Electronic Transactions on Numerical Analysis.
Volume 28, pp. 40-64, 2007.

ISSN 1068-9613.

ETNA
Kent State University
etna@mcs.kent.edu

CONVERGENCE THEORY FOR INEXACT INVERSE ITERATION APPLIED TO THE GENERALISED NONSYMMETRIC EIGENPROBLEM*

MELINA A. FREITAG† AND ALASTAIR SPENCE†

Dedicated to Gene Golub on the occasion of his 75th birthday

Abstract. In this paper we consider the computation of a finite eigenvalue and corresponding right eigenvector of a large sparse generalised eigenproblem $\mathbf{Ax} = \lambda\mathbf{Mx}$ using inexact inverse iteration. Our convergence theory is quite general and requires few assumptions on $\mathbf{A}$ and $\mathbf{M}$. In particular, there is no need for $\mathbf{M}$ to be symmetric positive definite or even nonsingular. The theory includes both fixed and variable shift strategies, and the bounds obtained are improvements on those currently in the literature. In addition, the analysis developed here is used to provide a convergence theory for a version of inexact simplified Jacobi-Davidson. Several numerical examples are presented to illustrate the theory: including applications in nuclear reactor stability, with $\mathbf{M}$ singular and nonsymmetric, the linearised Navier-Stokes equations and the bounded finline dielectric waveguide.

Key words. Inexact inverse iteration, nonsymmetric generalised eigenproblem

AMS subject classifications. Primary 65F15. Secondary 15A18, 65F50.

1. Introduction.

Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ and $\mathbf{M} \in \mathbb{C}^{n\times n}$ be large, sparse and complex. We consider the computation of a simple, finite eigenvalue and corresponding eigenvector of the generalised eigenvalue problem

$$\mathbf{Ax} = \lambda\mathbf{Mx}, \quad \mathbf{x} \neq \mathbf{0}, \tag{1.1}$$

using inverse iteration with iterative solves of the resulting linear systems

$$(\mathbf{A} - \sigma\mathbf{M})\mathbf{y} = \mathbf{Mx}.$$

Here σ is a complex shift chosen to be close to the desired eigenvalue. We call this method "inexact inverse iteration", and consider the case where the linear system is solved to some prescribed tolerance only. It is well known that, using exact solves, inverse iteration achieves linear convergence with a fixed shift and quadratic convergence for a Rayleigh quotient shift; see [18] and [17]. For more information about inverse iteration we refer to the classic articles [7] and [19], and the more recent survey [11]. In this paper, we shall explore, under minimal assumptions, convergence rates attained by inexact inverse iteration, illustrate the theory with reference to some physical examples, and obtain a convergence result for a version of the inexact Jacobi-Davidson method.

The paper by Golub and Ye [8] provides a convergence theory of inexact inverse iteration for a fixed shift strategy for nonsingular $\mathbf{M}$ with $\mathbf{M}^{-1}\mathbf{A}$ diagonalisable. Linear convergence is proved if a suitable solve tolerance is chosen to decrease linearly. An early paper, which also considers inexact inverse iteration applied to a diagonalisable problem is the one by Lai et al. [12]. They provide a theory for the standard eigenproblem with a fixed shift strategy and obtain linear convergence for both the eigenvalue and the eigenvector if the solve tolerance decreases depending on a quantity containing unknown parameters. They also give numerical results on a transformed generalised eigenvalue problem. In [3] a convergence theory is given for Rayleigh quotient shifts assuming $\mathbf{M}$ is symmetric positive definite. Following [8], the convergence theory in [3] uses a decomposition in terms of the right eigenvectors. One result

*Received December 11, 2006. Accepted for publication March 26, 2007. Recommended by A. Wathen.
†Department of Mathematical Sciences, University of Bath, Claverton Down, BA2 7AY, United Kingdom ({m.freitag,as}@maths.bath.ac.uk).

ETNA
Kent State University
etna@mcs.kent.edu

in [3] is that for a variable shift strategy, the linear systems need not be solved accurately to obtain a convergent method.

An alternative approach to the convergence theory of inexact inverse iteration for general $\mathbf{A}$ and $\mathbf{M}$ is presented in [5], where it is shown that inexact inverse iteration is a modified Newton method if a certain normalisation for the eigenvector and a special update of the shift is used. The only assumptions are that the desired eigenvalue is simple and finite. It is then shown that inexact inverse iteration converges linearly for close enough starting values, and that for a decreasing tolerance quadratic convergence is attained, as would be expected from a theory based on Newton's method. The advantage of this approach is that an eigenvector expansion is not used and so error bounds do not contain a term involving the norm of the inverse of the matrix of eigenvectors, as appears in [8] and [3]. The disadvantage is that the convergence rate itself depends on the norm of the inverse of the Jacobian, which is hard to estimate in practice.

In this paper we consider a quite general setting, where $\mathbf{A}$ and $\mathbf{M}$ are nonsymmetric matrices with both $\mathbf{A}$ and $\mathbf{M}$ allowed to be singular, but without a common null vector. We only assume that the sought eigenpair $(\lambda_1, \mathbf{x}_1)$ is simple, well-separated and finite. We provide a convergence theory for inexact inverse iteration applied to this generalised eigenproblem for both fixed and variable shifts. This theory extends the results of [5], since this new theory holds for any shift, not just the shift that gives the equivalence of Newton's method to inverse iteration. Also, the convergence rate is seen to depend on how close the sought eigenvalue is to the rest of the spectrum, a natural result that is somewhat hidden in the theory in [5]. We use a decomposition that allows us to consider nondiagonalisable problems where $\mathbf{M}$ may be singular. To be precise, we use a splitting of the approximate right eigenvector in terms of the exact right eigenvector and a basis of a right invariant subspace. This is an approach used by Stewart [27] to provide a perturbation theory of invariant subspaces, and allows us to overcome the theoretical dependence of the allowed solve tolerance on the basis of eigenvectors, which appears in [8] and [3]. If a decreasing solve tolerance is required, then we take it to be proportional to the eigenvalue residual, as done in [3].

Inexact inverse iteration applied to the symmetric standard eigenvalue problem has been considered by [25] and [2]. Both approaches use a natural orthogonal splitting and consider fixed and Rayleigh quotient shifts. Linear convergence for the fixed shift and locally cubic convergence for the Rayleigh quotient shift is obtained if the solve tolerance is chosen to decrease in a certain way. The approach in [2] is more natural, since the solve tolerance is chosen to decrease in proportion to the eigenvalue residual. Simoncini and Eldén observe in [20], that inexact Rayleigh-quotient iteration is equivalent to a Newton method on a unit sphere and also discuss a reformulation for efficient iterative solves. Notay [15] considers the computation of the smallest eigenvalue and associated eigenvector for a Hermitian positive definite generalised eigenproblem using inexact Rayleigh quotient iteration. In practice, subspace methods like shift-invert (restarted) Arnoldi and Jacobi-Davidson are more likely to be used in eigenvalue computations, though inexact inverse iteration has proved to be a useful tool in improving estimates obtained from inexact shift-invert Arnoldi's method with very coarse tolerances; see [9].

It is well known that there is a close connection between inverse iteration and the Jacobi-Davidson method; see [21, 22, 24]. We shall use the convergence theory developed here for inexact inverse iteration applied to (1.1) to provide a convergence theory for a version of inexact simplified Jacobi-Davidson.

The paper is organised as follows. In section 2 standard results on the generalised eigenproblem are summarised and a generalised Rayleigh quotient is discussed. Section 3 provides the main result of the paper; a new convergence measure is introduced and the main conver-

ETNA
Kent State University
etna@mcs.kent.edu

gence result for inexact inverse iteration applied to the generalised nonhermitian eigenproblem is stated and proved. Section 4 contains some additional convergence results. In section 5 we give numerical tests on examples arising from modelling of a nuclear reactor and the linearised incompressible Navier-Stokes equations. Section 6 presents a convergence analysis for the inexact simplified Jacobi-Davidson method and provides some numerical results to illustrate the theory.

Throughout this paper we use $\|\cdot\| = \|\cdot\|_2$.

2. Standard results on the generalised eigenproblem. In order to state convergence results for Algorithm 3.1 we need some results about the generalised eigenproblem. First recall that the eigenvalues of (1.1) are given by $\lambda(\mathbf{A},\mathbf{M}) := \{z \in \mathbb{C} :\ \det(\mathbf{A} - z\mathbf{M}) = 0\}$.

We use the following theorem for a canonical form of (1.1), which is a generalisation of the Schur Decomposition of the standard eigenproblem.

THEOREM 2.1 (Generalised Schur Decomposition). *If $\mathbf{A} \in \mathbb{C}^{n\times n}$ and $\mathbf{M} \in \mathbb{C}^{n\times n}$, then there exist unitary matrices $\mathbf{Q}$ and $\mathbf{Z}$ such that $\mathbf{Q}^H\mathbf{A}\mathbf{Z} = \mathbf{T}$ and $\mathbf{Q}^H\mathbf{M}\mathbf{Z} = \mathbf{S}$ are upper triangular. If for some j, t_{jj} and s_{jj} are both zero, then $\lambda(\mathbf{A},\mathbf{M}) = \mathbb{C}$. If $s_{jj} \neq 0$, then $\lambda(\mathbf{A},\mathbf{M}) = \{t_{jj}/s_{jj}\}$, otherwise, the jth eigenvalue of problem (1.1) is an infinite eigenvalue.*

Proof. See [6, page 377]. □

Using this theorem, together with the fact that $\mathbf{Q}$ and $\mathbf{Z}$ can be chosen such that s_{jj} and t_{jj} appear in any order along the diagonal, we can introduce the following partition of the eigenproblem in canonical form:

$$\mathbf{Q}^H\mathbf{A}\mathbf{Z} = \begin{bmatrix} t_{11} & \mathbf{t}_{12}^H \\ \mathbf{0} & \mathbf{T}_{22} \end{bmatrix} \quad \text{and} \quad \mathbf{Q}^H\mathbf{M}\mathbf{Z} = \begin{bmatrix} s_{11} & \mathbf{s}_{12}^H \\ \mathbf{0} & \mathbf{S}_{22} \end{bmatrix}, \tag{2.1}$$

where $\mathbf{T}_{22} \in \mathbb{C}^{(n-1)\times(n-1)}$ and $\mathbf{S}_{22} \in \mathbb{C}^{(n-1)\times(n-1)}$. If λ_1, the desired eigenvalue, is finite, then $s_{11} \neq 0$ and $\lambda_1 = t_{11}/s_{11}$. The factorisation (2.1) provides an orthogonal similarity transform, but in order to decompose the problem for the convergence analysis, we make a further transformation to block diagonalise the problem. To this end we define the linear transformation $\Phi :\ \mathbb{C}^{(n-1)\times 2} \to \mathbb{C}^{2\times(n-1)}$ by

$$\Phi(\mathbf{h},\mathbf{g}) := (t_{11}\mathbf{h}^H - \mathbf{g}^H\mathbf{T}_{22}, s_{11}\mathbf{h}^H - \mathbf{g}^H\mathbf{S}_{22}), \tag{2.2}$$

where $\mathbf{g} \in \mathbb{C}^{(n-1)\times 1}$ and $\mathbf{h} \in \mathbb{C}^{(n-1)\times 1}$. (This transformation is a simplification of that suggested by Stewart in [26].)

LEMMA 2.2. *The operator Φ from (2.2) is nonsingular if and only if*

$$\lambda_1 = \frac{t_{11}}{s_{11}} \notin \lambda(\mathbf{T}_{22},\mathbf{S}_{22}).$$

Proof. See [26, Theorem 4.1]. □

Hence Φ is nonsingular if and only if λ_1 is a simple eigenvalue of (1.1). With Lemma 2.2 we can prove the following result.

LEMMA 2.3. *If the operator Φ from (2.2) is nonsingular and $\mathbf{G}$, $\mathbf{H}$ are defined by*

$$\mathbf{G} = \begin{bmatrix} 1 & \mathbf{g}_{12}^H \\ \mathbf{0} & \mathbf{I}_{n-1} \end{bmatrix} \quad \textit{and} \quad \mathbf{H} = \begin{bmatrix} 1 & \mathbf{h}_{12}^H \\ \mathbf{0} & \mathbf{I}_{n-1} \end{bmatrix},$$

where $\Phi(\mathbf{h}_{12},\mathbf{g}_{12}) = (-\mathbf{t}_{12}^H, -\mathbf{s}_{12}^H)$, then, with $\mathbf{T}$ and $\mathbf{S}$ defined in Theorem 2.1,

$$\mathbf{G}^{-1}\mathbf{T}\mathbf{H} = \textit{diag}(t_{11},\mathbf{T}_{22}) \quad \textit{and} \quad \mathbf{G}^{-1}\mathbf{S}\mathbf{H} = \textit{diag}(s_{11},\mathbf{S}_{22}).$$

Furthermore,

$$\|\mathbf{H}\|^2 = \|\mathbf{H}^{-1}\|^2 = C_{\|\mathbf{h}_{12}\|}, \quad C_{\|\mathbf{h}_{12}\|} := (\|\mathbf{h}_{12}\|^2 + \sqrt{\|\mathbf{h}_{12}\|^4 + 4\,\|\mathbf{h}_{12}\|^2} + 2)/2, \tag{2.3}$$

with similar results for $\|\mathbf{G}\|^2$ *and* $\|\mathbf{G}^{-1}\|^2$.

Proof. If Φ is nonsingular, then the vectors $\mathbf{g}_{12}$ and $\mathbf{h}_{12}$ exist and a simple calculation gives $\mathbf{G}^{-1}\mathbf{T}\mathbf{H} = \mathrm{diag}(t_{11}, \mathbf{T}_{22})$ and $\mathbf{G}^{-1}\mathbf{S}\mathbf{H} = \mathrm{diag}(s_{11}, \mathbf{S}_{22})$. Result (2.3) follows by direct calculation of the spectral radius of $\mathbf{H}^H\mathbf{H}$. ◻

Note that $C_{\|\mathbf{h}_{12}\|}$ and $C_{\|\mathbf{g}_{12}\|}$ measure the conditioning of the eigenvalue λ_1, with large values of $C_{\|\mathbf{h}_{12}\|}$ and $C_{\|\mathbf{g}_{12}\|}$ implying a poorly conditioned problem. We shall see in section 3 that $\|\mathbf{g}_{12}\|$ and $\|\mathbf{h}_{12}\|$ appear in the bounds in the convergence theory.

Combining Theorem 2.1 and Lemma 2.3 gives the following corollary.

