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 $\mathbf{Av} = \lambda\mathbf{v}$. Our analysis is designed to apply to the case when $\mathbf{A}$ is large and sparse and where iterative methods are used to solve the shifted linear systems $(\mathbf{A} - \sigma\mathbf{I})\mathbf{y} = \mathbf{x}$ which arise. We present a convergence theory that is independent of the nature of the inexact solver used, and, though the use of the Rayleigh quotient is emphasised, our analysis also extends to quite general choices for shift and inexact solver strategies. Additionally, the convergence framework allows us to treat both standard preconditioning and to present a new analysis of the variation introduced by Simoncini and Eldén (BIT, Vol 42, pp.159-182, 2002). Also, we provide an analysis of the performance of inner iteration solves when preconditioned MINRES is used as the inexact solver. This analysis provides descriptive bounds which are shown to predict well the actual behaviour observed in practice. Also, it explains the improvement in performance of the modification introduced by Simoncini and Eldén over the standard preconditioned form. Importantly, our analysis shows that letting the shift tend to the eigenvalue, as is the case if the Rayleigh quotient is used, does not harm significantly the performance of the iterative method for the shifted systems. Throughout the paper numerical results are given to illustrate the theory.

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

$$\mathbf{Av} \quad = \quad \lambda\mathbf{v}, \tag{1}$$

*Fachbereich Mathematik, JWG- Universität Frankfurt, Postfach 11 19 32, D-60054 Frankfurt, Germany

†Department of Mathematical Sciences, University of Bath, Claverton Down, BA2 7AY, United Kingdom

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

$$(\mathbf{A} - \sigma \mathbf{I})\mathbf{y} \quad = \quad \mathbf{x}, \tag{2}$$

where σ is the shift. If $\mathbf{A}$ 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 inverse iteration algorithm requiring the solve of (2) at each step, with the inner iteration being the inexact 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 a variant of 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 [19]. Inexact inverse iteration for symmetric matrices was discussed in [22] 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 [20] where a version of inexact Rayleigh quotient iteration is discussed. Several new ideas are introduced especially with regard to the appropriate linear system to be solved when Cholesky preconditioning is applied to (2), and with regard to the stopping condition in the inner iteration. We shall discuss some of these ideas in detail in this paper. Also [20] contains a theoretical discussion on the equivalence of inexact inverse iteration and Jacobi-Davidson (or projected Newton's Method). For nonsymmetric matrices an inexact inverse iteration algorithm with fixed shift 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 [13]. 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], [12] and [14]. Particularly successful for extreme eigenvalues is the LOBPCG method discussed in [11]

There are several new features in the present paper. In Section 2 we present a convergence theory, independent of the details of the inexact iterative solver. This allows us to recover and extend existing results of [22] on inexact Rayleigh quotient iteration, and also to obtain a quite general convergence result. Next we extend the theory to include the alteration in the right hand side of (2) introduced in [20], but note that our analysis is valid for any preconditioner and is not restricted to Cholesky preconditioners as in [20]. In Section 3 we use some standard results for MINRES (see, for example, [16, 8]) to provide new bounds on the number of inner iterations at each step of the outer iteration. These bounds are seen in our numerical examples to provide qualitatively correct information about the performance of both unpreconditioned and preconditioned inner solves. We also present an analysis that confirms the superiority of the system introduced by [20] for preconditioned Rayleigh quotient iteration over the standard preconditioned system. Our analysis also shows that we need not be concerned that the Krylov solver is applied to a matrix which is becoming more and more singular. The explanation lies in the interplay between the shift tending towards the eigenvalue and the right hand side of the shifted system 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 [19].

A key feature of this paper is that we use a residual stopping condition in the convergence theory of §2, in our numerical experiments using MINRES, and in the analysis of the performance of the inner solve in §3. This allows a unified account of both the theory and practice.

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

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v}, \quad \| \mathbf{v} \| = 1, \tag{3}$$

where $\mathbf{A}$ is a real symmetric $n \times n$ matrix, with eigenvalues λ_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 σ. 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 σ which are close enough to λ_1 in the sense that

$$0 < |\lambda_1 - \sigma| < \frac{1}{2} \min_{j=2,\ldots,n} |\lambda_1 - \lambda_j| \; . \tag{4}$$

Then we have an induced ordering on the eigenvalues

$$0 \, < |\lambda_1 - \sigma| < |\lambda_2 - \sigma| \leq \ldots \leq |\lambda_n - \sigma| \; . \tag{5}$$

Note that this ordering depends on σ, and as σ varies it is possible this this ordering might change. However, nothing essential in the theory is lost by assuming this ordering.

We are interested in the case when $\mathbf{A}$ is large and sparse and so the shifted systems will be solved (inexactly) by some iterative algorithm. In Section 3 we will consider in detail the case when the iterative solver is MINRES (see, for example, [8]), which is appropriate since $\mathbf{A}-\sigma\mathbf{I}$ is symmetric but is likely to be indefinite. However, in this section we will present a convergence theory that is independent of the solver.

The inexact inverse iteration algorithm is given in Algorithm 1.

Algorithm 1: Inexact Inverse Iteration alg1

```
Given $\mathbf{x}^{(0)}$ with $\| \mathbf{x}^{(0)} \|= 1$. For $i = 0, 1, 2, \ldots$
    (1) Choose $\sigma^{(i)}$ and $\tau^{(i)}$
    (2) Solve $(\mathbf{A}-\sigma^{(i)}\mathbf{I})\mathbf{y}^{(i)} = \mathbf{x}^{(i)}$ inexactly, that is,
        $\| (\mathbf{A}-\sigma^{(i)}\mathbf{I})\mathbf{y}^{(i)} - \mathbf{x}^{(i)} \| \leq \tau^{(i)}$
    (3) Update $\mathbf{x}^{(i+1)} = \mathbf{y}^{(i)}/ \| \mathbf{y}^{(i)} \|$
    (4) Test for convergence
```

We refer to the iteration in Algorithm 1 as the *outer* iteration and the iteration implicit in the inexact solve as the *inner* iteration, and so inexact inverse iteration is an example of an inner-outer iterative algorithm (see, for example, [2, 3, 21, 5]).

