Inexact Inverse Iteration for Symmetric Matrices

Jörg Berns-Müller*  Ivan G. Graham*  Alastair Spence*

Abstract

In this paper we analyse inexact inverse iteration for the real symmetric eigenvalue problem $Av = \lambda v$. Our analysis is designed to apply to the case when $A$ is large and sparse and where iterative methods are used to solve the shifted linear systems $(A - \sigma I)y = x$ which arise. We first present a general convergence theory that is independent of the nature of the inexact solver used. Next we consider in detail the performance of preconditioned MINRES as the solver for $(A - \sigma I)y = x$. In particular we analyse the cost of one solve and then the overall cost of the whole algorithm. Also we present a new treatment of the approach discussed by [18] to set up an appropriate right hand side for the preconditioned system. Our analysis shows that, if MINRES is the inexact solver then the optimal strategy is to let the shift tend to the desired eigenvalue as quickly as possible. In most practical situations that will mean that the shift should be taken as the Rayleigh quotient, just as if direct solves were used. Numerical results are given to illustrate the theory in the paper.

AMS subject classification: Primary 65F15, Secondary 65F10

Keywords: Inverse iteration, iterative methods, preconditioned MINRES, Rayleigh quotient iteration

1 Introduction

In this paper we discuss the effect of inexact (iterative) solves on inverse iteration for the eigenvalue problem

$$Av = \lambda v,$$  \hfill (1)

where $A$ is a large, sparse, symmetric real matrix. Inverse iteration requires the solution of shifted linear systems of the form

$$(A - \sigma I)y = x,$$  \hfill (2)

where $\sigma$ is the shift. If $A$ is large and sparse, say arising from a discretised partial differential equation in 3D, direct methods become impractical and iterative
methods with preconditioning become necessary to solve (2). In this setting we arrive at an inner-outer iterative method for (1): the outer iteration is the basic iterative inverse iteration algorithm requiring the solve of (2) at each step, with the inner iteration being the inexact iterative solution of (2).

Here we are thinking of inverse iteration as a technique in its own right for finding an eigenvalue and eigenvector (and which can be interpreted as Newton’s method, see, for example, [4]) rather than the standard technique of finding an eigenvector given a very accurate estimate for the eigenvalue. Of course nowadays one would almost certainly use a Lanczos-type algorithm, perhaps in the shift-invert mode, to solve (1), but we believe that an in-depth understanding of the basic inexact inverse iteration algorithm for a simple eigenvalue is required before we can hope to understand properly the performance of more sophisticated algorithms if inexact solves are used for shifted systems.

A very early paper on the use of iterative methods to solve (2) is [17]. Inexact Inverse Iteration for symmetric matrices was discussed in [20] where a general theory, independent of the details of the solver was presented along with some new eigenvalue bounds. An important recent paper on inexact inverse iteration is [18] where a version of inexact Rayleigh quotient iteration is discussed. Several key ideas are introduced especially with regard to the derivation of the appropriate linear system to be solved when Cholesky preconditioning is applied to (2), and the determination of an appropriate stopping condition in the inner iteration. We shall discuss these ideas in detail in this paper. Also [18] contains a theoretical discussion on the equivalence of inexact inverse iteration and Jacobi-Davidson (or projected Newton’s Method). For nonsymmetric matrices a fixed shift inexact inverse iteration algorithm is discussed in [6]. A convergence theory is given along with an analysis of the choice of tolerance used in the inner solves. Convergence results for non-symmetric matrices are also given in [11] Other related work on the use of inexact Rayleigh quotient iteration to compute the smallest eigenvalue of generalised Hermitian eigenvalue problems is discussed in [10] and [12].

There are several new features in this paper. In Section 2 we present a very general convergence theory, independent of the details of the inexact iterative solver. This allows us to recover and extend existing results of [20]. Also we include in the theory the option to change the right hand side of (2) to allow for flexibility when considering preconditioned solves. In Section 3 we use some standard results for MINRES (see, for example, [14, 8]) to provide a new “a priori” lower bound on the number of inner iterations at each step of the outer iteration. This is seen in our numerical examples to provide qualitatively correct information about the performance of both unpreconditioned and preconditioned inner solves. In contrast to the lower bound in Section 3, we derive in Section 4 a new “a posteriori” upper bound on the number of inner iterations at each step of the outer iteration and hence obtain an upper bound on the total number of inner iterations for both unpreconditioned and preconditioned iterations. The usefulness of the “a posteriori” upper bound is that it allows us to provide an analysis of the overall efficiency of the inexact inverse iteration algorithm. Numerical experiments show that this analysis describes well both
the behaviour of the inner iterations and the total number of inner iterations
needed to achieve a desired accuracy. One important consequence of this “a
posteriori” bound analysis is that the most efficient methods are seen to be
those which choose the shifts to converge as quickly as possible to the desired
eigenvalues. In practice this will probably mean that the shifts should be chosen
to be Rayleigh quotients. This is consistent with the best strategy when direct
solves are used and shows that we need not be concerned that the Krylov solver
is applied to a matrix which is becoming more and more singular. The explana-
tion lies in the interplay between the shift tending towards the eigenvalue and
the right hand side tending to the corresponding eigenvector, together with the
fact that Krylov solvers handle very well nearly singular systems with only a
small number of critical eigenvalues. Similar ideas were explored in [17]. An-
other feature of our paper is that we provide a new analysis of the approach in
[18] and confirm that their choice of right hand side for preconditioned Rayleigh
quotient iteration is superior to the standard preconditioned system. However,
our analysis is valid for any preconditioner and is not restricted to Cholesky
preconditioners based on the the factorisation as in [18].

Finally in Section 5 we present several numerical examples to illustrate the
theory in the paper. We also give a brief comparison of some stopping condi-
tions.

We mention that a more detailed account of the material in this paper,
including more extensive numerical tests, is contained in Chapters 2 and 3 of
[1].

2 Inexact Inverse Iteration

2.1 Preliminaries

Consider the solution of the eigenvalue problem

\[ \mathbf{A} \mathbf{v} = \lambda \mathbf{v}, \quad \| \mathbf{v} \|_2 = 1, \]  

(3)

where \( \mathbf{A} \) is a real symmetric \( n \times n \) matrix, with eigenvalues \( \lambda_j, j = 1, \ldots, n \) and
corresponding orthonormalised eigenvectors \( \mathbf{v}_j, j = 1, \ldots, n \). Inverse Iteration
for (3) requires the solution of shifted systems of the form \((\mathbf{A} - \sigma \mathbf{I}) \mathbf{y} = \mathbf{x}\) for
some chosen real shift \( \sigma \). Let \( (\lambda_1, \mathbf{v}_1) \) denote a simple eigenpair of (3) which we
wish to compute. Throughout this paper we will be interested in shifts \( \sigma \) which
are close enough to \( \lambda_1 \) in the sense that

\[ 0 < |\lambda_1 - \sigma| < \frac{1}{2} \min_{j=2,\ldots,n} |\lambda_1 - \lambda_j|. \]  

(4)

Then we have an induced ordering on the eigenvalues

\[ 0 < |\lambda_1 - \sigma| < |\lambda_2 - \sigma| \leq \ldots \leq |\lambda_n - \sigma|. \]

This ordering depends on \( \sigma \) but for convenience we ignore this in the notation.
We are interested in the case when $A$ is large and sparse and so the shifted systems will be solved (inexactly) by some iterative algorithm. In Section 3 we will assume that the iterative solver is MINRES (see, for example, [8]), which is appropriate since $A - \sigma I$ is symmetric. However, in this section we will present a convergence theory that is independent of the solver. To this end we introduce the following definition.

**Definition 2.1** For any $B \in \mathbb{R}^{n \times n}, b \in \mathbb{R}^n$, any norm $\| \cdot \|$ on $\mathbb{R}^n$, and any tolerance $\tau \geq 0$, we define the set

$$\mathcal{A}(B, b, \| \cdot \|, \tau) := \{ z \in \mathbb{R}^n : \| b - Bz \| \leq \tau \}.$$

This is the set of inexact solutions of $Bz = b$ such that the corresponding residual is less than the absolute tolerance $\tau$. Obviously when $B$ is nonsingular and $\tau = 0$ the only element of $\mathcal{A}$ is the unique solution $z$.

Unless indicated otherwise throughout this paper, we will assume that $\| \cdot \| = \| \cdot \|_2$, the standard Euclidean norm in $\mathbb{R}^n$.

The inexact inverse iteration algorithm is given in Algorithm 1.

---

**Algorithm 1: Inexact Inverse Iteration**

Given $x^0$ with $\| x^0 \| = 1$. For $i = 0, 1, 2, \ldots$

- Choose $\sigma^i$ and $\tau^i$
- Calculate $y^i \in \mathcal{A}(A - \sigma^i I, x^i, \| \cdot \|, \tau^i)$
- Update $x^{i+1} = y^i / \| y^i \|
- Test for convergence

We refer to the iteration in Algorithm 1 as the *outer* iteration and the iteration implicit in Definition 2.1 as the *inner* iteration, and so inexact inverse iteration is an example of an inner-outer iterative algorithm (see, for example, [2, 3, 19, 5]).

Whatever iterative algorithm is used to solve $(A - \sigma^i I)y^i = x^i$ in Algorithm 1 it will almost certainly be applied to a preconditioned system, and since $A$ is symmetric it is common to use a positive definite preconditioner. Let $P$ be a positive definite matrix that approximates $(A - \sigma^i I)$ in some way. Then, at least in theory, there exists a factorisation $P = P_1 P_1^T$. To respect the symmetry in the system it is normal to consider the usual symmetrically preconditioned system

$$P_1^{-1}(A - \sigma^i I)P_1^{-T}\tilde{y}^i = P_1^{-1}x^i; \quad y^i = P_1^{-T}\tilde{y}^i. \quad (5)$$
Of course the preconditioner $P$ or the factorisation $P = P_1 P_T^T$ are often not needed in practice and implementation of (5) will require only the action of $P^{-1}$. An alternative to (5) is discussed in [18], which we shall see in Section 4 is beneficial for the performance of the preconditioned iterative solver. Their idea is to replace (5) by

$$P_1^{-1} (A - \sigma^i I) P_1^{-T} \tilde{y}^i = P_T^T x^i; \quad y^i = P_1^{-T} \tilde{y}^i,$$

where the right hand side in the shifted linear system differs from that of (5). In order to analyse both (5) and (6) we consider the convergence of inexact inverse iteration using the inexact solve of

$$P_1^{-1} (A - \sigma^i I) P_1^{-T} \tilde{y}^i = T x^i; \quad y^i = P_1^{-T} \tilde{y}^i,$$

for a general matrix $T$, though $T = P_T^T$ recovers (6) and $T = P_1^{-1}$ recovers (5). If $P_1 = I$, and $T = I$ then we return to the original unpreconditioned system $(A - \sigma^i I) y^i = x^i$.