COROLLARY 2.4. *Define*

$$\mathbf{U} = \mathbf{Q}\mathbf{G} \tag{2.4}$$

and

$$\mathbf{X} = \mathbf{Z}\mathbf{H}. \tag{2.5}$$

Then both $\mathbf{U}$ *and* $\mathbf{X}$ *are nonsingular and we can block factorise* $\mathbf{A} - \lambda\mathbf{M}$ *as*

$$\mathbf{U}^{-1}(\mathbf{A} - \lambda\mathbf{M})\mathbf{X} = \begin{bmatrix} t_{11} & \mathbf{0}^H \\ \mathbf{0} & \mathbf{T}_{22} \end{bmatrix} - \lambda \begin{bmatrix} s_{11} & \mathbf{0}^H \\ \mathbf{0} & \mathbf{S}_{22} \end{bmatrix}. \tag{2.6}$$

For our purposes, decomposition (2.6) has advantages over the Schur factorisation (2.1), since (2.6) allows the eigenvalue problem $\mathbf{A}\mathbf{x} = \lambda\mathbf{M}\mathbf{x}$ to be split into two problems. The first problem is the trivial $\lambda\, t_{11} = s_{11}$. The second problem arising from the $(n-1)\times(n-1)$ block is that of finding $\lambda(\mathbf{T}_{22}, \mathbf{S}_{22})$, which contains the $(n-1)$ eigenvalues excluding λ_1. From (2.6) we have

$$(\mathbf{A} - \lambda_1\mathbf{M})\mathbf{x}_1 = \mathbf{0} \quad \text{and} \quad \mathbf{u}_1^H(\mathbf{A} - \lambda_1\mathbf{M}) = \mathbf{0}, \tag{2.7}$$

where $\lambda_1 = \dfrac{t_{11}}{s_{11}}$ is an eigenvalue of (1.1), with corresponding right and left eigenvectors, $\mathbf{x}_1 = \mathbf{X}\mathbf{e}_1$ and $\mathbf{u}_1 = \mathbf{U}^{-H}\mathbf{e}_1$, where $\mathbf{e}_1$ is the first canonical vector.

Note that $\lambda_1 = \dfrac{t_{11}}{s_{11}}$ is a finite eigenvalue if and only if

$$\mathbf{u}_1^H\mathbf{M}\mathbf{x}_1 \neq 0, \tag{2.8}$$

since, by (2.6) and the special structure of $\mathbf{G}$ and $\mathbf{H}$ in Lemma 2.3, we have

$$s_{11} = \mathbf{q}_1^H\mathbf{M}\mathbf{z}_1 = \mathbf{e}_1^H\mathbf{Q}^H\mathbf{M}\mathbf{Z}\mathbf{e}_1 = \mathbf{e}_1^H\mathbf{G}^{-1}\mathbf{Q}^H\mathbf{M}\mathbf{Z}\mathbf{H}\mathbf{e}_1 = \mathbf{e}_1^H\mathbf{U}^{-1}\mathbf{M}\mathbf{X}\mathbf{e}_1 = \mathbf{u}_1^H\mathbf{M}\mathbf{x}_1.$$

Next, for $\mathbf{x} \in \mathbb{C}^n$, with $\mathbf{x}^H\mathbf{M}\mathbf{x} \neq 0$, we define the Rayleigh quotient by $\dfrac{\mathbf{x}^H\mathbf{A}\mathbf{x}}{\mathbf{x}^H\mathbf{M}\mathbf{x}}$. Note that $\mathbf{x}^H\mathbf{M}\mathbf{x} \neq 0$ does not generally hold, unless $\mathbf{M}$ is positive definite. Therefore, instead of the Rayleigh quotient, we consider the related generalised Rayleigh quotient

$$\frac{\mathbf{c}^H\mathbf{A}\mathbf{x}}{\mathbf{c}^H\mathbf{M}\mathbf{x}}, \tag{2.9}$$

where $\mathbf{c} \in \mathbb{C}^n$ is some known vector, such that $\mathbf{c}^H\mathbf{M}\mathbf{x} \neq 0$. In our computations we take $\mathbf{c} = \mathbf{M}\mathbf{x}$, which yields

$$\rho(\mathbf{x}) := \frac{\mathbf{x}^H\mathbf{M}^H\mathbf{A}\mathbf{x}}{\mathbf{x}^H\mathbf{M}^H\mathbf{M}\mathbf{x}}, \tag{2.10}$$

and has the desirable minimisation property: for any given $\mathbf{x}$,

$$\|\mathbf{A}\mathbf{x} - \rho(\mathbf{x})\mathbf{M}\mathbf{x}\| = \min_{z\in\mathbb{C}} \|\mathbf{A}\mathbf{x} - z\mathbf{M}\mathbf{x}\|. \tag{2.11}$$

(This property can be verified using simple least-squares approximation as in [29, page 203].) If we normalise $\mathbf{x}$ such that $\mathbf{x}^H\mathbf{M}^H\mathbf{M}\mathbf{x} = 1$, then $\rho(\mathbf{x}) = \mathbf{x}^H\mathbf{M}^H\mathbf{A}\mathbf{x}$.

3. Inexact inverse iteration. We assume that the generalised nonsymmetric eigenproblem (1.1) has a simple, well-separated eigenvalue (λ_1 satisfying (2.7) and(2.8)). This section contains the convergence theory for inexact inverse iteration described by Algorithm 3.1.

ALGORITHM 3.1 (Inexact Inverse Iteration).
Input: $\mathbf{x}^{(0)}$, i_{max}.
For $i = 1, \ldots, i_{max}$
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* $\mathbf{x}^{(i+1)} = \mathbf{y}^{(i)}/\phi(\mathbf{y}^{(i)})$.
4. *Set* $\lambda^{(i+1)} = \rho(\mathbf{x}^{(i+1)})$.
5. *Evaluate* $\mathbf{r}^{(i+1)} = (\mathbf{A} - \lambda^{(i+1)}\mathbf{M})\mathbf{x}^{(i+1)}$.
6. *Test for convergence.*

Output: $\mathbf{x}^{(i_{max})}$.

Note that we choose $\lambda^{(i+1)} = \rho(\mathbf{x}^{(i+1)})$ to make use of the minimisation property (2.11). Also, in Algorithm 3.1 the function $\phi(\mathbf{y}^{(i)})$ is a scalar normalisation. Common choices for this normalisation are $\phi(\mathbf{y}^{(i)}) = {\mathbf{z}^{(i)}}^H\mathbf{y}^{(i)}$, for some $\mathbf{z}^{(i)} \in \mathbb{C}^n$, or a norm of $\mathbf{y}^{(i)}$, such as $\phi(\mathbf{y}^{(i)}) = \|\mathbf{y}^{(i)}\|$ or, if $\mathbf{M}$ is positive definite, $\phi(\mathbf{y}^{(i)}) = \|\mathbf{y}^{(i)}\|_{\mathbf{M}}$.

We introduce a new convergence measure in section 3.1, provide a one step bound in section 3.2 and finally give convergence results for both fixed and variable shifts in section 3.3. In section 4 we discuss some properties of the function $\phi(\mathbf{y})$.

3.1. The measure of convergence. In order to analyse the convergence of inexact inverse iteration we use a different approach to the one used in [3, 8], where the splitting is done in terms of the right eigenvectors of the problem. We split the approximate right eigenvector into two components: the first is in the direction of the exact right eigenvector, and the second lies in the right invariant subspace not containing the exact eigenvector. This decomposition is based on that used in [27] for the perturbation theory of invariant subspaces. However, we introduce a scaling, namely $\alpha^{(i)}$ as in [3], which turns out to be advantageous in the analysis. Let us decompose $\mathbf{x}^{(i)}$, the vector approximating $\mathbf{x}_1$, as

$$\mathbf{x}^{(i)} = \alpha^{(i)}(\mathbf{x}_1 q^{(i)} + \mathbf{X}_2\mathbf{p}^{(i)}), \tag{3.1}$$

where $q^{(i)} \in \mathbb{C}$, $\mathbf{p}^{(i)} \in \mathbb{C}^{(n-1)\times 1}$ and $\mathbf{X}_2 = \mathbf{X}\bar{\mathbf{I}}_{n-1}$, where $\mathbf{X}$ is given by (2.5) and

$$\bar{\mathbf{I}}_{n-1} = \begin{bmatrix} \mathbf{0}^H \\ \mathbf{I}_{n-1} \end{bmatrix} \in \mathbf{C}^{n\times(n-1)},$$

with $\mathbf{I}_{n-1}$ being the identity matrix of size $(n-1)$. The scalar $\alpha^{(i)}$ is chosen so that $\mathbf{x}^{(i)}$ is normalised as $\phi(\mathbf{x}^{(i)}) = 1$. For the convergence theory we leave the scaling of the eigenvector approximate and exact right eigenvector $\mathbf{x}^{(i)}$ and $\mathbf{x}_1$ open. However, in sections 4 and 5 we will use $\|\mathbf{M}\mathbf{x}^{(i)}\| = 1$.

Clearly, $q^{(i)}$ and $\mathbf{p}^{(i)}$ measure how close the approximate eigenvector $\mathbf{x}^{(i)}$ is to the sought eigenvector $\mathbf{x}_1$. As we shall see in the following analysis, the advantage of this splitting is that we need not be concerned about any highly nonnormal behaviour in the matrix pair $(\mathbf{T}_{22}, \mathbf{S}_{22})$. This is in contrast to the approach in [3], where the splitting only exists for positive definite $\mathbf{M}$ and involved a bound on the condition number of the matrix of eigenvectors. Now set

$$\alpha^{(i)} := \|\mathbf{U}^{-1}\mathbf{M}\mathbf{x}^{(i)}\|,$$

and multiply (3.1) from the left by $\mathbf{U}^{-1}\mathbf{M}$. Using

$$\mathbf{U}^{-1}\mathbf{M}\mathbf{x}_1 = s_{11}\mathbf{e}_1 \quad \text{and} \quad \mathbf{U}^{-1}\mathbf{M}\mathbf{X}_2 = \left[\begin{array}{ccc} \mathbf{e}_2 & \dots & \mathbf{e}_n \end{array}\right] \mathbf{S}_{22} = \bar{\mathbf{I}}_{n-1}\mathbf{S}_{22}, \tag{3.2}$$

from (2.6), where $\mathbf{e}_i$ is the ith canonical vector, we have

$$\begin{aligned} 1 = \frac{\|\mathbf{U}^{-1}\mathbf{M}\mathbf{x}^{(i)}\|}{\alpha^{(i)}} &= \|s_{11}q^{(i)}\mathbf{e}_1 + \bar{\mathbf{I}}_{n-1}\mathbf{S}_{22}\mathbf{p}^{(i)}\| \\ &= ((s_{11}q^{(i)})^2 + \|\mathbf{S}_{22}\mathbf{p}^{(i)}\|^2)^{\frac{1}{2}}. \end{aligned} \tag{3.3}$$

Thus $|s_{11}q^{(i)}|$ and $\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|$ can be interpreted as generalisations of the cosine and sine functions as used in the orthogonal decomposition for the symmetric eigenproblem, cf. [18]. Also, from (3.3), we have $|s_{11}q^{(i)}| \leq 1$ and $\|\mathbf{S}_{22}\mathbf{p}^{(i)}\| \leq 1$. Note that (3.3) also indicates why $\alpha^{(i)}$ was introduced in (3.1). This scaling is not used in [27] or [28]. It is now natural to introduce

$$T^{(i)} := \frac{\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|}{|s_{11}q^{(i)}|}$$

as our measure for convergence. Equation (3.3) shows that $T^{(i)}$ can be interpreted as a generalised tangent. Using (3.1) we have, for $\alpha^{(i)}q^{(i)} \neq 0$,

$$\left\|\frac{\mathbf{x}^{(i)}}{\alpha^{(i)}q^{(i)}} - \mathbf{x}_1\right\| = \frac{\|\mathbf{X}_2\mathbf{p}^{(i)}\|}{|q^{(i)}|} \leq \frac{\|\mathbf{X}_2\|\|\mathbf{p}^{(i)}\|}{|q^{(i)}|} \leq \frac{\|\mathbf{X}\|\|\mathbf{p}^{(i)}\|}{|q^{(i)}|},$$

and also

$$\|\mathbf{X}_2\mathbf{p}^{(i)}\| = \left\|\mathbf{X}\left[\begin{array}{c} 0 \\ \mathbf{p}^{(i)} \end{array}\right]\right\| \geq \frac{\|\mathbf{p}^{(i)}\|}{\|\mathbf{X}^{-1}\|}.$$

Using the last two bounds together with (2.5) we obtain

$$\frac{1}{\|\mathbf{H}^{-1}\|}\frac{\|\mathbf{p}^{(i)}\|}{|q^{(i)}|} \leq \left\|\frac{\mathbf{x}^{(i)}}{\alpha^{(i)}q^{(i)}} - \mathbf{x}_1\right\| \leq \|\mathbf{H}\|\frac{\|\mathbf{p}^{(i)}\|}{|q^{(i)}|}, \tag{3.4}$$

with expressions on $\|\mathbf{H}\|$ and $\|\mathbf{H}^{-1}\|$ given by (2.3).

Hence (3.4) yields that $\dfrac{\|\mathbf{p}^{(i)}\|}{|q^{(i)}|} \to 0$ if and only if $\text{span}\{\mathbf{x}^{(i)}\} \to \text{span}\{\mathbf{x}_1\}$. Further we have

$$T^{(i)} \leq \frac{\|\mathbf{S}_{22}\|\|\mathbf{p}^{(i)}\|}{|s_{11}||q^{(i)}|},$$

and hence, since $\mathbf{s}_{11}$ and $\mathbf{S}_{22}$ are constant, $T^{(i)} \to 0$ if $\dfrac{\|\mathbf{p}^{(i)}\|}{|q^{(i)}|} \to 0$, and so the function $T^{(i)}$ measures the quality of the approximation of $\mathbf{x}^{(i)}$ to $\mathbf{x}_1$. Note that this measure is only of theoretical interest, since both $\mathbf{S}_{22}$ and s_{11} are not available.

The following lemma provides bounds on the absolute error in the eigenvalue approximation $|\rho(\mathbf{x}^{(i)}) - \lambda_1|$ and on the eigenvalue residual, defined by

$$\mathbf{r}^{(i)} := (\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbf{x}^{(i)}. \tag{3.5}$$

LEMMA 3.2. *The generalised Rayleigh quotient $\rho(\mathbf{x}^{(i)})$ given in (2.10) satisfies*

$$|\rho(\mathbf{x}^{(i)}) - \lambda_1| \le C_{\|\mathbf{g}_{12}\|} \|\mathbf{T}_{22} - \lambda_1 \mathbf{S}_{22}\| \|\mathbf{p}^{(i)}\|, \tag{3.6}$$

and the eigenvalue residual (3.5) satisfies

$$\|\mathbf{r}^{(i)}\| \le C_{\|\mathbf{g}_{12}\|} \|\mathbf{T}_{22} - \lambda_1 \mathbf{S}_{22}\| \|\mathbf{p}^{(i)}\|, \tag{3.7}$$

where $\mathbf{p}^{(i)}$ is given in (3.1) and $C_{\|\mathbf{g}_{12}\|}$ is given in (2.3).