Before we analyse the convergence of Algorithm 1 we recall some definitions and notation. A key concept is the orthogonal splitting used in [17, p. 63]. If $\mathbf{x}^{(i)}$ is any unit vector approximating $\mathbf{v}_1$, we introduce the splitting

$$\mathbf{x}^{(i)} = \cos\theta^{(i)}\,\mathbf{v}_1 + \sin\theta^{(i)}\,\mathbf{u}^{(i)}, \quad \mathbf{u}^{(i)} \perp \mathbf{v}_1, \tag{6}$$

with $\| \mathbf{v}_1 \| = \| \mathbf{u}^{(i)} \| = 1$ and $\theta^{(i)} = \angle(\mathbf{x}^{(i)}, \mathbf{v}_1)$, the error angle. For convenience we usually write

$$c^{(i)} = \cos\theta^{(i)}, \quad s^{(i)} = \sin\theta^{(i)}, \quad \text{and} \quad t^{(i)} = |s^{(i)}|/|c^{(i)}| = |\tan\theta^{(i)}|. \tag{7}$$

From (6), $\|\mathbf{x}^{(i)} - c^{(i)}\mathbf{v}_1\| = |s^{(i)}| \leq t^{(i)}$ and normally we use $|s^{(i)}|$ or $t^{(i)}$ as a measure of the convergence of $\mathbf{x}^{(i)}$ to span $\{\mathbf{v}_1\}$. However, we note that convergence occurs if we can prove one of $|\theta^{(i)}| \to 0, |s^{(i)}| \to 0, t^{(i)} \to 0$, or $|c^{(i)}| \to 1$. Also, recall that, for $\mathbf{x}^{(i)}$ given by (6), the Rayleigh quotient, $\varrho(\mathbf{x}^{(i)}) = \mathbf{x}^{(i)^T}\mathbf{A}\mathbf{x}^{(i)}$, satisfies

$$\lambda_1 - \varrho(\mathbf{x}^{(i)}) = (s^{(i)})^2[\lambda_1 - \varrho(\mathbf{u}^{(i)})], \tag{8}$$

and the eigenvalue residual, $\mathbf{r}^{(i)}$, defined by

$$\mathbf{r}^{(i)} := (\mathbf{A} - \varrho(\mathbf{x}^{(i)})\mathbf{I})\mathbf{x}^{(i)} \tag{9}$$

satisfies ([17], Theorem 11.7.1)

$$|s^{(i)}|\,|\lambda_2 - \varrho(\mathbf{x}^{(i)})| \leq \| \mathbf{r}^{(i)} \| \leq |s^{(i)}|\,|\lambda_n - \lambda_1|. \tag{10}$$

2.2 Convergence Theory

We now present a convergence theory for inexact inverse iteration. Various choices for $\sigma^{(i)}$ and $\tau^{(i)}$ in Algorithm 1 are possible and our analysis allows us to obtain a full understanding of the effects of the different options. However, the most important choice for $\sigma^{(i)}$ is the Rayleigh quotient, and we shall emphasise this case throughout. For example, it is a classical result, proved in [15], that inverse iteration with Rayleigh quotient shifts and exact linear solves converges cubically when applied to symmetric matrices (see [17] 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 Theorem 2.1 how to do this.

We start with a bound for the error after one step of Algorithm 1. Similar results are to be found in [22] and, for nonsymmetric matrices with fixed shift, [6]. First we provide some notation. Because of step (2) of Algorithm 1 we have a residual

$$\mathbf{res}^{(i)} \quad := \quad \mathbf{x}^{(i)} - (\mathbf{A} - \sigma^{(i)}\mathbf{I})\mathbf{y}^{(i)} \tag{11}$$

which satisfies

$$\| \, \mathbf{res}^{(i)} \, \|_2 \quad \leq \quad \tau^{(i)}. \tag{12}$$

Note that the "inner linear solve" residual $\mathbf{res}^{(i)}$ should not be confused with the outer eigenvalue residual $\mathbf{r}^{(i)}$ which was defined in (9). Now we can state the following Lemma.

Lemma 2.1 *If $\mathbf{x}^{(i)}$ is such that*

$$|c^{(i)}| > \tau^{(i)}, \tag{13}$$

then one step of Algorithm 1 yields $\mathbf{x}^{(i+1)}$ with

$$\frac{|\lambda_1 - \sigma^{(i)}|}{|\lambda_n - \sigma^{(i)}|} \, \frac{\big|\, |s^{(i)}| - \, \| \top\!\!\top \, \mathbf{res}^{(i)} \| \big|}{|c^{(i)}| + \tau^{(i)}} \leq t^{(i+1)} \leq \frac{|\lambda_1 - \sigma^{(i)}|}{|\lambda_2 - \sigma^{(i)}|} \, \frac{|s^{(i)}| + \tau^{(i)}}{|c^{(i)}| - \tau^{(i)}}. \tag{14}$$

where $\mathbf{res}^{(i)}$ satisfies (11) and (12), and $\top\!\!\top := \mathbf{I} - \mathbf{v}_1\mathbf{v}_1^T$.

Proof: First use step (3) in Algorithm 1 and (6) to rewrite (11) as

$$\|\mathbf{y}^{(i)}\| \, (\mathbf{A} - \sigma^{(i)}\mathbf{I})\mathbf{x}^{(i+1)} \quad = \quad \cos\theta^{(i)} \, \mathbf{v}_1 + \sin\theta^{(i)} \, \mathbf{u}^{(i)} - \mathbf{res}^{(i)},$$

and using (6) again gives

$$\begin{aligned} \|\mathbf{y}^{(i)}\| \, \{c^{(i+1)}(\lambda_1 - \sigma^{(i)})\mathbf{v}_1 + s^{(i+1)}(\mathbf{A} - \sigma^{(i)}\mathbf{I})\mathbf{u}^{(i+1)}\} \\ = c^{(i)}\mathbf{v}_1 + s^{(i)}\mathbf{u}^{(i)} - \mathbf{res}^{(i)}. \end{aligned} \tag{15}$$

