SIAM J. MATRIX ANAL. APPL.
Vol. 28, No. 4, pp. 1069–1082 

# 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 $\mathbf{Ax} = \lambda\mathbf{Mx}$, where $\mathbf{M}$ is symmetric positive definite and the problem is diagonalizable. Our analysis is designed to apply to the case when $\mathbf{A}$ and $\mathbf{M}$ are large and sparse and preconditioned iterative methods are used to solve shifted linear systems with coefficient matrix $\mathbf{A} - \sigma\mathbf{M}$. 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

$$\mathbf{Ax} = \lambda\mathbf{Mx}, \tag{1.1}$$

where $\mathbf{A}$ is an $n \times n$ nonsymmetric matrix, and $\mathbf{M}$ is an $n \times n$ symmetric positive definite matrix with $\mathbf{x} \in \mathbb{C}^n$, $\lambda \in \mathbb{C}$. In our analysis we restrict ourselves to the case where $\mathbf{M}^{-1}\mathbf{A}$ is diagonalizable; that is, (1.1) has a full set of eigenvectors. Here $n$ is large and $\mathbf{A}$ and $\mathbf{M}$ 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* $\mathbf{x}^{(0)}$*, then iterate:*
(1) *Choose* $\sigma^{(i)}$*.*

*Received by the editors January 25, 2005; accepted for publication (in revised form) by I. C. F. Ipsen December 22, 2005; published electronically December 18, 2006. This work was supported by the Engineering and Physical Sciences Research Council, UK, grant GR/M59075.
http://www.siam.org/journals/simax/28-4/62325.html

†Fachbereich Mathematik, JWG-Universität Frankfurt, Postfach 11 19 32, D-60054 Frankfurt, Germany (berns@math.uni-frankfurt.de).
‡Department of Mathematical Sciences, University of Bath, Bath, United Kingdom (as@maths.bath.ac.uk).

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

Here $\varphi(\mathbf{y}^{(i)})$ denotes a scalar normalizing function. Common choices for $\varphi$ are $\varphi(\mathbf{y}^{(i)}) = \| \mathbf{y}^{(i)} \|_{\mathbf{M}}$ and $\varphi(\mathbf{y}^{(i)}) = \mathbf{z}^H\mathbf{y}^{(i)}$ for some fixed vector $\mathbf{z}$. Often the choice $\mathbf{z} = \mathbf{e}_k$ is made, where $\mathbf{e}_k$ 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 $\sigma^{(i+1)} = \sigma^{(i)} + 1/(\mathbf{z}^H\mathbf{M}\mathbf{y}^{(i)})$ if $\varphi(\mathbf{y}^{(i)}) = \mathbf{z}^H\mathbf{M}\mathbf{y}^{(i)}$; 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* $\mathbf{x}^{(0)}$*, then iterate:*
(1) *Choose* $\sigma^{(i)}$ *and* $\tau^{(i)}$.
(2) *Find* $\mathbf{y}^{(i)}$ *such that* $\|(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} - \mathbf{M}\mathbf{x}^{(i)}\| \leq \tau^{(i)}$.
(3) *Set* $\mathbf{x}^{(i+1)} = \mathbf{y}^{(i)}/\varphi(\mathbf{y}^{(i)})$.

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 $(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{M}\mathbf{x}^{(i)}$ to a prescribed accuracy. Since most iterative linear solvers have stopping conditions based on the residual we use the residual condition $\|(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} - \mathbf{M}\mathbf{x}^{(i)}\| \leq \tau^{(i)}$. 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 discussion.

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 $\mathbf{Ax} = \lambda \mathbf{Mx}$ is diagonalizable; that is, there exist an invertible matrix $\mathbf{V}$ and a diagonal matrix $\mathbf{\Lambda}$ (both possibly complex) such that

$$\mathbf{AV} = \mathbf{MV\Lambda}, \tag{2.1}$$

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

$$\mathbf{UA} = \mathbf{\Lambda UM}, \tag{2.2}$$

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

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

$$\begin{aligned} & (\mathbf{A} - \sigma\mathbf{M})\mathbf{V} = \mathbf{MV}(\mathbf{\Lambda} - \sigma\mathbf{I}) \\ \Leftrightarrow \quad & \mathbf{V}(\mathbf{\Lambda} - \sigma\mathbf{I})^{-1} = (\mathbf{A} - \sigma\mathbf{M})^{-1}\mathbf{MV}. \end{aligned} \tag{2.3}$$

Similarly we can use (2.2) to obtain

$$\begin{aligned} & \mathbf{U}(\mathbf{A} - \sigma\mathbf{M}) = (\mathbf{\Lambda} - \sigma\mathbf{I})\mathbf{UM} \\ \Leftrightarrow \quad & (\mathbf{\Lambda} - \sigma\mathbf{I})^{-1}\mathbf{U} = \mathbf{UM}(\mathbf{A} - \sigma\mathbf{M})^{-1}. \end{aligned} \tag{2.4}$$

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

$$\mathbf{x}^{(i)} = \alpha^{(i)}(c^{(i)}\mathbf{v}_1 + s^{(i)}\mathbf{w}^{(i)}), \tag{2.5}$$

where $\mathbf{w}^{(i)} \in \operatorname{span}(\mathbf{v}_2, \ldots, \mathbf{v}_n)$ and $\| \mathbf{UMw}^{(i)} \| = 1$. The splitting implies that $\mathbf{V}^{-1}\mathbf{w}^{(i)} \in \operatorname{span}(\mathbf{e}_2, \ldots, \mathbf{e}_n)$ and scaling implies that $\| \mathbf{V}^{-1}\mathbf{w}^{(i)} \| = \| \mathbf{UMw}^{(i)} \| = 1$. Defining