Proof. Since $(\mathbf{A} - \lambda_1\mathbf{M})\mathbf{x}^{(i)} = \alpha^{(i)}(\mathbf{A} - \lambda_1\mathbf{M})\mathbf{X}_2\mathbf{p}^{(i)}$, using (3.1) we have

$$\begin{aligned} |\rho(\mathbf{x}^{(i)}) - \lambda_1| &= \frac{|{\mathbf{x}^{(i)}}^H \mathbf{M}^H (\mathbf{A} - \lambda_1\mathbf{M})\mathbf{x}^{(i)}|}{\|\mathbf{M}\mathbf{x}^{(i)}\|^2} \\ &= \frac{|\alpha^{(i)}|\,|{\mathbf{x}^{(i)}}^H \mathbf{M}^H \mathbf{U}\mathbf{U}^{-1} (\mathbf{A} - \lambda_1\mathbf{M})\mathbf{X}_2\mathbf{p}^{(i)}|}{\|\mathbf{M}\mathbf{x}^{(i)}\|^2}. \end{aligned}$$

Hence, using (2.6) and the definition of $\alpha^{(i)}$, we get

$$\begin{aligned} |\rho(\mathbf{x}^{(i)}) - \lambda_1| &= \frac{\|\mathbf{U}^{-1}\mathbf{M}\mathbf{x}^{(i)}\|\,|{\mathbf{x}^{(i)}}^H \mathbf{M}^H \mathbf{U}\bar{\mathbf{I}}_{n-1} (\mathbf{T}_{22} - \lambda_1\mathbf{S}_{22})\mathbf{p}^{(i)}|}{\|\mathbf{M}\mathbf{x}^{(i)}\|^2} \\ &\le \|\mathbf{U}^{-1}\|\|\mathbf{U}\|\|(\mathbf{T}_{22} - \lambda_1\mathbf{S}_{22})\mathbf{p}^{(i)}\|. \end{aligned} \tag{3.8}$$

Now we have

$$\|\mathbf{U}\| = \|\mathbf{Q}\mathbf{G}\| = \|\mathbf{G}\| \quad \text{and} \quad \|\mathbf{U}^{-1}\| = \|\mathbf{G}^{-1}\mathbf{Q}^H\| = \|\mathbf{G}^{-1}\|,$$

since $\mathbf{Q}$ is unitary. Hence, from equation (3.8), we obtain

$$\begin{aligned} |\rho(\mathbf{x}^{(i)}) - \lambda_1| &\le \|\mathbf{G}\|\|\mathbf{G}^{-1}\|\|(\mathbf{T}_{22} - \lambda_1\mathbf{S}_{22})\mathbf{p}^{(i)}\| \\ &\le C_{\|\mathbf{g}_{12}\|}\|(\mathbf{T}_{22} - \lambda_1\mathbf{S}_{22})\|\|\mathbf{p}^{(i)}\|, \end{aligned}$$

as required. The eigenvalue residual can be written as

$$\mathbf{r}^{(i)} = (\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbf{x}^{(i)} = (\mathbf{A} - \lambda_1\mathbf{M})\mathbf{x}^{(i)} + (\lambda_1 - \rho(\mathbf{x}^{(i)}))\mathbf{M}\mathbf{x}^{(i)}$$

and hence, using the same idea as in the first part of the proof, we obtain

$$\begin{aligned} \mathbf{r}^{(i)} &= \alpha^{(i)}(\mathbf{A} - \lambda_1\mathbf{M})\mathbf{X}_2\mathbf{p}^{(i)} - \frac{\alpha^{(i)}({\mathbf{x}^{(i)}}^H\mathbf{M}^H(\mathbf{A} - \lambda_1\mathbf{M})\mathbf{X}_2\mathbf{p}^{(i)})\mathbf{M}\mathbf{x}^{(i)}}{{\mathbf{x}^{(i)}}^H\mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}} \\ &= \left(\mathbf{I} - \frac{\mathbf{M}\mathbf{x}^{(i)}{\mathbf{x}^{(i)}}^H\mathbf{M}^H}{{\mathbf{x}^{(i)}}^H\mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}}\right)\alpha^{(i)}(\mathbf{A} - \lambda_1\mathbf{M})\mathbf{X}_2\mathbf{p}^{(i)}. \end{aligned}$$

ETNA
Kent State University
etna@mcs.kent.edu

This yields $\|\mathbf{r}^{(i)}\| \leq \alpha^{(i)}\|(\mathbf{A}-\lambda_1\mathbf{M})\mathbf{X}_2\mathbf{p}^{(i)}\|$ and proceeding as in the first part of the proof gives the required result. $\square$

Lemma 3.2 shows that the generalised Rayleigh quotient $\rho(\mathbf{x}^{(i)})$ defined by (2.10) converges linearly in $\|\mathbf{p}^{(i)}\|$ to λ_1 and the eigenvalue residual $\mathbf{r}^{(i)}$ converges linearly in $\|\mathbf{p}^{(i)}\|$ to zero. This observation leads to more practical measures of convergence than the generalised tangent $T^{(i)}$, which is only of theoretical nature. Nonetheless, one must recognise the limitations of this approach: if $C_{\|\mathbf{g}_{12}\|}$ is large, then the error in the generalised Rayleigh quotient and the residual may be large, even if $\|\mathbf{p}^{(i)}\|$ is small.

The lemma in the following subsection provides a bound on the generalised tangent $T^{(i)}$ after one step of inexact inverse iteration, and is a generalisation of Lemma 3.1 proved in [3] for a diagonalisable problem with symmetric positive definite $\mathbf{M}$.

3.2. A one step bound. In this subsection we provide the main lemma used in the convergence theory for inexact inverse iteration. Let the sought eigenvalue λ_1 be simple, finite and well separated. Furthermore let the starting vector $\mathbf{x}^{(0)}$ be neither the solution $\mathbf{x}_1$ itself, that is, $\mathbf{p}^{(0)} \neq \mathbf{0}$, nor deficient in the sought eigendirection, that is, $q^{(0)} \neq 0$. (This is the same as assuming that $0 < \|\mathbf{S}_{22}\mathbf{p}^{(i)}\| < 1$.) We have the following lemma.

LEMMA 3.3. *Let the generalised eigenproblem $\mathbf{Ax} = \lambda\mathbf{Mx}$ have a simple finite eigenpair $(\lambda_1, \mathbf{x}_1)$ and let (3.1) hold for the approximate eigenpair. Assume the shift satisfies $\sigma^{(i)} \notin \lambda(\mathbf{T}_{22}, \mathbf{S}_{22})$. Further let*

$$\mathbf{Mx}^{(i)} - (\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{d}^{(i)},$$

with $\|\mathbf{d}^{(i)}\| \leq \tau^{(i)}\|\mathbf{Mx}^{(i)}\|$ in Algorithm 3.1 and

$$\tau^{(i)} < \beta\alpha^{(i)}\frac{|s_{11}q^{(i)}|}{\|\mathbf{u}_1\|\|\mathbf{Mx}^{(i)}\|}, \tag{3.9}$$

with $\beta \in (0,1)$. Then

$$T^{(i+1)} = \frac{\|\mathbf{S}_{22}\mathbf{p}^{(i+1)}\|}{|s_{11}q^{(i+1)}|} \leq \frac{|\lambda_1 - \sigma^{(i)}|\|\mathbf{S}_{22}\|}{\|(\mathbf{T}_{22} - \sigma^{(i)}\mathbf{S}_{22})^{-1}\|^{-1}}\frac{\left(\|\alpha^{(i)}\mathbf{S}_{22}\mathbf{p}^{(i)}\| + \|\mathbf{d}^{(i)}\|\right)}{(1-\beta)|\alpha^{(i)}s_{11}q^{(i)}|}. \tag{3.10}$$

Proof. Using

$$(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{Mx}^{(i)} - \mathbf{d}^{(i)} \quad \text{and} \quad \mathbf{x}^{(i+1)} = \frac{\mathbf{y}^{(i)}}{\phi(\mathbf{y}^{(i)})}$$

from Algorithm 3.1 together with the splitting (3.1) for $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(i+1)}$ we obtain

$$\begin{aligned} \phi(\mathbf{y}^{(i)})(\mathbf{A} - \sigma^{(i)}\mathbf{M})(\alpha^{(i+1)}\mathbf{x}_1 q^{(i+1)} + \alpha^{(i+1)}\mathbf{X}_2\mathbf{p}^{(i+1)}) \\ = \mathbf{M}(\alpha^{(i)}\mathbf{x}_1 q^{(i)} + \alpha^{(i)}\mathbf{X}_2\mathbf{p}^{(i)}) - \mathbf{d}^{(i)}. \end{aligned} \tag{3.11}$$

Using (2.6) we get that

$$\begin{aligned} \mathbf{U}^{-1}(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{x}_1 &= (t_{11} - \sigma^{(i)}s_{11})\mathbf{e}_1, \\ \mathbf{U}^{-1}(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{X}_2 &= \begin{bmatrix} \mathbf{0} \\ \mathbf{T}_{22} - \sigma^{(i)}\mathbf{S}_{22} \end{bmatrix} = \bar{\mathbf{I}}_{n-1}(\mathbf{T}_{22} - \sigma^{(i)}\mathbf{S}_{22}), \end{aligned}$$

where $\bar{\mathbf{I}}_{n-1}$ is defined in (3.2). Thus, multiplying (3.11) by $\mathbf{U}^{-1}$ from the left we obtain

$$\begin{aligned} \phi(\mathbf{y}^{(i)})\left(\alpha^{(i+1)}(t_{11} - \sigma^{(i)}s_{11})q^{(i+1)}\mathbf{e}_1 + \alpha^{(i+1)}\bar{\mathbf{I}}_{n-1}(\mathbf{T}_{22} - \sigma^{(i)}\mathbf{S}_{22})\mathbf{p}^{(i+1)}\right) \\ = \alpha^{(i)}s_{11}q^{(i)}\mathbf{e}_1 + \alpha^{(i)}\bar{\mathbf{I}}_{n-1}\mathbf{S}_{22}\mathbf{p}^{(i)} - \mathbf{U}^{-1}\mathbf{d}^{(i)}. \end{aligned} \tag{3.12}$$

Multiplying (3.12) by $\mathbf{e}_1^H$ and $\bar{\mathbf{I}}_{n-1}^H$ from the left we split (3.12) into two equations, namely,

$$\phi(\mathbf{y}^{(i)})\alpha^{(i+1)}(t_{11} - \sigma^{(i)} s_{11})q^{(i+1)} = \alpha^{(i)} s_{11} q^{(i)} - \mathbf{e}_1^H \mathbf{U}^{-1}\mathbf{d}^{(i)}$$

in the direction of $\mathbf{e}_1$ and

$$\phi(\mathbf{y}^{(i)})\alpha^{(i+1)}(\mathbf{T}_{22} - \sigma^{(i)}\mathbf{S}_{22})\mathbf{p}^{(i+1)} = \alpha^{(i)}\mathbf{S}_{22}\mathbf{p}^{(i)} - \bar{\mathbf{I}}_{n-1}^H \mathbf{U}^{-1}\mathbf{d}^{(i)}$$

in span$\{\mathbf{e}_1\}^\perp$. With the left eigenvector $\mathbf{u}_1^H = \mathbf{e}_1^H\mathbf{U}^{-1}$ and the left invariant subspace $\mathbf{U}_2^H := \begin{bmatrix} \mathbf{e}_2 & \dots & \mathbf{e}_n \end{bmatrix}^H \mathbf{U}^{-1}$, and assuming that $\sigma^{(i)}$ is not an eigenvalue of $(\mathbf{T}_{22}, \mathbf{S}_{22})$, as well as $s_{11} \neq 0$, we get

$$\begin{aligned} T^{(i+1)} &= \frac{\|\mathbf{S}_{22}\mathbf{p}^{(i+1)}\|}{|s_{11}q^{(i+1)}|} \\ &\leq \frac{|\lambda_1 - \sigma^{(i)}|\|\mathbf{S}_{22}\|\|(\mathbf{T}_{22} - \sigma^{(i)}\mathbf{S}_{22})^{-1}\| \left(\|\alpha^{(i)}\mathbf{S}_{22}\mathbf{p}^{(i)}\| + \|\mathbf{U}_2^H\mathbf{d}^{(i)}\|\right)}{|\alpha^{(i)}s_{11}q^{(i)}| - |\mathbf{u}_1^H\mathbf{d}^{(i)}|}. \end{aligned}$$

Using (3.9) we obtain

$$T^{(i+1)} = \frac{\|\mathbf{S}_{22}\mathbf{p}^{(i+1)}\|}{|s_{11}q^{(i+1)}|} \leq \frac{|\lambda_1 - \sigma^{(i)}|\|\mathbf{S}_{22}\|}{\|(\mathbf{T}_{22} - \sigma^{(i)}\mathbf{S}_{22})^{-1}\|^{-1}} \frac{\left(\|\alpha^{(i)}\mathbf{S}_{22}\mathbf{p}^{(i)}\| + \|\mathbf{U}_2^H\mathbf{d}^{(i)}\|\right)}{(1-\beta)|\alpha^{(i)}s_{11}q^{(i)}|}. \tag{3.13}$$

Now $\|\mathbf{U}_2\| = 1$, since, using equation (2.6) we may write

$$\mathbf{U}_2^H = \bar{\mathbf{I}}_{n-1}^H\mathbf{U}^{-1} = \bar{\mathbf{I}}_{n-1}^H\mathbf{G}^{-1}\mathbf{Q}^H,$$

and with the special form of $\mathbf{G}$ (see Lemma 2.3) we obtain

$$\mathbf{U}_2^H = \bar{\mathbf{I}}_{n-1}^H\mathbf{U}^{-1} = \bar{\mathbf{I}}_{n-1}^H \begin{bmatrix} 1 & -\mathbf{g}_{12}^H \\ \mathbf{0} & \mathbf{I}_{n-1} \end{bmatrix} \mathbf{Q}^H = \begin{bmatrix} \mathbf{0} & \mathbf{I}_{n-1} \end{bmatrix} \mathbf{Q}^H.$$

Since $\mathbf{Q}^H$ is unitary we have $\|\mathbf{U}_2^H\| = 1$. Hence,

$$T^{(i+1)} = \frac{\|\mathbf{S}_{22}\mathbf{p}^{(i+1)}\|}{|s_{11}q^{(i+1)}|} \leq \frac{|\lambda_1 - \sigma^{(i)}|\|\mathbf{S}_{22}\|}{\|(\mathbf{T}_{22} - \sigma^{(i)}\mathbf{S}_{22})^{-1}\|^{-1}} \frac{\left(\|\alpha^{(i)}\mathbf{S}_{22}\mathbf{p}^{(i)}\| + \|\mathbf{d}^{(i)}\|\right)}{(1-\beta)|\alpha^{(i)}s_{11}q^{(i)}|}, \tag{3.14}$$

as required. □

This bound is a significant improvement over the corresponding results in [8, Lemma 2.2] and [3, Lemma 3.1] which have a bound involving the norm of the unknown eigenvector basis matrix. This matrix may be arbitrarily ill-conditioned, and hence may result in an unnecessarily severe restriction on the solve tolerance in the later theory.