**Algorithm 2: Preconditioned Inexact Inverse Iteration**

Choose $P = P_1 P_T^T$ and $T$. Given $x^0$ with $\|x^0\| = 1$

For $i = 0, 1, 2, \ldots$

- Choose $\sigma^i$ and $\tau^i$
- Calculate $\tilde{y}^i \in A(P_1^{-1} (A - \sigma^i I) P_1^{-T}, T x^i, \|\cdot\|, \tau^i)$
- $y^i = P_1^{-T} \tilde{y}^i$
- Update $x^{i+1} = y^i / \|y^i\|$
- Test for convergence

In the next subsection we analyse the convergence of Algorithm 2, a general symmetrically preconditioned inverse iteration algorithm. To do this we use the orthogonal splitting approach in [15, p. 63]. If $x^i$ is an approximation to $v_1$ we introduce the splitting

$$x^i = \cos \theta^i v_1 + \sin \theta^i u^i, \quad u^i \perp v_1,$$

with $\|v_1\| = \|u^i\| = 1$ and $\theta^i = \angle(x^i, v_1)$, the error angle. For convenience we usually write

$$c^i = \cos \theta^i, \quad s^i = \sin \theta^i, \quad \text{and} \quad t^i = \|s^i|/|c^i|,$$
with $t^i = |\tan \theta^i|$. From (8), $||x^i - c^i v_1|| = |s^i| \leq t^i$ and we use $|s^i|$ or $t^i$ as a measure of the convergence of $x^i$ to span $\{v_1\}$. Also, recall that for $x^i$ given by (8) the Rayleigh quotient, $\varrho(x^i) = x^i^T A x^i$, satisfies

$$\lambda_1 - \varrho(x^i) = (s^i)^2 [\varrho(u^i) - \lambda_1],$$

(10)

and the eigenvalue residual, $r^i$, satisfies

$$r^i := (A - \varrho(x^i) I)x^i = s^i (A - \lambda_1 I)u^i + (s^i)^2 [\varrho(u^i) - \lambda_1] x^i$$

(11)

and so $||r^i|| = O(|s^i|)$ for small enough $|s^i|$.

### 2.2 Convergence Theory

In this subsection we present the theory for the convergence of inexact inverse iteration. Various choices for $\sigma^i$ and $\tau^i$ in Algorithm 2 are possible and our analysis allows us to obtain a full understanding of the different possible options. For example, it is a classical result proved by [13] that inverse iteration with Rayleigh quotient shifts and exact linear solves converges cubically when applied to symmetric matrices (see [15] for an elegant treatment). It is natural to ask how the tolerance $\tau^i$ should be chosen so that inexact inverse iteration with Rayleigh quotient shifts can recover cubic convergence. We shall see in Corollary 2.1, part c), how to do this.

We start with a Lemma that provides a bound for the error in one step of inexact inverse iteration in Algorithm 2.

**Lemma 2.1** If $x^i$, $T$ and $P_1$ are such that $|v_1^T P_1 T x^i| > \|P_1\| \tau^i$ then, using the notation in (9), one step of preconditioned inexact iteration using Algorithm 2 yields $x^{i+1}$ and

$$t^{i+1} \leq \frac{\lambda_1 - \sigma^i}{\lambda_2 - \sigma^i} \frac{\| P_1 T x^i \| + \| P_1 \| \tau^i}{\| v_1^T P_1 T x^i \| - \| P_1 \| \tau^i}$$

(12)

where $P = I - v_1 v_1^T$.

**Proof:** First with $\tilde{y}^i \in A(P_1^{-1}(A - \sigma^i I)P_1^{-T}, T x^i, \| \cdot \|_2, \tau^i)$ the inexact solve step yields a residual

$$\text{res}^i := T x^i - P_1^{-1} (A - \sigma^i I)P_1^{-T} \tilde{y}^i$$

(13)

with $y^i = P_1^{-T} \tilde{y}^i$ and

$$\| \text{res}^i \|_2 \leq \tau^i.$$  

(14)

Rewrite (13) as

$$\|y^i\| (A - \sigma^i I)x^{i+1} = P_1 T x^i - P_1 \text{res}^i$$

and use the orthogonal splitting (8) to give

$$\|y^i\| \{ c^{i+1}(\lambda_1 - \sigma^i) v_1 + s^{i+1}(A - \sigma^i I)u^{i+1} \} = P_1 T x^i - P_1 \text{res}^i.$$
We now resolve into two equations in $\text{span}\{v_1\}$ and $\{v_1\}^\perp$, respectively. Along $v_1$ we have
\[
\|y^i\| c^{i+1} (\lambda_1 - \sigma^i) = v_1^T P_1 T x^i - v_1^T P_1 \text{res}^i. \tag{15}
\]
Also let $(A - \sigma^i I)\perp$ denote the restriction of $(A - \sigma^i I)$ to $\{v_1\}^\perp$. This is invertible on $\{v_1\}^\perp$ and satisfies $\|(A - \sigma^i I)\perp^{-1}\| \leq |\lambda_2 - \sigma^i|^{-1}$, and so in $\{v_1\}^\perp$ we have
\[
\|y^i\| c^{i+1} \leq \|(A - \sigma^i I)\perp^{-1}\| \{\|\Pi T x^i\| + \|\Pi P_1 \text{res}^i\|\} \tag{16}
\]
where $\Pi := (I - v_1 v_1^T)$. Hence using (15) and (16) we obtain
\[
t^{i+1} \leq \frac{\|y^i\| c^{i+1}}{\lambda_1 - \sigma^i} \frac{|\Pi P_1 T x^i\| + \|\Pi P_1 \text{res}^i\|}{|v_1^T P_1 T x^i| - |v_1^T P_1 \text{res}^i|},
\]
from which (12) follows. Therefore
\[
t^{i+1} \leq \frac{\|y^i\| c^{i+1}}{\lambda_2 - \sigma^i} \frac{\|\Pi P_1 T x^i\| + \|\Pi P_1 \text{res}^i\|}{|v_1^T P_1 T x^i| - |v_1^T P_1 \text{res}^i|},
\]
(17)
In some cases there will be an option to stop the inner iteration when $\|P_1 T x^i - (A - \sigma^i I)y^i\| \leq \tau^i$, rather than that assumed above. In this case $\|P_1 \text{res}^i\| \leq \tau^i$ and we obtain
\[
t^{i+1} \leq \frac{\|y^i\| c^{i+1}}{\lambda_1 - \sigma^i} \frac{\|\Pi P_1 T x^i\| + \|\Pi P_1 \text{res}^i\|}{|v_1^T P_1 T x^i| - |v_1^T P_1 \text{res}^i|},
\] instead of (12).

In the following subsection we look at three special cases providing convergence results in each case.

### 2.3 Convergence Theory for Unpreconditioned Solves

In this subsection we provide a convergence theory for the case of unpreconditioned Inexact Inverse Iteration, that is Algorithm 1. In this case $P_1 = I = T$ and recalling (8) the one step bound (12) reduces to
\[
t^{i+1} \leq \frac{|\lambda_1 - \sigma^i|}{|\lambda_2 - \sigma^i|} \frac{\|\Pi P_1 T x^i\| + \|\Pi \text{res}^i\|}{|v_1^T P_1 T x^i| - |v_1^T P_1 \text{res}^i|}, \tag{18}
\]
which provides the standard error bound for exact inverse iteration if $\tau^i = 0$ (see, for example, [15]). Equation (18) is sufficient for our needs in this paper, but a more refined analysis is possible using (15) and (16), to give
\[
t^{i+1} \leq \frac{|\lambda_1 - \sigma^i|}{|\lambda_2 - \sigma^i|} \frac{\|\Pi \text{res}^i\|}{|v_1^T P_1 \text{res}^i|},
\]
which would be useful if more were known about the size of certain components of $\text{res}^i$ (see [1] for more on this theory). Bounds similar to (18) can be found in [20], and [6] (for non-symmetric problems). We use (18) to give a general convergence result after which we look at three different choices for $\sigma^i$ and $\tau^i$. 7
Theorem 2.1  Let $A$ be a real $n \times n$ symmetric matrix and consider the application of inexact inverse iteration defined by Algorithm 1 to find the simple eigenpair $(\lambda_1, v_1)$. With $x^i = c^iv_1 + s^iu^i$, assume $\sigma^i$ and $\tau^i$ satisfy
\[ |\lambda_1 - \sigma^i| \leq \min\{C_1|s^i|^\alpha, \frac{1}{2}|\lambda_2 - \lambda_1|\} \]
\[ \tau^i \leq \min\{C_2|s^i|^\beta, C_3|c^i|\} \]
for some positive constants $\alpha, \beta, C_1, C_2, C_3$ with $0 \leq \beta \leq 1$, $0 \leq C_3 < 1$ and $\alpha + \beta \geq 1$. If $c^0 \neq 0$, and the initial approximation $x^0$ is such that
\[ C_4 := |s^0|^{\alpha + \beta - 1} \frac{2C_1(1 + C_2)}{|\lambda_2 - \lambda_1|(1 - C_3)} < 1 \]
then
\[ t^{i+1} \leq C_4 t^i \forall i, \]
and the method converges. Hence, $t^{i+1} \leq C(t^i)^{\alpha + \beta}$ for some positive constant $C$ independent of $i$, and so convergence is of order $(\alpha + \beta)$.

Proof: From (18) we have
\[ t^{i+1} \leq \frac{2C_1|s^i|^\alpha |s^i| + C_2|s^i|^\beta}{|\lambda_2 - \lambda_1| |c^i| - C_3|c^i|} \]
\[ \leq t^i \frac{|s^i|^{\alpha + \beta - 1}}{|s^0|^{\alpha + \beta - 1}} C_4 \]
Convergence follows by induction on $t^i$. Finally
\[ t^{i+1} \leq \frac{C_4}{|s^0|^{\alpha + \beta - 1}} (t^i)^{\alpha + \beta}. \] (19)

As a consequence of Theorem 2.1 we obtain the convergence of $x^i$ to $\pm v_1$ and $g(x^i)$ to $\lambda_1$. Note that the condition $c^0 \neq 0$ ensures that the starting vector is not orthogonal to the required direction $v_1$.

For the symmetric eigenvalue problem under discussion here, the natural choice for $\sigma^i$ is the Rayleigh Quotient given by (10), so that $|\lambda_1 - \sigma^i| \leq C_1|s^i|^2$. Possible choices for the solver tolerance would be to keep $\tau^i$ fixed ($\beta = 0$ in Theorem 2.1) or choose $\tau^i$ proportional to the eigenvalue residual $r^i$, so that $\tau^i \leq C_2|s^i|$ using (11). Alternatively one could also keep $\sigma^i$ fixed and choose $\tau^i \leq C_2|s^i|$. Note the theory indicates that there is likely to be little gain in choosing $\tau^i = o(|s^i|)$. For convenience, convergence rates for three likely options are summarised in the following Corollary.

Corollary 2.1  Suppose $x^0, \sigma^0$ and $\tau^0$ satisfy the conditions of Theorem 2.1 for $i = 0$. 

8
a) If \( \sigma^i = \sigma^0 \) and \( \tau^i = \min \{ \tau^0, C_2 \| r^i \| \} \) then \( \alpha = 0, \beta = 1 \) and Algorithm 1 will converge linearly.

b) If \( \sigma^i = g(x^i) \) and \( \tau^i = \tau^0 \) then \( \alpha = 2, \beta = 0 \) and Algorithm 1 will converge quadratically.

c) If \( \sigma^i = g(x^i) \) and \( \tau^i = \min \{ \tau^0, C_2 \| r^i \| \} \) then \( \alpha = 2, \beta = 1 \) and Algorithm 1 will converge cubically.

