INEXACT INVERSE ITERATION WITH VARIABLE SHIFT FOR NONSYMMETRIC GENERALIZED EIGENVALUE PROBLEMS∗

JÖRG BERNS-MÜLLER† AND ALASTAIR SPENCE‡

Abstract. In this paper we analyze inexact inverse iteration for the nonsymmetric generalized eigenvalue problem $Ax = \lambda Mx$, where $M$ is symmetric positive definite and the problem is diagonalizable. Our analysis is designed to apply to the case when $A$ and $M$ are large and sparse and preconditioned iterative methods are used to solve shifted linear systems with coefficient matrix $A - \sigma M$. We prove a convergence result for the variable shift case (for example, where the shift is the Rayleigh quotient) which extends current results for the case of a fixed shift. Additionally, we consider the approach from [V. Simoncini and L. Eldén, BIT, 42 (2002), pp. 159–182] to modify the right-hand side when using preconditioned solves. Several numerical experiments are presented that illustrate the theory and provide a basis for the discussion of practical issues.

Key words. eigenvalue approximation, inverse iteration, iterative methods

AMS subject classifications. 65F10, 65F15

DOI. 10.1137/050623255

1. Introduction. Consider the generalized eigenvalue problem

$$(1.1) \quad Ax = \lambda Mx,$$

where $A$ is an $n \times n$ nonsymmetric matrix, and $M$ is an $n \times n$ symmetric positive definite matrix with $x \in \mathbb{C}^n$, $\lambda \in \mathbb{C}$. In our analysis we restrict ourselves to the case where $M^{-1}A$ is diagonalizable; that is, (1.1) has a full set of eigenvectors. Here $n$ is large and $A$ and $M$ are assumed to be sparse.

Large-scale eigenvalue problems arise in many applications, such as the determination of linearized stability of a three-dimensional fluid flow. Typically only a few eigenvalues are of interest to the user, and therefore iterative projection methods such as Arnoldi’s method [1] and its modern variants [11, 7], or Davidson-type methods [13, 22], and subspace iteration [8, 24, 12] are applied. However, to speed up the convergence (see [2, section 3.3]), often these methods are applied to a “shift-invert” form of (1.1) with the resulting large, sparse linear systems solved iteratively. To obtain a reliable and efficient eigenvalue solver one requires a good understanding of the interaction between the iterative linear solver and the iterative eigenvalue solver. In this paper we study inexact inverse iteration, the simplest inexact iterative method, as a first step in helping to understand more sophisticated inexact eigenvalue techniques.

The classical inverse iteration algorithm to find a single eigenvalue of (1.1) is given as follows.

Algorithm 1. inverse iteration.

Given $x^{(0)}$, then iterate:

1. Choose $\sigma^{(i)}$.
(2) Solve \((A - \sigma^{(i)} M)y^{(i)} = Mx^{(i)}\).
(3) Set \(x^{(i+1)} = y^{(i)}/\varphi(y^{(i)})\).

Here \(\varphi(y^{(i)})\) denotes a scalar normalizing function. Common choices for \(\varphi\) are \(\varphi(y^{(i)}) = \|y^{(i)}\|_M\) and \(\varphi(y^{(i)}) = z^H y^{(i)}\) for some fixed vector \(z\). Often the choice \(z = e_k\) is made, where \(e_k\) denotes the \(k\)th canonical unit vector and \(k\) corresponds to a component of large modulus in the desired eigenvector. One can keep \(\sigma^{(i)}\) fixed, so that \(\sigma^{(i)} = \sigma^{(0)}\), to obtain a fixed shift method. Alternatively, one can obtain a variable shift method by updating \(\sigma^{(i)}\), typically by the Rayleigh quotient or by \(\sigma^{(i+1)} = \sigma^{(i)} + 1/(z^H My^{(i)})\) if \(\varphi(y^{(i)}) = z^H My^{(i)}\); see [25, p. 637], [6]. An early fundamental paper on Rayleigh quotient iteration for nonsymmetric problems with exact solves is [16].

We consider the following inexact version of inverse iteration.

**Algorithm 2. inexact inverse iteration.**

Given \(x^{(0)}\), then iterate:

1. Choose \(\sigma^{(i)}\) and \(\tau^{(i)}\).
2. Find \(y^{(i)}\) such that \(\|(A - \sigma^{(i)} M)y^{(i)} - Mx^{(i)}\| \leq \tau^{(i)}\).
3. Set \(x^{(i+1)} = y^{(i)}/\varphi(y^{(i)})\).

Algorithm 2 is an example of an “inner-outer” iterative algorithm; see, for example, [5]. Here the outer iteration being indexed by \(i\) is the standard step in inverse iteration, and the inner iteration refers to the iterative solution of the linear system \((A - \sigma^{(i)} M)y^{(i)} = Mx^{(i)}\) to a prescribed accuracy. Since most iterative linear solvers have stopping conditions based on the residual we use the residual condition \(\|(A - \sigma^{(i)} M)y^{(i)} - Mx^{(i)}\| \leq \tau^{(i)}\). In practice there are various ways to formulate the inner iteration stopping condition (usually as a relative condition). Here we use an absolute stopping condition to simplify the analysis.

An early paper on inexact inverse iteration for the standard symmetric eigenvalue problem is [19]. More recently [23, 21, 14, 9, 3] various aspects of inexact inverse iteration for the symmetric eigenvalue problem have been considered, usually with the shift chosen as the Rayleigh quotient. It is known (see [10, 6]) that with a fixed and not too accurate shift one needs to solve the shifted linear equations more and more accurately. Additionally, for nonsymmetric generalized eigenvalue problems, the analysis in [6] shows how the accuracy of the inner solves affects the convergence of the outer iteration. Here we extend the convergence theory to the case of variable shifts, for example, when the Rayleigh quotient is used. In this case we show that the tolerance for the inexact solve need not decrease, provided the shift tends towards the desired eigenvalue. The analysis in this paper will be independent of a specific linear solver; we assume only that the residual of the inexact linear solve can be controlled.