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

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

$$\mathbf{UMx}^{(i)} = \alpha^{(i)}c^{(i)}\mathbf{UMv}_1 + \alpha^{(i)}s^{(i)}\mathbf{UMw}^{(i)}, \tag{2.6}$$

and so

$$\begin{aligned} 1 = \frac{\|\mathbf{UMx}^{(i)}\|}{\alpha^{(i)}} &= \|c^{(i)}\mathbf{e}_1 + s^{(i)}\mathbf{UMw}^{(i)}\| \\ &= \left(|c^{(i)}|^2 + |s^{(i)}|^2\right)^{\frac{1}{2}} \end{aligned}$$

since $\mathbf{e}_1 \perp \mathbf{UMw}^{(i)}$. 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 $\mathbf{F}$, defined by

$$\mathbf{F} := (\mathbf{I} - \mathbf{e}_1\mathbf{e}_1^T)\mathbf{UM} \;=\; \mathbf{UM}(\mathbf{I} - \mathbf{v}_1\mathbf{u}_1^T\mathbf{M}), \tag{2.7}$$

and note that $\mathbf{Fv}_1 = \mathbf{0}$ and $\mathbf{Fv}_j = \mathbf{e}_j$, so that

$$(\mathbf{UM} - \mathbf{F})\mathbf{x}^{(i)} = \alpha^{(i)}c^{(i)}\mathbf{e}_1, \tag{2.8}$$

and

$$\mathbf{Fx}^{(i)} = \alpha^{(i)}s^{(i)}\mathbf{UMw}^{(i)}. \tag{2.9}$$

Hence $\|(\mathbf{UM} - \mathbf{F})\mathbf{x}^{(i)}\|$ measures the length of the component of $\mathbf{x}^{(i)}$ in the direction of $\mathbf{v}_1$ and $\mathbf{Fx}^{(i)}$ picks out the second term in (2.6). So it is natural to introduce as a measure for convergence of $\mathbf{x}^{(i)}$ to $\mathrm{span}(\mathbf{v}_1)$ the generalized tangent (cf. [6, section 2.1])

$$t^{(i)} \;:=\; \frac{|s^{(i)}|}{|c^{(i)}|} \;=\; \frac{\|\mathbf{Fx}^{(i)}\|}{\|(\mathbf{UM} - \mathbf{F})\mathbf{x}^{(i)}\|}. \tag{2.10}$$

Clearly $\| \frac{1}{c^{(i)}\alpha^{(i)}}\mathbf{x}^{(i)} - \mathbf{v}_1 \| = t^{(i)} \; \| \mathbf{w}^{(i)} \|$, and so $t^{(i)}$ measures the quality of the approximation of $\mathbf{x}^{(i)}$ to $\mathbf{v}_1$. Note that $t^{(i)}$ is independent of the factor $\alpha^{(i)}$ and that in the inverse iteration algorithm $\mathbf{x}^{(i)}$ is scaled so that $\varphi(\mathbf{x}^{(i)}) = 1$.

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

$$\varrho(\mathbf{x}) := \frac{\mathbf{x}^H\mathbf{Ax}}{\mathbf{x}^H\mathbf{Mx}} \tag{2.11}$$

and that

$$\varrho(\mathbf{x}^{(i)}) - \lambda_1 = \frac{(\mathbf{x}^{(i)})^H(\mathbf{A} - \lambda_1\mathbf{M})\mathbf{x}^{(i)}}{(\mathbf{x}^{(i)})^H\mathbf{Mx}^{(i)}} \;=\; O(|s^{(i)}|) \tag{2.12}$$

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

$$(\mathbf{A} - \varrho(\mathbf{x}^{(i)})\mathbf{M})\mathbf{x}^{(i)} = (\mathbf{A} - \lambda_1\mathbf{M})\mathbf{x}^{(i)} + (\lambda_1 - \varrho(\mathbf{x}^{(i)}))\mathbf{Mx}^{(i)} \tag{2.13}$$

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

$$\mathbf{r}^{(i)} := \big(\mathbf{A} - \varrho(\mathbf{x}^{(i)})\mathbf{M}\big)\mathbf{x}^{(i)} \tag{2.14}$$

satisfies

$$\|\mathbf{r}^{(i)}\| = O(|s^{(i)}|). \tag{2.15}$$

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

$$\sigma^{(i+1)} = \sigma^{(i)} + 1/\varphi(\mathbf{y}^{(i)}), \tag{3.1}$$

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

$$\sigma^{(i)} = \frac{\mathbf{z}^H \mathbf{A}\mathbf{x}^{(i)}}{\mathbf{z}^H \mathbf{M}\mathbf{x}^{(i)}}, \tag{3.2}$$

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

$$\sigma^{(i+1)} = \sigma^{(i)} + \frac{1}{\mathbf{z}^H\mathbf{M}\mathbf{y}^{(i)}} \;=\; \frac{\mathbf{z}^H\mathbf{A}\mathbf{x}^{(i+1)}}{\mathbf{z}^H\mathbf{M}\mathbf{x}^{(i+1)}}, \tag{3.3}$$

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 $\mathbf{x}^{(0)}$ 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

$$|\lambda_1 - \sigma^{(i)}| \leq \frac{1}{2}\,|\lambda_2 - \lambda_1| \quad \forall i, \tag{3.4}$$