Condition (3.9) asks that $\tau^{(i)}$ is small enough and bounded in terms of $|\alpha^{(i)}s_{11}q^{(i)}|$, which can be considered as a generalised cosine. In practice this means that if the eigenvector approximation $\mathbf{x}^{(i)}$ is coarse, $|s_{11}q^{(i)}|$ is close to zero and hence $\tau^{(i)}$ has to be chosen small enough.

Note that in the case of $\tau^{(i)} = 0$, that is, we solve the inner system exactly, we have $\beta = 0$ as well as $\mathbf{d}^{(i)} = \mathbf{0}$ and hence

$$T^{(i+1)} \leq \frac{|\lambda_1 - \sigma^{(i)}|\|\mathbf{S}_{22}\|}{\|(\mathbf{T}_{22} - \sigma^{(i)}\mathbf{S}_{22})^{-1}\|^{-1}} T^{(i)}.$$

As in [28], we introduce the function $\mathrm{sep}(\lambda_1, (\mathbf{T}_{22}, \mathbf{S}_{22}))$, which measures the separation of the sought simple eigenvalue λ_1 from the eigenvalues $\lambda(\mathbf{T}_{22}, \mathbf{S}_{22})$, as follows,

$$\begin{aligned}\mathrm{sep}(\lambda_1, (\mathbf{T}_{22}, \mathbf{S}_{22})) &:= \inf_{\|\mathbf{a}\|=1} \|(\mathbf{T}_{22} - \lambda_1 \mathbf{S}_{22})\mathbf{a}\| \\ &= \begin{cases} \|(\mathbf{T}_{22} - \lambda_1 \mathbf{S}_{22})^{-1}\|^{-1}, & \lambda_1 \notin \lambda(\mathbf{T}_{22}, \mathbf{S}_{22}) \\ 0, & \lambda_1 \in \lambda(\mathbf{T}_{22}, \mathbf{S}_{22}) \end{cases}.\end{aligned} \tag{3.15}$$

Using this definition we get

$$\begin{aligned}\mathrm{sep}(\sigma^{(i)}, (\mathbf{T}_{22}, \mathbf{S}_{22})) &= \inf_{\|\mathbf{a}\|=1} \|(\mathbf{T}_{22} - \sigma^{(i)} \mathbf{S}_{22})\mathbf{a}\| \\ &\geq \mathrm{sep}(\lambda_1, (\mathbf{T}_{22}, \mathbf{S}_{22})) - |\lambda_1 - \sigma^{(i)}|\|\mathbf{S}_{22}\|,\end{aligned}$$

and also

$$T^{(i+1)} \leq \frac{|\lambda_1 - \sigma^{(i)}|\|\mathbf{S}_{22}\|}{\mathrm{sep}(\sigma^{(i)}, (\mathbf{T}_{22}, \mathbf{S}_{22}))} T^{(i)}$$

for the case of exact solves. Since $\mathrm{sep}(\sigma^{(i)}, (\mathbf{T}_{22}, \mathbf{S}_{22}))$ is a measure for the separation of the shift $\sigma^{(i)}$ from the rest of the spectrum, this means that the convergence rate depends on the ratio $\dfrac{|\lambda_1 - \sigma^{(i)}|\|\mathbf{S}_{22}\|}{\mathrm{sep}(\sigma^{(i)}, (\mathbf{T}_{22}, \mathbf{S}_{22}))}$. For diagonalisable systems, where $\mathbf{T}_{22}$ is diagonal and $\mathbf{S}_{22} = \mathbf{I}_{n-1}$, this ratio becomes $\dfrac{|\lambda_1 - \sigma^{(i)}|}{|\lambda_2 - \sigma^{(i)}|}$, the familiar ratio obtained for inverse iteration. In the next subsection we give the convergence rate for inexact inverse iteration for certain choices of the shift and the solve tolerance, using Lemma 3.3.

3.3. Convergence rate for inexact inverse iteration. Assume that the shift $\sigma^{(i)}$ in Algorithm 3.1 satisfies

$$|\lambda_1 - \sigma^{(i)}| < \frac{\mathrm{sep}(\lambda_1, (\mathbf{T}_{22}, \mathbf{S}_{22}))}{2\|\mathbf{S}_{22}\|}, \tag{3.16}$$

that is, $\sigma^{(i)}$ is close to λ_1 and certainly closer to λ_1 than to any other eigenvalue. Then, using (3.16), for the first factor on the right hand side of (3.13),

$$\begin{aligned}\frac{|\lambda_1 - \sigma^{(i)}|\|\mathbf{S}_{22}\|}{\|(\mathbf{T}_{22} - \sigma^{(i)}\mathbf{S}_{22})^{-1}\|^{-1}} &\leq \frac{|\lambda_1 - \sigma^{(i)}|\|\mathbf{S}_{22}\|}{\mathrm{sep}(\lambda_1, (\mathbf{T}_{22}, \mathbf{S}_{22})) - |\lambda_1 - \sigma^{(i)}|\|\mathbf{S}_{22}\|} \\ &< \frac{|\lambda_1 - \sigma^{(i)}|\|\mathbf{S}_{22}\|}{|\lambda_1 - \sigma^{(i)}|\|\mathbf{S}_{22}\|} = 1\end{aligned}$$

holds. Note that for diagonalisable systems with $\mathbf{S}_{22} = \mathbf{I}_{n-1}$ condition (3.16) becomes $|\lambda_1 - \sigma^{(i)}| < \frac{1}{2}|\lambda_2 - \lambda_1|$, where $|\lambda_2 - \lambda_1| = \min_{j \neq 1} |\lambda_j - \lambda_1|$ and hence $|\lambda_1 - \sigma^{(i)}| < |\lambda_2 - \sigma^{(i)}|$, a familiar condition for the choice of the shift.

Using Lemma 3.3 we can prove convergence results for variable and fixed shifts $\sigma^{(i)}$ and for different choices of the tolerances $\tau^{(i)}$.

THEOREM 3.4 (Convergence of Algorithm 3.1). *Let (1.1) be a generalised eigenproblem and consider the application of Algorithm 3.1 to find a simple eigenvalue λ_1 with corresponding right eigenvector $\mathbf{x}_1$. Let the assumptions of Lemma 3.3 hold and let $0 < \|\mathbf{S}_{22}\mathbf{p}^{(0)}\| < 1$, that is, $\mathbf{x}^{(0)}$ is neither the solution itself nor deficient in the sought eigendirection.*

1. *Assume $\sigma^{(i)}$ also satisfies*

$$|\lambda_1 - \sigma^{(i)}| < \frac{sep(\lambda_1, (\mathbf{T}_{22}, \mathbf{S}_{22}))}{2\|\mathbf{S}_{22}\|}\|\mathbf{S}_{22}\mathbf{p}^{(i)}\| \tag{3.17}$$

and $\|\mathbf{d}^{(i)}\| \leq \tau^{(i)}\|\mathbf{M}\mathbf{x}^{(i)}\|$, *where* $\tau^{(i)} < \dfrac{\alpha^{(i)}}{\|\mathbf{M}\mathbf{x}^{(i)}\|\|\mathbf{u}_1\|}\beta|s_{11}q^{(i)}|$ *with* $0 \leq 2\beta < 1 - T^{(0)}$. *Then Algorithm 3.1 converges linearly, that is,*

$$T^{(i+1)} \leq \left(\frac{T^{(0)} + \beta}{1 - \beta}\right) T^{(i)} \leq \left(\frac{T^{(0)} + \beta}{1 - \beta}\right)^{i+1} T^{(0)}.$$

If, in addition, $\tau^{(i)} < \alpha^{(i)}\eta\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|/\|\mathbf{M}\mathbf{x}^{(i)}\|$ *for some constant* $\eta > 0$, *then the convergence is quadratic, that is,* $T^{(i+1)} \leq q{T^{(i)}}^2$ *for some* $q > 0$, *and for large enough* i.

2. *If* $\tau^{(i)} < \alpha^{(i)}\eta\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|/\|\mathbf{M}\mathbf{x}^{(i)}\|$ *for some positive constant* η *and furthermore*

$$|\lambda_1 - \sigma^{(i)}| < \frac{1-\beta}{2 - \beta + \eta + \delta}\frac{sep(\lambda_1, (\mathbf{T}_{22}, \mathbf{S}_{22}))}{\|\mathbf{S}_{22}\|}, \tag{3.18}$$

where $\delta > 0$, *then Algorithm 3.1 converges linearly, that is,*

$$T^{(i+1)} \leq qT^{(i)} \leq q^{i+1}T^{(0)}$$

for some constant $q < 1$, *and for large enough* i.

Proof.

1. If (3.17) holds, then

$$\frac{|\lambda_1 - \sigma^{(i)}|\|\mathbf{S}_{22}\|}{\|(\mathbf{T}_{22} - \sigma^{(i)}\mathbf{S}_{22})^{-1}\|^{-1}} < \frac{|\lambda_1 - \sigma^{(i)}|\|\mathbf{S}_{22}\|\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|}{2|\lambda_1 - \sigma^{(i)}|\|\mathbf{S}_{22}\| - |\lambda_1 - \sigma^{(i)}|\|\mathbf{S}_{22}\|\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|}$$
$$\leq \|\mathbf{S}_{22}\mathbf{p}^{(i)}\|,$$

since $\|\mathbf{S}_{22}\mathbf{p}^{(i)}\| < 1$. Thus, from (3.14),

$$\begin{aligned} T^{(i+1)} &\leq \|\mathbf{S}_{22}\mathbf{p}^{(i)}\|\frac{\|\alpha^{(i)}\mathbf{S}_{22}\mathbf{p}^{(i)}\| + \tau^{(i)}\|\mathbf{M}\mathbf{x}^{(i)}\|}{(1-\beta)|\alpha^{(i)}s_{11}q^{(i)}|} \\ &\leq \|\mathbf{S}_{22}\mathbf{p}^{(i)}\|\frac{\|\mathbf{S}_{22}\mathbf{p}^{(i)}\| + \beta}{(1-\beta)|s_{11}q^{(i)}|}, \end{aligned}$$

where we have used $\dfrac{\tau^{(i)}\|\mathbf{M}\mathbf{x}^{(i)}\|}{\alpha^{(i)}} \leq \dfrac{\beta|s_{11}q^{(i)}|}{\|\mathbf{u}_1\|} \leq \beta$. Now $\|\mathbf{S}_{22}\mathbf{p}^{(i)}\| \leq T^{(i)}$ gives

$$T^{(i+1)} \leq T^{(i)}\frac{T^{(i)} + \beta}{1 - \beta},$$

which yields linear convergence by induction, if $T^{(0)} < 1 - 2\beta$. Quadratic convergence follows for large enough i and for $\tau^{(i)}$ linearly decreasing in $\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|$, since

$$\begin{aligned} T^{(i+1)} &\leq \|\mathbf{S}_{22}\mathbf{p}^{(i)}\|\frac{\|\alpha^{(i)}\mathbf{S}_{22}\mathbf{p}^{(i)}\| + \tau^{(i)}\|\mathbf{M}\mathbf{x}^{(i)}\|}{(1-\beta)|\alpha^{(i)}s_{11}q^{(i)}|} \\ &\leq \|\mathbf{S}_{22}\mathbf{p}^{(i)}\|\frac{\|\mathbf{S}_{22}\mathbf{p}^{(i)}\| + \eta\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|}{(1-\beta)|s_{11}q^{(i)}|} \\ &= \frac{\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|}{|s_{11}q^{(i)}|}\frac{\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|(1+\eta)}{(1-\beta)|s_{11}q^{(i)}|} = q{T^{(i)}}^2, \end{aligned}$$

for $q = (1+\eta)/(1-\beta)$. We have used $|s_{11}q^{(i)}| < 1$.

2. If (3.18) holds, then

$$\begin{aligned}\frac{|\lambda_1 - \sigma^{(i)}|\|\mathbf{S}_{22}\|}{\|(\mathbf{T}_{22} - \sigma^{(i)}\mathbf{S}_{22})^{-1}\|^{-1}} &\le \frac{|\lambda_1 - \sigma^{(i)}|\|\mathbf{S}_{22}\|}{\operatorname{sep}(\lambda_1, (\mathbf{T}_{22}, \mathbf{S}_{22})) - |\lambda_1 - \sigma^{(i)}|\|\mathbf{S}_{22}\|} \\ &< \frac{|\lambda_1 - \sigma^{(i)}|\|\mathbf{S}_{22}\|(1-\beta)}{((2-\beta+\eta+\delta) - (1-\beta))|\lambda_1 - \sigma^{(i)}|\|\mathbf{S}_{22}\|} \\ &= \frac{1-\beta}{1+\eta+\delta} < 1.\end{aligned}$$

Further, if $\tau^{(i)} < \alpha^{(i)}\eta\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|/\|\mathbf{M}\mathbf{x}^{(i)}\|$ in (3.14), then (with the results from the first part of the proof)

$$T^{(i+1)} < \frac{1-\beta}{1+\eta+\delta}\frac{1+\eta}{1-\beta}T^{(i)} = \frac{1+\eta}{1+\eta+\delta}T^{(i)},$$

and hence $T^{(i+1)} \le qT^{(i)}$ holds with $q = (1+\eta)/(1+\eta+\delta) < 1$.