The plan of the paper is as follows. Section 2 gives some basic results and notation. Section 3 contains a convergence analysis for inexact inverse iteration. In particular, if Rayleigh quotient shifts are chosen, we see how to regain the quadratic convergence that is achieved using exact linear solves. Alternatively, we show that if the linear systems are solved to a fixed tolerance, we can still achieve a convergent method but with the rate of convergence being only linear. In section 4 we extend the approach of [21] based on modifying the right-hand side of the standard inverse iteration formulation with the aim of reducing the number of inner iterations needed per outer iteration but maintaining the variable shift. This idea is motivated by the work in [20] and has proven to be effective for the symmetric eigenvalue problem. We give a convergence theory and compare it with more standard approaches. In the paper several numerical examples are given to both illustrate the theory and aid the
Throughout this paper we use $\| \cdot \|$ for $\| \cdot \|_2$; however, most results are norm independent.

2. Some basic results. We restrict our attention to the case where the generalized eigenvalue problem $Ax = \lambda Mx$ is diagonalizable; that is, there exist an invertible matrix $V$ and a diagonal matrix $\Lambda$ (both possibly complex) such that

$$AV = M\Lambda,$$

and so the eigenvalues of $A$ lie on the diagonal of $\Lambda$ and the columns of $V$ are the right eigenvectors, that is, $Av_j = \lambda_j Mv_j$, $j = 1, \ldots, n$. The corresponding decomposition in terms of the left eigenvectors is

$$UA = \Lambda U,$$

where $U$ can be chosen as $U = V^{-1}M^{-1}$ and so $UMV = I$. Hence the rows of $U$ are the left eigenvectors, that is, $u_j = U^T e_j$ with $u_j^T A = \lambda_j u_j^T M$, $j = 1, \ldots, n$. Note that for the theory we leave the scaling of the eigenvectors open, but we could ask that $\|v_j\| = 1$ or $\|v_j\|_M = 1$. In either case $UMV = I$ provides the corresponding scaling for $u_j$.

Using the decomposition (2.1) and assuming that $\sigma$ is not an eigenvalue of (1.1) we can write

$$\left( A - \sigma M \right) V = MV(\Lambda - \sigma I)$$

$$\Leftrightarrow V(\Lambda - \sigma I)^{-1} = (A - \sigma M)^{-1}MV.$$  

(2.3)

Similarly we can use (2.2) to obtain

$$U(A - \sigma M) = (\Lambda - \sigma I)UM$$

$$\Leftrightarrow (\Lambda - \sigma I)^{-1} U = UM(A - \sigma M)^{-1}.$$  

(2.4)

2.1. The generalized tangent. In order to analyze the convergence of inexact inverse iteration described in Algorithm 2 we use the following splitting:

$$x^{(i)} = a^{(i)}(c^{(i)}v_1 + s^{(i)}w^{(i)}),$$

where $w^{(i)} \in \text{span}(v_2, \ldots, v_n)$ and $\|UMw^{(i)}\| = 1$. The splitting implies that $V^{-1}w^{(i)} \in \text{span}(e_2, \ldots, e_n)$ and scaling implies that $\|V^{-1}w^{(i)}\| = \|UMw^{(i)}\| = 1$. Defining

$$a^{(i)} := \|UMx^{(i)}\|$$

gives $|s^{(i)}|^2 + |c^{(i)}|^2 = 1$, since from (2.5) we have

$$UMx^{(i)} = a^{(i)}c^{(i)}UMv_1 + a^{(i)}s^{(i)}UMw^{(i)},$$

(2.6)

and so

$$1 = \|UMx^{(i)}\| / \alpha^{(i)} = \|c^{(i)}e_1 + s^{(i)}UMw^{(i)}\|$$

$$= \left( |c^{(i)}|^2 + |s^{(i)}|^2 \right)^{1/2}.$$
since \(e_1 \perp UMw^{(i)}\). Thus we interpret \(s^{(i)}\) as a generalized sine and \(c^{(i)}\) as a generalized cosine, which is in the spirit of the orthogonal decomposition in [17] used for the symmetric eigenvalue problem analysis. For convenience we introduce the matrix \(F\), defined by

\[
F := (I - e_1e_1^T)UM = UM(I - v_1u_1^TM),
\]

and note that \(Fv_1 = 0\) and \(Fv_j = e_j\), so that

\[
(UM - F)x^{(i)} = \alpha^{(i)}c^{(i)}e_1,
\]

and

\[
Fx^{(i)} = \alpha^{(i)}s^{(i)}UMw^{(i)}.
\]

Hence \(\| (UM - F)x^{(i)} \| \) measures the length of the component of \(x^{(i)}\) in the direction of \(v_1\) and \(Fx^{(i)}\) picks out the second term in (2.6). So it is natural to introduce as a measure for convergence of \(x^{(i)}\) to \(\text{span}(v_1)\) the generalized tangent (cf. [6, section 2.1])

\[
t^{(i)} := \frac{s^{(i)}}{c^{(i)}} = \frac{\|Fx^{(i)}\|}{\| (UM - F)x^{(i)} \|}.
\]

Clearly \(\frac{1}{\alpha^{(i)}s^{(i)}}x^{(i)} - v_1 = t^{(i)} \|w^{(i)}\|\), and so \(t^{(i)}\) measures the quality of the approximation of \(x^{(i)}\) to \(v_1\). Note that \(t^{(i)}\) is independent of the factor \(\alpha^{(i)}\) and that in the inverse iteration algorithm \(x^{(i)}\) is scaled so that \(\varphi(x^{(i)}) = 1\).