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

$$\mathbf{d}^{(i)} := \mathbf{M}\mathbf{x}^{(i)} - (\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)}. \tag{3.5}$$

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

$$\varphi(\mathbf{y}^{(i)})\mathbf{x}^{(i+1)} = (\mathbf{A} - \sigma^{(i)}\mathbf{M})^{-1}(\mathbf{M}\mathbf{x}^{(i)} - \mathbf{d}^{(i)}). \tag{3.6}$$

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*

$$\|\mathbf{d}^{(i)}\| \quad\leq\quad \tau^{(i)} \quad<\quad \beta\,|\mathbf{u}_1^T\mathbf{M}\mathbf{x}^{(i)}|\;/\;\|\mathbf{u}_1\| \tag{3.7}$$

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

$$t^{(i+1)} \leq \frac{|\lambda_1 - \sigma^{(i)}|}{|\lambda_2 - \sigma^{(i)}|}\;\frac{|\alpha^{(i)}s^{(i)}| + \|\mathbf{U}\mathbf{d}^{(i)}\|}{(1-\beta)\;|\mathbf{u}_1^T\mathbf{M}\mathbf{x}^{(i)}|}. \tag{3.8}$$

*Proof.* Recall that $\mathbf{u}_1^T\mathbf{M}\mathbf{x}^{(i+1)} = \alpha^{(i+1)}c^{(i+1)}$, and $\mathbf{u}_1^T = \mathbf{e}_1^T\mathbf{U}$. Hence premultiplying the update equation (3.6) by $\mathbf{u}_1^T\mathbf{M}$ and using $\mathbf{U}\mathbf{M}(\mathbf{A}-\sigma^{(i)}\mathbf{M})^{-1} = (\mathbf{\Lambda}-\sigma^{(i)}\mathbf{I})\mathbf{U}$ (see (2.4)), we obtain

$$\begin{aligned}\varphi(\mathbf{y}^{(i)})\alpha^{(i+1)}c^{(i+1)} &= \mathbf{e}_1^T(\mathbf{\Lambda}-\sigma^{(i)}\mathbf{I})^{-1}\mathbf{U}(\mathbf{M}\mathbf{x}^{(i)}-\mathbf{d}^{(i)})\\ &= (\lambda_1-\sigma^{(i)})^{-1}\mathbf{u}_1^T(\mathbf{M}\mathbf{x}^{(i)}-\mathbf{d}^{(i)}).\end{aligned} \tag{3.9}$$

Further, using (3.7)

$$|\mathbf{u}_1^T\mathbf{M}\mathbf{x}^{(i)}| - |\mathbf{u}_1^T\mathbf{d}^{(i)}| \geq (1-\beta)\,|\mathbf{u}_1^T\mathbf{M}\mathbf{x}^{(i)}|\,. \tag{3.10}$$

Hence

$$\begin{aligned}|\varphi(\mathbf{y}^{(i)})||\alpha^{(i+1)}c^{(i+1)}| &\geq \frac{|\mathbf{u}_1^T\mathbf{M}\mathbf{x}^{(i)}| - |\mathbf{u}_1^T\mathbf{d}^{(i)}|}{|\lambda_1-\sigma^{(i)}|}\\ &\geq (1-\beta)\frac{|\mathbf{u}_1^T\mathbf{M}\mathbf{x}^{(i)}|}{|\lambda_1-\sigma^{(i)}|}.\end{aligned} \tag{3.11}$$

To obtain an upper bound on $|s^{(i+1)}|$ we apply $\mathbf{F}$, defined by (2.7), to (3.6) to obtain

$$\varphi(\mathbf{y}^{(i)})\mathbf{F}\mathbf{x}^{(i+1)} = (\mathbf{I}-\mathbf{e}_1\mathbf{e}_1^T)\mathbf{U}\mathbf{M}(\mathbf{A}-\sigma^{(i)}\mathbf{M})^{-1}(\mathbf{M}\mathbf{x}^{(i)}-\mathbf{d}^{(i)}) \tag{3.12}$$

and using (2.4),

$$\begin{aligned}\varphi(\mathbf{y}^{(i)})\mathbf{F}\mathbf{x}^{(i+1)} &= (\mathbf{I}-\mathbf{e}_1\mathbf{e}_1^T)(\mathbf{\Lambda}-\sigma^{(i)}\mathbf{I})^{-1}\mathbf{U}(\mathbf{M}\mathbf{x}^{(i)}-\mathbf{d}^{(i)})\\ &= (\mathbf{\Lambda}-\sigma^{(i)}\mathbf{I})^{-1}(\mathbf{I}-\mathbf{e}_1\mathbf{e}_1^T)\mathbf{U}(\mathbf{M}\mathbf{x}^{(i)}-\mathbf{d}^{(i)}).\end{aligned} \tag{3.13}$$

Taking norms we obtain

$$\begin{aligned}\|\varphi(\mathbf{y}^{(i)})\mathbf{F}\mathbf{x}^{(i+1)}\| &= \|(\mathbf{\Lambda}-\sigma^{(i)}\mathbf{I})^{-1}(\mathbf{I}-\mathbf{e}_1\mathbf{e}_1^T)\mathbf{U}(\mathbf{M}\mathbf{x}^{(i)}-\mathbf{d}^{(i)})\|\\ &\leq \|(\mathbf{\Lambda}-\sigma^{(i)}\mathbf{I})^{-1}(\mathbf{I}-\mathbf{e}_1\mathbf{e}_1^T)\|\;\|(\mathbf{I}-\mathbf{e}_1\mathbf{e}_1^T)\mathbf{U}(\mathbf{M}\mathbf{x}^{(i)}-\mathbf{d}^{(i)})\|\\ &\leq \frac{1}{|\lambda_2-\sigma^{(i)}|}\Big(\,|\alpha^{(i)}s^{(i)}| + \|(\mathbf{I}-\mathbf{e}_1\mathbf{e}_1^T)\mathbf{U}\mathbf{d}^{(i)}\|\,\Big).\end{aligned} \tag{3.14}$$