Numerical experiments illustrating these convergence results are given in Example 5.1.

Note that case c) shows that it is possible to recover with inexact solves the cubic convergence achieved by exact solves as proved in [13]. This result is implicit in [20], but the present theory is more general than [20]. Also we see that a strategy based on a fixed shift, \( \sigma^0 \) and a fixed tolerance, \( \tau^0 \neq 0 \) is unlikely to converge as can be verified by the following simple example.

Example 2.1 Suppose that \( x^i = c^i v_1 + s^i v_2 \). If we construct a particular \( y^i \) via the formula, \( y^i = c^i \lambda_1 - \sigma^0 v_1 + s^i \), then computing \( x^{i+1} \) and representing it in the form (8) we obtain \( \tan \theta^{i+1} = \frac{\lambda_1 - \sigma^i}{\lambda_2 - \sigma^i} - \frac{s^i}{c^i} \). Repeating this formula for each \( i \) we obtain a fixed point iteration. The fixed point satisfies \( \sin \theta = \frac{\lambda_1 - \sigma^0}{\lambda_2 - \lambda_1} \tau^0 \), which is nonzero unless \( \sigma^0 = \lambda_1 \) or \( \tau^0 = 0 \). This non-convergence is often referred to as stagnation, see, for example, [16].

2.4 Convergence Theory for Preconditioned Solves

In this section we first discuss briefly the convergence theory for standard preconditioned inverse iteration. Next we present a new convergence analysis of the variation introduced by [18].

Consider the theory for the standard form of preconditioned inexact inverse iteration, that is we take \( T = P_1^{-1} \) in Algorithm 2 and solve \( Bz^i = b^i \) where \( B = P_1^{-1}(A - \sigma^i I)P_1^{-T} \) and \( b^i = P_1^{-1} x^i \). Thus the onestep bound (12) gives

\[
t^{i+1} \leq \frac{|\lambda_1 - \sigma^i|}{|\lambda_2 - \sigma^i|} \left( \| s^i \| + \| b^i \| \right) \tau^i
\]

where the preconditioned solve stops using the residual condition \( \| b^i - Bz^i \| \leq \tau^i \). Here the only difference between the theory for the preconditioned and unpreconditioned cases is that \( \tau^i \) terms on the right hand side of (18) are replaced by \( \| P_1 \| \tau^i \) and this factor can be included in the constants \( C_2 \) and \( C_3 \) in the bounds for \( \tau^i \) in Theorem 2.1. Hence Theorem 2.1 and Corollary 2.1 can be applied. For completeness we note this in the following Corollary.

Corollary 2.2 Assume Algorithm 2 with \( T = P_1^{-1} \) is used to find the simple eigenpair \( (\lambda_1, v_1) \) and that the conditions of Theorem 2.1 hold.
a) Then the conclusions of Theorem 2.1 and Corollary 2.1 hold.

b) In particular, assume \( \sigma^i \) is chosen as the Rayleigh quotient. Then if \( \tau^i = \tau^0 \), quadratic convergence would be attained. If, however, \( \tau^i \) is chosen proportional to the eigenvalue residual, then cubic convergence would be attained.

The two variations of the Rayleigh quotient iteration outlined in b) are used in Example 5.2 of Section 5 and the numerical results are given in Table 3.

Now we discuss the case where \( \sigma^i \) is taken to be the Rayleigh quotient and Algorithm 2 is used with \( T = P^T I \) so that instead of solving \((A - \sigma^i I)y^i = x^i\) by the standard preconditioned iterative method, we are in fact solving \((A - \sigma^i I)y^i = Px^i\) as suggested implicitly in [18]. In this case the one step bound (12) gives

\[
t^i + 1 \leq \frac{|\lambda_1 - \sigma^i|}{|\lambda_2 - \sigma^i|} \left[ \|P\| \|x^i\| + \|P_1\| \tau^i \right].
\]

For a general preconditioner we will be required to have available a subroutine for computing the action of \( P^{-1} \). However in general it is not required to know the action of \( P \) and in some cases the action of \( P \) is not available. Thus in general all that can be said is that, under suitable conditions, the second quotient on the right hand side of (20) will be bounded. So we can only ensure quadratic convergence of the outer iteration by forcing quadratic convergence of the shift \( \sigma^i \) to \( \lambda_1 \) such as is achieved by the Rayleigh quotient shift. Hence we have the following theorem.

**Theorem 2.2** Let \( A \) be a real \( n \times n \) matrix and consider the application of preconditioned inexact inverse iteration defined by Algorithm 2 using (6) to find the simple eigenpair. With \( x^i = c^i v_1 + s^i u^i \), and \( c^0 \neq 0 \), assume \( \sigma^i \) and \( \tau^i \) satisfy

\[
|\lambda_1 - \sigma^i| \leq \min\{C_1 |s^i|^2, \frac{1}{2} |\lambda_2 - \lambda_1|\}
\]

\[
\tau^i \leq \min\{C_2, C_3 |v_1^T Px^i|\}
\]

for some positive constants \( C_1, C_2 \) and \( C_3 \). If the initial approximation \( x^0 \) is such that

\[
C_4 := \frac{|s^0|^2 |\lambda_1 - \sigma^i|}{|\lambda_2 - \lambda_1| |v_1^T Px^0| (1 - C_3)} < 1
\]

then the method is convergent with \( t^{i+1} \leq C_4 (t^i)^2 \).

**Proof:** From (20)

\[
t^{i+1} \leq \frac{2C_1 |s^i|^2}{|\lambda_2 - \lambda_1|} \left[ \|P\| + \|P_1\| \right] \tau^i \leq \frac{2C_1 (\|P\| + \|P_1\| C_2)}{|\lambda_2 - \lambda_1| |v_1^T Px^i| (1 - C_3)} |s^i|^2
\]

(21)
As $C_4 < 1$, then $t^{i+1} \leq C_4 t^i$ and convergence occurs. From (21) the convergence is quadratic. \(\square\)

Numerical results illustrating this theory are given in Example 5.3. For theoretical interest only, we note that if $P = A$ then (12) reduces to

$$t^{i+1} \leq \frac{|\lambda_1 - \sigma|}{|\lambda_2 - \sigma|} |s^i| |\lambda_1| + \|P_1\| \tau^i$$

and now it is possible to recover a cubically convergent method if $\tau^i$ is chosen proportional to $|s^i|$.

However there is the potential of a significant reduction in the domain of convergence when we decide to solve $(A - \sigma I)y_i = Px^i$ rather than $(A - \sigma I)y^i = x^i$ as the following simple example shows.

**Example 2.2 (Reduced domain of convergence using (6))** Assume $A$ is symmetric positive definite and let $P = A$. Take the factorisation $P_1 = A^{\frac{1}{2}}$. Then $P_1^{-1}(A - \sigma I)P_1^{-1}y^i = P_1^T x^i$, which is the preconditioned system (6), reduces to $(I - \sigma A^{-1})y^i = x^i$. Assume now $x^i = c^i v_1 + s^i v_2$ and that exact solves are used to obtain $y^i$. We readily obtain the one step bound

$$t^{i+1} \leq \frac{|\lambda_1 - \sigma|}{|\lambda_2 - \sigma|} |\lambda_2| |\lambda_1| - \|P_1\| \tau^i$$

Now the factor $|\lambda_2/\lambda_1|$ may be large enough to significantly reduce the domain of convergence. The domain of convergence would be reduced further if one added the effects of an inexact solver and a less good preconditioner.

With such a possible drawback there must be some gain in considering this option. As we see in Section 4.4 the benefit comes when one considers the number of inner iterations needed to solve the linear system when using a Krylov method.

### 3 The Iterative Solver

Since $(A - \sigma I)$ is symmetric but probably indefinite the natural Krylov method to solve $(A - \sigma I)y^i = x^i$ is MINRES (see, for example, [7]). As discussed in the previous section the linear system will normally be solved using a preconditioner, $P$ say, where $P$ is positive definite and approximates $(A - \sigma I)$ in some way. If $A$ arises from a discretised partial differential equation $P^{-1}$ may be constructed for example using a domain decomposition or multigrid technique. Alternatively $P$ may be obtained using a Cholesky factorisation of $A$ (or of a shifted $A$).

Let us first review some standard results on MINRES applied to a general symmetric linear system

$$Bz = b. \tag{22}$$

Define the Krylov space $K_k(B, b)$ by

$$K_k(B, b) = \text{span} \{b, Bb, \ldots, B^{k-1}b\}.$$
Throughout this paper we take an initial guess \( z_0 = 0 \), though, other choices are possible. MINRES seeks a solution \( z_k \in K_k(B, b) \) characterised by the property \( \| b - Bz_k \|_2 = \min_{z \in K_k} \| b - Bz \|_2 \). Assume \( B \) has an eigenvalue decomposition of the form \( B = W\Lambda W^T \), where \( W \) is orthogonal and \( \Lambda = \text{diag}(\mu_1, \ldots, \mu_n) \). Then

\[
\| b - Bz_k \|_2 = \min_{q \in P_k} \| q(B)b \|_2
\]

where \( P_k \) denotes the space of polynomials of degree \( k \) with \( q(0) = 1 \). Thus

\[
\| b - Bz_k \|_2 = \min_{q \in P_k - |J|} \| q(\Lambda)w^Tb \|_2
\]

where \( \kappa = \frac{\max |\mu_j|}{\min |\mu_j|} \). However, this bound may not reflect the performance of MINRES in our application, where because of the particular distribution of the spectrum, MINRES performs significantly better than indicated by (24).