For future reference we recall that for \(x \in \mathbb{C}^n\) the Rayleigh quotient for (1.1) is defined by

\[
\varphi(x) := \frac{x^HAx}{x^HMx}
\]

and that

\[
\varphi(x^{(i)}) - \lambda_1 = \frac{(x^{(i)})^H(A - \lambda_1M)x^{(i)}}{(x^{(i)})^HMx^{(i)}} = O(|s^{(i)}|)
\]

since \((A - \lambda_1M)x^{(i)} = \alpha^{(i)}s^{(i)}(A - \lambda_1M)w^{(i)}\), using (2.5). Thus, the Rayleigh quotient converges linearly in \(|s^{(i)}|\) to \(\lambda_1\). Also, since

\[
(A - \varphi(x^{(i)})M)x^{(i)} = (A - \lambda_1M)x^{(i)} + (\lambda_1 - \varphi(x^{(i)}))Mx^{(i)}
\]

we have that the eigenvalue residual \(r^{(i)}\) defined by

\[
r^{(i)} := (A - \varphi(x^{(i)})M)x^{(i)}
\]

satisfies

\[
\|r^{(i)}\| = O(|s^{(i)}|).
\]

Note that while both (2.12) and (2.15) indicate that convergence is linear in \(|s^{(i)}|\), it is often the case that convergence to an eigenvalue is faster than convergence to the corresponding eigenvalue residual.
3. Convergence of inexact inverse iteration. In this section we provide the convergence analysis for inexact inverse iteration using a variable shift strategy. In section 3.1 we provide a lemma which gives a bound on the generalized tangent \( t^{(i+1)} \). This bound is then used in the convergence theorem in section 3.2. Numerical experiments are presented to illustrate the theory.

Practical choices for \( \sigma^{(i)} \) are the update technique

\[
\sigma^{(i+1)} = \sigma^{(i)} + 1/\varphi(y^{(i)}),
\]

the Rayleigh quotient given by (2.11), or the related

\[
\sigma^{(i)} = \frac{z^H A x^{(i)}}{z^H M x^{(i)}},
\]

where \( z \) is some fixed vector chosen to maximize \( |z^H M x^{(i)}| \). For \( M = I \) it is common to take \( z = e_k \), where \( k \) corresponds to the component of maximum modulus of \( x^{(i)} \) (for example, see [18]). If the choice \( \varphi(y^{(i)}) = z^H M y^{(i)} \) is made, then for exact solves it is easily shown that

\[
\sigma^{(i+1)} = \sigma^{(i)} + \frac{1}{z^H M y^{(i)}} = \frac{z^H A x^{(i+1)}}{z^H M x^{(i+1)}},
\]

so that (3.1) and (3.2) are equivalent. For inexact solves we use (3.2), and it is easily shown that \( \lambda_1 - \sigma^{(i)} = O(t^{(i)}) \) (cf. (2.12)).

3.1. One step bound. Let us assume that the sought eigenvalue, say \( \lambda_1 \), is simple and well separated. Next, we assume the starting vector \( x^{(0)} \) is neither the solution itself nor is it deficient in the sought eigendirection, that is, \( 0 < |s^{(i)}| < 1 \).

Further, we assume that the shift \( \sigma^{(i)} \) satisfies

\[
|\lambda_1 - \sigma^{(i)}| \leq \frac{1}{2} |\lambda_2 - \lambda_1| \quad \forall i,
\]

where \( |\lambda_2 - \lambda_1| = \min_{j \neq 1} |\lambda_j - \lambda_1| \). Hence \( |\lambda_1 - \sigma^{(i)}| < |\lambda_2 - \sigma^{(i)}| \).

Now consider step (2) of inexact inverse iteration, given by Algorithm 2, and define

\[
d^{(i)} := M x^{(i)} - (A - \sigma^{(i)} M) y^{(i)}.\]

Rearranging this equation and using the scaling of \( x^{(i+1)} \) from step (3) in Algorithm 2 together with the fact that \( A - \sigma^{(i)} M \) is invertible we obtain the update equation

\[
\varphi(y^{(i)}) x^{(i+1)} = (A - \sigma^{(i)} M)^{-1} (M x^{(i)} - d^{(i)}).
\]

This is the equation on which the following analysis is based.

**Lemma 3.1.** Assume the shifts satisfy (3.4) and that the bound on the residual \( \tau^{(i)} \) in Algorithm 2 satisfies

\[
||d^{(i)}|| \leq \tau^{(i)} < \beta ||u_1^T M x^{(i)}|| / ||u_1||
\]