With $t^{(i+1)}$ defined by (2.9), and using (2.8), we have

$$\begin{aligned}t^{(i+1)} &= \frac{\|\varphi(\mathbf{y}^{(i)})\mathbf{F}\mathbf{x}^{(i+1)}\|}{\|\varphi(\mathbf{y}^{(i)})(\mathbf{U}\mathbf{M}-\mathbf{F})\mathbf{x}^{(i+1)}\|}\\ &\leq \frac{\|\varphi(\mathbf{y}^{(i)})\mathbf{F}\mathbf{x}^{(i+1)}\|}{|\varphi(\mathbf{y}^{(i)})\alpha^{(i+1)}c^{(i+1)}|}.\end{aligned}$$

Hence, using (3.10), (3.11), and (3.14),

$$t^{(i+1)} \leq \frac{|\lambda_1-\sigma^{(i)}|}{|\lambda_2-\sigma^{(i)}|}\,\frac{|\alpha^{(i)}s^{(i)}| + \|(\mathbf{I}-\mathbf{e}_1\mathbf{e}_1^T)\mathbf{U}\mathbf{d}^{(i)}\|}{|\mathbf{u}_1^T\mathbf{M}\mathbf{x}^{(i)}| - |\mathbf{u}_1^T\mathbf{d}^{(i)}|}. \qquad \square$$

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 $|\mathbf{u}_1^T\mathbf{M}\mathbf{x}^{(i)}| = \alpha^{(i)}\ |c^{(i)}|$ which is related to the cosine of the angle between $\mathbf{v}_1$ and $\mathbf{x}^{(i)}$, 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 $\mathbf{x}^{(i)}$.

In the case where $\mathbf{d}^{(i)} = \mathbf{0}$, we can take $\beta = 0$ in (3.7), and (3.8) reduces to $t^{(i+1)} \le |\frac{\lambda_1-\sigma^{(i)}}{\lambda_2-\sigma^{(i)}}|t^{(i)}$, the familiar expression when exact solves are employed. If (3.4) holds and $\|\mathbf{d}^{(i)}\| \le \tau^{(i)} \le C\ |s^{(i)}|$, as is the case if the solve tolerance is bounded by $\|\mathbf{r}^{(i)}\|$ defined in (2.14), then (3.8) indicates that we can expect Algorithm 2 to achieve quadratic convergence, the same asymptotic rate of convergence as the exact solves case. However, if (3.4) holds and $\|\mathbf{d}^{(i)}\| \le \tau^{(i)} \le$ *constant*, then we would expect a reduced rate of convergence in Algorithm 2. These expectations about the (outer) convergence rate of Algorithm 2 are made precise in the following section.

**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* $\mathbf{A}$, $\mathbf{M} \in \mathbb{R}^{n\times n}$ *with* $\mathbf{M}$ *symmetric positive definite. Let the generalized eigenvalue problem* $\mathbf{A}\mathbf{x} = \lambda\mathbf{M}\mathbf{x}$ *be diagonalizable and have simple eigenpair* $(\lambda_1, \mathbf{v}_1)$. *Further let* $\mathbf{x}^{(i)} = \alpha^{(i)}(c^{(i)}\mathbf{v}_1 + s^{(i)}\mathbf{w}^{(i)})$ *with* $|s^{(0)}| < 1$ *and let the shift updates satisfy*

$$|\lambda_1 - \sigma^{(i)}| \le \frac{|\lambda_1-\lambda_2|}{2}\ |s^{(i)}| \quad \forall i. \tag{3.15}$$

*Assume that, for* $\mathbf{d}^{(i)}$ *defined by* (3.5), $\|\mathbf{d}^{(i)}\| \le \tau^{(i)}$ *with*

$$\tau^{(i)} < \alpha^{(i)}\beta c^{(i)} /\ \|\mathbf{U}\|, \tag{3.16}$$

*where*

$$0 \quad \le \quad \beta \quad < \quad \frac{1-|s^{(0)}|}{2}. \tag{3.17}$$

*Then inexact inverse iteration as given in Algorithm* 2 *using a variable shift converges (at least) linearly,* $t^{(i+1)} \le qt^{(i)} \le q^{i+1}t^{(0)}$, *where*

$$q := \frac{|s^{(0)}| + \beta}{1-\beta} \quad < \quad 1. \tag{3.18}$$

*Proof.* With $|\lambda_1 - \sigma^{(i)}| \le \frac{1}{2}\ |\lambda_1-\lambda_2|\,|s^{(i)}|$ and hence $|\lambda_2 - \sigma^{(i)}| > \frac{1}{2}\ |\lambda_2 - \lambda_1|$, we have

$$\frac{|\lambda_1-\sigma^{(i)}|}{|\lambda_2-\sigma^{(i)}|} < |s^{(i)}|\ . \tag{3.19}$$