Assume that the spectrum of \( B \) contains a small number of successive eigenvalues which are in some way distinguished from the remaining eigenvalues of \( B \). If \( J \subset \mathbb{N}_n := \{1, 2, \ldots, n\} \) then \( \{\mu_j\}_{j \in J} \) is the distinguished set and \(|J|\) denotes the number of elements in it. Set \( J^c := \mathbb{N}_n - J \) and \( Q_J := \text{diag}(\delta_1, \ldots, \delta_n) \) where \( \delta_j = 0 \) if \( j \in J \) and \( \delta_j = 1 \) otherwise. Further, as in [7, §3.1] or [9, §7.3.6], introduce the polynomial

\[
p_J(t) := \prod_{j \in J} \frac{\mu_j - t}{\mu_j}
\]

which is zero for \( t \in \{\mu_j\}_{j \in J} \). Rewrite (23) as

\[
\begin{align*}
\| b - Bz_k \|_2 &\leq \min_{q \in P_k - |J|} \| q(\Lambda)p_J(\Lambda)W^Tb \|_2 \\
&\leq \min_{q \in P_k - |J|} \| q(\Lambda)p_J(\Lambda)Q_JW^Tb \|_2 \\
&\leq \min_{q \in P_k - |J|} \| q(\Lambda)p_J(\Lambda)Q_JW^Tb \|_2 \\
&\leq \min_{q \in P_k - |J|} \| q(\Lambda)p_J(\Lambda)Q_JW^Tb \|_2 \\
&\leq \left\{ \min_{q \in P_k - |J|} \| q(\mu_j) \| \right\} \left\{ \max_{j \in J'} \left| \frac{\mu_i - \mu_j}{\mu_i} \right| \right\} \| Q_JW^Tb \|_2 .
\end{align*}
\]

Using (26) and standard results on Chebyshev polynomials (see, for example, [8, §3.1] or [9, §7.3.4]) we obtain the following theorem.
Theorem 3.1  Suppose that the symmetric matrix $B$ has eigenvalues $\mu_1, \ldots, \mu_n$ with corresponding eigenvectors $w_1, \ldots, w_n$. Let $\{\mu_j\}_{j \in J}$ be $|J|$ successive eigenvalues of $B$ and introduce the reduced condition number

$$\kappa_J(B) := \max_{j \in J^c} |\mu_j| / \min_{j \in J^c} |\mu_j|.$$ 

With $p_J(t)$ and $Q_J$ defined as above then

$$\|b - Bz\|_2 \leq 2 \left( \max_{\mu_j \in J^c} |p_J(\mu_j)| \right)^{|J|} \left( \frac{\sqrt{\kappa_J(B)} - 1}{\sqrt{\kappa_J(B)} + 1} \right)^{k-|J|} \|Q_J W^T b\|_2$$

when $\{\mu_j\}_{j \in J^c}$ contains only elements of the same sign, and

$$\|b - Bz\|_2 \leq 2 \left( \max_{\mu_j \in J^c} |p_J(\mu_j)| \right)^{|J|} \left( \frac{\sqrt{\kappa_J(B)} - 1}{\sqrt{\kappa_J(B)} + 1} \right)^{k-|J|} \|Q_J W^T b\|_2$$

otherwise.

In the remainder of this paper we shall assume that the distinguished set of eigenvalues of $B$ consists of the simple eigenvalue $\mu_1$. In this case we write

$$\kappa_J(B) = \kappa_1(B) = \max_{j=2, \ldots, n} |\mu_j| / \min_{j=2, \ldots, n} |\mu_j|, \text{ and } p_1(t) = (\mu_1 - t)/\mu_1.$$ 

Also we define the two quantities

$$q_e := \frac{\sqrt{\kappa_1(B)} - 1}{\sqrt{\kappa_1(B)} + 1}, \quad (27)$$

and

$$q_i := \frac{\sqrt{\kappa_1(B)} - 1}{\sqrt{\kappa_1(B)} + 1}, \quad (28)$$

where $q_e$ refers to the case where $\mu_1$ is an extreme eigenvalue and $\mu_2, \ldots, \mu_n$ are of the same sign, and $q_i$ covers all other situations. For this case we have the following key result.

Corollary 3.1  Assume the conditions of Theorem 3.1 hold with $J = \{1\}$. Then

$$\|b - Bz_k\|_2 \leq 2 \left( \max_{j=2, \ldots, n} \frac{|\mu_1 - \mu_j|}{|\mu_1|} \right) (q)^{k-\delta} \|Q_1 W^T b\|_2, \quad (29)$$

where $q = q_e, \delta = 1$ if $\mu_1$ is an extreme eigenvalue and $\mu_2, \ldots, \mu_n$ have all the same sign, and $q = q_i, \delta = 2$ otherwise.

Note that if $q = q_e$ then the residual reduction indicated by (29) is the same as that achieved by the Conjugate Gradient method applied to a positive definite symmetric matrix.

3.1 Eigenvalue bounds

In later analysis we shall assume bounds independent on $i$ for $|\lambda_1 - \sigma^i| / |\mu_1|$, where $\mu_1$ is the smallest eigenvalue (in modulus) of $B$ where $B = P_1^{-1}(A - \sigma^i I) P_1^T$. In many practical applications it may be hard to obtain rigorous bounds, but here we examine two cases where bounds are possible.
Example 3.1 (Domain Decomposition Preconditioners) First, if $A$ is symmetric positive definite and $P^{-1}$ arises from Domain Decomposition then a typical assumption is that there is a bound on the condition number of $P^{-1}A$. Thus if we denote the eigenvalues of any matrix $M$ by $\lambda_j(M)$ we assume
\[ \gamma_L \leq \lambda_j(P^{-1}A) \leq \gamma_U \]
for some positive constants $\gamma_L, \gamma_U$. Now with $B = P^{-1}1(A - \sigma^iI)P_1^T$ we have
\[ \lambda_j(B) = \lambda_j(P_1^{-\frac{1}{2}}A^{\frac{1}{2}}(I - \sigma^iA^{-1})A^{\frac{1}{2}}P_1^{-T}) \]
and Sylvester’s Inertia Theorem can be used to provide bounds. For example, if $\lambda_1 < \sigma^i < \lambda_2$ then $\mu_1 < 0$ (using Sylvester’s Inertia Theorem) and
\[ (1 - \frac{\sigma^i}{\lambda_1})\lambda_n(P^{-1}A) \leq \mu_1 \leq (1 - \frac{\sigma^i}{\lambda_1})\lambda_1(P^{-1}A), \]
so that
\[ \frac{\lambda_1}{\gamma_U} \leq \frac{\lambda_1}{\lambda_n(P^{-1}A)} \leq \frac{|\lambda_1 - \sigma^i|}{|\mu_1|} \leq \frac{\lambda_1}{\lambda_1(P^{-1}A)} \leq \frac{\lambda_1}{\gamma_L}. \tag{30} \]

Example 3.2 (Cholesky Preconditioners) Second if $A$ is symmetric and an incomplete Cholesky factorisation of $A$ is used to find $P$, i.e. $A = P_1P_1^T + E$ with $E$ ‘small’, then using ideas in [18] we write
\[ P_1^{-1}(A - \sigma^iI)P_1^{-T}w_1 = \mu_1w_1, \]
as
\[ (A - \sigma^iI)\tilde{w} = \mu_1P_1\tilde{w_1} = \mu_1(A + E)\tilde{w}, \]
and so
\[ (A - \frac{\mu_1}{1 - \mu_1}E)\tilde{w} = \frac{\sigma^i}{1 - \mu_1}\tilde{w}. \]
Now comparing with $Av_1 = \lambda_1v_1$ the Bauer-Fike Theorem gives
\[ \left| \frac{\sigma^i}{1 - \mu_1} - \lambda_1 \right| = \left| \frac{\mu_1}{1 - \mu_1} \right| \|E\| \]
and hence
\[ \lambda_1 - \|E\| \leq \frac{|\lambda_1 - \sigma^i|}{|\mu_1|} \leq \lambda_1 + \|E\|. \tag{31} \]
Thus in both examples considered here we can say that
\[ |\lambda_1| C_6' \leq \frac{|\lambda_1 - \sigma^i|}{|\mu_1|} \leq |\lambda_1| C_6 \]
for some positive constants $C_6'$ and $C_6$ independent of $i$.\]
3.2 An “a priori” Lower Bound for MINRES Iterations

Consider now the application of the bound (29) to the problem of computing the smallest eigenvalue of \( A \), i.e. we seek \( \lambda_1 \), where \( \lambda_1 < \lambda_2 \leq \lambda_3 \ldots \leq \lambda_n \). Assume also that any shift \( \sigma^i \) satisfies (4) so that we can regard \( \lambda_1 - \sigma^i \) as well-separated from \( \lambda_j - \sigma^i, j = 2, \ldots, n \).

For unpreconditioned MINRES we have \( P_1 = I = T \) so that for the MINRES theory we take \( B = (A - \sigma^i I) \). Thus \( \mu_j = \lambda_j - \sigma^i, j \in \mathbb{N}_n \) and

\[
\kappa_1(B) = \kappa_1 = \frac{|\lambda_n - \sigma^i|}{|\lambda_2 - \sigma^i|}. \tag{32}
\]

we have that

\[
\|Q_1 W^T b\| = \|Q_1 W^T x^i\| = |s^i|.
\]

Denote the number of inner iterations in MINRES at the \( i \)th outer iteration by \( k^i \). Then, from (29), to achieve

\[
\|(A - \sigma^i I)y^i - x^i\|_2 \leq \tau^i, \tag{33}
\]

it is sufficient that \( k^i \) satisfies

\[
2 \left( \frac{|\lambda_1 - \lambda_n|}{|\lambda_1 - \sigma^i|} \right) q^i_e := \tau^i, \tag{34}
\]

where \( q_e \) is as in (27), or

\[
2|\lambda_1 - \lambda_n| \frac{|s^i|}{|\lambda_1 - \sigma^i|} \leq (q_e^{-1})^{k^i - 1}.
\]

Taking logarithms, to achieve (33) it is sufficient that \( k^i \) satisfies

\[
k^i \geq 1 + \left\{ \log 2|\lambda_1 - \lambda_n| + \log \frac{|s^i|}{|\lambda_1 - \sigma^i|} \right\} / \log q_e^{-1}.
\]

The theory in Section 2 assumed the upper bounds \( \tau^i \leq C_1 |s^i|^\beta \) and \( |\lambda_1 - \sigma^i| \leq C_2 |s^i|^\alpha \). Now, in addition, assume the two sided bounds

\[
C'_1 |s^i|^\beta \leq \tau^i \leq C_1 |s^i|^\beta, \quad C'_2 |s^i|^\alpha \leq |\lambda_1 - \sigma^i| \leq C_2 |s^i|^\alpha, \tag{35}
\]

where the constants are positive and independent of \( i \). These are not unrealistic in practice. For example they are satisfied by all the choices in Corollary 2.1.

Using the lower bounds for \( \tau^i \) and \( |\lambda_1 - \sigma^i| \) in (35) we see that (33) is satisfied if

\[
k^i \geq 1 + \left\{ \log \frac{2|\lambda_1 - \lambda_n|}{C'_1 C_2} + (\alpha + \beta - 1) \log |s^i|^{-1} \right\} / \log q_e^{-1}. \tag{36}
\]

Thus the right hand side is an “a priori” lower bound on the number of inner iterations needed to achieve (33). We have proved the following Lemma.
Lemma 3.1 Assume Algorithm 1 is used to compute the smallest eigenvalue of \( A \) using unpreconditioned MINRES for the inexact solves. Here \( k^i \) denotes the number of inner solves needed at each outer iteration. Assume in addition that the lower bounds given by (35) hold. Then if

\[
k^i \geq 1 + \left\{ \log \frac{2|\lambda_1 - \lambda_n|}{C_1 C_2} + (\alpha + \beta - 1) \log |s^i|^{-1} \right\} / \log(q_e^{-1})
\]

Algorithm 1 converges with a convergence rate predicted by Theorem 2.1. We note that if \( \alpha + \beta = 1 \) (i.e. linear convergence in Theorem 2.1) then the lower bound on \( k^i \) does not grow, whereas if \( \alpha + \beta = 2 \) which arises, say, when the shift is chosen to be the Rayleigh quotient and \( \tau^i \) is held constant, then the lower bound grows with \( \log |s^i|^{-1} \). These conclusions are supported by the numerical results discussed in Example 5.1. However it is well known that the bound given by (29) is, at best, descriptive and is unlikely to be sharp. As a consequence (36) is unlikely to provide a realistic estimate for \( k^i \) and should only be used in a “a priori” descriptive sense as above. Nevertheless, descriptive bounds such as this play a key role in the appraisal of the practical value of various iterative methods.