Thus we have proved Theorem 3.4. □

Note that if β is chosen close to zero, that is, more accurate solves are used for the inner iteration (see (3.9)), then according to Theorem 3.4, which requires $\beta < (1 - T^{(0)})/2$, $T^{(0)}$ is allowed to be close to one, and hence the initial eigenvector approximation is allowed to be coarse. In contrast, for a larger value of β, which allows the solve tolerance $\tau^{(i)}$ to be larger, we require that $T^{(0)}$ is very small and hence the initial eigenvector approximation $\mathbf{x}^{(0)}$ has to be very close to the sought eigenvector. Also, note that $\|\mathbf{u}_1\| = 1 + \|\mathbf{g}_{12}\|$, so that if $\|\mathbf{g}_{12}\|$ is large, then $\|\mathbf{u}_1\|$ is large and the solve tolerance satisfying (3.9) may be small. Note also that condition (3.9) is the same condition as $\tau^{(i)} < \beta|\mathbf{u}_1^H\mathbf{M}\mathbf{x}^{(i)}|/\ \|\mathbf{u}_1\|$ in Lemma 3.1 of [3].

REMARK 3.5. One way of choosing $\tau^{(i)} < \alpha^{(i)}\eta\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|/\|\mathbf{M}\mathbf{x}^{(i)}\|$ is to use

$$\tau^{(i)} = C\|\mathbf{r}^{(i)}\|,$$

where $\mathbf{r}^{(i)}$ is the eigenvalue residual which is given by (3.5) and satisfies $\|\mathbf{r}^{(i)}\| := \mathcal{O}(\|\mathbf{p}^{(i)}\|)$, and C is a small enough constant.

REMARK 3.6. We point out two shift strategies:

- *Fixed shift*: With a decreasing tolerance $\tau^{(i)} = C_1\|\mathbf{r}^{(i)}\|$ for small enough $\tau^{(0)}$ and C_1 the second case in Theorem 3.4 arises. If the shift satisfies (3.18), that is, the shift is close enough to the sought eigenvalue, then Algorithm 3.1 exhibits linear convergence.
- *Rayleigh quotient shift*: A generalised Rayleigh quotient shift $\sigma^{(i)} = \rho(\mathbf{x}^{(i)})$ chosen as in (2.9) satisfies (see (3.6)) $|\sigma^{(i)} - \lambda_1| = C_2\|\mathbf{p}^{(i)}\|$ for some constant C_2. Hence, for small enough C_2 it will also satisfy (3.17). Therefore, with a decreasing tolerance $\tau^{(i)} = C_1\|\mathbf{r}^{(i)}\|$ quadratic convergence is achieved for small enough $\tau^{(0)}$.

Finally we would like to discuss the application of Theorem 3.4 to the case of positive definite $\mathbf{M}$ and diagonalisable $\mathbf{M}^{-1}\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$; see [3]. In this case $\mathbf{S}$ is the identity matrix, and $\mathbf{T}$ can be represented by a diagonal matrix. Condition (3.17) then becomes

$$|\lambda_1 - \sigma^{(i)}| < \frac{|\lambda_1 - \lambda_2|}{2}\|\mathbf{p}^{(i)}\|,$$

which is the same condition as that in [3].

4. Further convergence results. This section contains some additional convergence results including an analysis of the behaviour of the normalisation function $\phi(\mathbf{y})$ from Algorithm 3.1 during inexact inverse iteration.

First we give an extension of Lemma 3.2 which provides a lower bound on the eigenvalue residual in terms of $\mathbf{p}^{(i)}$.

LEMMA 4.1. *Let the assumptions of Lemma 3.2 be satisfied. Then the following bound holds:*

$$\|\mathbf{p}^{(i)}\| \leq \frac{1}{\alpha^{(i)}} \frac{1}{sep(\rho(\mathbf{x}^{(i)}), (\mathbf{T}_{22}, \mathbf{S}_{22}))} \|\mathbf{r}^{(i)}\| \leq \frac{\|\mathbf{G}\|}{sep(\rho(\mathbf{x}^{(i)}), (\mathbf{T}_{22}, \mathbf{S}_{22}))} \|\mathbf{r}^{(i)}\|.$$

Proof. With $\|\mathbf{U}_2^H\| = 1$ (see remarks after Lemma 3.3) and $\mathbf{U}_2^H = \bar{\mathbf{I}}_{n-1}^H \mathbf{G}^{-1}\mathbf{Q}^H$ we have

$$\begin{aligned}
\|\mathbf{r}^{(i)}\| &\geq \|\mathbf{U}_2^H \mathbf{r}^{(i)}\| = \|\bar{\mathbf{I}}_{n-1}^H \mathbf{G}^{-1}\mathbf{Q}^H (\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbf{x}^{(i)}\| \\
&= \|\bar{\mathbf{I}}_{n-1}^H \mathbf{G}^{-1}\mathbf{Q}^H (\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbf{Z}\mathbf{Z}^H\mathbf{x}^{(i)}\| \\
&= \|\bar{\mathbf{I}}_{n-1}^H \mathbf{G}^{-1}(\mathbf{T} - \rho(\mathbf{x}^{(i)})\mathbf{S})\mathbf{Z}^H\mathbf{x}^{(i)}\| \\
&= \|\bar{\mathbf{I}}_{n-1}^H \mathbf{G}^{-1}(\mathbf{T} - \rho(\mathbf{x}^{(i)})\mathbf{S})\mathbf{H}\mathbf{H}^{-1}\mathbf{Z}^H\mathbf{x}^{(i)}\| \\
&= \left\| \bar{\mathbf{I}}_{n-1}^H \begin{bmatrix} t_{11} - \rho(\mathbf{x}^{(i)})s_{11} & \mathbf{0}^H \\ \mathbf{0} & \mathbf{T}_{22} - \rho(\mathbf{x}^{(i)})\mathbf{S}_{22} \end{bmatrix} \mathbf{H}^{-1}\mathbf{Z}^H\mathbf{x}^{(i)} \right\|.
\end{aligned}$$

With $\mathbf{H}^{-1}\mathbf{Z}^H = \mathbf{X}^{-1}$ and using (3.1) as well as the special structure of $\bar{\mathbf{I}}_{n-1}^H$, we then obtain

$$\begin{aligned}
\|\mathbf{r}^{(i)}\| &\geq \left\| \alpha^{(i)} \begin{bmatrix} \mathbf{0}^H \\ \mathbf{I}_{n-1} \end{bmatrix}^H \begin{bmatrix} t_{11} - \rho(\mathbf{x}^{(i)})s_{11} & \mathbf{0}^H \\ \mathbf{0} & \mathbf{T}_{22} - \rho(\mathbf{x}^{(i)})\mathbf{S}_{22} \end{bmatrix} \mathbf{X}^{-1}(\mathbf{x}_1 q^{(i)} + \mathbf{X}_2\mathbf{p}^{(i)}) \right\| \\
&= \alpha^{(i)} \left\| \begin{bmatrix} \mathbf{0}^H \\ \mathbf{I}_{n-1} \end{bmatrix}^H \begin{bmatrix} t_{11} - \rho(\mathbf{x}^{(i)})s_{11} & \mathbf{0}^H \\ \mathbf{0} & \mathbf{T}_{22} - \rho(\mathbf{x}^{(i)})\mathbf{S}_{22} \end{bmatrix} (q^{(i)}\mathbf{e}_1 + \bar{\mathbf{I}}_{n-1}\mathbf{p}^{(i)}) \right\| \\
&= \alpha^{(i)} \left\| \mathbf{I}_{n-1}(\mathbf{T}_{22} - \rho(\mathbf{x}^{(i)})\mathbf{S}_{22})\mathbf{p}^{(i)} \right\|.
\end{aligned}$$

The definition of the separation (3.15) yields

$$\|\mathbf{r}^{(i)}\| \geq \alpha^{(i)} \frac{\|(\mathbf{T}_{22} - \rho(\mathbf{x}^{(i)})\mathbf{S}_{22})\mathbf{p}^{(i)}\|}{\|\mathbf{p}^{(i)}\|} \|\mathbf{p}^{(i)}\| \geq \alpha^{(i)} \mathrm{sep}(\rho(\mathbf{x}^{(i)}), (\mathbf{T}_{22}, \mathbf{S}_{22}))\|\mathbf{p}^{(i)}\|.$$

Finally, using $1 = \|\mathbf{U}\mathbf{U}^{-1}\mathbf{M}\mathbf{x}^{(i)}\| \leq \|\mathbf{U}\|\alpha^{(i)}$ and $\|\mathbf{U}\| = \|\mathbf{G}\|$ gives the bound on $\alpha^{(i)}$ and the desired result. ☐

Lemmas 4.1 and 3.2 show that the eigenvalue residual is equivalent to $\|\mathbf{p}^{(i)}\|$ as a measure of convergence, provided λ_1 is a well-separated eigenvalue, though, of course, in practice, if $\|\mathbf{G}\|$ is large, then a small residual does not necessarily imply a small error. The following proposition gives upper and lower bounds on $\dfrac{1}{\phi(\mathbf{y}^{(i)})}$ in terms of $\|\mathbf{r}^{(i)}\|$.

PROPOSITION 4.2. *Let* $(\lambda^{(i)}, \mathbf{x}^{(i)})$ *with* $\|\mathbf{M}\mathbf{x}^{(i)}\| = 1$ *be the current approximation to* $(\lambda_1, \mathbf{x}_1)$. *Assume that* $\mathbf{y}^{(i)}$ *is such that*

$$\mathbf{M}\mathbf{x}^{(i)} - (\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{d}^{(i)}, \quad \textit{where} \quad \|\mathbf{d}^{(i)}\| \leq \tau^{(i)} < 1.$$

Then

$$\|\mathbf{r}^{(i+1)}\| \leq \frac{1 + \tau^{(i)}}{\phi(\mathbf{y}^{(i)})} \tag{4.1}$$

ETNA
Kent State University
etna@mcs.kent.edu

and

$$\frac{1-\tau^{(i)}}{\phi(\mathbf{y}^{(i)})} \leq \|\mathbf{r}^{(i+1)}\| + |\rho(\mathbf{x}^{(i+1)}) - \sigma^{(i)}|, \tag{4.2}$$

where $\mathbf{r}^{(i+1)} = \mathbf{A}\mathbf{x}^{(i+1)} - \rho(\mathbf{x}^{(i+1)})\mathbf{M}\mathbf{x}^{(i+1)}$.

Proof. We have $(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{M}\mathbf{x}^{(i)} - \mathbf{d}^{(i)}$ and, since $\mathbf{x}^{(i+1)} = \dfrac{\mathbf{y}^{(i)}}{\phi(\mathbf{y}^{(i)})}$,

$$\mathbf{A}\mathbf{x}^{(i+1)} - \sigma^{(i)}\mathbf{M}\mathbf{x}^{(i+1)} = \frac{1}{\phi(\mathbf{y}^{(i)})}(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)}.$$

Hence

$$\frac{\|\mathbf{A}\mathbf{x}^{(i+1)} - \sigma^{(i)}\mathbf{M}\mathbf{x}^{(i+1)}\|}{\|\mathbf{M}\mathbf{x}^{(i)} - \mathbf{d}^{(i)}\|} = \frac{1}{\phi(\mathbf{y}^{(i)})}. \tag{4.3}$$

Finally, $\|\mathbf{M}\mathbf{x}^{(i)} - \mathbf{d}^{(i)}\| \leq 1 + \tau^{(i)}$ together with the minimising property of $\rho(\mathbf{x}^{(i+1)})$ (see (2.11)) yields the first bound (4.1). In order to obtain the second bound, equality (4.3) gives

$$\begin{aligned}\frac{1}{\phi(\mathbf{y}^{(i)})} &\leq \frac{\|\mathbf{A}\mathbf{x}^{(i+1)} - \rho(\mathbf{x}^{(i+1)})\mathbf{M}\mathbf{x}^{(i+1)}\| + |\rho(\mathbf{x}^{(i+1)}) - \sigma^{(i)}|\|\mathbf{M}\mathbf{x}^{(i+1)}\|}{\|\mathbf{M}\mathbf{x}^{(i)}\| - \|\mathbf{d}^{(i)}\|} \\ &\leq \frac{\|\mathbf{r}^{(i+1)}\| + |\rho(\mathbf{x}^{(i+1)}) - \sigma^{(i)})|}{1-\tau^{(i)}},\end{aligned}$$

which yields (4.2). □

Proposition 4.2 provides the following result. If we chose the shift to be $\sigma^{(i)} := \rho(\mathbf{x}^{(i)})$, then

$$\frac{1-\tau^{(i)}}{\phi(\mathbf{y}^{(i)})} - |\rho(\mathbf{x}^{(i+1)}) - \rho(\mathbf{x}^{(i)})| \leq \|\mathbf{r}^{(i+1)}\| \leq \frac{1+\tau^{(i)}}{\phi(\mathbf{y}^{(i)})}.$$

From section 3, convergence of inexact inverse iteration yields $\|\mathbf{p}^{(i)}\| \to 0$. By Lemmas 3.2 and 4.1, convergence of inexact inverse iteration implies $\|\mathbf{r}^{(i)}\| \to 0$ as well as $|\rho(\mathbf{x}^{(i)}) - \lambda_1| \to 0$. The last property also yields $|\rho(\mathbf{x}^{(i+1)}) - \rho(\mathbf{x}^{(i)})| \to 0$, if inexact inverse iteration converges. Therefore Proposition 4.2 shows that inexact inverse iteration converges if and only if $\phi(\mathbf{y}^{(i)}) \to \infty$ as $i \to \infty$. Note that $\phi(\mathbf{y}^{(i)}) := \|\mathbf{M}\mathbf{y}^{(i)}\|$ in Proposition 4.2.

We end this section with an application of inexact inverse iteration to block structured systems of the form $\mathbf{A}\mathbf{x} = \lambda\mathbf{M}\mathbf{x}$, where

$$\mathbf{A} = \begin{bmatrix} \mathbf{K} & \mathbf{C} \\ \mathbf{C}^H & \mathbf{0} \end{bmatrix} \quad \text{and} \quad \mathbf{M} = \begin{bmatrix} \mathbf{M}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

and $\mathbf{M}_1$ is symmetric positive definite. Matrices with this block structure arise after a mixed finite element discretisation of the linearised incompressible Navier-Stokes equations. If the desired eigenvector is written in terms of the velocity and pressure components $\mathbf{x} = [\mathbf{x}_u^H \ \mathbf{x}_p^H]^H$, the incompressibility condition $\mathbf{C}^H\mathbf{x}_u = 0$ holds. If the system $(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{M}\mathbf{x}^{(i)}$ is solved inexactly, we cannot guarantee that $\mathbf{C}^H\mathbf{x}_u^{(i)} = 0$, even if the starting guess satisfies $\mathbf{C}^H\mathbf{x}_u^{(0)} = 0$: we only know that $\|\mathbf{C}^H\mathbf{x}_u^{(i)}\| \leq \tau^{(i)}$. The following corollary shows that inexact inverse iteration need not enforce the incompressibility condition at each outer iteration.