Thus, from (3.8),

$$
\begin{aligned}
t^{(i+1)} &\leq |s^{(i)}| \; \frac{|\alpha^{(i)} s^{(i)}| + \|\mathbf{U}\| \; \tau^{(i)}}{(1-\beta) \; |\alpha^{(i)} c^{(i)}|} \\
&\leq t^{(i)} \, \frac{|s^{(i)}| + \beta \, |c^{(i)}|}{1-\beta} \\
&\leq t^{(i)} \, \frac{|s^{(0)}| + \beta}{1-\beta}. \qquad (3.20)
\end{aligned}
$$

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*

$$
\tau^{(i)} \leq \alpha^{(i)} \min(\beta c^{(0)}/\|\mathbf{U}\|, \gamma \, |s^{(i)}|) \qquad (3.21)
$$

*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 $\|\mathbf{U}\|$ 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

$$
\begin{aligned}
-\Delta u + 5u_x + 5u_y = \lambda u \quad &\text{in } D := [0,1] \times [0,1], \\
u = 0 \quad &\text{on } \Gamma := \partial D,
\end{aligned}
$$

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 $Re(\lambda_j) > 61.8$. Note that the eigenvalue residual $\mathbf{r}^{(i)}$ 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

TABLE 3.1
*Generalized tangent $t^{(i)}$ and number of inner iterations $k^{(i)}$ for* **RQIf** (a) *and* (b) *and* **RQId** (c). *In* (a) $\tau_0 = 0.1$, *in* (b) $\tau_0 = 0.001$, *and in* (c) $\tau_0 = 0.2$ *and* $\tau_1 = 0.5$.

| | (a) | | (b) | | (c) | |
|---|---|---|---|---|---|---|
| | $t^{(i)}$ | $k^{(i-1)}$ | $t^{(i)}$ | $k^{(i-1)}$ | $t^{(i)}$ | $k^{(i-1)}$ |
| 0 | 5.0e-02 | | 5.0e-02 | | 5.0e-02 | |
| 1 | 9.0e-03 | 11 | 4.4e-04 | 23 | 1.6e-02 | 13 |
| 2 | 2.4e-04 | 19 | 8.0e-08 | 36 | 2.8e-05 | 35 |
| 3 | 4.6e-06 | 29 | 7.7e-12 | 51 | 2.9e-10 | 54 |
| 4 | 2.6e-08 | 37 | | | 6.8e-12 | 51 |
| 5 | 4.7e-11 | 47 | | | | |
| 6 | 1.0e-11 | 52 | | | | |
| $\sum k^{(i-1)}$ | | 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 $\mathbf{v}_1$), 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, $\sigma^{(i)} = \varrho(\mathbf{x}^{(i)})$ and $\tau^{(i)} = \tau_0\|\mathbf{M}\mathbf{x}^{(i)}\|$.

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

As $\|\mathbf{r}^{(i)}\| \,/\, |\varrho^{(i)}|$ is proportional to $|s^{(i)}|$ and $\|\mathbf{M}\mathbf{x}^{(i)}\|$ is proportional to $\alpha^{(i)}$ we expect according to Theorem 3.2 linear convergence for **RQIf** and according to Corollary 3.3 quadratic convergence for **RQId**.

In Table 3.1, cases (a) and (b) illustrate the behavior of **RQIf** with $\tau_0 = 0.1$ and 0.001, respectively. Case (c) gives results for **RQId**, that is, Rayleigh quotient shifts and a decreasing tolerance based on the eigenvalue residual (2.14). We present results for the approximation of $(\lambda_1, \mathbf{v}_1)$ and stop the entire calculation once the relative eigenvalue residual $\|\mathbf{r}^{(i)}\| \,/\varrho^{(i)}$ is smaller than $\tau_{outer} = 10^{-14}$.

*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, superlinear 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

$$(\mathbf{A} - \sigma\mathbf{M})\mathbf{y}^{(i)} = \mathbf{M}\mathbf{x}^{(i)} \tag{4.1}$$

[20] used the system

$$(\mathbf{A} - \sigma\mathbf{M})\mathbf{y}^{(i)} = \mathbf{x}^{(i)} \tag{4.2}$$

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 $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ using a preconditioner $\mathbf{P} \approx (\mathbf{A} - \sigma\mathbf{I})$, Simoncini and Eldén [21] solve

$$\mathbf{P}^{-1}(\mathbf{A} - \sigma\mathbf{I})\mathbf{y}^{(i)} = \mathbf{x}^{(i)} \tag{4.3}$$

rather than the obvious system

$$\mathbf{P}^{-1}(\mathbf{A} - \sigma\mathbf{I})\mathbf{y}^{(i)} = \mathbf{P}^{-1}\mathbf{x}^{(i)}. \tag{4.4}$$

The motivation for this alteration is that in (4.3) the right-hand side $\mathbf{x}^{(i)}$ is both close to a null vector of $\mathbf{P}^{-1}(\mathbf{A} - \sigma\mathbf{I})$ and close to a scaled version of the solution. The vector $\mathbf{P}^{-1}\mathbf{x}^{(i)}$ has neither of these properties. Here we combine the two ideas. Let $\mathbf{P} \approx (\mathbf{A} - \sigma\mathbf{M})$ be a preconditioner for use within GMRES. Given an approximate eigenvector $\mathbf{x}^{(i)}$ to obtain an improved eigendirection using preconditioned GMRES we solve

$$\mathbf{P}^{-1}(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{x}^{(i)} \tag{4.5}$$

rather than the obvious $\mathbf{P}^{-1}(\mathbf{A}-\sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{P}^{-1}\mathbf{M}\mathbf{x}^{(i)}$. As we shall show below, by changing the right-hand side from $\mathbf{P}^{-1}\mathbf{M}\mathbf{x}^{(i)}$ to $\mathbf{x}^{(i)}$ 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* $\mathbf{x}^{(0)}$, *then iterate:*
(1) *Choose* $\tau^{(i)}$, *and set* $\sigma^{(i)} = \varrho(\mathbf{x}^{(i)})$.
(2) *Find* $\mathbf{y}^{(i)}$ *such that* $\|\mathbf{x}^{(i)} - \mathbf{P}^{-1}(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)}\| \leq \tau^{(i)}$.
(3) *Set* $\mathbf{x}^{(i+1)} = \mathbf{y}^{(i)}/\varphi(\mathbf{y}^{(i)})$.