A similar analysis can be carried out if one wishes to calculate an interior eigenvalue of a general matrix \( A \). We do not repeat this analysis here but a bound similar to (36) is obtained with the main difference that the factor \( q_e \) is replaced by \( q_i \) given by (28). However the comments about the likely increase in the number of inner iterations remain valid. Numerical examples to illustrate this theory are given in Section 5 where an interior eigenvalue is computed.

Consider now a similar analysis to find a lower bound for the number of inner iterations for the preconditioned case given by (5), namely, \( B = P_1^{-1}(A - \sigma I)P_1^{-T} \) and \( b^i = P_1^{-1}x^i \). The analysis proceeds as before except that \( \|Q_i W^T b\| = \|Q_i W^T P_1^{-1} x^i \| \leq \|P_1^{-1}\| \). From (29) we see that, in order to achieve \( \|Bz_k - b\| \leq \tau^i \) it is sufficient that

\[
2 \left\{ \max_{j=2,\ldots,n} \frac{|\mu_1 - \mu_j|}{|\mu_1|} \right\} q_e^{k^i} \|P_1^{-1}\| \leq \tau^i,
\]

where \( q_e \) refers to \( B \). Note that we no longer have an \( |s^i| \) term on the left hand side of the bound (cf. (34)). Thus we ask that

\[
2 \left\{ \max_{j=2,\ldots,n} \frac{|\mu_1 - \mu_j|}{|\mu_1|} \|P_1^{-1}\| \right\} \left\{ \frac{1}{|\lambda_1 - \sigma^i| \tau^i} \right\} \leq (q_e^{-1})^{k^i-1}.
\]

Now repeat the above analysis assuming (35) and an upper bound on \( |(\lambda_1 - \sigma^i)/\mu_1| \) as given in Section 3.1, to obtain

\[
k^i \geq 1 + \left\{ \log C + (\alpha + \beta) \log |s^i|^{-1} \right\} / \log q_e^{-1}
\]

(38) for some positive constant \( C \) independent if \( i \). This lower bound indicates that in the standard preconditioned case we expect growth in the number of inner iterations.
iterations for any convergent method. In particular there will be growth for a linearly convergent method, and this is observed in the results for Example 5.2 given in Table 2.

In the following section we consider the efficiency of the overall process in Algorithm 2. We wish to minimise the work done, which we take to be the number of matrix vector multiplications, knowing that a faster converging Algorithm will lead to fewer outer iterations but at the cost of more inner iterations per step.

4 Efficiency Analysis of MINRES

In this section we obtain a measure for the work done in terms of the total number of inner iterations in inexact inverse iteration required to compute a simple, well-separated eigenvalue of $A$. Let $k^i$ denote the number of inner iterations (i.e. matrix vector multiplications) required to achieve the required residual tolerance (39) below, and assume the outer iteration converges in $N$ iterations. Then the work done in terms of total number of inner iterations will be $\sum_{i=0}^{N-1} k^i$. It is this quantity that we seek to minimise. To do this we first present an analysis which provides an “a posteriori” upper bound on $k^i$ as distinct from the “a priori” lower bound given in Lemma 3.1.

4.1 An “a posteriori” Upper Bound for $k^i$ in MINRES

In this subsection we assume that the inexact solve of $(A - \sigma^i I)y^i = x^i$ is achieved by solving the preconditioned system $Bz^i = b^i$ to a required tolerance $\tau^i$ using MINRES with $k^i$ iterations, where $k^i$ is defined by the relation

$$\|\text{res}_{ki}^i\| \leq \tau^i < \|\text{res}_{k_{i-1}}^i\|,$$

with $\text{res}_{ki}^i = b^i - Bz_{ki}$. Such a $k^i$ must exist since from Theorem 3.1,

$$\|\text{res}_{ki}^i\| \leq 2\frac{|\mu_1 - \mu_n|}{|\mu_1|} q^{k^i-\delta} \chi^i,$$

for $\delta$ and $q$ (with $0 < q < 1$) as in Corollary 3.1. Here we have introduced the quantity $\chi^i$ defined by

$$\chi^i := \|Q_1 W^T b^i\|,$$

and $\chi^i$ will play a key role in later analysis. Note that with $Q_1$ defined by (32) and $W$ the matrix of eigenvectors of $B$ then $Q_1 W^T b = (0, w_2^T b, \ldots, w_m^T b)^T$ and so $Q_1 W^T b$ represents the projection of $b$ onto $\{w_1\}^\perp$. If $B = A - \sigma^i I$ and $b^i = x^i$ then from (8)

$$Q_1 W^T b^i = \sin \theta^i u^i$$

and so $\chi^i = \sin \theta^i$. 

17
Now the general one step bound given in Lemma 2.1 gives

\[ t^{i+1} \leq \frac{|\lambda_1 - \sigma^i|}{|\lambda_2 - \sigma^i|} \left( \| P_1 T x^i \| + \| P_1 \| \tau^i \right) \| v_1^TP_1 Tx^i \| \| \lambda_1 - \sigma^i \| \| P_1 \| \tau^i, \]  

(43)

and we now combine (39), (40) and (43) in the following analysis. Our first result is a lemma which provides an upper bound on \( k^i \) which depends on \( \log(\chi^i / t^{i+1}) \).

In this Lemma we make two assumptions on \( \text{res}^i_k \), namely (44) and (45) below.

The first just says that \( \text{res}^i_k \) shouldn’t be “too large”. The second is more subtle. We ask that \( \text{res}^i_k \) shouldn’t be “too small”, that is, the inexact solver should not produce a much smaller residual than expected. We’ll look at this again when we consider the unpreconditioned case in the next subsection.

**Lemma 4.1** Assume (39), and that \( \text{res}^i_k \) satisfy the following two bounds

\[
| v_1^TP_1 \text{res}^i_k | \leq \frac{1}{2} | v_1^TP_1 T x^i | \]  

(44)

\[
\| v_1^TP_1 T x^i \| \leq C_5 \| P_1 \| \| \text{res}^i_k \|, \]  

(45)

for some positive \( C_5 \) independent of \( i \). Also assume

\[
\frac{|\lambda_1 - \sigma^i|}{|\mu_1|} \leq \lambda_1 C_6, \]  

(46)

for some positive constant \( C_6 \) independent of \( i \). Then we have the following upper bound for \( k^i \)

\[
k^i \leq 1 + \delta + \left( \log C_7 + \log \frac{\chi^i}{\tau^{i+1}} \right) / \log(q^{-1}) \]  

(47)

where \( \chi^i \) is defined by (41) and

\[
C_7 := \frac{4(1 + C_5)}{|\lambda_2 - \lambda_1| | v_1^TP_1 T x^i |} \frac{|\lambda_1 - \sigma^i|}{|\mu_1|} |\mu_1 - \mu_n|. \]  

(48)

Here \( \delta \) and \( q = q_e \) are as in Corollary 3.1.

**Proof:** Using (39) and assumptions (44) and (45) we have

\[
t^{i+1} \leq \frac{4(1 + C_5)}{|\lambda_2 - \lambda_1| | v_1^TP_1 T x^i |} \frac{|\lambda_1 - \sigma^i|}{|\mu_1|} \tau^i
\]

and assuming the right hand side bound in (39), and (40) with \( k^i \) replaced by \( k^i - 1 \), we obtain

\[
t^{i+1} \leq C_7(q)^{k^i - 1 - \delta} \chi^i, \]  

(49)

where

\[
C_7 := \frac{4(1 + C_5)}{|\lambda_2 - \lambda_1| | v_1^TP_1 T x^i |} \frac{|\lambda_1 - \sigma^i|}{|\mu_1|} |\mu_1 - \mu_n|. \]  

(48)
Now rearranging and taking logs in (49) gives

\[ k^i \leq 1 + \delta + (\log C_7 + \log \frac{\chi^i}{t^{i+1}}) / \log(q^{-1}). \]

\[ \square \]

We shall consider the first two assumptions later when we discuss specific choices for \( P_1 \) and \( T \). The bound (46) was discussed in Section 3.1. For this paper the key term in the bound is \( \log(\chi^i/t^{i+1}) \) and in particular, we will see later, it is desirable that \( \chi^i \leq C t^i \) as in the case of unpreconditioned MINRES in Section 3.2. Finally we note that (39) implies that the convergence theory in Section 2.2 will apply.

We have the following Theorem.

**Theorem 4.1** Assume Algorithm 1 using unpreconditioned MINRES is used to compute \( \lambda_1 \) with \( \sigma^i \) and \( \tau^i \) choosen to provide an algorithm that converges in \( N \) outer iterations. Assume \( k^i \) is defined by \( \| \text{res}^i \| \leq \tau^i < \| \text{res}^i_{k^i} \| \) where \( \text{res}^i := x^i - (A - \sigma^i I)y^i \). Then with \( x^i = c^i v_1 + s^i u_i \) and \( t^i = |s^i / c^i| \),

\[ k^i \leq 1 + \delta + (\log C_7 + \log \frac{t^i}{t^{i+1}}) / \log(q^{-1}). \]

Here \( \delta \) and \( q \) are as in Corollary 3.1.

We shall now apply Theorem 4.1 to the three versions of Algorithm 2.

### 4.2 No Preconditioning

In this case \( B = A - \sigma^i I, b^i = x^i \) and \( P_1 = T = I \). Then, with \( x^i = c^i v_1 + s^i u^i \) we have \( v_1^T P_1 T x^i = c^i \), while \( \Pi P_1 T x^i = s^i u^i \) and so assumptions (44) and (45) reduce to

\[ v_1^T \text{res}^i \leq \frac{1}{2} |c^i|, \]

\[ \sin \theta^i \leq C_5 \tau^i, \]

and

\[ C_7 = \frac{4(1 + C_5)}{|c^i|} \frac{|\lambda_2 - \lambda_1|}{|\lambda_2 - \lambda_1|}, \]

which near convergence is essentially independent of \( i \) since \( |c^i| \to 1 \) as \( i \to \infty \). Assumption (51) asks that the residual shouldn’t be any smaller than a constant multiple of \( \sin \theta^i \). Typically this will be the case but there may be situations where the inexact solver does much better than expected and in such cases our theory does not apply. Finally \( \chi^i = ||Q_1 W^T x^i|| = s^i \) and so \( \chi^i \leq t^i \). Thus (47) proves the following theorem.

**Theorem 4.2** Consider the use of unpreconditioned MINRES to solve \( (A - \sigma^i I)y^i = x^i \). Assuming the conditions of Lemma 4.1 hold, then an upper
bound for the number of matrix vector multiplications in each outer iteration is given by

\[ k^i \leq 1 + \delta + (\log C_7 + \log \frac{t^i}{t^{i+1}}) / \log(q^{-1}), \quad (53) \]

where \( C_7 \) is given by (52) and is independent of \( i \). If we seek an improvement in the error angle by a factor of \( 10^{-\gamma} \), \( \gamma > 0 \), so that \( t^N = 10^{-\gamma} t^0 \), then

\[ \sum_{i=0}^{N-1} k^i \leq \left[ N\{1 + \delta + \log C_7\} + \gamma \log 10 \right] / \log(q^{-1}). \quad (54) \]

Here \( \delta \) and \( q \) are given in Corollary 3.1.