Because $\mathbf{v}_1$ is orthogonal to both $\mathbf{u}^{(i)}$ and $\mathbf{u}^{(i+1)}$ we can resolve (15) into two equations in span$\{\mathbf{v}_1\}$ and $\{\mathbf{v}_1\}^\perp$, respectively. Along $\mathbf{v}_1$ we have

$$\|\mathbf{y}^{(i)}\| \, c^{(i+1)}(\lambda_1 - \sigma^{(i)}) \quad = \quad c^{(i)} - \mathbf{v}_1^T\mathbf{res}^{(i)}, \tag{16}$$

and in $\{\mathbf{v}_1\}^\perp$ we have

$$\|\mathbf{y}^{(i)}\| \ s^{(i+1)}(\mathbf{A}-\sigma^{(i)}\mathbf{I})\mathbf{u}^{(i+1)} \quad = \quad s^{(i)}\mathbf{u}^{(i)} - \top\!\!\top\,\mathbf{res}^{(i)}. \tag{17}$$

Now let $(\mathbf{A}-\sigma^{(i)}\mathbf{I})_\perp$ denote the restriction of $(\mathbf{A}-\sigma^{(i)}\mathbf{I})$ to $\{\mathbf{v}_1\}^\perp$. This linear operator is invertible on $\{\mathbf{v}_1\}^\perp$ and satisfies $\|(\mathbf{A}-\sigma^{(i)}\mathbf{I})_\perp^{-1}\|=|\lambda_2-\sigma^{(i)}|^{-1}$, and $\|(\mathbf{A}-\sigma^{(i)}\mathbf{I})_\perp\|=|\lambda_n-\sigma^{(i)}|$. Thus from (17) we obtain

$$\|\mathbf{y}^{(i)}\| \ |s^{(i+1)}| \quad \leq \quad |\lambda_2-\sigma^{(i)}|^{-1}\ \{|s^{(i)}|+\|\top\!\!\top\,\mathbf{res}^{(i)}\|\}. \tag{18}$$

Combining this with a trivial lower bound on $\|\mathbf{y}^{(i)}\|\ c^{(i+1)}$ from (16) we obtain

$$t^{(i+1)} = \frac{\|\mathbf{y}^{(i)}\|\ |s^{(i+1)}|}{\|\mathbf{y}^{(i)}\|\ |c^{(i+1)}|} \quad \leq \quad \left|\frac{\lambda_1-\sigma^{(i)}}{\lambda_2-\sigma^{(i)}}\right| \quad \frac{|s^{(i)}|+\|\top\!\!\top\,\mathbf{res}^{(i)}\|}{|c^{(i)}|-|\mathbf{v}_1^T\mathbf{res}^{(i)}|}, \tag{19}$$

from which the right hand side of (14) follows on applying (12).

A similar approach is used to obtain the left hand side of (14). First observe

$$\begin{aligned} \|\mathbf{y}^{(i)}\|\ |s^{(i+1)}| \quad &= \quad \|(\mathbf{A}-\sigma^{(i)}\mathbf{I})_\perp^{-1}(s^{(i)}\mathbf{u}^{(i)}+\top\!\!\top\,\mathbf{res}^{(i)})\| && (20)\\ &\geq \quad \frac{1}{|\lambda_n-\sigma^{(i)}|}\ \|s^{(i)}\mathbf{u}^{(i)}+\top\!\!\top\,\mathbf{res}^{(i)})\| && (21)\\ &\geq \quad \frac{1}{|\lambda_n-\sigma^{(i)}|}\ \Big||s^{(i)}|-\|\top\!\!\top\,\mathbf{res}^{(i)}\|\Big|. && (22) \end{aligned}$$

Then combine this with a trivial upper bound on $\|\,\mathbf{y}^{(i)}\,\|\ c^{(i+1)}$ from (16), to obtain

$$\left|\frac{\lambda_1-\sigma^{(i)}}{\lambda_n-\sigma^{(i)}}\right| \quad \frac{\big||s^{(i)}|-\|\top\!\!\top\,\mathbf{res}^{(i)}\|\big|}{|c^{(i)}|+|\mathbf{v}_1^T\mathbf{res}^{(i)}|} \leq t^{(i+1)}, \tag{23}$$

from which the left-hand side of (14) follows. □

Note that if $\tau^{(i)}=0$, as is the case for exact solves, then we recover the standard results for exact solves (see, for example, [17]).

Since we restrict attention to symmetric matrices there are two main choices for $\sigma^{(i)}$, namely the natural Rayleigh quotient shift (popular because of its ease of computation and its quadratic approximation property when the approximate eigenvector is accurate enough), and a fixed shift (employed in cases where the approximate eigenvector may be poor). We restrict to the Rayleigh quotient case here but return to the fixed shift case in §2.3. Then with $\sigma^{(i)}=\varrho(\mathbf{x}^{(i)})$ we have from (8)

$$|\lambda_1-\sigma^{(i)}| = |\lambda_1-\varrho(\mathbf{u}^{(i)})||s^{(i)}|^2. \tag{24}$$

There are also two practical options for the choice of tolerance $\tau^{(i)}$. Either we can choose to decrease $\tau^{(i)}$ as the outer iteration proceeds, or $\tau^{(i)}$ can be held fixed. Let us first consider the case of a decreasing tolerance, so assume in addition to (13) that there is a constant C_2, independent of i, such that

$$\tau^{(i)} \leq C_2|s^{(i)}| < |c^{(i)}| \tag{25}$$

as would be the case if $\tau^{(i)}$ were chosen bounded by a multiple of the eigenvalue residual $\|\mathbf{r}^{(i)}\|$, see (9) and (10). Then the right hand inequality in (14) combined with (24) and (25) gives

$$\frac{t^{(i+1)}}{(t^{(i)})^3} \leq \frac{|\lambda_1 - \varrho(\mathbf{u}^{(i)})|(c^{(i)})^2}{|\lambda_2 - \varrho(\mathbf{x}^{(i)})|} \; \frac{1+C_2}{1-C_2 t^{(i)}}, \tag{26}$$

(cf. [17] eqn(4.22)). Equation (26) shows that in the asymptotic regime we achieve cubic convergence, just as would be attained if exact solves were used. This result is implicit in [22].