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

$$\mathbf{d}^{(i)} := \mathbf{x}^{(i)} - \mathbf{P}^{-1}(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} \tag{4.6}$$

so that the inexact solve step can be written as

$$(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{P}\mathbf{x}^{(i)} - \mathbf{P}\mathbf{d}^{(i)}, \tag{4.7}$$

which should be compared with the inexact solve step

$$(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{M}\mathbf{x}^{(i)} - \mathbf{d}^{(i)} \tag{4.8}$$

in section 3. From (4.8) we obtain

$$\varphi(\mathbf{y}^{(i)})\mathbf{x}^{(i+1)} = (\mathbf{A} - \sigma^{(i)}\mathbf{M})^{-1}\mathbf{P}(\mathbf{x}^{(i)} - \mathbf{d}^{(i)}) \tag{4.9}$$

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

$$\|\mathbf{d}^{(i)}\| \quad \leq \quad \tau^{(i)} \quad \leq \quad \beta' \, |\mathbf{u}_1^T\mathbf{P}\mathbf{x}^{(i)}| \,/\, \|\mathbf{U}\mathbf{P}\| \tag{4.10}$$

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

$$|\mathbf{u}_1^T\mathbf{P}\mathbf{x}^{(i)}| - |\mathbf{u}_1^T\mathbf{P}\mathbf{d}^{(i)}| \geq (1-\beta') \, |\mathbf{u}_1^T\mathbf{P}\mathbf{x}^{(i)}| \, . \tag{4.11}$$

Next, we introduce the expression

$$T_P(\mathbf{z}) := \frac{\|(\mathbf{I} - \mathbf{e}_1\mathbf{e}_1^T)\mathbf{U}\mathbf{P}\mathbf{z}\|}{|\mathbf{u}_1^T\mathbf{P}\mathbf{z}|}, \tag{4.12}$$

where $\mathbf{z} \in \mathbb{C}^n$. By analogy with (2.7) and (2.10), $T_P(\mathbf{z})$ looks like a generalized tangent with respect to $\mathbf{P}$ rather than $\mathbf{M}$. However, for a general preconditioner $T_P(\mathbf{v}_1) \neq 0$. In fact, $T_P(\mathbf{v}_1)$ measures the effect of $\mathbf{P}$ on the eigenvector $\mathbf{v}_1$, and we shall see in Theorem 4.2 that large values of $T_P(\mathbf{v}_1)$ will slow down or possibly destroy the convergence of Algorithm 3. Note that, under (4.11),

$$T_P(\mathbf{x}^{(i)} - \mathbf{d}^{(i)}) \leq \frac{1}{1-\beta'} \left( T_P(\mathbf{x}^{(i)}) + \frac{\|\mathbf{U}\mathbf{P}\mathbf{d}^{(i)}\|}{|\mathbf{u}_1^T\mathbf{P}\mathbf{x}^{(i)}|} \right). \tag{4.13}$$

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*

$$\begin{aligned} t^{(i+1)} &\leq \frac{|\lambda_1 - \sigma^{(i)}|}{|\lambda_2 - \sigma^{(i)}|} \, T_P(\mathbf{x}^{(i)} - \mathbf{d}^{(i)}) \\ &\leq |s^{(i)}| \; T_P(\mathbf{x}^{(i)} - \mathbf{d}^{(i)}), \end{aligned} \tag{4.14}$$

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

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

$$\begin{aligned} t^{(i+1)} &= \frac{\|\mathbf{F}\varphi(\mathbf{y}^{(i)})\mathbf{x}^{(i+1)}\|}{\|(\mathbf{U}\mathbf{M} - \mathbf{F})\varphi(\mathbf{y}^{(i)})\mathbf{x}^{(i+1)}\|} \\ &= \frac{\|\mathbf{F}(\mathbf{A} - \sigma^{(i)}\mathbf{M})^{-1}(\mathbf{P}\mathbf{x}^{(i)} - \mathbf{P}\mathbf{d}^{(i)})\|}{\|(\mathbf{U}\mathbf{M} - \mathbf{F})(\mathbf{A} - \sigma^{(i)}\mathbf{M})^{-1}(\mathbf{P}\mathbf{x}^{(i)} - \mathbf{P}\mathbf{d}^{(i)})\|} \\ &= \frac{\|(\mathbf{I} - \mathbf{e}_1\mathbf{e}_1^T)\mathbf{U}\mathbf{M}(\mathbf{A} - \sigma^{(i)}\mathbf{M})^{-1}(\mathbf{P}\mathbf{x}^{(i)} - \mathbf{P}\mathbf{d}^{(i)})\|}{|\mathbf{e}_1^T\mathbf{U}\mathbf{M}(\mathbf{A} - \sigma^{(i)}\mathbf{M})^{-1}(\mathbf{P}\mathbf{x}^{(i)} - \mathbf{P}\mathbf{d}^{(i)})|} \\ &= \frac{\|(\mathbf{I} - \mathbf{e}_1\mathbf{e}_1^T)(\mathbf{\Lambda} - \sigma^{(i)}\mathbf{I})^{-1}\mathbf{U}(\mathbf{P}\mathbf{x}^{(i)} - \mathbf{P}\mathbf{d}^{(i)})\|}{|\mathbf{e}_1^T(\mathbf{\Lambda} - \sigma^{(i)}\mathbf{I})^{-1}\mathbf{U}(\mathbf{P}\mathbf{x}^{(i)} - \mathbf{P}\mathbf{d}^{(i)})|} \\ &\leq \frac{|\lambda_1 - \sigma^{(i)}|}{|\lambda_2 - \sigma^{(i)}|} \, \frac{\|(\mathbf{I} - \mathbf{e}_1\mathbf{e}_1^T)\mathbf{U}\mathbf{P}(\mathbf{x}^{(i)} - \mathbf{d}^{(i)})\|}{|\mathbf{e}_1^T\mathbf{U}\mathbf{P}(\mathbf{x}^{(i)} - \mathbf{d}^{(i)})|}, \end{aligned}$$