for some \( \beta \in (0, 1) \). Then

\[
t^{(i+1)} \leq \frac{|\lambda_1 - \sigma^{(i)}|}{|\lambda_2 - \sigma^{(i)}|} \frac{||U d^{(i)}||}{(1 - \beta) ||u_1^T M x^{(i)}||}.
\]
Proof. Recall that \( u_I^T M x^{(i+1)} = (i+1)c^{(i+1)} \), and \( u_I^T = e_1^T U \). Hence premultiplying the update equation (3.6) by \( u_I^T M \) and using \( U(M - \sigma(I)M)^{-1} = (\Lambda - \sigma(I)I)U \) (see (2.4)), we obtain
\[
\varphi(y^{(i)})\alpha^{(i+1)}c^{(i+1)} = e_1^T (\Lambda - \sigma(I)I)^{-1}U(Mx^{(i)} - d^{(i)})
\]
(3.9)
\[
= (\Lambda - \sigma(I))^{-1}u_I^T (Mx^{(i)} - d^{(i)}).
\]
Further, using (3.7)
\[
|u_I^T Mx^{(i)}| - |u_I^T d^{(i)}| \geq (1 - \beta) |u_I^T Mx^{(i)}|.
\]
Hence
\[
|\varphi(y^{(i)})|\alpha^{(i+1)}c^{(i+1)} \geq \frac{|u_I^T Mx^{(i)}| - |u_I^T d^{(i)}|}{|\Lambda - \sigma(I)|}
\]
(3.11)
\[
\geq (1 - \beta) \frac{|u_I^T Mx^{(i)}|}{|\Lambda - \sigma(I)|}.
\]
To obtain an upper bound on \( |s^{(i+1)}| \) we apply \( F \), defined by (2.7), to (3.6) to obtain
\[
\varphi(y^{(i)})Fx^{(i+1)} = (I - e_1e_1^T)UM(\Lambda - \sigma(I)M)^{-1}(Mx^{(i)} - d^{(i)})
\]
and using (2.4),
\[
\varphi(y^{(i)})Fx^{(i+1)} = (I - e_1e_1^T)(\Lambda - \sigma(I)I)^{-1}U(Mx^{(i)} - d^{(i)})
\]
(3.13)
\[
= (\Lambda - \sigma(I)I)^{-1}(I - e_1e_1^T)U(Mx^{(i)} - d^{(i)}).
\]
Taking norms we obtain
\[
\|\varphi(y^{(i)})Fx^{(i+1)}\| = \|(\Lambda - \sigma(I)I)^{-1}(I - e_1e_1^T)U(Mx^{(i)} - d^{(i)})\|
\]
\[
\leq \|(\Lambda - \sigma(I)I)^{-1}(I - e_1e_1^T)\| \|(I - e_1e_1^T)U(Mx^{(i)} - d^{(i)})\|
\]
(3.14)
\[
\leq \frac{1}{|\alpha(I)|} \left( |\alpha(I)s^{(i)}| + \|(I - e_1e_1^T)Ud^{(i)}\| \right).
\]
With \( t^{(i+1)} \) defined by (2.9), and using (2.8), we have
\[
t^{(i+1)} = \frac{\|\varphi(y^{(i)})Fx^{(i+1)}\|}{\|\varphi(y^{(i)})U - F\|}\|x^{(i+1)}\|
\]
\[
\leq \frac{\|\varphi(y^{(i)})Fx^{(i+1)}\|}{\|\varphi(y^{(i)})U - F\|}\|x^{(i+1)}\|.
\]
Hence, using (3.10), (3.11), and (3.14),
\[
t^{(i+1)} \leq \frac{|\alpha(I)s^{(i)}| + \|(I - e_1e_1^T)Ud^{(i)}\|}{|\alpha(I)|}. \]

This result is similar to results in [23, 21, 3] in the symmetric case and [6, 15] in the unsymmetric case. One advantage of our approach over that in [6, 15] is that it
can be applied to both fixed and variable shift strategies, though here we concentrate on the variable shift analysis.

Condition (3.7) asks that \( \tau^{(i)} \) be bounded in terms of \(|u_i^T M x^{(i)}| = \alpha^{(i)} |c^{(i)}|\) which is related to the cosine of the angle between \(v_1\) and \(x^{(i)}\), the exact and the approximate eigenvectors. In Algorithm 2 we used an absolute tolerance criteria for the inexact solves involving \(\tau^{(i)}\). Now Lemma 3.1 shows that this constraint naturally should be relative to the scaling of \(x^{(i)}\).

In the case where \(d^{(i)} = 0\), we can take \(\beta = 0\) in (3.7), and (3.8) reduces to
\[
|\lambda_1 - \sigma^{(i)}| \leq \frac{|\lambda_1 - \lambda_2|}{2} |s^{(i)}| \quad \forall i.
\]

(3.15)

Assume that, for \(d^{(i)}\) defined by (3.5), \(\|d^{(i)}\| \leq \tau^{(i)}\) with
\[
\tau^{(i)} < \alpha^{(i)} \beta c^{(i)} / \|U\|,
\]

(3.16)

where
\[
0 \leq \beta < \frac{1 - |s^{(0)}|}{2}.
\]

(3.17)

Then inexact inverse iteration as given in Algorithm 2 using a variable shift converges (at least) linearly, \(t^{(i+1)} \leq q t^{(i)} \leq q^{i+1} t^{(0)}\), where
\[
q := \frac{|s^{(0)}| + \beta}{1 - \beta} < 1.
\]

(3.18)

Proof. With \(|\lambda_1 - \sigma^{(i)}| \leq \frac{1}{2} |\lambda_1 - \lambda_2| |s^{(i)}|\) and hence \(|\lambda_2 - \sigma^{(i)}| > \frac{1}{2} |\lambda_2 - \lambda_1|\), we have
\[
\frac{|\lambda_1 - \sigma^{(i)}|}{|\lambda_2 - \sigma^{(i)}|} < |s^{(i)}|.
\]

(3.19)

Thus, from (3.8),

3.2. Convergence theorem for variable shifts.

The following theorem provides sufficient conditions under which an inexact inverse iteration algorithm with linearly converging shifts achieves linear convergence, even if the residual tolerance is fixed.

**Theorem 3.2.** Given \(A, M \in \mathbb{R}^{n \times n}\) with \(M\) symmetric positive definite. Let the generalized eigenvalue problem \(Ax = \lambda Mx\) be diagonalizable and have simple eigenpair \((\lambda_1, v_1)\). Further let \(x^{(i)} = \alpha^{(i)} (c^{(i)} v_1 + s^{(i)} w^{(i)})\) with \(|s^{(0)}| < 1\) and let the shift updates satisfy

\[
|\lambda_1 - \sigma^{(i)}| \leq \frac{|\lambda_1 - \lambda_2|}{2} |s^{(i)}| \quad \forall i.
\]

(3.15)

Assume that, for \(d^{(i)}\) defined by (3.5), \(\|d^{(i)}\| \leq \tau^{(i)}\) with
\[
\tau^{(i)} < \alpha^{(i)} \beta c^{(i)} / \|U\|,
\]

(3.16)

where
\[
0 \leq \beta < \frac{1 - |s^{(0)}|}{2}.
\]

(3.17)

Then inexact inverse iteration as given in Algorithm 2 using a variable shift converges (at least) linearly, \(t^{(i+1)} \leq q t^{(i)} \leq q^{i+1} t^{(0)}\), where
\[
q := \frac{|s^{(0)}| + \beta}{1 - \beta} < 1.
\]

(3.18)