The more interesting and practical case is when the tolerances $\tau^{(i)}$ are not required to decrease as the outer iteration proceeds so that the inner solves are implemented with a fixed tolerance and hence are potentially cheaper. So, in contrast to (25), let us assume that

$$\tau^{(i)} = \tau^{(0)}, \quad \forall i; \quad \tau^{(0)} \leq C_3 |c^{(i)}|, \quad C_3 < 1. \tag{27}$$

With this choice for $\tau^{(i)}$, (14) with (24) gives

$$\frac{t^{(i+1)}}{(t^{(i)})^2} \leq \frac{|\varrho(\mathbf{u}^{(i)}) - \lambda_1|(c^{(i)})^2}{|\lambda_2 - \varrho(\mathbf{x}^{(i)})|} \quad \frac{t^{(i)}+C_3}{1-C_3}. \tag{28}$$

Thus with a fixed tolerance cubic convergence is lost but quadratic convergence is maintained because of the quadratic convergence of the Rayleigh quotient. We gather together these results in the following Theorem.

Theorem 2.1 *Let* $\mathbf{A}$ *be a real* $n \times n$ *symmetric matrix and consider the application of Algorithm 1 with* $\sigma^{(i)}$ *chosen to be the Rayleigh quotient* $\varrho(\mathbf{x}^{(i)})$*. Assume* $c^{(i)}$ *and* $s^{(i)}$ *are given by* $\mathbf{x}^{(i)} = c^{(i)}\mathbf{v}_1 + s^{(i)}\mathbf{u}^{(i)}$ *and (12) is satisfied.*

(a) (Decreasing tolerance) If $\tau^{(i)}$ *in (12) satisfies (25) then Algorithm 1 converges cubically.*

(b) (Fixed tolerance) If $\tau^{(i)}$ *in (12) is chosen to satisfy (27) then Algorithm 1 converges quadratically.*

This theorem follows from bounds (26) and (28) above. It is also a corollary of Theorem 2.2 that allows more general choices for $\sigma^{(i)}$ and $\tau^{(i)}$, and which is proved in the next subsection.

Now we look at some numerical results from a simple model problem to illustrate Theorem 2.1.

Example 2.1 *Consider the eigenvalue problem for the 2-D Laplacian,* $-\nabla^2 u = \lambda u$*, with homogenous Dirichlet boundary conditions on the rectangle* $0 \leq x \leq 1$*,* $0 \leq y \leq 1.3$*, which is discretised using finite differences with the 5-point Laplacian approximation on a* 12×12 *regular grid. In Table 1 we present numerical results obtained when calculating* λ_1 *(*$\simeq 15.6$*) the smallest (simple) eigenvalue of the discretised matrix. Using unpreconditioned MINRES as the inexact solver we apply the two versions of inexact inverse iteration discussed in Theorem 2.1, namely*

	RQId		RQIf	
	$\tau^{(0)} = 0.1, C_2 = 0.1$		$\tau^{(0)} = 0.1$	
i	$\log_{10} s^{(i)}$	$k^{(i-1)}$	$\log_{10} s^{(i)}$	$k^{(i-1)}$
0	-0.14		-0.12	
1	-1.62	15	-1.41	19
2	-4.33	24	-3.85	19
3	-12.90	45	-9.03	33
4	-36.19	78	-19.46	50
5	-82.66	113	-40.72	76
6			-82.96	108
$\sum k^{(i)}$		275		305

Table 1: Numerical results for unpreconditioned MINRES applied to Example 2.1 using the methods in Theorem 2.1. Here $s^{(i)}$ denotes $\sin\theta^{(i)}$ defined by (6), $k^{(i)}$ denotes the number of inner iterations at the ith step, and C_2 is the constant in (25).

RQId: *Rayleigh quotient shift, decreasing tolerance,*

RQIf: *Rayleigh quotient shift, fixed tolerance,*

(cases a) and b) in Theorem 2.1 respectively). Each row in Table 1 provides the outer iteration number, $\log_{10} s^{(i)}$ *(calculated using the exact* $\mathbf{v}_1$*) and* $k^{(i-1)}$ *the number of inner iterations needed to satisfy the residual condition in step (2) of the Algorithm. To illustrate accurately the convergence rates attained, the experiment was carried out using MATLAB with variable precision arithmetic using 128 decimal digit arithmetic. In both experiments we used* $\|(\mathbf{A}-\varrho(\mathbf{x}^{(i)})\mathbf{I})\mathbf{x}^{(i)}\| \leq 10^{-80}|\varrho(\mathbf{x}^{(i)})|$ *as the test for convergence in step (4) of Algorithm 1.*

As predicted in Theorem 2.1, we observe in Table 1 quadratic convergence for **RQIf** and cubic convergence for **RQId**. At a practical level (that is, to double precision on most of the current computers) our experiments have shown that there is little difference between the two methods with both needing roughly speaking about the same number of inner iterations in total.

2.3 Theory for a General Method

In this subsection we provide the details of the convergence theory for Algorithm 1 for general choices for $\sigma^{(i)}$and $\tau^{(i)}$. We shall use the right hand bound in (14)

$$t^{(i+1)} \quad \leq \quad \frac{|\lambda_1 - \sigma^{(i)}|}{|\lambda_2 - \sigma^{(i)}|} \, \frac{|s^{(i)}| + \tau^{(i)}}{|c^{(i)}| - \tau^{(i)}}, \tag{29}$$

though a more refined analysis would be possible using the right-hand side of (19), namely,

$$t^{(i+1)} \quad \leq \quad \frac{|\lambda_1 - \sigma^{(i)}|}{|\lambda_2 - \sigma^{(i)}|} \, \frac{|s^{(i)}| + \|\Pi \, \mathbf{res}^{(i)}\|}{|c^{(i)}| - |\mathbf{v}_1^T \mathbf{res}^{(i)}|}, \tag{30}$$

if more were known about the size of certain components of $\mathbf{res}^{(i)}$ (see [1]). Similar expressions are found in [22] and [20], where a bound is derived directly using the properties of a Krylov solver. In the latter paper the observation is made that for a Krylov solver the term $\mathbf{v}_1^T \mathbf{res}^{(i)}$ is likely to reduce significantly in the first few iterations since $\mathbf{x}^{(i)}$ is rich in the direction $\mathbf{v}_1$. Hence, since $|c^{(i)}|$ tends to 1, the factor $|c^{(i)}| - |\mathbf{v}_1^T \mathbf{res}^{(i)}|$ is unlikely to cause any problems in practice.