TABLE 4.1
*Generalized tangent $t^{(i)}$ and number of inner iterations $k^{(i)}$ for* **RQIf** (a) *and* **RQImodrhs** (b) *with $\tau_0 = 0.05$ for both methods.*

| | (a) | | (b) | |
|---|---|---|---|---|
| | $t^{(i)}$ | $k^{(i-1)}$ | $t^{(i)}$ | $k^{(i-1)}$ |
| 0 | 2.0e-02 | | 2.0e-02 | |
| 1 | 1.6e-02 | 30 | 1.6e-02 | 30 |
| 2 | 2.9e-05 | 41 | 1.2e-04 | 37 |
| 3 | 4.7e-08 | 47 | 9.8e-07 | 37 |
| 4 | 2.0e-08 | 47 | 4.4e-08 | 36 |
| 5 | | | 1.7e-08 | 24 |
| $\sum k^{(i-1)}$ | | 165 | | 164 |

from which the required result follows. $\square$

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 $T_P(\mathbf{v}_1) \neq 0$ and*

$$q \ := \ \frac{1}{1-\beta'}\,(2T_P(\mathbf{v}_1) + \beta') \ < \ 1. \tag{4.15}$$

*Then, for $\mathbf{x}^{(0)}$ close enough to $\mathbf{v}_1$, Algorithm* 3 *converges linearly with $t^{(i+1)} \leq q t^{(i)}$.*

*Proof.* Due to the condition on $\tau^{(i)}$, (4.10), we can use (4.13) and (4.10) (again) to give $T_P(\mathbf{x}^{(i)} + \mathbf{d}^{(i)}) \leq (1-\beta')^{-1}(T_P(\mathbf{x}^{(i)}) + \beta')$. Hence it remains to show that $T_P(\mathbf{x}^{(i)}) \leq 2T_P(\mathbf{v}_1)$, which is valid for $\mathbf{x}^{(0)}$ close enough to $\mathbf{v}_1$ as $T_P(\mathbf{v}_1) \neq 0$. $\square$

Lemma 4.1 and Theorem 4.2 show that the quantity $T_P(\mathbf{v}_1)$ plays an important role in the convergence of Algorithm 3, and ideally $T_P(\mathbf{v}_1)$ should be small. In practical situations we will have little knowledge of the effect of $\mathbf{P}$ on $\mathbf{v}_1$, but it is clear that if $\mathbf{u}_1^T\mathbf{P}\mathbf{v}_1$ is small, and hence $T_P(\mathbf{v}_1)$ is large; then Algorithm 3 may converge slowly or may possibly fail to converge. Note that we ignore the unlikely case $T_P(\mathbf{v}_1) = 0$ in Theorem 4.2, though in this case one could recover quadratic convergence using a decreasing tolerance. We present numerical values for $T_P(\mathbf{v}_1)$ 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 $\sigma^{(i)} = \varrho(\mathbf{x}^{(i)})$ and tolerance $\tau^{(i)} = \tau_0\,\|\mathbf{P}\mathbf{x}^{(i)}\|$.

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

TABLE 4.2
*Generalized tangent $t^{(i)}$ for* **RQImodrhs** *with $\tau_0 = 0.01$ using two different preconditioners. In* (a) *milu$(\mathbf{A}, 0.1)$, where $T_P(\mathbf{v}_1) = 0.34$, and in* (b) *milu$(\mathbf{A} - 320\mathbf{M}, 10^{-4})$, where $T_P(\mathbf{v}_1) = 0.045$.*

| | (a) | (b) |
|---|---|---|
| | $t^{(i)}$ | $t^{(i)}$ |
| 0 | 2.0e-02 | 2.0e-02 |
| 1 | 3.1e-04 | 1.9e-04 |
| 2 | 4.7e-05 | 5.8e-07 |
| 3 | 2.6e-06 | 1.5e-09 |
| 4 | 1.3e-07 | |
| 5 | 1.1e-08 | |

has been modified **RQImodrhs** still provides a linearly converging method as stated in Theorem 4.2. Further, the number of inner iterations used at each outer iteration by **RQImodrhs** does not increase with $i$, which leads to an efficient iteration process. (The link between the outer convergence and the cost of the inner solves using GMRES is discussed further in [4].) We also observe, however, that **RQImodrhs** requires more outer iterations. This is to be expected from the convergence theory because of the nonzero term $T_P(\mathbf{v}_1)$ in (4.15) and is observed in other experiments; see Table 8.6 in [4]. Note that the choices for $\tau_0$ in Example 2 are not optimal for either method. For **RQImodrhs** the optimal value (that is, the value producing the smallest total number of inner iterations) is $\tau_0 = 0.1$, and for **RQIf** the optimal value is $\tau_0 = 0.001$. However, there was little difference in the performance of the methods. In both cases the total number of inner iterations was around 130.