COROLLARY 4.3. *Let the assumptions of Proposition 4.2 be satisfied and consider inexact inverse iteration applied to the block structured system*

$$\begin{bmatrix} \mathbf{M}_1\mathbf{x}_u^{(i)} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{K}-\rho(\mathbf{x}^{(i)})\mathbf{M}_1 & \mathbf{C} \\ \mathbf{C}^H & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{y}_u^{(i)} \\ \mathbf{y}_p^{(i)} \end{bmatrix} = \begin{bmatrix} \mathbf{d}_u^{(i)} \\ \mathbf{d}_p^{(i)} \end{bmatrix}, \quad \textit{where} \quad \|\mathbf{d}^{(i)}\| \leq \tau^{(i)}.$$

Then $\|\mathbf{C}^H\mathbf{x}_u^{(i)}\| \to 0$ *as* $i \to \infty$.

Proof. From Algorithm 3.1 and Proposition 4.2 we have

$$\|\mathbf{C}^H\mathbf{x}_u^{(i+1)}\| \leq \frac{\|\mathbf{C}^H\mathbf{y}_u^{(i)}\|}{\phi(\mathbf{y}^{(i)})} \leq \frac{\tau^{(i)}}{\phi(\mathbf{y}^{(i)})} \to 0 \quad \text{as} \quad i \to \infty. \qquad \square$$

5. Two numerical examples. Finally, we give two test problems for our theory. We chose problems $\mathbf{A}\mathbf{x} = \lambda\mathbf{M}\mathbf{x}$ which are not necessarily diagonalisable and with singular $\mathbf{M}$, since problems with positive definite $\mathbf{M}$ (including the standard problem $\mathbf{M} = \mathbf{I}$) have been extensively investigated by other authors; see, for example, [2, 3]. The paper by Smit and Paardekooper [25] contains examples for the standard symmetric eigenproblem and that of Golub and Ye [8] discusses the standard diagonalisable problem $\mathbf{M}^{-1}\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$. A nuclear reactor problem similar to the one in the following example with $\mathbf{M}$ singular is considered in [12]. However, in [12] the problem is first transformed into a standard eigenproblem.

EXAMPLE 5.1 (Nuclear Reactor Problem). The standard model to describe the neutron balance in a 2D model of a nuclear reactor is given by the two-group neutron equations

$$\begin{aligned} -\text{div}(K_1\nabla u_1) + (\Sigma_{a,1} + \Sigma_s)u_1 &= \frac{1}{\mu_1}(\Sigma_{f,1}u_1 + \Sigma_{f,2}u_2), \\ -\text{div}(K_2\nabla u_2) + \Sigma_{a,2}u_1 - \Sigma_s u_2 &= 0, \end{aligned}$$

where u_1 and u_2 are defined on $[0,1]\times[0,1]$ and represent the density distributions of fast and thermal neutrons respectively. K_1 and K_2 are diffusion coefficients and $\Sigma_{a,1}, \Sigma_{a,2}, \Sigma_s, \Sigma_{f,1}$ and $\Sigma_{f,2}$ measure interaction probabilities and take different piecewise constant values in different regions of the reactor, which for this example are given in Figure 5.1 and Table 5.1.

The largest μ_1 such that $1/\mu_1$ is an eigenvalue of the system equation is a measure for the criticality of a reactor, with $\mu_1 < 1$ representing subcriticality and $\mu_1 > 1$ representing supercriticality. The aim is to maintain the reactor in the critical phase with $\mu_1 = 1$. The boundary conditions for $g = 1, 2$ are

$$\begin{aligned} u_g &= 0 \quad \text{if} \quad x_1 = 0 \quad \text{or} \quad x_2 = 0, \\ K_g\frac{\partial u_g}{\partial x_i} &= 0 \quad \text{if} \quad x_i = 1, \quad \text{for} \quad i = 1, 2. \end{aligned}$$

Discretising the problem using a finite difference approximation on an $h \times h$ grid, where $h = 1/m$, we obtain a $2m^2 \times 2m^2$ discrete eigenproblem $\mathbf{A}\mathbf{u} = \lambda\mathbf{M}\mathbf{u}$, where $\mathbf{A}$ and $\mathbf{M}$ are both nonsymmetric and $\mathbf{M}$ is singular. We seek the smallest eigenvalue $\lambda_1(= 1/\mu_1)$, which determines the criticality of the reactor. We choose $m = 32$, which leads to a system of size $n = 2048$. For initial conditions, we take $\mathbf{x}^{(0)} = [1, \ldots, 1]^H/\sqrt{n}$. In fact, the exact eigenvalue is given by $\lambda_1 = 0.9707$ and $\cos(\mathbf{x}_1, \mathbf{x}^{(0)}) \approx 0.44$.

We use a fixed shift and a variable shift strategy. The vector $\mathbf{x}^{(i)}$ is normalised such that $\|\mathbf{M}\mathbf{x}^{(i)}\| = 1$, that is, $\phi(\mathbf{y}^{(i)}) = \sqrt{{\mathbf{y}^{(i)}}^H\mathbf{M}^H\mathbf{M}\mathbf{y}^{(i)}}$ in Algorithm 3.1. For the inner solver we use right-preconditioned GMRES with an incomplete LU-factorisation as preconditioner. We perform three different numerical experiments.

ETNA
Kent State University
etna@mcs.kent.edu

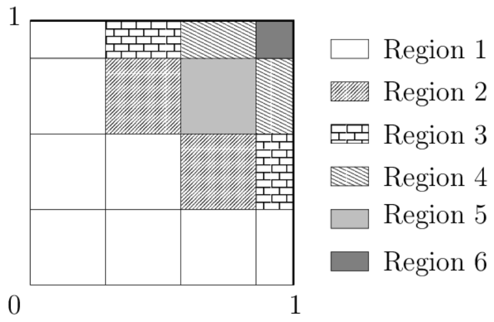

FIG. 5.1. *Nuclear reactor problem geometry.*

TABLE 5.1
Data for the nuclear reactor problem.

	K_1	K_2	$\Sigma_{a,1}$	$\Sigma_{a,12}$	Σ_s	$\Sigma_{f,1}$	$\Sigma_{f,2}$
Region 1	2.939e-5	1.306e-5	0.0089	0.109	0.0	0.0	0.0079
Region 2	4.245e-5	1.306e-5	0.0105	0.025	0.0	0.0	0.0222
Region 3	4.359e-5	1.394e-5	0.0092	0.093	0.0066	0.140	0.0156
Region 4	4.395e-5	1.355e-5	0.0091	0.083	0.0057	0.109	0.0159
Region 5	4.398e-5	1.355e-5	0.0097	0.098	0.0066	0.124	0.0151
Region 6	4.415e-5	1.345e-5	0.0093	0.085	0.0057	0.107	0.0157

(a) Inexact inverse iteration using a fixed shift $\sigma^{(i)} = \sigma = 0.9$ and a decreasing solve tolerance $\tau^{(i)}$ for the inner solver which satisfies

$$\tau^{(i)} = \min\{0.1, \|\mathbf{r}^{(i)}\|\}, \tag{5.1}$$

where $\mathbf{r}^{(i)}$ is defined by (3.5). The iteration stops once the eigenvalue residual satisfies $\|\mathbf{r}^{(i)}\| < 10^{-9}$.

(b) Inexact inverse iteration using a variable shift given by $\rho(\mathbf{x}^{(i)})$ from (2.10) and a decreasing solve tolerance $\tau^{(i)}$ for the inner solver which satisfies (5.1). The iteration stops once the eigenvalue residual satisfies $\|\mathbf{r}^{(i)}\| < 10^{-14}$.

(c) Inexact inverse iteration using a variable shift given by $\rho(\mathbf{x}^{(i)})$ from (2.10) with a fixed solve tolerance, which we chose to be $\tau^{(i)} = \tau^{(0)} = 0.4$. This iteration also stops once the eigenvalue residual satisfies $\|\mathbf{r}^{(i)}\| < 10^{-9}$.

Figure 5.2 illustrates the convergence history of the eigenvalue residuals for the three different experiments described in (a), (b) and (c) above. The choice of (5.1) to provide a solve tolerance $\tau^{(i)}$ is consistent with the discussion in Remark 3.5 and the assumption in Theorem 3.4. We have used this decreasing tolerance throughout our computations. As proved in Theorem 3.4, case (2), inexact inverse iteration with a decreasing solve tolerance and with a fixed shift, chosen to be close enough to the desired eigenvalue, exhibits linear convergence, as shown in Figure 5.2, case (a) (see also the discussion on the fixed shift in Remark 3.6). If we use a generalised Rayleigh quotient as a shift (where the Rayleigh quotient is close enough to the sought eigenvalue) and a fixed solve tolerance $\tau^{(0)}$ the Algorithm 3.1 converges linearly (case (c)), whereas for a decreasing tolerance quadratic convergence is readily observed (case (b)). This covers case (1) in Theorem 3.4, we also refer to the discussion on the Rayleigh quotient shift in Remark 3.6.

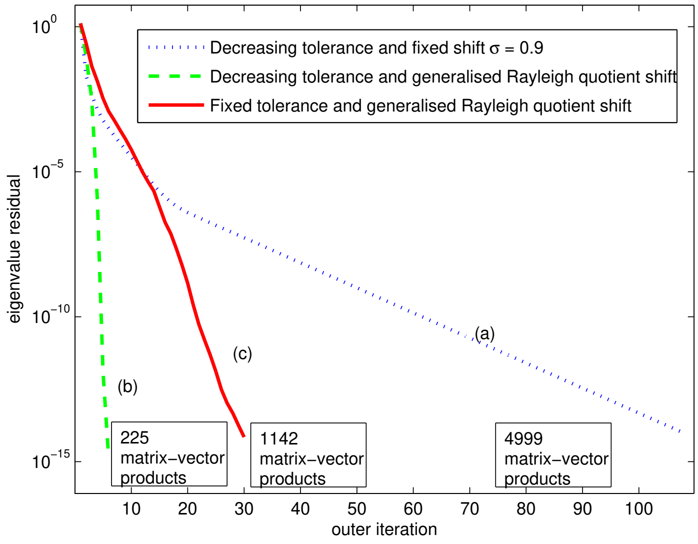

FIG. 5.2. *Convergence history of the eigenvalue residuals for Example 5.1 using* fixed shift $\sigma = 0.9$ *and* variable shift *and fixed or decreasing tolerances (see tests (a), (b) and (c)).*

We would like to note that all three methods have the same initial eigenvalue residual. Both methods (a) and (c) exhibit linear convergence, but the method with a variable shift and fixed solve tolerance performs better than the fixed shift method with a decreasing solve tolerance. This improvement in the behaviour of method (c) over (a) may be explained by close examination of the asymptotic constants in the expressions for linear convergence in Theorem 3.4. For a good starting guess (that is a $T^{(0)}$ close to zero) and a small enough β with $\beta < (1 - T^{(0)})/2$, the constant of linear convergence for method (c) may be much smaller than one, and hence smaller than the convergence rate for method (a). In our particular computations the constants for linear convergence are about 0.82 for method (a) and about 0.32 for method (c).

The total amount of work is measured by the number of matrix-vector multiplications given in Figure 5.2. We can observe that method (b), inexact Rayleigh quotient iteration with a decreasing solve tolerance, achieves the fastest convergence rate with smallest amount of work.

EXAMPLE 5.2 (The Linearised Steady Navier-Stokes Equations). For the stability analysis of the steady state solutions of the Navier-Stokes equations generalised eigenproblems of the form $\mathbf{A}\mathbf{x} = \lambda\mathbf{M}\mathbf{x}$ arise, where $\mathbf{A}$ and $\mathbf{M}$ have a special block structure, that is,

$$\mathbf{A} = \begin{bmatrix} \mathbf{K} & \mathbf{C} \\ \mathbf{C}^H & \mathbf{0} \end{bmatrix} \quad \text{and} \quad \mathbf{M} = \begin{bmatrix} \mathbf{M}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

Of particular interest for the stability analysis are the leftmost eigenvalues of the system. (The right half-plane is the stable region in our formulation.) We consider incompressible fluid flow past a cylinder with Reynolds number equal to 1. Using a mixed finite element discretisation of the Navier-Stokes equations the above block structured systems arises, where $\mathbf{K}$ is 1406×1406 and nonsymmetric, $\mathbf{C}$ is 1406×232 and has full rank, and $\mathbf{M}_1$ is 1406×1406 and symmetric positive definite. The system has 1638 degrees of freedom. The leftmost

ETNA
Kent State University
etna@mcs.kent.edu

TABLE 5.2
Incompressibility condition $\|\mathbf{C}^H\mathbf{x}_u^{(i)}\|$ in the course of inexact inverse iteration without the application of π.

Outer it. i	$\|\mathbf{r}^{(i)}\|$	$\|\mathbf{C}^H\mathbf{x}_u^{(i)}\|$	$\|\mathbf{C}^H\mathbf{y}_u^{(i)}\|$
1	3.2970e-01	0	1.2446e-02
2	1.9519e-02	1.3454e-04	4.7833e-03
3	1.1518e-02	2.0178e-04	7.3705e-03
4	7.3977e-03	4.4779e-04	1.6494e-02
5	3.5684e-03	2.8949e-04	1.2807e-02
6	1.0365e-03	1.6762e-04	1.3858e-02
7	1.1658e-04	3.3947e-05	1.1832e-02
8	7.1789e-06	2.8401e-07	3.2990e-03
9	1.3820e-06	1.0094e-07	5.9614e-03
10	5.2651e-07	6.0768e-08	1.0112e-02
11	1.6630e-07	1.6899e-08	8.9196e-03
12	5.3896e-08	3.1178e-09	3.8395e-03

eigenvalues of the problem correct to two decimal places are given by

$$\lambda_{1/2} = 0.21 \pm 0.16i,$$

and we aim to find the complex eigenvalue λ_1 nearest to $0.21+0.16i$. We normalise $\mathbf{x}^{(i)}$ such that $\|\mathbf{M}\mathbf{x}^{(i)}\| = 1$, that is, $\phi(\mathbf{y}^{(i)}) = \sqrt{{\mathbf{y}^{(i)}}^H\mathbf{M}^H\mathbf{M}\mathbf{y}^{(i)}}$ as in the first example. The convergence performance of the three methods considered in the previous example is repeated in this example and we do not reproduce the results here. Rather, we look at the incompressibility condition $\mathbf{C}^H\mathbf{x}_u^{(i)} = \mathbf{0}$ and examine how it behaves under inexact inverse iteration. In particular we ask if there is any advantage to be gained by imposing the incompressibility condition after each inexact solve. To this end we carry out inexact inverse iteration using a variable shift given by $\rho(\mathbf{x}^{(i)})$ from (2.10) and a close enough starting guess. We use a fixed solve tolerance $\tau^{(i)} = \tau^{(0)} = 0.1$. The iteration stops once the eigenvalue residual satisfies $\|\mathbf{r}^{(i)}\| < 10^{-7}$. To impose the incompressibility condition after an inner iteration we replace $\mathbf{x}_u^{(i)}$ by $\pi\mathbf{x}_u^{(i)}$ where the projection π is defined by

$$\pi := \mathbf{I} - \mathbf{C}(\mathbf{C}^H\mathbf{C})^{-1}\mathbf{C}^H.$$

We compare two methods: the projection π is not applied at the start of each outer iteration i; and π is applied at the beginning of each outer iteration. In this case, after each inner solve we apply π to $\mathbf{y}_u^{(i)}$, such that

$$\mathbf{C}^H\mathbf{x}_u^{(i+1)} = \mathbf{C}^H\frac{\mathbf{y}_u^{(i)}}{\phi(\mathbf{y}_u^{(i)})} = \mathbf{0}.$$

For both experiments we take the initial condition such that $\mathbf{C}^H\mathbf{x}_u^{(0)} = \mathbf{0}$.