Equation (29) is used in the proof of the following general convergence result.

Theorem 2.2 *Let $\mathbf{A}$ be a real $n \times n$ symmetric matrix and consider the application of Algorithm 1 to find a simple eigenpair $(\lambda_1, \mathbf{v}_1)$. Assume $\mathbf{x}^{(i)}$ is given by (6) and that $\sigma^{(i)}$ satisfy (4). Additionally, assume $\sigma^{(i)}$ and $\tau^{(i)}$ are chosen to satisfy*

$$\begin{aligned} |\lambda_1 - \sigma^{(i)}| &\leq C_1 |s^{(i)}|^{\alpha} & (31) \\ \tau^{(i)} &\leq \min\{C_2 |s^{(i)}|^{\beta}, C_3 |c^{(i)}|\} & (32) \end{aligned}$$

for some constants $\alpha, \beta, C_1, C_2, C_3$ independent of i 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 $\mathbf{x}^{(0)}$ 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

$$t^{(i+1)} \leq C_4 t^{(i)},$$

with C_4 independent of i, and the method converges. In this case,

$$t^{(i+1)} \leq C (t^{(i)})^{\alpha+\beta}$$

with $C = C_4 / \, |s^{(0)}|^{\alpha+\beta-1}$, and so convergence is of order $(\alpha + \beta)$.

Proof: Inserting the bounds (31) and (32) into (29) produces

$$t^{(i+1)} \leq \frac{2C_1 |s^{(i)}|^{\alpha}}{|\lambda_2 - \lambda_1|} \, \frac{|s^{(i)}| + C_2 |s^{(i)}|^{\beta}}{|c^{(i)}| - C_3 |c^{(i)}|},$$

where we have used (4) and (5) to bound the term $|\lambda_2 - \sigma^{(i)}|$ in the denominator of (29). Now, rearranginging we have

$$t^{(i+1)} \leq C_4 t^{(i)} \frac{(1 + C_2 |s^{(i)}|^{\beta-1}) |s^{(i)}|^{\alpha}}{(1 + C_2) |s^0|^{\alpha+\beta-1}} \leq C_4 t^{(i)} \frac{|s^{(i)}|^{\alpha+\beta-1}}{|s^0|^{\alpha+\beta-1}}.$$

Convergence follows by induction on i. Finally

$$t^{(i+1)} \leq \frac{C_4}{|s^{(0)}|^{\alpha+\beta-1}} (t^{(i)})^{\alpha+\beta}. \tag{33}$$

□

As a consequence of Theorem 2.2 we obtain the convergence of $\mathbf{x}^{(i)}$ to $\pm\mathbf{v}_1$ and $\varrho(\mathbf{x}^{(i)})$ to λ_1 as $i \to \infty$. Note that the condition $c^{(0)} \neq 0$ ensures that the starting vector is not orthogonal to the required direction $\mathbf{v}_1$.

Theorem 2.1 is a special case of Theorem 2.2. For Rayleigh quotient shifts we have $\alpha = 2$, and cases (a) and (b) in Theorem 2.1 correspond to $\beta = 1$ and $\beta = 0$ respectively.

In addition to the two strategies examined in Theorem 2.1 a third strategy could be based on keeping $\sigma^{(i)}$ fixed and choosing $\tau^{(i)} \leq C_2|s^{(i)}|$ as would be the case if $\tau^{(i)} = O(\|\mathbf{r}^{(i)}\|)$ (see (10)). Theorem 2.2 shows that this approach would attain at least linear convergence for accurate enough $\sigma^{(0)}$ and $\mathbf{x}^{(0)}$. Also the theory indicates that there is likely to be little gain in ever choosing $\tau^{(i)} = o(|s^{(i)}|)$.

Finally 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 2.2 *Suppose that* $\mathbf{x}^{(i)} = c^{(i)}\mathbf{v}_1 + s^{(i)}\mathbf{v}_2$. *If we construct a particular* $\mathbf{y}^{(i)}$ *via the formula,* $\mathbf{y}^{(i)} = \frac{c^{(i)}}{\lambda_1-\sigma^{(0)}}\mathbf{v}_1 + \frac{s^{(i)}+\tau^{(0)}}{\lambda_2-\sigma^{(0)}}\mathbf{v}_2$ *then* $\mathbf{x}^{(i)}$ *and* $\mathbf{y}^{(i)}$ *satisfy* $\|\mathbf{x}^{(i)} - (\mathbf{A} - \sigma^{(0)}\mathbf{I})\mathbf{y}^{(i)}\| = \tau^{(0)}$. *Then computing* $\mathbf{x}^{(i+1)}$ *and representing it in the form (6) we obtain* $\tan\theta^{(i+1)} = \frac{\lambda_1-\sigma^{(0)}}{\lambda_2-\sigma^{(0)}} \frac{\sin\theta^{(i)}+\tau^{(0)}}{\cos\theta^{(i)}}$. *Repeating this formula for each i we obtain a fixed point iteration for* $\theta^{(i)}$. *If* $\theta^{(i)} \to \theta$ *as* $i \to \infty$ *then* θ *satisfies* $\sin\theta = \frac{\lambda_1-\sigma^{(0)}}{\lambda_2-\lambda_1}\tau^{(0)}$, *which is nonzero unless* $\sigma^{(0)} = \lambda_1$ *or* $\tau^{(0)} = 0$. *Such non-convergence is often refered to as stagnation, see, for example, [18].*

2.4 Convergence Theory for Preconditioned Solves

In this subsection we first discuss briefly the convergence theory for standard inverse iteration with preconditioned inner solves. Then we shall use this to derive a new convergence analysis of the variant of preconditioned inverse iteration introduced by [20].