Proof. With \(|\lambda_1 - \sigma^{(i)}| \leq \frac{1}{2} |\lambda_1 - \lambda_2| |s^{(i)}|\) and hence \(|\lambda_2 - \sigma^{(i)}| > \frac{1}{2} |\lambda_2 - \lambda_1|\), we have
\[
\frac{|\lambda_1 - \sigma^{(i)}|}{|\lambda_2 - \sigma^{(i)}|} < |s^{(i)}|.
\]

(3.19)
\[ \begin{align*}
    t^{(i+1)} & \leq |s^{(i)}| \left\{ \frac{|\alpha^{(i)} s^{(i)}| + \|U\| \tau^{(i)}}{(1 - \beta) |\alpha^{(i)} c^{(i)}|} \right\} \\
    & \leq t^{(i)} \frac{|s^{(i)}| + \beta |c^{(i)}|}{1 - \beta} \\
    & \leq t^{(i)} \frac{|s^{(0)}| + \beta}{1 - \beta}.
\end{align*} \]

(3.20)

Set \( q = (|s^{(0)}| + \beta)/(1 - \beta) \). If \( \beta \) satisfies (3.17), then \( q < 1 \), and linear convergence is proved by induction.

This theorem shows that for a close enough starting guess, namely \( |s^{(0)}| < 1 - 2\beta \), and for a shift converging linearly, say using (3.2) or (2.11), then we obtain a linearly converging method, provided the inner iteration is solved to a strict enough tolerance (which itself does not tend to zero).

Not surprisingly, if we ask that the bound on the tolerances \( \tau^{(i)} \) is linear in \( |s^{(i)}| \) instead of being held fixed as allowed by (3.16), then one achieves quadratic convergence. This is stated in the following corollary.

Corollary 3.3. Assume the conditions of Theorem 3.2 are satisfied but that (3.16) is replaced by

\[ \tau^{(i)} \leq \alpha^{(i)} \min(\beta c^{(0)}/\|U\|, \gamma |s^{(i)}|) \]

(3.21)

for some constant \( \gamma \geq 0 \); then the convergence is (at least quadratic), that is, \( t^{(i+1)} \to 0 \) (monotonically) with \( t^{(i+1)} \leq q(t^{(i)})^2 \) for some \( q > 0 \).

Conditions (3.16), (3.17), and (3.18) make precise statements such as “\( \tau^{(i)} \) is small enough” and “\( \mathbf{x}^{(0)} \) is close enough to \( \mathbf{v}_1 \).” Those are unlikely to be of any quantitative use since they are probably too restrictive and contain quantities that are unknown (for example \( \|U\| \) and \( |\lambda_2 - \lambda_1| \)). Of course, the conditions (3.16), (3.18), and (3.21) are not necessary, and in our experiments considerably larger values for \( \tau^{(i)} \) have been used successfully. Condition (3.15) is easily satisfied if \( \sigma^{(i)} \) is given by (3.2) and if \( \mathbf{z} \) is sufficiently close to the left eigenvector \( \mathbf{u}_1 \). However, this is a theoretically sufficient condition, and as is the case in many practical situations convergence occurs without this condition being fulfilled.

We now present some numerical results to illustrate the theory given in Theorem 3.2 and Corollary 3.3. In our experiments different choices of shift produced no significant changes in the results, so we present numerical results for the Rayleigh quotient shift only.

Example 1. Consider \( \mathbf{A} \) and \( \mathbf{M} \) derived by discretizing

\[-\Delta u + 5u_x + 5u_y = \lambda u \quad \text{in } D := [0, 1] \times [0, 1],
\]

\[ u = 0 \quad \text{on } \Gamma := \partial D, \]

using the Galerkin FEM on regular triangular elements with piecewise linear functions. This eigenvalue problem is also discussed in [6]. Here we use a 32 by 32 grid which leads to 961 degrees of freedom. For the discrete eigenvalue problem it is known that \( \lambda_1 \approx 32.2 \) and \( \lambda_2 \approx 61.7 \) with all other eigenvalues satisfying \( \text{Re}(\lambda_j) > 61.8 \). Note that the eigenvalue residual \( r^{(i)} \) defined by (2.14) is proportional to \( |s^{(i)}| \) (using (2.15)), and so this provides a practical way to implement a decreasing tolerance. As inexact linear solver we use preconditioned full GMRES (that is, without restarts), where the preconditioner \( \mathbf{P} \approx \mathbf{A} \) is obtained by an incomplete modified LU decomposition.
Rayleigh quotient iteration with decreasing tolerance, that is, $\text{RQId}$, with drop tolerance $= 0$ $\lambda$
calculating fixed tolerance $\tau$ $\parallel$ eigenvalue residual $RQIf$, Rayleigh quotient iteration with fixed tolerance, that is, versions of Algorithm 2.

As predicted by Corollary 3.3. Thus we recover the convergence rate attained for nonsymmetric problems if the Rayleigh quotient iteration is used with exact solves. We point out that the last iteration in (c) is stopped due to the fact that the relative inverse iteration where the right-hand side is altered with the aim of improving the performance of the preconditioned iterative solver at the risk of slowing down the

Discussion of results. Case (a) shows that the Rayleigh quotient iteration with fixed tolerance $\tau_0 = 0.1$ achieves linear convergence (indeed, in this experiment, super-linear convergence). Case (c) shows that the Rayleigh quotient iteration with linearly decreasing tolerance based on the eigenvalue residual achieves quadratic convergence as predicted by Corollary 3.3. Thus we recover the convergence rate attained for nonsymmetric problems if the Rayleigh quotient iteration is used with exact solves. We point out that the last iteration in (c) is stopped due to the fact that the relative outer tolerance condition is satisfied within GMRES, and so quadratic convergence is lost in the final step. Case (b) shows results obtained using the Rayleigh quotient iteration with a small fixed tolerance. First, we note that since $\tau_0$ is small the method behaves very similarly to the exact solves case. Further, case (b) exhibits initially quadratic convergence as the $s(i)$ dominates $\tau(i)$ in the numerator of (3.8). However, this quadratic convergence is lost when the tangent, $t(i)$, has reduced to the order of the stopping tolerance, and then $\tau(i)$ dominates $s(i)$.