Remarks:

a) Though the quantities \( C_7 \) and \( q^{-1} \) play an important role in the overall number of inner iterations, this analysis indicates that the main factor in determining the total number of inner iterations is the rate of convergence of the outer iteration. For \( \gamma \) large enough the discrete nature of \( N \) may be neglected, and we see that \( \sum_{i=0}^{N-1} k^i \) is linear in \( N \). Now Theorem 2.1 shows that the outer convergence rate is of order \( \alpha + \beta \), and it can be shown [1] that \( N \) is decreasing in \( \alpha + \beta \). Thus the bound on \( \sum_{i=0}^{N-1} k^i \) is minimised for maximal \( \alpha + \beta \). Since \( \beta \in [0, 1] \) we take \( \beta = 1 \) and ask that \( \alpha \) is maximised. In practise this means take \( \alpha = 2 \), that is, choose \( \sigma \) to be the Rayleigh quotient. Bearing in mind, however, that in practical situations cubic convergence will rarely improve over quadratic, we would expect a method based on choosing \( \sigma \) as the Rayleigh quotient and \( \tau = \tau^0 \), a constant, which has quadratic convergence to be the most efficient method. This claim is confirmed by the numerical results in Example 5.1. One might think that this strategy will provide numerical problems to the Krylov solver. However the above analysis shows that this not the case because of the interplay between the shift and the right hand side which is tending to the desired eigenvector. The fact that the form of the right hand side in (2) is important when the system is solved using a Krylov method was was recognised in the early paper [17]. Here the bound given in Corollary 3.1 shows that MINRES copes well with a single eigenvalue near zero in \((A - \sigma I)\).

b) Note from (53) that the bound on the number of inner iterations required at the \( i \)th outer iteration depends on the improvement in the (tangent of) the error angle. We shall discuss this in Example 5.1.

4.3 Standard Preconditioning

In this case we assume the system \((A - \sigma I)y^i = x^i\) is solved by the usual symmetrically preconditioned system

\[ P_1^{-1}(A - \sigma I)P_1^{-T}\tilde{y}^i = P_1^{-1}x^i; \quad y^i = P_1^{-T}\tilde{y}^i \]
for some $P_1$. In this case $T = I$. Then with $x^i = c^i v_1 + s^i u^i$ we have $P_1 x^i = c^i P_1 v_1 + s^i P_1 u^i$. And since we are unlikely to know much about the action of $P_1$ on $v_1$ we may be reduced to using the simple bound $\| P_1 x^i \| \leq \| P_1 \|$. Note, however for convergence we require that $v_1^T P x^0$ is bounded away from zero (see definition of $C_4$ in Theorem 2.2). There is little simplification in the theory and crucially $\chi^i = \| Q_1 W^T b^i \| = \| Q_1 W^T P_1^{-1} x^i \| \leq \| P_1 \|$ whereas in the unpreconditioned case $\chi^i < t^i$ (see (42)). We now have the following theorem.

**Theorem 4.3** Assume that in Algorithm 2 the preconditioned system

$$ P_1^{-1} (A - \sigma^i I) P_1^{-T} y^i = P_1^{-1} x^i, \quad y^i = P_1^{-T} \tilde{y}^i $$

is used to solve $(A - \sigma^i I) y^i = x^i$. Assuming also that the conditions of Lemma 4.1 hold then an upper bound for the number of inner iterations at the $i$th step of the outer iteration is given by

$$ k^i \leq 1 + \delta + \{ \log C_7 + \log(1/t^{i+1}) \} / \log(q^{-1}), $$

(55)

for some positive constant $C_7$ independent of $i$. If there are $N$ outer iterations then the total number of inner iterations is bounded above by

$$ \sum_{i=0}^{N-1} k^i \leq \left[ N + \{ \log C_7 + \sum_{i=0}^{N-1} \log(1/t^{i+1}) \} / \log(q^{-1}) \right]. $$

(56)

where $\delta$ and $q$ are as in Corollary 3.1.

**Remarks:**

a) The upper bound given by (55) indicates that $k^i$ depends on the (tangent of the) achieved error angle no matter how accurate the starting guess is. This is confirmed by the numerical experiments in Example 5.2.

b) In (54) the sum $\sum_{i=0}^{N-1} \log(t^i/t^{i+1})$ reduces to $\log(t^0/t^N)$. No such cancellation occurs in (56) and typically we would expect the term $\sum_{i=0}^{N-1} \log(1/t^{i+1})$ to be considerably larger than the corresponding term in (54). (Of course the $q$ and $C_7$ quantities will be smaller in the preconditioned case, so the overall work in terms of inner iterations will be smaller.) In the next section we see that the modification in [18] recovers the $\log(t^i/t^{i+1})$ term that appears in (54).

### 4.4 Simoncini-Elden Preconditioning

In this section we analyse the case when the system

$$ (A - \sigma^i I) y^i = P x^i $$

(57)

is solved to provide a new estimate for $v_1$ the desired eigendirection. We shall restrict attention to the case when $\sigma^i$ is chosen to be the Rayleigh quotient and $\tau^i =$constant, so that the convergence result given by Theorem 2.2 applies. With $P = P_1 P_1^T$ the symmetrically preconditioned form of (57) is

$$ P_1^{-1} (A - \sigma^i I) P_1^{-T} \tilde{y}^i = P_1^{-T} x^i; \quad y^i = P_1^{-T} \tilde{y}^i, $$

21
so in the MINRES convergence theory we take $B = P_i^{-1}(A - \sigma^i I)P_i^{-T}$, and $b^i = P_i^T x^i$. With $x^i = c^i v_1 + s^i u^i$ and $T = P_i^T$ we have

$$v_i^T P_i T x^i = v_i^T P x^i = c^i v_i^T P v_1 + s^i v_1 P u^i$$

and

$$\Pi P_i T x^i = c^i \Pi P v_1 + s^i \Pi P u^i.$$ 

Normally $P$ approximates $A$ and so assumptions (44) and (45) in Lemma 4.1 are not unreasonable if (50) and (51) hold in the unpreconditioned case. For example, if $P = A$ and $P_i = A^\frac{1}{2}$ then $v_i P_i T x^i = v_i^T A x^i = \lambda_i c^i$ and $\Pi P_i T x^i = s^i \{ A_i \} u^i$. Thus (44) is satisfied if $2\lambda_i^{-1}\sqrt{\lambda_i} \tau^i \leq c^i$, which will hold for small enough $\tau^i$, and (45) is satisfied if $|s^i| \sqrt{\lambda_i} \leq C_5 \tau^i$ which will hold for small enough $|s^i|$.

Now we consider the determination of the bound on $\chi^i$ given by (41). Write $B$ in the form

$$B = P_i^{-1}(A - \lambda_1 I)P_i^{-T} + (\lambda_1 - \sigma^i)P_i^{-1}P_i^{-T}$$

where we note that $P_i^{-1}(A - \lambda_1 I)P_i^{-T}(P_i^T v_1) = 0$. Standard perturbation theory for simple eigenvalues shows that $w_1$, the eigenvector of $B$ corresponding to $\mu_1$, satisfies $P_i v_1 = \zeta w_1 + O(\lambda_1 - \sigma^i)$ for some real constant $\zeta$. Hence if $\sigma^i$ is the Rayleigh quotient,

$$P_i^T v_1 = \zeta w_1 + O((t^i)^2).$$

(58)

Thus,

$$Q_1 W^T b^i = Q_1 W^T P_i^T x^i = c^i Q_1 W^T P_i^T v_1 + s^i Q_1 W^T P_i u^i = c^i Q_1 W^T (\zeta w_1 + O((t^i)^2)) + s^i Q_1 W^T P_i u^i$$

(59)

and using (58) we easily obtain, for small enough $t^i$,

$$\|Q_1 W^T b^i\| \leq C_8 t^i$$

(61)

for some positive constant $C_8$ independent of $i$. We thus have the following theorem.

**Theorem 4.1** Consider the use of Algorithm 2 to find $\lambda_1$ where the system $(A - \sigma^i I)y^i = P_i x^i$ is solved by preconditioned MINRES. With $k^i$ such that (41) holds we have the following upper bound on the number of inner iterations at the $i$th outer iteration

$$k^i \leq 1 + \delta + (\log C_9 + \log \frac{t^i}{\mu_{i+1}})/\log(q^{-1})$$

(62)

for some positive $C_9$ independent of $i$. If $N$ outer iterations are needed to reduce the error angle by $10^{-\gamma}$, for some $\gamma$ then

$$\sum_{i=0}^{N-1} k^i \leq N\{1 + \delta + \log C_9\} + \gamma \log 10)/\log(q^{-1}),$$

(63)
where $\delta$ and $q$ are as in Corollary 3.1.

Remarks:

a) We see that the upper bound given by (63) is an improvement on (54) because of the reduced value of $q$ because of preconditioning. Also it is an improvement on (56) because the term $\log(t^0/t^N)$, which arises from summing the $\log(t^i/t^{i+1})$ terms, is typically much less than $\sum_{i=0}^{N-1} \log(1/t^{i+1})$.

b) Example 5.3 provides numerical results for this version of inverse iteration. There we see the superiority of this approach in terms of total number of inner iterations required to achieve a given accuracy in $\lambda_1$.

c) The change of the right hand side reduces, at least in theory, the convergence area. However this is not observed in our experience.

d) We remark that for Algorithm 2 using (6) with fixed tolerance (i.e. as in Theorem 2.2) we obtain the following “a priori” lower bound for the number of inner iterations

$$k^i \geq 1 + \{\log C + (\alpha - 1) \log |s^i|^{-1}\} / \log q^{-1}$$

for some positive constant $C$ independent of $i$ (cf. (36)). The proof is similar to that of Lemma 3.1 and is omitted. For the case of interest here, $\alpha = 2$, and so (64) indicates growth with $\log |s^i|^{-1}$ in the number of inner iterations as the outer iteration converges. This is confirmed by results in Tables 4 and 5 for Example 5.3.

5 Numerical Examples

In this section we provide some numerical examples to illustrate the various theoretical results in this paper. We mainly use the test example ‘bcsstk09’ from matrix market (http://math.nist.gov/MatrixMarket), though similar results are obtained for other examples. Note that we do not claim that inexact inverse iteration is the best method to compute the eigenvalues of this matrix but it is a sufficiently realistic example to test the theory presented here.