Whatever iterative algorithm is used to solve

$$(\mathbf{A} - \sigma^{(i)}\mathbf{I})\mathbf{y}^{(i)} = \mathbf{x}^{(i)} \tag{34}$$

in Algorithm 1 it will almost certainly be applied to a preconditioned system, and since $\mathbf{A}$ is symmetric it is common to use a symmetric positive definite preconditioner. If $\mathbf{P}$ is a positive definite symmetric matrix that approximates $(\mathbf{A} - \sigma^{(i)}\mathbf{I})$ in some way, then, at least for the theory, we may introduce a factorisation $\mathbf{P} = \mathbf{P}_1\mathbf{P}_1^T$, and to preserve the symmetry in the system we may consider the symmetrically preconditioned system

$$\mathbf{P}_1^{-1}(\mathbf{A} - \sigma^{(i)}\mathbf{I})\mathbf{P}_1^{-T}\widetilde{\mathbf{y}}^{(i)} \;=\; \mathbf{P}_1^{-1}\mathbf{x}^{(i)}; \qquad \mathbf{y}^{(i)} \;=\; \mathbf{P}_1^{-T}\widetilde{\mathbf{y}}^{(i)}. \tag{35}$$

Of course the preconditioner $\mathbf{P}$ or the factorisation $\mathbf{P} = \mathbf{P}_1\mathbf{P}_1^T$ is often not needed in practice and implementation of iterative methods for (35) may only require the action of $\mathbf{P}^{-1}$.

We shall assume that the stopping condition for (35) is based on the residual of the original system, that is, although $(\mathbf{A}-\sigma^{(i)}\mathbf{I})\mathbf{y}^{(i)} = \mathbf{x}^{(i)}$ is solved using (35) the iteration stops when $\|\mathbf{res}^{(i)}\| \leq \tau^{(i)}$ where $\mathbf{res}^{(i)}$ is defined by (11). Hence the convergence theory for the outer iteration given in the previous subsections applies, and is not repeated here.

An alternative to (35) is discussed in [20]. As we shall see in Section 3, this alternative is beneficial for the performance of the preconditioned iterative solver. The idea is to replace (35) by

$$\mathbf{P}_1^{-1}(\mathbf{A}-\sigma^{(i)}\mathbf{I})\mathbf{P}_1^{-T}\widetilde{\mathbf{y}}^{(i)} \;=\; \mathbf{P}_1^T\mathbf{x}^{(i)}; \qquad \mathbf{y}^{(i)} \;=\; \mathbf{P}_1^{-T}\widetilde{\mathbf{y}}^{(i)}. \tag{36}$$

Here the right hand side in the shifted linear system differs from that of (35). Multiplying by $\mathbf{P}_1$ shows that (36) is equivalent to

$$(\mathbf{A}-\sigma^{(i)}\mathbf{I})\mathbf{y}^{(i)} \;=\; \mathbf{P}\mathbf{x}^{(i)}, \tag{37}$$

so that the basic step in inverse iteration is changed by replacing $\mathbf{x}^{(i)}$ by $\mathbf{P}\mathbf{x}^{(i)}$ on the right hand side. (Note, for this formulation, if $\mathbf{P} = (\mathbf{A}-\sigma^{(i)}\mathbf{I})$, then $\mathbf{y}^{(i)} = \mathbf{x}^{(i)}$ and no progress is made by the outer iteration.) Not surprisingly, our analysis will show that this alteration removes the possibility of attaining cubic convergence (except in a very special case, see Example 2.4). If we assume that the iterative solution of (36) is stopped by a relative residual test on (37), namely,

$$\|(\mathbf{A}-\sigma^{(i)}\mathbf{I})\mathbf{y}^{(i)} - \mathbf{P}\mathbf{x}^{(i)}\| \leq \tau^{(i)}\ \|\mathbf{P}\mathbf{x}^{(i)}\|, \tag{38}$$

then we obtain Algorithm 2. Note that putting $\mathbf{P} = \mathbf{I}$ in Algorithm 2 recovers Algorithm 1.

To analyse Algorithm 2 let us introduce

$$\mathbf{res}^{(i)} := \mathbf{P}\mathbf{x}^{(i)} - (\mathbf{A}-\sigma^{(i)}\mathbf{I})\mathbf{y}^{(i)}, \tag{39}$$

where, as indicated by step (3) of Algorithm 2,

$$\|\mathbf{res}^{(i)}\| \leq \tau^{(i)}\ \|\mathbf{P}\mathbf{x}^{(i)}\|\ . \tag{40}$$

If we assume (cf. (13))

$$|\mathbf{v}_1^T\mathbf{P}\mathbf{x}^{(i)}| > |\mathbf{v}_1^T\mathbf{res}^{(i)}|, \tag{41}$$

then by carrying out an analysis similar to the proof of Lemma 2.1 we obtain the one step bound

$$t^{(i+1)} \quad\leq\quad \frac{|\lambda_1-\sigma^{(i)}|}{|\lambda_2-\sigma^{(i)}|}\ \frac{\|\top\!\top\,\mathbf{P}\mathbf{x}^{(i)}\| + \|\top\!\top\,\mathbf{res}^{(i)}\|}{|\mathbf{v}_1^T\mathbf{P}\mathbf{x}^{(i)}| - |\mathbf{v}_1^T\mathbf{res}^{(i)}|}, \tag{42}$$

Algorithm 2: Preconditioned Inexact Rayleigh Quotient Iteration $\boxed{\text{alg2}}$

```
Choose $\mathbf{P}$. Given $\mathbf{x}^{(0)}$ with $\| \mathbf{x}^{(0)} \|= 1$.
For $i = 0, 1, 2, \dots$
    (1) Choose $\sigma^{(i)}$ as the Rayleigh quotient
    (2) Choose $\tau^{(i)}$
    (3) Solve $(\mathbf{A} - \sigma^{(i)}\mathbf{I})\mathbf{y}^{(i)} = \mathbf{P}\mathbf{x}^{(i)}$ inexactly using the preconditioner $\mathbf{P}$, such that
        $\| (\mathbf{A} - \sigma^{(i)}\mathbf{I})\mathbf{y}^{(i)} - \mathbf{P}\mathbf{x}^{(i)} \| \leq \tau^{(i)} \, \|\mathbf{P}\mathbf{x}^{(i)}\|$
    (4) Update $\mathbf{x}^{(i+1)} = \mathbf{y}^{(i)} / \, \|\mathbf{y}^{(i)}\|$
    (5) Test for convergence
```

which reduces to (19) if $\mathbf{P} = \mathbf{I}$.