4. Modified right-hand side. In this section we analyze a variation of inexact inverse iteration where the right-hand side is altered with the aim of improving the performance of the preconditioned iterative solver at the risk of slowing down the

Table 3.1

Generalized tangent $t^{(i)}$ and number of inner iterations $k^{(i)}$ for $\text{RQIf}$ (a) and (b) and $\text{RQId}$ (c). In (a) $\tau_0 = 0.1$, in (b) $\tau_0 = 0.001$, and in (c) $\tau_0 = 0.2$ and $\tau_1 = 0.5$.

<table>
<thead>
<tr>
<th></th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\varphi(i)$</td>
<td>$k^{(i-1)}$</td>
<td>$\varphi(i)$</td>
</tr>
<tr>
<td>0</td>
<td>5.0e-02</td>
<td>5.0e-02</td>
<td>5.0e-02</td>
</tr>
<tr>
<td>1</td>
<td>9.0e-03</td>
<td>11</td>
<td>4.4e-04</td>
</tr>
<tr>
<td>2</td>
<td>2.4e-04</td>
<td>19</td>
<td>8.0e-08</td>
</tr>
<tr>
<td>3</td>
<td>4.6e-06</td>
<td>29</td>
<td>7.7e-12</td>
</tr>
<tr>
<td>4</td>
<td>2.6e-08</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4.7e-11</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1.0e-11</td>
<td>52</td>
<td></td>
</tr>
</tbody>
</table>

$\sum k^{(i-1)}$ 195 110 153

with drop tolerance $= 0.1$. In Table 3.1 we present numerical results obtained when calculating $\lambda_1$. Each row in Table 3.1 provides the generalized tangent, $t^{(i)}$ (calculated knowing the exact solution $v_1$), and $k^{(i-1)}$ the number of inner iterations used by preconditioned GMRES to satisfy the residual condition. We use the following two versions of Algorithm 2.

**RQIf.** Rayleigh quotient iteration with fixed tolerance, that is, $\sigma(i) = \varphi(x^{(i)})$ and $\tau(i) = \tau_0 \|Mx^{(i)}\|$.

**RQId.** Rayleigh quotient iteration with decreasing tolerance, that is, $\sigma(i) = \varphi(x^{(i)})$ and $\tau(i) = \min\{\tau_0, \tau_1 \|r^{(i)}\|/\varphi(i) \} \|Mx^{(i)}\|$.

As $\|r^{(i)}\|/\varphi(i)$ is proportional to $s(i)$ and $\|Mx^{(i)}\|$ is proportional to $\alpha(i)$ we expect according to Theorem 3.2 linear convergence for $\text{RQIf}$ and according to Corollary 3.3 quadratic convergence for $\text{RQId}$.

In Table 3.1, cases (a) and (b) illustrate the behavior of $\text{RQIf}$ with $\tau_0 = 0.1$ and 0.001, respectively. Case (c) gives results for $\text{RQId}$, that is, Rayleigh quotient shifts and a decreasing tolerance based on the eigenvalue residual (2.14). We present results for the approximation of $(\lambda_1, v_1)$ and stop the entire calculation once the relative eigenvalue residual $\|r^{(i)}\|/\varphi(i)$ is smaller than $\tau_{\text{outer}} = 10^{-14}$.

Discussion of results. Case (a) shows that the Rayleigh quotient iteration with fixed tolerance $\tau_0 = 0.1$ achieves linear convergence (indeed, in this experiment, super-linear convergence). Case (c) shows that the Rayleigh quotient iteration with linearly decreasing tolerance based on the eigenvalue residual achieves quadratic convergence as predicted by Corollary 3.3. Thus we recover the convergence rate attained for nonsymmetric problems if the Rayleigh quotient iteration is used with exact solves. We point out that the last iteration in (c) is stopped due to the fact that the relative outer tolerance condition is satisfied within GMRES, and so quadratic convergence is lost in the final step. Case (b) shows results obtained using the Rayleigh quotient iteration with a small fixed tolerance. First, we note that since $\tau_0$ is small the method behaves very similarly to the exact solves case. Further, case (b) exhibits initially quadratic convergence as the $s(i)$ dominates $\tau(i)$ in the numerator of (3.8). However, this quadratic convergence is lost when the tangent, $t(i)$, has reduced to the order of the stopping tolerance, and then $\tau(i)$ dominates $s(i)$.

4. Modified right-hand side. In this section we analyze a variation of inexact inverse iteration where the right-hand side is altered with the aim of improving the performance of the preconditioned iterative solver at the risk of slowing down the
outer convergence rate. This idea has been used in [20] and [21]. Instead of solving
(4.1) \[(A - \sigma M)y^{(i)} = Mx^{(i)}\]
[20] used the system
(4.2) \[(A - \sigma M)y^{(i)} = x^{(i)}\]
with no theoretical justification but with the remark that computational time is saved
with the modified right-hand side. Also, for the solution of the standard symmetric
eigenvalue problem \(Ax = \lambda x\) using a preconditioner \(P \approx (A - \sigma I)\), Simoncini and
Eldén [21] solve
(4.3) \[P^{-1}(A - \sigma I)y^{(i)} = x^{(i)}\]
rather than the obvious system
(4.4) \[P^{-1}(A - \sigma I)y^{(i)} = P^{-1}x^{(i)}\]
The motivation for this alteration is that in (4.3) the right-hand side \(x^{(i)}\) is both close
to a null vector of \(P^{-1}(A - \sigma I)\) and close to a scaled version of the solution. The
vector \(P^{-1}x^{(i)}\) has neither of these properties. Here we combine the two ideas. Let
\(P \approx (A - \sigma M)\) be a preconditioner for use within GMRES. Given an approximate
eigenvector \(x^{(i)}\) to obtain an improved eigendirection using preconditioned GMRES
we solve
(4.5) \[P^{-1}(A - \sigma^{(i)}M)y^{(i)} = x^{(i)}\]
rather than the obvious \(P^{-1}(A - \sigma^{(i)}M)y^{(i)} = P^{-1}Mx^{(i)}\). As we shall show below,
by changing the right-hand side from \(P^{-1}Mx^{(i)}\) to \(x^{(i)}\) the convergence theory changes.
The expected gain is that (4.5) will prove to be significantly cheaper to solve in terms
of inner iterations. For the standard symmetric eigenvalue problem where the shift
was chosen as the Rayleigh quotient this was indeed the case. We shall see that
for nonsymmetric problems the situation is not so clear-cut. In this paper we shall
concentrate on the outer convergence theory. The algorithm derived from solving
(4.5) which uses the Rayleigh quotient shift is defined as follows.