In all our experiments we normally use the residual stopping condition

$$\|b^i - Bz_k\| \leq \tau^i,$$  \hspace{1cm} (65)

to stop the inner iterations. We measure convergence of the outer iteration by the relative eigenvalue residual

$$\|(A - \vartheta(x^i))x^i\| / |\vartheta(x^i)| \leq \text{etol}$$  \hspace{1cm} (66)

where $\text{etol}$ is prescribed in advance. By contrast in [18] an approximation to the eigenvalue residual is used to stop the inner iteration. To be precise they use the inner test

$$\frac{\|z^i_k\| - \|z^i_{k-1}\|}{\|z^i_k\|} \leq \tau^{i}_{SE}.$$  \hspace{1cm} (67)
which is derived from the eigenvalue residual test (66), in combination with
\[ \| z_i^k \| > \|(A - \rho(x^i)I)x^i\|^{-1}, \] (68)
where \( \tau_{SE} \) is a prescribed tolerance (see [18] for full details). We take advantage of the nice idea in [18] of calculating the eigenvalue residual \( \|(A - \rho(x^i)I)x^i\|^{-1} \) within the MINRES inner iteration (see [18, Appendix A]) to monitor the eigenvalue residual and stop the outer iteration when (66) is satisfied.

**Example 5.1** First consider the eigenvalue problem for the 2-D Laplacian,
\[-\nabla^2 u = \lambda u, \]
with homogenous boundary conditions on the rectangle \( 0 \leq x \leq 1, 0 \leq y \leq 1 \), which is discretised using finite differences on a \( 12 \times 12 \) regular grid. In Table 1 we present numerical results obtained when calculating \( \lambda_1 \) (\( \approx 15.6 \)) the smallest eigenvalue of the discretised matrix. Using unpreconditioned MINRES as the inexact solver we apply the three versions of inexact inverse iteration discussed in Corollary 2.1, namely

- **FSd:** Fixed shift, decreasing tolerance,
- **RQIf:** Rayleigh quotient shift, fixed tolerance,
- **RQId:** Rayleigh quotient shift, decreasing tolerance,

(cases a),b) and c) in Corollary 2.1 respectively). Each row in Table 1 provides the outer iteration number, \( \log_{10} t^i \) (calculated knowing the exact \( v_1 \)) and \( k^{i-1} \) the number of inner iterations needed satisfy the residual condition that gives the error angle \( t^i \). To illustrate several ideas in our theory the experiment was carried out using MATLAB with variable precision arithmetic, \( vpa = 128 \). For RQIf and RQId we use \( etol = 10^{-80} \) as the outer stopping condition.

**Conclusions 5.1:**

a) We note that as predicted in Corollary 2.1, we observe linear convergence for FSd, quadratic convergence for RQIf and cubic convergence for RQId.

b) As suggested in Lemma 3.1 the number of inner iterations at the \( i \)th step in the linearly convergent method remains approximately constant even though the tolerance is decreasing. For RQIf and RQId the number of inner iterations at the \( i \)th step increases with \( i \) with the quadratically convergent method needing fewer iterations per step than the cubically convergent method.

c) The quadratically convergent method requires more outer iterations steps and (eventually) needs more inner iterations as predicted by Theorem 4.2. However at a practical level (that is, to double precision on most of the current computers) it is clear that to all intents and purposes the quadratic method will perform as well as the cubic and will most likely need fewer inner iterations in total.
d) From Table 1 we see that RQIf needs 76 iterations to make an improvement of \(10^{-20}\), whereas RQId needs 78 to make an improvement of \(10^{-23}\) which is in good agreement with (54) (see Remark b) after Theorem 4.2).

**Example 5.2** Next we consider the matrix market example ‘bcsstk09’, an eigenvalue problem of dimension 1083. In Tables 2 and 3 we present results to find \(\lambda_{20} = 4.06 \times 10^5\) the 20th smallest eigenvalue. The numerical method used Algorithm 2 with \(T = P^{-1}\), that is the system \((A - \sigma I)y^i = x^i\) is solved using the standard preconditioned version (5). The preconditioner was taken to be an incomplete Cholesky decomposition of \(A\) using the matlab routine ‘cholinc’ with droptol=1e-2. We use the methods FSd (which exhibits linear convergence), RQIf (which exhibits quadratic convergence), and RQId (which in theory exhibits cubic convergence) as given by Corollary 2.1. In all runs we stopped when the relative eigenvalue residual satisfied \(\| (A - \phi(x^i)I)x^i \| / |\phi(x^i)| \leq 10^{-10}\). Tables 2 and 3 present results using preconditioned MINRES. For FSd we present results for two choices of \(\tau^0\). Here, as in following examples, we take the starting vector \(x^0 = c^0v_1 + s^0u_0\) where \(t^0 = s^0/c^0 = 0.02\). Other starting guesses were taken but no qualitative differences in the results were observed.

**Conclusions 5.2:**

a) The methods FSd and RQIf achieve linear and quadratic convergence respectively as predicted by Corollary 2.1.

b) As predicted by (38) we observe growth in the values of \(k^i\) as the outer iteration converges (in contrast to the behavior in the unpreconditioned case shown in the left hand column of Table 1).

c) The upper bound (55) indicates that, whatever the outer rate of convergence of the method the number of inner iterations \(k^i\) to produce an error angle \(\theta^i\) depends on \(\log(1/\theta^i)\) independent of the previous error angle. This is confirmed in Table 2 at the entries with the superscript 1: from two different applications to find \(\lambda_{20}\) the method FSd achieved an error angle of \(2 \times 10^{-10}\) after 80 iterations even though the previous error angles were different. Also from Tables 2 and 3 at the entries with superscript 2, we see that both FSd and RQIf took 65 iterations to achieve an error angle around \(10^{-7}\) even though FSd started from an error angle of \(8 \times 10^{-7}\) and RQIf started from an error angle \(1 \times 10^{-2}\).

d) Comparing the total number of inner iterations in the two runs of the linearly converging FSd with the total number for RQIf we see that the quadratically convergent method needs many fewer inner iterations (183 compared with 667) even though the starting error angle is twice as large. This behaviour is typical of what we have observed in many experiments (see [1]). This is in agreement with (56) in Theorem 4.3, which shows that the upper bound on \(\sum_{i=0}^{N-1} k^i\) is minimised by making \(N\) as small as possible. The best way to achieve this for the symmetric eigenvalue problem is to choose \(\sigma^i\) as the Rayleigh quotient.
Table 1: Numerical Results for unpreconditioned MINRES applied to Example 5.1 using the methods in Corollary 2.1. Here \( s^i \) denotes \( \sin \theta^i \) defined by (8), \( k^i \) denotes the number of inner iterations at the \( i \)th step, and \( C_2 \) is the constant in Corollary 2.1.

<table>
<thead>
<tr>
<th>( \log_{10} s^i )</th>
<th>( k^{i-1} )</th>
<th>( \log_{10} s^i )</th>
<th>( k^{i-1} )</th>
<th>( \log_{10} s^i )</th>
<th>( k^{i-1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.12</td>
<td>-0.12</td>
<td>-0.14</td>
<td></td>
<td>-1.62</td>
<td>15</td>
</tr>
<tr>
<td>-1.48</td>
<td>18</td>
<td>-1.41</td>
<td>19</td>
<td>-3.33</td>
<td>24</td>
</tr>
<tr>
<td>-2.26</td>
<td>9</td>
<td>-3.85</td>
<td>19</td>
<td>-12.90</td>
<td>45</td>
</tr>
<tr>
<td>-3.29</td>
<td>9</td>
<td>-9.03</td>
<td>33</td>
<td>-36.19</td>
<td>78</td>
</tr>
<tr>
<td>-4.20</td>
<td>10</td>
<td>-19.46</td>
<td>50</td>
<td>-82.66</td>
<td>113</td>
</tr>
<tr>
<td>-5.18</td>
<td>11</td>
<td>-40.72</td>
<td>76</td>
<td>-82.66</td>
<td>113</td>
</tr>
<tr>
<td>-6.07</td>
<td>8</td>
<td>-82.96</td>
<td>108</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ \sum k^i \]

<table>
<thead>
<tr>
<th>( \tau^0 = 0.1 )</th>
<th>( \tau^0 = 0.1, C_2 = 0.1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C_2 = 0.5 )</td>
<td>( C_2 = 0.1 )</td>
</tr>
</tbody>
</table>

Table 2: Numerical results for preconditioned MINRES applied to Example 5.2 using method a) of Corollary 2.1 (FSd) for the 20th smallest eigenvalue. Here \( t^i = |\sin \theta^i / \cos \theta^i| \) given by (8), \( k^i \) denotes the number of inner iterations at the \( i \)th step and \( ||r^i|| \) denotes the eigenvalue residual given by (11). \( \tau^0 \) and \( C_2 \) are as given in Corollary 2.1.

<table>
<thead>
<tr>
<th>( \tau^0 = 0.2 )</th>
<th>( C_2 = 100 )</th>
<th>( \tau^0 = 0.02 )</th>
<th>( C_2 = 100 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( i )</td>
<td>(</td>
<td></td>
<td>r^i</td>
</tr>
<tr>
<td>0</td>
<td>2.5e+05</td>
<td>1.0e-02</td>
<td>2.5e+05</td>
</tr>
<tr>
<td>1</td>
<td>1.6e+02</td>
<td>2.6e-03</td>
<td>2.9e+01</td>
</tr>
<tr>
<td>2</td>
<td>4.6e+01</td>
<td>5.6e-04</td>
<td>6.8e+00</td>
</tr>
<tr>
<td>3</td>
<td>1.2e+01</td>
<td>5.9e-05</td>
<td>1.1e+00</td>
</tr>
<tr>
<td>4</td>
<td>4.3e+00</td>
<td>1.7e-05</td>
<td>4.2e-01</td>
</tr>
<tr>
<td>5</td>
<td>1.2e+00</td>
<td>2.8e-06</td>
<td>5.2e-02</td>
</tr>
<tr>
<td>6</td>
<td>4.2e+01</td>
<td>8.5e-07</td>
<td>1.2e-02</td>
</tr>
<tr>
<td>7</td>
<td>1.6e+01</td>
<td>3.4e-07</td>
<td>9.9e-04</td>
</tr>
<tr>
<td>8</td>
<td>2.1e+02</td>
<td>6.9e-08</td>
<td>3.4e-04</td>
</tr>
<tr>
<td>9</td>
<td>4.4e+03</td>
<td>5.0e-09</td>
<td>8.4e-05</td>
</tr>
<tr>
<td>10</td>
<td>1.1e+03</td>
<td>2.1e-09</td>
<td>3.7e-05</td>
</tr>
<tr>
<td>11</td>
<td>3.9e-04</td>
<td>1.4e-09</td>
<td>74</td>
</tr>
<tr>
<td>12</td>
<td>6.3e-05</td>
<td>2.1e-10</td>
<td>78</td>
</tr>
<tr>
<td>13</td>
<td>3.0e-05</td>
<td>1.2e-10</td>
<td>80^1</td>
</tr>
</tbody>
</table>

\[ \sum k^i \]

26
### Table 3: Numerical results for preconditioned MINRES applied to Example 5.2 using method b) and c) of Corollary 2.1 (RQIf and RQId) for the 20th smallest eigenvalue. Here $t^i = |\sin \theta^i / \cos \theta^i|$ given by (8). $k^i$ denotes the number of inner iterations at the $i$th step and $\|r^i\|$ denotes the eigenvalue residual given by (11). $\tau^0$ and $C_2$ are as given in Corollary 2.1.