Tables 5.2 and 5.3 show the eigenvalue residual $\|\mathbf{r}^{(i)}\|$, $\|\mathbf{C}^H\mathbf{x}_u^{(i)}\|$ and $\|\mathbf{C}^H\mathbf{y}_u^{(i)}\|$ at each outer iteration i. The second column of Table 5.3 shows $\|\mathbf{C}^H\mathbf{x}_u^{(i)}\|$ before projection is applied for the beginning of the next outer iteration step. We observe that there is almost no difference between performing inexact inverse iteration with or without projection at the beginning of each outer step. We also see $\|\mathbf{C}^H\mathbf{x}_u^{(i)}\| \to 0$ as i increases, as predicted by Corollary 4.3, and hence, the application of the projection π at every step is not necessary. Also note that in both tables $\|\mathbf{C}^H\mathbf{y}_u^{(i)}\| \leq \tau^{(0)} = 0.1$.

ETNA
Kent State University
etna@mcs.kent.edu

TABLE 5.3
Incompressibility condition $\|\mathbf{C}^H\mathbf{x}_u^{(i)}\|$ in the course of inexact inverse iteration with the application of π.

Outer it. i	$\|\mathbf{r}^{(i)}\|$	$\|\mathbf{C}^H\mathbf{x}_u^{(i)}\|$	$\|\mathbf{C}^H\mathbf{y}_u^{(i)}\|$
1	3.2970e-01	0	1.2446e-02
2	1.9631e-02	1.3454e-04	4.7833e-03
3	1.2169e-02	2.0592e-04	7.5205e-03
4	1.1431e-02	4.4542e-04	1.6396e-02
5	5.9688e-03	2.9315e-04	1.2954e-02
6	3.0500e-03	1.6095e-04	1.3298e-02
7	4.3488e-04	3.4289e-05	1.2147e-02
8	8.4934e-06	2.8349e-07	3.2432e-03
9	1.7348e-06	1.0312e-07	6.2898e-03
10	7.9410e-07	6.0285e-08	1.0026e-02
11	2.9405e-07	1.6987e-08	8.9189e-03
12	6.4187e-08	3.1543e-09	3.8886e-03

6. A convergence theory for inexact simple Jacobi-Davidson method. In this section we show how the convergence theory obtained in section 3 may be applied to a simplified version of the inexact Jacobi-Davidson method. The Jacobi-Davidson method was introduced by Sleijpen and van der Vorst, see [22] and [24] for the linear eigenproblem, and it has been applied to the generalised eigenproblem and matrix pencils; see [4] and [21]. A survey has been given in [10]; see also [1]. A convergence theory for Jacobi-Davidson applied to the Hermitian eigenproblem has been given in [30] and for a special inner solver in [14]. The relationship between a simplified version of Jacobi-Davidson method and Newton's method for exact solves has been established in several papers; see, for example, [22, 23, 24], and [15]. Here we provide a convergence theory for a version of an inexact simplified Jacobi-Davidson method for the generalised eigenvalue problem (1.1), and also present some numerical results to illustrate our theory.

6.1. A simplified Jacobi-Davidson method. First, we briefly describe one possible version of a simplified Jacobi-Davidson algorithm for the generalised eigenvalue problem (1.1); see [14, Algorithm 2.1] and [30, Algorithm 3.1] for similar algorithms for standard Hermitian eigenproblems.

Assume $(\rho(\mathbf{x}^{(i)}), \mathbf{x}^{(i)})$ approximates $(\lambda_1, \mathbf{x}_1)$, and introduce the orthogonal projections

$$\mathbf{P}^{(i)} = \mathbf{I} - \frac{\mathbf{M}\mathbf{x}^{(i)}\mathbf{x}^{(i)}{}^H\mathbf{M}^H}{\mathbf{x}^{(i)}{}^H\mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}} \quad \text{and} \quad \mathbf{Q}^{(i)} = \mathbf{I} - \frac{\mathbf{x}^{(i)}\mathbf{x}^{(i)}{}^H\mathbf{M}^H\mathbf{M}}{\mathbf{x}^{(i)}{}^H\mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}}.$$

With $\mathbf{r}^{(i)}$ defined by (3.5) solve the correction equation

$$\mathbf{P}^{(i)}(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbf{Q}^{(i)}\mathbf{s}^{(i)} = -\mathbf{r}^{(i)}, \quad \text{where} \quad \mathbf{s}^{(i)} \perp \mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}, \tag{6.1}$$

for $\mathbf{s}^{(i)}$. An improved guess for the eigenvector is given by a suitably normalised $\mathbf{x}^{(i)} + \mathbf{s}^{(i)}$. For other choices of projections and discussions on the correction equation (6.1) we refer to [21]. The motivation behind the Jacobi-Davidson algorithm is that for large systems which are solved iteratively, the form of the correction equation (6.1) is more amenable to efficient solution than the corresponding system for inverse iteration. Also, in practice, a subspace version of Jacobi-Davidson is used with each new direction being added to increase the dimension of a search space, but we do not consider this version here. Algorithm 6.1 provides a precise description of the method we discuss in this paper.

ETNA
Kent State University
etna@mcs.kent.edu

ALGORITHM 6.1 (Simplified Jacobi-Davidson (Jacobi-Davidson without subspace acceleration)).
Input: $\mathbf{x}^{(0)}$, i_{max}.
For $i = 1, \dots, i_{max}$
1. *Choose* $\tau^{(i)}$.
2. $\mathbf{r}^{(i)} = (\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbf{x}^{(i)}$.
3. *Find* $\mathbf{s}^{(i)}$ *such that* $\|\mathbf{P}^{(i)}(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbf{Q}^{(i)}\mathbf{s}^{(i)} + \mathbf{r}^{(i)}\| \le \tau^{(i)}\|\mathbf{r}^{(i)}\|$ *for* $\mathbf{s}^{(i)} \perp \mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}$.
4. *Set* $\mathbf{x}^{(i+1)} = (\mathbf{x}^{(i)} + \mathbf{s}^{(i)})/\phi(\mathbf{x}^{(i)} + \mathbf{s}^{(i)})$.
5. *Test for convergence.*

Output: $\mathbf{x}^{(i_{max})}$.

The function ϕ is a normalisation, which for both practical computations and theoretical comparisons between Rayleigh quotient iteration and Jacobi-Davidson, is taken to be the same as in Algorithm 3.1.

In this section we shall provide a convergence theory for the inexact simplified Jacobi-Davidson method given in Algorithm 6.1. To do this we shall first show the close connection of inexact simplified Jacobi-Davidson with inexact Rayleigh-quotient iteration and then apply the convergence theory in section 3. Though simplified Jacobi-Davidson is not used in practice its convergence may be considered as a worst-case scenario for the more usual subspace Jacobi-Davidson procedure, and the convergence results here can be similarly interpreted.

First, we point out the following well-known equivalence between the simplified Jacobi-Davidson method and Rayleigh quotient iteration for *exact* system solves, which has been proved in [14, 16, 24], and in [21] for the generalised eigenproblem.

LEMMA 6.2. *Suppose the correction equation in Algorithm 6.1 has a unique solution* $\mathbf{s}^{(i)}$. *Then the Jacobi-Davidson solution* $\mathbf{x}_{JD}^{(i+1)} = \mathbf{x}^{(i)} + \mathbf{s}^{(i)}$ *satisfies*

$$(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbf{z}^{(i+1)} = \mathbf{M}\mathbf{x}^{(i)},$$

where

$$\mathbf{z}^{(i+1)} = \frac{1}{\gamma^{(i)}}\mathbf{x}_{JD}^{(i+1)}, \quad \text{with} \quad \gamma^{(i)} = \frac{{\mathbf{x}^{(i)}}^H\mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}}{{\mathbf{x}^{(i)}}^H\mathbf{M}^H\mathbf{M}(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})^{-1}\mathbf{M}\mathbf{x}^{(i)}}. \tag{6.2}$$

From Lemma 6.2 it is clear that for exact solves one step of simplified Jacobi-Davidson produces an improved approximation to the desired eigenvector that has the same direction as that given by one step of Rayleigh quotient iteration. Hence, as observed in [24], if the correction equation is solved exactly, the method converges as fast as Rayleigh quotient iteration (that is quadratically for nonsymmetric systems). The next section shows how we can find a similar equivalence between inexact Rayleigh quotient iteration and the inexact Jacobi-Davidson method.

6.2. Inexact Jacobi-Davidson and Rayleigh quotient iterations. Assume we have an eigenvector approximation $\mathbf{x}^{(i)}$. We compare one step of inexact Rayleigh quotient iteration, that is,

$$(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbf{y}^{(i)} = \mathbf{M}\mathbf{x}^{(i)} - \mathbf{d}_I^{(i)}, \tag{6.3}$$

where $\|\mathbf{d}_I^{(i)}\| \le \tau_I^{(i)}\|\mathbf{M}\mathbf{x}^{(i)}\|$, with $\tau_I^{(i)} < 1$, with one step of inexact Jacobi-Davidson method, that is,

$$\mathbf{P}^{(i)}(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbf{Q}^{(i)}\mathbf{s}^{(i)} = -\mathbf{r}^{(i)} + \mathbf{d}_{JD}^{(i)}, \quad \text{for} \quad \mathbf{s}^{(i)} \perp \mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}, \tag{6.4}$$

ETNA
Kent State University
etna@mcs.kent.edu

where $\|\mathbf{d}_{JD}^{(i)}\| \leq \tau_{JD}^{(i)}\|\mathbf{r}^{(i)}\|$ and $\tau_{JD}^{(i)} < 1$. First, we transform (6.4) into a system of the form (6.3), as follows. Since $\mathbf{Q}\mathbf{s}^{(i)} = \mathbf{s}^{(i)}$ and $\mathbf{r}^{(i)} = \mathbf{P}^{(i)}\mathbf{r}^{(i)} = \mathbf{P}^{(i)}(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbf{x}^{(i)}$, we can write (6.4) as

$$\mathbf{P}^{(i)}(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})(\mathbf{x}^{(i)} + \mathbf{s}^{(i)}) = \mathbf{d}_{JD}^{(i)}, \quad \mathbf{s}^{(i)} \perp \mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)},$$

or

$$(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})(\mathbf{x}^{(i)} + \mathbf{s}^{(i)}) = \gamma^{(i)}\mathbf{M}\mathbf{x}^{(i)} + \mathbf{d}_{JD}^{(i)},$$

where $\gamma^{(i)}$ is chosen such that $\mathbf{s}^{(i)} \perp \mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}$. Finally we obtain

$$(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\frac{\mathbf{x}^{(i)} + \mathbf{s}^{(i)}}{\gamma^{(i)}} = \mathbf{M}\mathbf{x}^{(i)} + \frac{\mathbf{d}_{JD}^{(i)}}{\gamma^{(i)}}, \tag{6.5}$$

where (see (6.2))

$$\gamma^{(i)} = \frac{\mathbf{x}^{(i)H}\mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)} - \mathbf{x}^{(i)H}\mathbf{M}^H\mathbf{M}(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})^{-1}\mathbf{d}_{JD}^{(i)}}{\mathbf{x}^{(i)H}\mathbf{M}^H\mathbf{M}(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})^{-1}\mathbf{M}\mathbf{x}^{(i)}}.$$

The linear system (6.5) is of the form (6.3), and under the assumption that

$$\frac{\|\mathbf{d}_{JD}^{(i)}\|}{|\gamma^{(i)}|} \leq \tau_I^{(i)}\|\mathbf{M}\mathbf{x}^{(i)}\|,$$

we can apply the theory of section 3. Thus, we obtain the following corollary from Theorem 3.4.

COROLLARY 6.3. *Let the assumptions and definitions of Theorem 3.4 hold and let*

$$\tau_I^{(i)} := \tau_{JD}^{(i)}\frac{\|\mathbf{r}^{(i)}\|}{|\gamma^{(i)}|}. \tag{6.6}$$

Then Algorithm 6.1 converges

- *linearly, if* $\tau_I^{(i)} < \dfrac{\alpha^{(i)}}{\|\mathbf{M}\mathbf{x}^{(i)}\|\|\mathbf{u}_1\|}\beta|s_{11}q^{(i)}|$, *with* $0 \leq 2\beta < 1 - T^{(0)}$, *and*
- *quadratically, if, in addition,* $\tau_I^{(i)} < \alpha^{(i)}\eta\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|/\|\mathbf{M}\mathbf{x}^{(i)}\|$ *for some constant* $\eta > 0$.