**Algorithm 3.** Inexact inverse iteration with modified right-hand side.

*Given* \(x^{(0)}\), *then iterate*:

1. **Choose** \(\tau^{(i)}\), and set \(\sigma^{(i)} = \phi(x^{(i)})\).
2. **Find** \(y^{(i)}\) such that \[\|x^{(i)} - P^{-1}(A - \sigma^{(i)}M)y^{(i)}\| \leq \tau^{(i)}\].
3. **Set** \(x^{(i+1)} = y^{(i)}/\varphi(y^{(i)})\).

Note that we use a standard residual condition rather than the stopping condition
used in [21, section 7]. We define the residual obtained by solving (4.5) approximately
as
(4.6) \[d^{(i)} := x^{(i)} - P^{-1}(A - \sigma^{(i)}M)y^{(i)}\]
so that the inexact solve step can be written as
(4.7) \[(A - \sigma^{(i)}M)y^{(i)} = Px^{(i)} - Pd^{(i)}\],
which should be compared with the inexact solve step
(4.8) \[(A - \sigma^{(i)}M)y^{(i)} = Mx^{(i)} - d^{(i)}\]
in section 3. From (4.8) we obtain

\[ \varphi(y^{(i)})x^{(i+1)} = (A - \sigma^{(i)}M)^{-1}P(x^{(i)} - d^{(i)}) \]  

(cf. (3.6)), which is used in the following analysis. First, assume the residual \(d^{(i)}\) satisfies the bound

\[ \|d^{(i)}\| \leq \tau^{(i)} \leq \beta^{'\prime} |u_i^TPx^{(i)}| / \|UP\| \]

for some \(\beta^{'\prime} \in ([0, 1)\) (cf. (3.7)), and hence it is easily shown that

\[ |u_i^TPx^{(i)}| - |u_i^TPd^{(i)}| \geq (1 - \beta^{'\prime}) |u_i^TPx^{(i)}| . \]

Next, we introduce the expression

\[ T_P(z) := \frac{|(I - e_1e_1^TP)UPz|}{|u_i^TPz|}, \]

where \(z \in \mathbb{C}^n\). By analogy with (2.7) and (2.10), \(T_P(z)\) looks like a generalized tangent with respect to \(P\) rather than \(M\). However, for a general preconditioner \(T_P(v_1) \neq 0\). In fact, \(T_P(v_1)\) measures the effect of \(P\) on the eigenvector \(v_1\), and we shall see in Theorem 4.2 that large values of \(T_P(v_1)\) will slow down or possibly destroy the convergence of Algorithm 3. Note that, under (4.11),

\[ T_P(x^{(i)} - d^{(i)}) \leq \frac{1}{1 - \beta^{'\prime}} \left( T_P(x^{(i)}) + \frac{|UPd^{(i)}|}{|u_i^TPx^{(i)}|} \right). \]

Now we give a one step bound for Algorithm 3 using a variable shift \(\sigma^{(i)}\).

**Lemma 4.1.** Assume \(\sigma^{(i)}\) satisfies (3.4) and (3.15). Further assume that (4.11) holds. Then

\[ t^{(i+1)} \leq \frac{|\lambda_1 - \sigma^{(i)}|}{|\lambda_2 - \sigma^{(i)}|} T_P(x^{(i)} - d^{(i)}) \]

\[ \leq |s^{(i)}| T_P(x^{(i)} - d^{(i)}), \]

where \(T_P(\cdot)\) is given by (4.12).

**Proof.** With the notation in sections 2 and 3 we have

\[ t^{(i+1)} = \frac{\|F\varphi(y^{(i)})x^{(i+1)}\|}{\|UM - F\varphi(y^{(i)})x^{(i+1)}\|} \]

\[ = \frac{\|F(A - \sigma^{(i)}M)^{-1}(Px^{(i)} - Pd^{(i)})\|}{\|(UM - F)(A - \sigma^{(i)}M)^{-1}(Px^{(i)} - Pd^{(i)})\|} \]

\[ = \frac{|(I - e_1e_1^TP)UM(A - \sigma^{(i)}M)^{-1}(Px^{(i)} - Pd^{(i)})|}{|e_1^TPM(A - \sigma^{(i)}M)^{-1}(Px^{(i)} - Pd^{(i)})|} \]

\[ = \frac{|(I - e_1e_1^T)(A - \sigma^{(i)}I)^{-1}U(Px^{(i)} - Pd^{(i)})|}{|e_1^T(A - \sigma^{(i)}I)^{-1}U(Px^{(i)} - Pd^{(i)})|} \]

\[ \leq \frac{|\lambda_1 - \sigma^{(i)}|}{|\lambda_2 - \sigma^{(i)}|} \frac{|(I - e_1e_1^T)UP(x^{(i)} - d^{(i)})|}{|e_1^TUP(x^{(i)} - d^{(i)})|}, \]
from which the required result follows. □

Clearly a formal statement of the convergence of Algorithm 3 merely requires conditions that ensure the second term on the right-hand side of (4.14) remains bounded below 1 for all $i$. For completeness we present such a theorem.

