
<th>( \rho_2 )</th>
<th>( \rho_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>exact Harmonic</td>
<td>6.8572788458</td>
<td>7.1222009509</td>
<td>7.18570255184</td>
<td>7.5126962627829</td>
</tr>
<tr>
<td>inexact Harmonic</td>
<td>6.8571791230</td>
<td>7.1222009690</td>
<td>7.18570254871</td>
<td>7.5126962627823</td>
</tr>
</tbody>
</table>

The same considerations as for the Arnoldi process apply here. The magnitude of the perturbation at the last iteration was \( 9 \cdot 10^{-5} \).

6. Exact and Inexact Lanczos methods. In this section we show that our analysis carries over to the case of the inexact Lanczos method. In the standard nonsymmetric Lanczos method the following relations hold,

\[
(6.1) \quad AQ_k = Q_kT_k + t_{k+1,k} q_{k+1} e_k^*, \quad A^* P_k = P_k T_k^* + t_{k,k+1} p_{k+1} e_k^*, \quad P_k^* Q_k = I_k.
\]

Writing \( Q_{k+1} = [q_1, q_2, \ldots, q_{k+1}] \), we assume \( \| q_i \| = 1, i = 1, \ldots, k+1 \). This condition, together with the biorthogonality of \( P_k, Q_k \) completely defines the column vectors.
in $P_k, Q_k$. Different normalizations would be possible, such as $\|q_i\| = 1 = \|p_i\|$, $i = 1, \ldots , k$. In (6.1), $T_k$ is a $k \times k$ nonsymmetric tridiagonal matrix; see e.g. [2].

When the matrix-vector products with $A$ and $A^*$ are performed inexactly, that is

$$q = Ap_i + f_i, \quad p = A^* p_i + g_i, \quad i = 1, \ldots , k,$$

the original Lanczos relations (6.1) transform as follows,

(6.2) $AQ_k = Q_k H_k + h_{k+1,k} q_{k+1} e_k^T + F_k,$
(6.3) $A^* P_k = P_k K_k + h_{k,k+1} p_{k+1} e_k^T + G_k,$ \hspace{1cm} $P_k^* Q_k = I,$

where we used $F_k = [f_1, \ldots , f_k]$ and $G_k = [g_1, \ldots , g_k]$. Here matrices $H_k$ and $K_k^*$ are upper Hessenberg and no longer tridiagonal. Their diagonal and next to diagonal elements are the same, whereas the remaining upper part of the two matrices differs. Clearly, the special properties of the Lanczos iteration are lost. Indeed, the inexact Lanczos iteration is a paired long-term recurrence, as opposed to the paired three-term recurrence of the Lanczos iteration (6.1). Moreover, while in the exact recurrence the matrix $T_k$ provides approximations to both right and left eigenvectors of $A$, this is no longer the case in the inexact method. Since $H_k$ and $K_k^*$ differ, both eigenvalue problems with $H_k$ and $K_k^*$ need to be solved to obtain right and left Ritz vectors.

Another property not inherited by the inexact process concerns convergence. Under certain conditions, the exact Lanczos recurrence determines quadratically converging Ritz values [1] [20]. Since neither of $H_k$ or $K_k$ alone provides right and left eigenvector approximations, convergence is only linear in the inexact case.

Proposition 2.2 can be applied to each of the two matrices $H_k$ and $K_k^*$, to show that the components of converging Ritz vectors have a decreasing pattern. Owing to the similarity between the inexact relations of (6.2), (6.3) with (3.1), we can thus apply the result on the dynamic accuracy stated in Theorem 3.1 to each of the two inexact Lanczos recurrences. This provides us with a way to monitor the gap $\| (A Q_k u - \theta Q_k u) - r_k \|$, where $(\theta , u)$ is a right eigenpair of $H_k$ in (6.2) and $r_k$ is the associated computed residual. An analogous result holds for left Ritz pairs.

**Example 6.1.** We consider the $900 \times 900$ matrix arising from the centered finite difference approximation of the operator $Lu = -u_{xx} - u_{yy} + (x + y)u_x$ on the unit square. The starting vector is the normalized vector of all ones, for both the $Q_k$ and $P_k$ sequences. We are interested in the approximation of the smallest (real) eigenvalue, $\lambda \approx 2.0276 \cdot 10^{-2}$, with a final residual tolerance of $\varepsilon = 10^{-8}$. The whole spectrum is depicted in the left plot of Figure 6.1. The convergence of the exact and inexact Lanczos right residual norms is reported in the right plot of Figure 6.1. The perturbations in the matrix-vector products were enforced by adding random vectors $f_i, g_i, i = 1, \ldots , k$. Their norms were monitored by means of (3.6), applied distinctly to the eigenpairs of the matrices $H_{i-1}, K_{i-1}$. The convergence curve in the inexact case agrees with that of the exact procedure until final accuracy is reached.

**7. Further comments.** It was empirically observed in the literature that when approximating the eigenpairs of a matrix by means of the Arnoldi method, the accuracy in the application of the operator may in some cases be relaxed while maintaining the convergence to the sought after eigenpairs. In this paper we have presented the theoretical foundation for the justification of this phenomenon, and provided a more robust relaxation criterion. Our results indicate that flexible accuracy can be safely

---

2We thank D. Day for pointing to this paper.
employed when the approximate eigenpair is sufficiently close to the target eigenpair, depending on the sensitivity of the given matrix.

Our analysis highlights that reasonably accurate results can be obtained in spite of large perturbations. The inexact Arnoldi relation (3.1) could be written as follows,

\[(A + E_m)V_m = V_m H_m + h_{m+1,m} v_{m+1}^e, \quad E_m = \sum_{k=1}^{m} f_k v_k^e.\]

A backward error analysis would suggest the rather pessimistic picture that a Ritz pair \((\theta, V_m u)\) would be an approximation to an eigenpair of \(A + E_m\) but not of \(A\). We have shown that the perturbation is performed in a way so that the approximation to the target eigenpairs of \(A\) is not affected. However, other inexact Ritz pairs may be perturbed by a quantity fully influenced by \(\|E_m\|\).

A related question that we have not answered is that the rate of convergence could be affected by the perturbation. More precisely, even though our theory ensures that the norm of the true residual of some selected Ritz pairs still converges to a small quantity, it is not clear whether it does so with the same convergence rate as in the unperturbed process. A similar problem is encountered in the inexact linear system setting [22, 30, 31]. In [23] it was shown that no convergence delay is observed in the inexact linear system case, unless the coefficient matrix and the right-hand side are very sensitive to perturbations. Although we expect these conclusions to carry over to the eigenvalue setting, an ad-hoc analysis remains to be done.

Practical implementations require (implicit) restarting and locking of converged eigenpairs [10]. As already mentioned in the introduction, we have not addressed these important issues which need special attention, since our theory predicts that flexible accuracy should take into account the occurrence of Ritz pairs converging to target eigenpairs of \(A\) at different rates. In the linear system setting, the problem of restarting has been recently addressed in [25].

**Acknowledgement.** We thank Jasper van den Eshof for his insightful remarks, and the two anonymous referees for their comments.
Variable accuracy in inexact eigencomputation

REFERENCES