We remark that in our experience with several different examples for the generalized nonsymmetric eigenvalue problem the choice of the constant $\tau_0$ as used in the bound on the tolerance is important for both the convergence and efficiency of Algorithm 3. This is in contrast to the standard symmetric eigenvalue problem where the corresponding algorithms are less sensitive to the choice of $\tau_0$, as reported in [3].

Next, we provide an example to demonstrate the effect of $T_P(\mathbf{v}_1)$ on the rate of convergence.

*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 $T_P(\mathbf{v}_1)$ 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 $\mathbf{A}$ using a drop tolerance of 0.1; we denote this by milu$(\mathbf{A}, 0.1)$. The other preconditioner, which we use in case (b), is also a modified incomplete LU decomposition constructed now from the shifted system $\mathbf{A} - 320\mathbf{M}$ using a drop tolerance of $10^{-4}$ (milu$(\mathbf{A} - 320\mathbf{M}, 10^{-4})$). 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 $T_P(\mathbf{v}_1) = 0.34$ and in (b) the "shifted" preconditioner which has $T_P(\mathbf{v}_1) = 0.045$.

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 $\tau_0 \ll T_P(\mathbf{v}_1)$, 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. In Algorithms 1 and 2 the 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 $T_P(\mathbf{v}_1)$ 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


[1] W. E. Arnoldi, *The principle of minimized iteration in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.

[2] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, SIAM, Philadelphia, 2000.

[3] J. Berns-Müller, I. G. Graham, and A. Spence, *Inverse iteration and inexact solves*, Linear Algebra Appl., to appear.

[4] J. Berns-Müller and A. Spence, *Inexact Inverse Iteration and GMRES*, Tech. report maths0507, University of Bath, Bath, UK, 2005.

[5] G. H. Golub and Q. Ye, *Inexact preconditioned conjugate gradient method with inner-outer iteration*, SIAM J. Sci. Comput., 21 (1999), pp. 1305–1320.

[6] G. H. Golub and Q. Ye, *Inexact inverse iteration for generalized eigenvalue problems*, BIT, 40 (2000), pp. 671–684.

[7] Z. Jia, *A refined iterative algorithm based on the block Arnoldi process for large unsymmetric eigenproblems*, Linear Algebra Appl., 270 (1998), pp. 171–189.

[8] H.-J. Jung, M.-C. Kim, and I.-W. Lee, *An improved subspace iteration method with shifting*, Comput. & Structures, 70 (1999), pp. 625–633.

[9] A. V. Knyazev, *Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method*, SIAM J. Sci. Comput., 23 (2001), pp. 517–541.

[10] Y.-L. Lai, K.-Y. Lin, and L. Wen-Wei, *An inexact inverse iteration for large sparse eigenvalue problems*, Numer. Linear Algebra Appl., 1 (1997), pp. 1–13.

[11] R. B. Lehoucq, *Analysis and Implementation of an Implicitly Restarted Arnoldi Iteration*, Ph.D. thesis, Rice University, Houston, TX, 1995.

[12] R. B. Lehoucq, *Implicitly restarted Arnoldi methods and subspace iteration*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 551–562.

[13] R. B. Morgan and D. S. Scott, *Generalizations of Davidson's method for computing eigenvalues of sparse symmetric matrices*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 817–825.

[14] K. Neymeyr, *A geometric theory for preconditioned inverse iteration* I: *Extrema of the Rayleigh quotient*, Linear Algebra Appl., 322 (2001), pp. 61–85.

[15] K. Neymeyr, *A geometric theory for preconditioned inverse iteration* II: *Convergence estimates*, Linear Algebra Appl., 322 (2001), pp. 87–104.

[16] B. N. Parlett, *The Rayleigh quotient iteration and some generalizations for nonnormal matrices*, Math. Comp., 28 (1974), pp. 679–693.

[17] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.

[18] G. Peters and J. H. Wilkinson, *Inverse iteration, ill-conditioned equations and Newton's method*, SIAM Rev., 21 (1979), pp. 339–360.

[19] A. Ruhe and T. Wiberg, *The method of conjugate gradients used in inverse iteration*, BIT, 12 (1972), pp. 543–554.

[20] D. S. Scott, *Solving sparse symmetric generalized eigenvalue problems without factorization*, SIAM J. Numer. Anal., 18 (1981), pp. 102–110.

[21] V. Simoncini and L. Eldén, *Inexact Rayleigh quotient-type methods for eigenvalue computations*, BIT, 42 (2002), pp. 159–182.

[22] G. L. G. Sleijpen and H. A. van der Vorst, *A Jacobi–Davidson iteration method for linear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 401–425.

[23] P. Smit and M. H. C. Paardekooper, *The effects of inexact solvers in algorithms for symmetric eigenvalue problems*, Linear Algebra Appl., 287 (1999), pp. 337–357.

[24] X. Wang and J. Zhou, *An accelerated subspace iteration method for generalized eigenproblems*, Comput. & Structures, 71 (1999), pp. 293–301.

[25] J. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, UK, 1965.