Now it is likely that little will be known about the quantity $\| \top\!\!\!\top \, \mathbf{P}\mathbf{x}^{(i)} \|$ on the numerator of the right hand side of (42) except that (for general choices of $\mathbf{P}$) it is unlikely to tend to zero, and so there will be little point in choosing decreasing tolerances $\tau^{(i)}$. Hence Algorithm 2 will typically be used with a fixed tolerance. Now, if we assume that there exists a C_3, independent of i, such that

$$\tau^{(i)} = \tau^{(0)}, \quad \forall i; \quad \tau^{(0)} \, \|\mathbf{P}\mathbf{x}^{(i)}\| \leq C_3 \, |\mathbf{v}_1^T \mathbf{P}\mathbf{x}^{(i)}|, \quad C_3 < 1, \tag{43}$$

(which if $\mathbf{P} = \mathbf{I}$ reduces to (27)) then recalling $\sigma^{(i)}$ is the Rayleigh quotient

$$\frac{t^{(i+1)}}{(t^{(i)})^2} \quad \leq \quad \frac{|\lambda_1 - \varrho(\mathbf{u}^{(i)})| |c^{(i)}|^2}{|\lambda_2 - \sigma^{(i)}|} \, \frac{\|\mathbf{P}\mathbf{x}^{(i)}\| \, (1 + \tau^{(0)})}{|\mathbf{v}_1^T \mathbf{P}\mathbf{x}^{(i)}| \, (1 - C_3)}. \tag{44}$$

Hence we have the following Theorem:

Theorem 2.3 *Let* $\mathbf{A}$ *be a real* $n \times n$ *symmetric matrix and consider the application of Algorithm 2 to find a simple eigenpair* $(\lambda_1, \mathbf{v}_1)$*. Assume* $c^{(0)} \neq 0$, $\tau^{(i)} = \tau^{(0)}$ *for all* i*, and that (41) and (43) hold. Then the method is quadratically convergent for a sufficiently close starting guess.*

For a general preconditioner we will be required to have available a subroutine for computing the action of $\mathbf{P}^{-1}$, but little may be known about the action of $\mathbf{P}$. Thus in general one may not be able to say much about the second quotient on the right hand side of (44). However, we can interprete condition (43) as describing the quality of $\mathbf{P}$ as a preconditioner for the eigenvalue problem. In particular, if $\mathbf{x}^{(i)} \to \mathbf{v}_1$,

$$\frac{\|\mathbf{P}\mathbf{x}^{(i)}\|}{|\mathbf{v}_1^T\mathbf{P}\mathbf{x}^{(i)}|} \to \frac{\|\mathbf{P}\mathbf{v}_1\|}{|\mathbf{v}_1^T\mathbf{P}\mathbf{v}_1|}, \tag{45}$$

then (43) requires that $\mathbf{P}\mathbf{v}_1$ should be rich in the direction $\mathbf{v}_1$. In Example 2.3 for the calculation of the 20^{th} eigenvalue of a certain matrix, direct evaluation found that $\|\mathbf{P}\mathbf{x}^{(i)}\| < 1.5\ |\mathbf{v}_1^T\mathbf{P}\mathbf{x}^{(i)}|$ for all i, which is consistent with assumption (43). In these same experiments $|\mathbf{v}_1^T\mathbf{res}^{(i)}| \leq 2 \times 10^{-3}$ for all i, and (41) is satisfied. The fact that $\mathbf{v}_1^T\mathbf{res}^{(i)}$ is likely to be small when a Krylov method is used as inexact solver was noted by [20].

Comparing (44) with (28), the corresponding bound for unpreconditioned solves with fixed tolerance, we see that quadratic convergence of the outer iteration is again attained, but the asymptotic constant may be larger than in the standard preconditioned case with consequently more outer iterations required for convergence.

Bound (42) also shows that choosing $\mathbf{res}^{(i)}$ to decrease like $|s^{(i)}|$ will normally not produce cubic convergence because of the presence of the $\|\top\!\!\top\,\mathbf{P}\mathbf{x}^{(i)}\|$ term in the numerator of the right hand side.

Numerical results illustrating this theory are given in the following Example.

Example 2.3 To show the advantage of (36) over (35) *Consider the generalised eigenvalue problem* $\mathbf{K}\mathbf{x}' = \lambda\mathbf{M}\mathbf{x}'$ *obtained from the matrix market matrices 'bcsstk09' and 'bcsstm09'. Here* $n = 1093$*,* $\mathbf{K}$ *and* $\mathbf{M}$ *are sparse, and* $\mathbf{M}$ *is a diagonal matrix with positive diagonal elements. This problem is reduced to a standard symmetric eigenvalue problem* $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$*, where* $\mathbf{A} = \mathbf{M}^{-1/2}\mathbf{K}\mathbf{M}^{-1/2}$ *and so* $\mathbf{A}$ *retains the sparsity structure of* $\mathbf{K}$*. In Table 2 we present computations of* $\lambda_{20} = 2.9 \times 10^9$ *the 20th smallest eigenvalue. The inner solve was carried out by preconditioned MINRES, with preconditioner taken to be an incomplete Cholesky decomposition of* $\mathbf{A}$ *using the matlab routine 'cholinc' with droptol=2e-3. We compare two methods both using a Rayleigh quotient shift and a fixed tolerance, so that both have quadratic convergence.*

(a) **RQIf***: Solve (35) using MINRES with stopping condition* $\|\mathbf{res}^{(i)}\| \leq \tau^{(0)} = 0.9$ *where* $\mathbf{res}^{(i)} = \mathbf{x}^{(i)} - (\mathbf{A} - \sigma^{(i)}\mathbf{I})\mathbf{y}^{(i)}$*, so Theorem 2.1 applies.*

(b) **SEf***: Solve (36) using MINRES with stopping condition* $\|\mathbf{res}^{(i)}\| \leq \tau^{(0)} \|\mathbf{P}\mathbf{x}^{(i)}\|$ *with* $\tau^{(0)} = 0.1$*, so Theorem 2.3 applies.*