Proof. Note that

$$\frac{\|\mathbf{d}_{JD}^{(i)}\|}{|\gamma^{(i)}|} \leq \tau_{JD}^{(i)}\frac{\|\mathbf{r}^{(i)}\|}{|\gamma^{(i)}|} := \tau_I^{(i)}\|\mathbf{M}\mathbf{x}^{(i)}\| \tag{6.7}$$

and using $\tau^{(i)} := \tau_I^{(i)}$ in Theorem 3.4 gives the result. □

EXAMPLE 6.4 (Bounded Finline Dielectric Waveguide). Consider the generalised eigenproblem $\mathbf{A}\mathbf{x} = \lambda\mathbf{M}\mathbf{x}$, where $\mathbf{A}$ and $\mathbf{M}$ are given by `bfw782a.mtx` and `bfw782b.mtx` in the Matrix Market library [13]. These are matrices of size 782, where $\mathbf{A}$ is real nonsymmetric and has 7514 non-zero entries, $\mathbf{M}$ is real symmetric indefinite and has 5982 non-zero entries. We seek the smallest eigenvalue in magnitude which is given by $\lambda_1 = 564.6$. Our only interest in this paper is the *outer convergence rate*, (though, for information we use GMRES for the inner solves in Algorithm 6.1). We use a variable shift given by the generalised Rayleigh quotient $\rho(\mathbf{x}^{(i)})$, and either a decreasing tolerance which is given by $\tau^{(i)} = \min\{0.05, 0.05\,\|\mathbf{r}^{(i)}\|\}$ or a fixed tolerance given by $\tau = 0.05$.

ETNA
Kent State University
etna@mcs.kent.edu

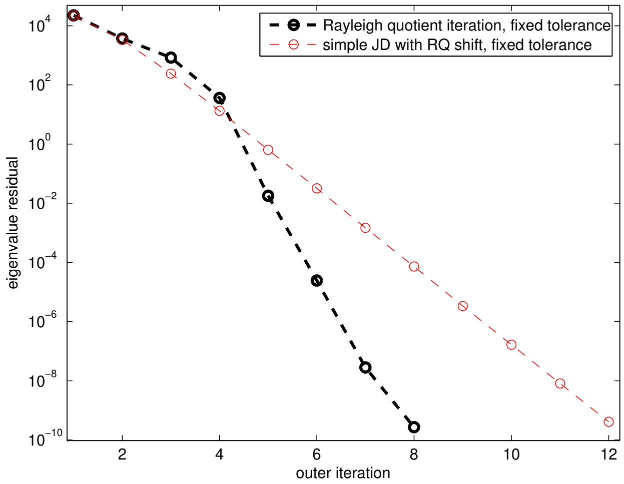

FIG. 6.1. *Convergence history of the eigenvalue residuals for Example 6.4 using* Rayleigh quotient shift *and inexact solves with fixed tolerance.*

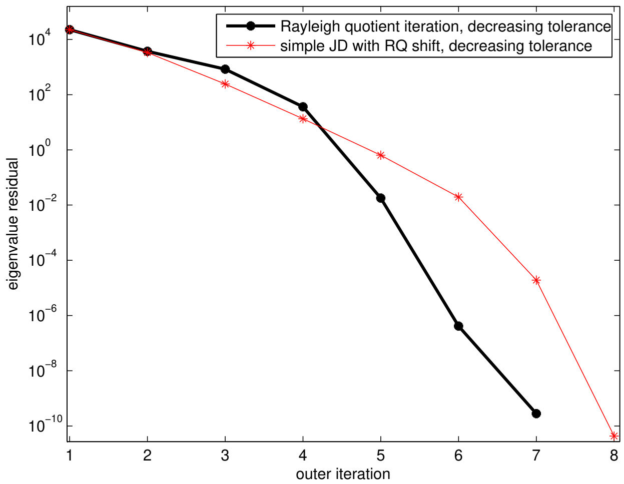

FIG. 6.2. *Convergence history of the eigenvalue residuals for Example 6.4 using* Rayleigh quotient shift *and inexact solves with decreasing tolerance.*

Figures 6.1 and 6.2 illustrate the convergence history for inexact Rayleigh quotient iteration and simple Jacobi-Davidson. We observe that a decreasing solve tolerance in the simple Jacobi-Davidson method with generalised Rayleigh quotient shift leads to quadratic convergence (Figure 6.2), whereas with a fixed solve tolerance only linear convergence may be achieved with a small enough tolerance (Figure 6.1). For comparison we have also plotted the results for inexact inverse iteration with a generalised Rayleigh quotient shift, where both the same decreasing tolerance $\tau^{(i)}$ and fixed tolerance τ were used as for the simple inexact Jacobi-Davidson method.

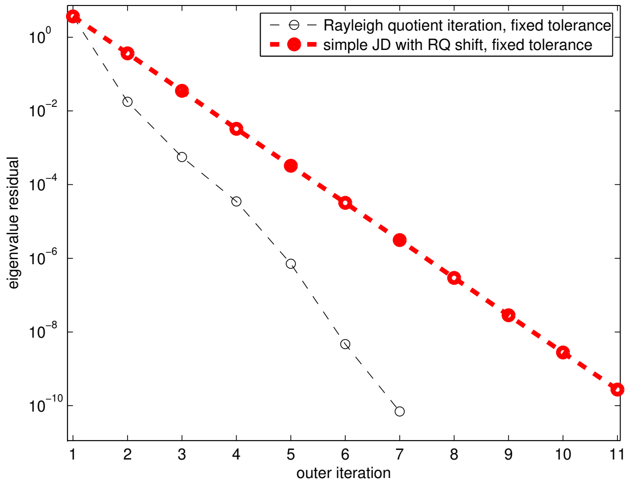

FIG. 6.3. *Convergence history of the eigenvalue residuals for Example 6.5 where* $\|\mathbf{r}^{(i)}\|/|\gamma^{(i)}| > 1$

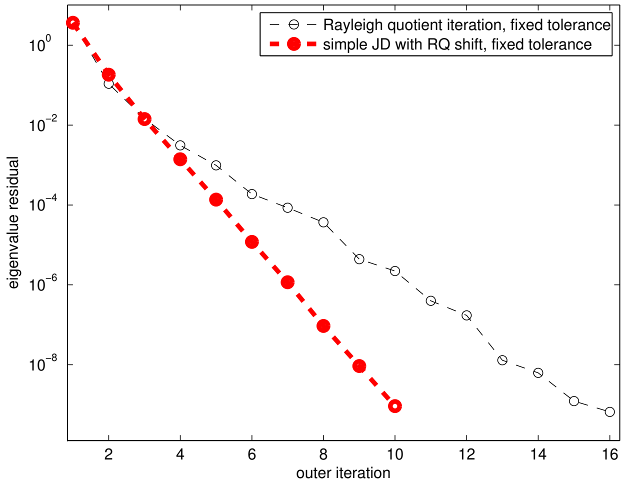

FIG. 6.4. *Convergence history of the eigenvalue residuals for Example 6.5 where* $\|\mathbf{r}^{(i)}\|/|\gamma^{(i)}| < 1$

Since, in this paper, we are only concerned about the outer convergence rate, from (6.7) we note that in theory the quantity $\|\mathbf{r}^{(i)}\|/|\gamma^{(i)}|$ is crucial for the comparison of the performance of the two methods. We note the following:

- If $\|\mathbf{r}^{(i)}\|/|\gamma^{(i)}| < 1$, then there is the potential that one step of the simple inexact Jacobi-Davidson method will perform better than one step of inexact Rayleigh quotient iteration.
- If $\|\mathbf{r}^{(i)}\|/|\gamma^{(i)}| > 1$, then there is the potential that one step of the inexact Rayleigh quotient iteration will perform better than one step of inexact simple Jacobi-Davidson method.

The following example illustrates this further.

ETNA
Kent State University
etna@mcs.kent.edu

TABLE 6.1
Values for $\|\mathbf{r}^{(i)}\|/|\gamma^{(i)}|$ in Figures 6.3 and 6.4 for fixed tolerance solves

It.	1	2	3	4	5	6	7	8	9	10
Fig. 6.3	27.4226	8.5952	4.0588	1.7692	1.3867	7.6525	1.2368	13.5016	1.2238	12.0983
Fig. 6.4	3.0399	0.7159	0.3132	0.1470	0.1706	0.4316	0.1368	0.7833	0.1401	

EXAMPLE 6.5. We construct two simple test examples, one for which the quantity $\|\mathbf{r}^{(i)}\|/|\gamma^{(i)}|$ turns out to be greater than one, and one for which this quantity is less than one. We use a standard eigenproblem $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ with $\mathbf{A} = \text{diag}(1, 2, \dots, 500)$ and set either $\mathbf{A}(1, 2 : 300) = 1$ (case (a)) or $\mathbf{A}(1, 2 : 300) = 10$ (case (b)). Clearly, in the second problem the nonnormality has been increased. We seek the smallest eigenvalue $\lambda_1 = 1$ and use GMRES for the inner solves. Further we use a variable shift given by the generalised Rayleigh quotient $\rho(\mathbf{x}^{(i)})$ and a fixed tolerance given by $\tau = 0.1$. We compare inexact Rayleigh quotient iteration and inexact simple Jacobi-Davidson. Both methods have linear convergence and stop once the eigenvalue residual satisfies $\|\mathbf{r}^{(i)}\| < 10^{-10}$.

Figure 6.3 illustrates the convergence history of the eigenvalue residuals for the two methods discussed above for case (a), the mildly nonnormal case. The corresponding values of $\|\mathbf{r}^{(i)}\|/|\gamma^{(i)}|$ are listed in the second row of Table 6.2 and turn out to be greater than one. As expected in this case, the convergence rate of inexact Rayleigh quotient iteration is better than the convergence rate of inexact simple Jacobi-Davidson with Rayleigh quotient shift. On the other hand, Figure 6.4 shows the convergence history of the eigenvalue residuals for case (b), there the nonnormality of the problem is larger. The corresponding values of $\|\mathbf{r}^{(i)}\|/|\gamma^{(i)}|$ are listed in the third row of Table 6.2 and are found to be less than one after the first iteration. As predicted, the convergence rate of inexact simple Jacobi-Davidson with Rayleigh quotient shift is better than inexact Rayleigh quotient iteration in this case.

Finally, we note that for Example 6.4 the quantity $\|\mathbf{r}^{(i)}\|/|\gamma^{(i)}|$ was greater than one throughout the computations, leading to a faster convergence rate for inexact Rayleigh quotient iteration. Further investigation onto this quantity is future research.

REFERENCES

[1] Z. BAI, J. DEMMEL, J. DONGARRA, A. RUHE, AND H. A. VAN DER VORST, *Templates for the Solution of Algebraic Eigenvalue Problems - A Practical Guide*, SIAM, Philadelphia, 2000.

[2] J. BERNS-MÜLLER, I. G. GRAHAM, AND A. SPENCE, *Inexact inverse iteration for symmetric matrices*, Linear Algebra Appl., 416 (2006), pp. 389–413.

[3] J. BERNS-MÜLLER AND A. SPENCE, *Inexact inverse iteration with variable shift for nonsymmetric generalized eigenvalue problems*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 1069–1082.

[4] D. R. FOKKEMA, G. L. G. SLEIJPEN, AND H. A. VAN DER VORST, *Jacobi-Davidson style QR and QZ algorithms for the reduction of matrix pencils*, SIAM J. Sci. Comput., 20 (1999), pp. 94–125.

[5] M. A. FREITAG AND A. SPENCE, *Convergence rates for inexact inverse iteration with application to preconditioned iterative solves*, BIT, 47 (2007), pp. 27–44.

[6] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 3rd edition, John Hopkins University Press, Baltimore, 1996.

[7] G. H. GOLUB AND J. H. WILKINSON, *Ill-conditioned eigensystems and the computation of the Jordan canonical form*, SIAM Rev., 18 (1976), pp. 578–619.

[8] G. H. GOLUB AND Q. YE, *Inexact inverse iteration for generalized eigenvalue problems*, BIT, 40 (2000), pp. 671–684.

[9] I. G. GRAHAM, A. SPENCE, AND E. VAINIKKO, *Parallel iterative methods for Navier-Stokes equations and application to eigenvalue computation*, Concurrency, Pract. Exp., 15 (2003), pp. 1151–1168.

[10] M. E. HOCHSTENBACH AND Y. NOTAY, *The Jacobi-Davidson method*, GAMM Mitt., 29 (2006), pp. 368–382.

[11] I. C. F. IPSEN, *Computing an eigenvector with inverse iteration*, SIAM Rev., 39 (1997), pp. 254–291.

[12] 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.

[13] *Matrix Market*, available online at `http://math.nist.gov/MatrixMarket/`.
[14] Y. NOTAY, *Combination of Jacobi-Davidson and conjugate gradients for the partial symmetric eigenproblem*, Numer. Linear Algebra Appl., 9 (2002), pp. 21–44.
[15] ———, *Convergence analysis of inexact Rayleigh quotient iteration*, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 627–644.
[16] ———, *Inner iterations in eigenvalue solvers*, Report GANMN 05-01, Université Libre de Bruxelles, Brussels, Belgium, 2005.
[17] B. N. PARLETT, *The Rayleigh quotient iteration and some generalizations for nonnormal matrices*, Math. Comp., 28 (1974), pp. 679–693.
[18] ———, *The Symmetric Eigenvalue Problem*, vol. 20 of Classics in Applied Mathematics, SIAM, Philadelphia, 1998.
[19] G. PETERS AND J. H. WILKINSON, *Inverse iteration, ill-conditioned equations and Newtons's method*, SIAM Rev., 21 (1979), pp. 339–360.
[20] V. SIMONCINI AND L. ELDÉN, *Inexact Rayleigh quotient-type methods for eigenvalue computations*, BIT, 42 (2002), pp. 159–182.
[21] G. L. G. SLEIJPEN, A. G. L. BOOTEN, D. R. FOKKEMA, AND H. A. VAN DER VORST, *Jacobi-Davidson type methods for generalized eigenproblems and polynomial eigenproblems*, BIT, 36 (1996), pp. 595–633.
[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] ———, *The Jacobi-Davidson method for eigenvalue problems and its relation with accelerated inexact Newton schemes*, in Iterative Methods in Linear Algebra, II., S. D. Margenov and P. S. Vassilevski, eds., vol. 3 of IMACS Series in Computational and Applied Mathematics, New Brunswick, NJ, U.S.A., 1996, pp. 377–389.
[24] ———, *A Jacobi-Davidson iteration method for linear eigenvalue problems*, SIAM Rev., 42 (2000), pp. 267–293.
[25] 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.
[26] G. W. STEWART, *On the sensitivity of the eigenvalue problem $Ax = \lambda Bx$*, SIAM J. Numer. Anal., 9 (1972), pp. 669–686.
[27] ———, *Matrix Algorithms II: Eigensystems*, SIAM, Philadelphia, 2001.
[28] G. W. STEWART AND J.-G. SUN, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.
[29] L. N. TREFETHEN AND D. I. BAU, *Numerical Linear Algebra*, SIAM, Philadelphia, 1997.
[30] J. VAN DEN ESHOF, *The convergence of Jacobi-Davidson for Hermitian eigenproblems*, Numer. Linear Algebra Appl., 9 (2002), pp. 163–179.