**Theorem 4.2.** Assume that the conditions of Lemma 4.1 hold, and let $\tau^{(i)}$ satisfy (4.10) with $\beta' \in [0, 1)$. Assume that $T_P(v_1) \neq 0$ and

$$q := \frac{1}{1 - \beta'} (2T_P(v_1) + \beta') < 1.$$  \hfill (4.15)

Then, for $x^{(0)}$ close enough to $v_1$, Algorithm 3 converges linearly with $t^{(i+1)} \leq qt^{(i)}$.

**Proof.** Due to the condition on $\tau^{(i)}$, (4.10), we can use (4.13) and (4.10) (again) to give $T_P(x^{(i)} + d^{(i)}) \leq (1 - \beta')^{-1} (T_P(x^{(i)} + \beta')$. Hence it remains to show that $T_P(x^{(i)}) \leq 2T_P(v_1)$, which is valid for $x^{(0)}$ close enough to $v_1$ as $T_P(v_1) \neq 0$. □

Lemma 4.1 and Theorem 4.2 show that the quantity $T_P(v_1)$ plays an important role in the convergence of Algorithm 3, and ideally $T_P(v_1)$ should be small. In practical situations we will have little knowledge of the effect of $P$ on $v_1$, but it is clear that if $u_1^T P v_1$ is small, and hence $T_P(v_1)$ is large; then Algorithm 3 may converge slowly or may possibly fail to converge. Note that we ignore the unlikely case $T_P(v_1) = 0$ in Theorem 4.2, though in this case one could recover quadratic convergence using a decreasing tolerance. We present numerical values for $T_P(v_1)$ in Table 4.2. First, we compare the performance of Algorithm 3 with the variable shift method RQIf discussed in Example 1.

**Example 2.** Again we consider the convection diffusion problem of Example 1; however, now we seek the interior eigenvalue $\lambda_{20} = 337.7$. Here we use preconditioned full GMRES with multigrid as preconditioner to solve the linear systems that arise. The preconditioner consists of one V-cycle and uses 3 Jacobi iterations for both pre- and postsmoothing on each grid. In case (a) of Table 4.1 we use RQIf with $\tau_0 = 0.05$ and in (b) we use RQImodrhs with $\tau_0 = 0.05$.

**RQImodrhs.** Algorithm 3 with $\sigma^{(i)} = q(x^{(i)})$ and tolerance $\tau^{(i)} = \tau_0 \|Px^{(i)}\|$. We present numerical results for calculating $\lambda_{20}$ up to a relative outer tolerance of $\tau_{outer} = 10^{-10}$ in Table 4.1.

**Discussion of results.** From case (a) we observe that the number of inner iterations $k^{(i)}$ increases as the outer process proceeds. This effect was already observed when calculating the eigenvalue $\lambda_1$ of the same example; see Table 3.1. However, the rate of increase here is not as substantial due to the fact that the multigrid preconditioner is a much better preconditioner than the one constructed by the incomplete LU decomposition. Case (b) shows that even though the right-hand side
Generalized tangent $t^{(i)}$ for RQImodrhs with $\tau_0 = 0.01$ using two different preconditioners. In (a) milu$(A, 0.1)$, where $T_P(v_1) = 0.34$, and in (b) milu$(A - 320M, 10^{-4})$, where $T_P(v_1) = 0.045$.

<table>
<thead>
<tr>
<th></th>
<th>(a)</th>
<th>(b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2.0e-02</td>
<td>2.0e-02</td>
</tr>
<tr>
<td>1</td>
<td>3.1e-04</td>
<td>1.9e-04</td>
</tr>
<tr>
<td>2</td>
<td>4.7e-05</td>
<td>5.8e-07</td>
</tr>
<tr>
<td>3</td>
<td>2.6e-06</td>
<td>1.5e-09</td>
</tr>
<tr>
<td>4</td>
<td>1.3e-07</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1.1e-08</td>
<td></td>
</tr>
</tbody>
</table>

Example 3. Again we consider the convection diffusion problem discussed in Example 2, and we seek the interior eigenvalue $\lambda_{20} = 337.7$. To demonstrate the effect of $T_P(v_1)$ on the convergence of RQImodrhs we consider two different preconditioner. In case (a) of Table 4.2 we use a modified incomplete LU decomposition constructed from the unshifted system $A$ using a drop tolerance of 0.1; we denote this by milu$(A, 0.1)$. The other preconditioner, which we use in case (b), is also a modified incomplete LU decomposition constructed now from the shifted system $A - 320M$ using a drop tolerance of $10^{-4}$ (milu$(A - 320M, 10^{-4})$). In Table 4.2 we present numerical results obtained using RQImodrhs with $\tau_0 = 0.01$ using in (a) the “unshifted” preconditioner which has for this example $T_P(v_1) = 0.34$ and in (b) the “shifted” preconditioner which has $T_P(v_1) = 0.045$.

Note that in our experience parameter values for $\tau_0$ smaller than 0.01 did not alter the outer convergence. This is not surprising since $\tau_0 \ll T_P(v_1)$, and hence according to Theorem 4.2 the effect of the inexact solves on the rate of convergence should not be significant.

Discussion of results. From Table 4.2 we observe that the outer convergence in case (a) is linear with a rate $t^{(i+1)}/t^{(i)} \approx 0.05$. Comparing this with the results for case (b) we observe a significant improvement in the outer rate of convergence, which results in a reduced number of outer iterations.
preconditioner merely makes the solution of the linear system more efficient, whereas in Algorithm 3 the preconditioner also affects the outer convergence rate, as seen by the presence of $T_R(v_1)$ term on the right-hand side in (4.15).

5. Conclusion. In this paper we provided a convergence theory for inexact inverse iteration with varying shifts applied to the nonsymmetric generalized eigenvalue problem. Additionally we extended the approach from [21] of modifying the right-hand side to the nonsymmetric generalized eigenvalue problem, presented a convergence theory, and showed that the preconditioner affects the outer convergence rate.

REFERENCES