<table>
<thead>
<tr>
<th></th>
<th>$\tau^0 = 0.1$</th>
<th></th>
<th>$\tau^0 = 0.2$</th>
<th></th>
<th>$C_2 = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$|r^i|$</td>
<td>$t^i$</td>
<td>$k^{i-1}$</td>
<td>$|r^i|$</td>
<td>$t^i$</td>
</tr>
<tr>
<td>0</td>
<td>5.0e+05</td>
<td>2.0e-02</td>
<td>0.0e+05</td>
<td>2.0e-02</td>
<td>5.0e+05</td>
</tr>
<tr>
<td>1</td>
<td>6.2e+02</td>
<td>1.3e-02</td>
<td>33</td>
<td>1.0e+03</td>
<td>1.6e-02</td>
</tr>
<tr>
<td>2</td>
<td>6.2e+02</td>
<td>4.2e-07</td>
<td>65</td>
<td>1.2e-01</td>
<td>3.1e-06</td>
</tr>
<tr>
<td>3</td>
<td>2.3e+05</td>
<td>9.7e-11</td>
<td>86</td>
<td>9.3e-06</td>
<td>3.1e-11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sum k^i$</td>
<td>186</td>
<td></td>
<td></td>
<td>186</td>
<td></td>
</tr>
</tbody>
</table>

### Table 4: Numerical Results for preconditioned MINRES to determine an extreme eigenvalue ($\lambda_1 = 7102$) and an interior eigenvalue ($\lambda_{20} = 4.06 \times 10^5$) using algorithm SEf. Here $r^i$, $t^i$ and $k^i$ are as in Table 2.

<table>
<thead>
<tr>
<th></th>
<th>$\tau^0 = 0.5$</th>
<th></th>
<th>$\tau^0 = 0.5$</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$|r^i|$</td>
<td>$t^i$</td>
<td>$k^{i-1}$</td>
<td>$|r^i|$</td>
<td>$t^i$</td>
</tr>
<tr>
<td>0</td>
<td>2.7e+05</td>
<td>1.0e-02</td>
<td>5.0e+05</td>
<td>2.0e-02</td>
<td>5.0e+05</td>
</tr>
<tr>
<td>1</td>
<td>1.5e+04</td>
<td>6.5e-03</td>
<td>2.2e+04</td>
<td>7.7e-03</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>1.7e+02</td>
<td>3.6e-04</td>
<td>7</td>
<td>4.7e+01</td>
<td>2.6e-04</td>
</tr>
<tr>
<td>3</td>
<td>3.9e-02</td>
<td>4.5e-08</td>
<td>15</td>
<td>1.5e+01</td>
<td>3.6e-04</td>
</tr>
<tr>
<td>4</td>
<td>4.1e-07</td>
<td>4.0e-12</td>
<td>17</td>
<td>3.4e-05</td>
<td>4.0e-10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sum k^i$</td>
<td>41</td>
<td></td>
<td></td>
<td>134</td>
<td></td>
</tr>
</tbody>
</table>

### Example 5.3
Again we consider ‘bcstk09’ and seek eigenvalues $\lambda_1 = 7102$ and $\lambda_{20} = 4.06 \times 10^5$ using preconditioned MINRES, with the same preconditioner as in Example 5.2. Here we use the Simoncini-Elden approach where the right hand side is given by (6). We take the shift to be the Rayleigh quotient and always use a fixed tolerance to stop the inner solve. Hence Theorem 4.4 applies. We also compare the use of the residual condition (65) to stop the inner iteration with the stopping condition (67) with (68). We shall use the following description for the methods:

**SEf** Rayleigh quotient shift, preconditioner as in (6), residual stopping condition given by (65) with $\tau^0$ fixed.

**SE** Rayleigh quotient shift, preconditioner as in (6), residual stopping condition given by (67) and (68) (see [18]).

Numerical results using SEf for $\lambda_1$ (called the extreme eigenvalue) and $\lambda_{20}$ (the interior eigenvalue) are given in Table 4, and for SE are given in Table 5. In this latter case the results are given for the optimal choice of $\tau_{SE}$ (see (67)).
$\text{extrem } \tau_{SE} = 0.1 \quad \text{interior } \tau_{SE} = 10^{-4}$

<table>
<thead>
<tr>
<th>(i)</th>
<th>(|r^i|)</th>
<th>(t^i)</th>
<th>(k^i)</th>
<th>(|r^i|)</th>
<th>(t^i)</th>
<th>(k^i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2.7e+05</td>
<td>1.0e-02</td>
<td>5.0e+05</td>
<td>2.0e-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.2e+04</td>
<td>4.6e-03</td>
<td>4</td>
<td>1.1e+04</td>
<td>7.0e-03</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>6.9e+01</td>
<td>5.1e-05</td>
<td>10</td>
<td>4.9e+00</td>
<td>2.1e-05</td>
<td>57</td>
</tr>
<tr>
<td>3</td>
<td>3.1e-03</td>
<td>2.0e-09</td>
<td>18</td>
<td>3.6e-05</td>
<td>9.6e-11</td>
<td>60</td>
</tr>
<tr>
<td>4</td>
<td>4.8e-07</td>
<td>2.7e-12</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\sum k^i)</td>
<td></td>
<td></td>
<td>45</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Numerical Results for preconditioned MINRES to determine an extreme eigenvalue \((\lambda_1)\) and an interior eigenvalue \((\lambda_{20})\) using algorithm SE. Here \(r^i\), \(t^i\) and \(k^i\) are as in Table 2.

$\text{extrem } \tau_{SE} = 10^{-3} \quad \text{interior } \tau_{SE} = 10^{-3}$

<table>
<thead>
<tr>
<th>(i)</th>
<th>(|r^i|)</th>
<th>(t^i)</th>
<th>(k^i)</th>
<th>(|r^i|)</th>
<th>(t^i)</th>
<th>(k^i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2.7e+05</td>
<td>1.0e-02</td>
<td>2.5e+05</td>
<td>1.0e-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.2e+04</td>
<td>5.1e-03</td>
<td>6</td>
<td>1.9e+03</td>
<td>2.4e-03</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>7.1e+01</td>
<td>3.9e-05</td>
<td>14</td>
<td>9.3e+01</td>
<td>9.2e-04</td>
<td>24</td>
</tr>
<tr>
<td>3</td>
<td>3.0e-03</td>
<td>1.6e-09</td>
<td>20</td>
<td>8.6e-03</td>
<td>3.6e-08</td>
<td>59</td>
</tr>
<tr>
<td>4</td>
<td>5.5e-07</td>
<td>3.0e-12</td>
<td>13</td>
<td>1.3e-03</td>
<td>1.2e-08</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td>1.3e-03</td>
<td>1.2e-08</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td>1.3e-03</td>
<td>1.2e-08</td>
<td>2</td>
</tr>
<tr>
<td>(\sum k^i)</td>
<td></td>
<td></td>
<td>53</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Numerical Results for preconditioned MINRES to determine an extreme eigenvalue \((\lambda_1)\) and an interior eigenvalue \((\lambda_{20})\) using algorithm SE. Here \(r^i\), \(t^i\) and \(k^i\) are as in Table 2. Note that the algorithm with \(\tau_{SE} = 10^{-3}\) has failed to converge for the interior eigenvalue and stagnates at \(\|r^i\| = 1.3 \times 10^{-3}\). (We do not reproduce any further iterates.)

Table 6 presents results to show the sensitivity of the performance of algorithm SE to the choice of \(\tau_{SE}\). In all cases the outer iteration was stopped using the relative eigenvalue residual test (66) with \(etol = 10^{-10}\), so that due to the sizes of \(\lambda_1\) and \(\lambda_{20}\) the eigenvalue residual values in Tables 4, 5, and 6 need to reduce below \(7 \times 10^{-6}\) for the extremal eigenvalue and \(4 \times 10^{-5}\) for the interior eigenvalue.

Conclusions 5.3:

a) All result in Tables 4 and 5 show quadratic convergence of the outer iteration as predicted by Theorem 2.2.

b) In Tables 4 and 5 the number of inner iterations needed to find the extreme eigenvalue is much less than the number needed to find the interior eigenvalue. For the extreme eigenvalue \(q = q_e\) in (63), and Conjugate
Gradient like behaviour is observed rather than the slower convergence predicted for the interior eigenvalue where $q = q_i$.

c) The results in Table 4 for the interior eigenvalue are directly comparable with the results of Table 3. The Simoncini-Elden modification of the right hand side of the preconditioned system produces a significant reduction in the total number of inner iterations (134 compared to 186) to achieve the same outer accuracy. This is in agreement with the theory: (63) has the term $\gamma \log 10$ arising from $\sum_{i=0}^{N-1} \log(t_i/t_{i+1})$, whereas (56) contains the much larger term $\sum_{i=0}^{N-1} \log(1/t_i+1)$.

d) In our experiment the choice for $\tau_{SE}$ (see (67)) was rather crucial. An illustration of this is Table 6. Here $\tau_{SE}$ was taken to be $10^{-3}$ for the computation of $\lambda_1$ and $\lambda_{20}$ rather than $10^{-1}$ and $10^{-4}$ in Table 5. There is a small increase for the total number of inner iterations for the extreme eigenvalue. However, for the interior eigenvalue the value of $\tau_{SE} = 10^{-3}$ proved to be not tight enough as the required accuracy is not achieved and stagnation occurred. Note that in Table 4 the corresponding fixed tolerance for SEf is set to $\tau^0 = 0.5$. In our experiments we found that the algorithm SEf was not very sensitive to the precise value of $\tau^0$ and in fact an even better choice for $\tau^0$ in terms of reducing $\sum k_i$ for the interior eigenvalue was $\tau^0 = 0.8$, giving $\sum k_i = 114$. A tightening of the stopping condition increases the number of inner iterations, using $\tau^0 = 0.01$ we gain $\sum k_i = 186$, this is in agreement with the theory as the tightening of the stopping conditions increases the number of inner iterations per outer iteration and the number of outer iterations cannot be improved.

e) Our experiments indicate that if the optimal value of $\tau_{SE}$ is known then the algorithm SE is almost always at least as quick as SEf. However at the moment there does not seem to be a reliable way to choose $\tau_{SE}$, and as we have seen in the previous paragraph failure can occur in circumstances where SEf succeeds. Further $\sum_{i=0}^{N-1} k_i$ can increase considerably if $\tau_{SE}$ is not optimal. However, both methods usually outperform RQIf.

Example 5.4 Finally it is helpful to compare the overall cost of finding an eigenvalue with that of carrying out a standard linear solve. Consider the problem of solving $(A - \sigma I)z = b$ where $A$ is the matrix ‘bsstk09’ as in previous examples, $\sigma = \frac{1}{2}(\lambda_{20} + \lambda_{21})$, and $b$ is the vector we used as starting vector in Example 5.2. From a starting value of $z_0 = 0$ the total number of matrix vector products needed to find $x$ to a residual accuracy of $10^{-10}$ was 86. This compares with 134 matrix vector products to find $\lambda_{20}$ for SEf see Table 4. Overall, we have found that on average it is 2-3 times more expensive to find an eigenvalue using RQIf or RQId than to solve a comparable linear system, even though there may be more than 3 outer iterations required.
References