In all runs we stopped the outer iteration when the relative eigenvalue residual satisfied $\|(\mathbf{A} - \varrho(\mathbf{x}^{(i)})\mathbf{I})\mathbf{x}^{(i)}\| \ / \ |\varrho(\mathbf{x}^{(i)})| \leq 10^{-9}$*. Table 2 presents the numerical results. Here we take a starting vector of the form* $\mathbf{x}^{(0)} = c^{(0)}\mathbf{v}_1 + s^{(0)}\mathbf{u}^{(0)}$ *where* $t^{(0)} = s^{(0)}/c^{(0)} = 0.001$*. Other starting guesses were taken but no qualitative differences in the results were observed.*

In Table 2 we observe that both methods exhibit quadratic convergence as predicted by the theory. Both took the same number of outer iterations, but

	RQIf $\tau^{(0)} = 0.9$			**SEf** $\tau^{(0)} = 0.1$		
	$\|\mathbf{r}^{(i)}\| \,/\, \|\rho^{(i)}\|$	$tan\theta^{(i)}$	$k^{(i-1)}$	$\|\mathbf{r}^{(i)}\| \,/\, \|\rho^{(i)}\|$	$tan\theta^{(i)}$	$k^{(i-1)}$
0	2.3e+00	1.0e-02		2.3e+00	1.0e-02	
1	1.1e+00	1.8e-03	56	1.0e-01	7.3e-04	77
2	6.5e-06	3.7e-09	102	1.3e-04	1.3e-08	84
3	8.5e-08	3.1e-10	124	9.5e-08	4.1e-10	65
$\sum k^{i-1}$			282			226

Table 2: Numerical results for Algorithm 2 using preconditioned MINRES applied to Example 2.3 using the methods in Theorem 2.1 and Theorem 2.3. The second and fifth columns give the respective relative eigenvalue residual.

SEf was more efficient in terms of the total number of inner iterations required. In fact, this is the key motivation for the modification of (35), namely that (36) is better suited for the application of the Krylov solver since it leads to reduced inner iteration counts. We explain this in detail in §3.

For theoretical interest only, we note that if $\mathbf{P} = \mathbf{A}$ in Algorithm 2 then (40) and (42) imply

$$t^{(i+1)} \quad \leq \quad \frac{|\lambda_1 - \sigma^{(i)}|}{|\lambda_2 - \sigma^{(i)}|} \, \frac{|s^{(i)}||\lambda_n| + \tau^{(i)}}{|c^{(i)}||\lambda_1| - \tau^{(i)}}. \tag{46}$$

Hence it is possible to recover a cubically convergent method if $\tau^{(i)}$ is choosen proportional to $|s^{(i)}|$ and this is obtained by preconditioning the shifted system (34) by $\mathbf{A}$.

The results in Table 2 show a significant reduction in the number of inner iterations taken by **SEf** compared to **RQIf**. The reason for this improvement is analysed in §3. However before doing this analysis we also note that the domain of convergence of the outer iteration may be reduced significantly when we decide to solve $(\mathbf{A} - \sigma^{(i)}\mathbf{I})\mathbf{y}^{(i)} = \mathbf{P}\mathbf{x}^{(i)}$ rather than $(\mathbf{A} - \sigma^{(i)}\mathbf{I})\mathbf{y}^{(i)} = \mathbf{x}^{(i)}$ as the following simple example shows.

Example 2.4 (Reduced domain of convergence using (36)) *Assume* $\mathbf{A}$ *is symmetric positive definite and let* $\mathbf{P} = \mathbf{A}$. *Take the factorisation* $\mathbf{P}_1 = \mathbf{A}^{\frac{1}{2}}$. *Then* $\mathbf{P}_1^{-1}(\mathbf{A} - \sigma^{(i)}\mathbf{I})\mathbf{P}_1^{-1}\tilde{\mathbf{y}}^{(i)} = \mathbf{P}_1^T\mathbf{x}^{(i)}$, *which is the preconditioned system (36), reduces to* $(\mathbf{I} - \sigma^{(i)}\mathbf{A}^{-1})\mathbf{y}^{(i)} = \mathbf{x}^{(i)}$. *Assume now* $\mathbf{x}^{(i)} = c^{(i)}\mathbf{v}_1 + s^{(i)}\mathbf{v}_2$ *and that exact solves are used to obtain* $\mathbf{y}^{(i)}$. *We readily obtain the one step bound*

$$\frac{t^{(i+1)}}{t^{(i)}} \quad \leq \quad \frac{|\lambda_1 - \sigma^{(i)}|}{|\lambda_2 - \sigma^{(i)}|} \, \frac{|\lambda_2|}{|\lambda_1|}.$$

Now the factor $|\lambda_2/\lambda_1|$ *may be large enough to significantly reduce the domain of convergence of the outer iteration. 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 prove in Section 3.3 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

Our aim in this section is to understand the performance of the inner iteration part of the inexact inverse iteration algorithm, and since $(\mathbf{A}-\sigma^{(i)}\mathbf{I})$ is symmetric but probably indefinite the natural Krylov method to solve $(\mathbf{A}-\sigma^{(i)}\mathbf{I})\mathbf{y}^{(i)} = \mathbf{x}^{(i)}$ is MINRES (see, for example, [7]).

First, we summarise some known results on MINRES in a form convenient for our use. In §3.1 we provide bounds on ratios of eigenvalues for two common preconditioners that are needed for §3.3. Then in §3.2 and §3.3 we provide an analysis of the number of iterations needed for unpreconditioned and preconditioned inner solves. Section 3.4 contains a discussion on 'a posteriori' bounds.

As discussed in the previous section the linear system will normally be solved iteratively using a preconditioner, $\mathbf{P}$ say, where $\mathbf{P}$ is positive definite and approximates $(\mathbf{A}-\sigma^{(i)}\mathbf{I})$ in some way. If $\mathbf{A}$ arises from a discretised partial differential equation $\mathbf{P}^{-1}$ may be constructed for example using a domain decomposition or multigrid technique. Alternatively $\mathbf{P}$ may be obtained using a Cholesky factorisation of $\mathbf{A}$ (or of a shifted $\mathbf{A}$). In this section we describe the theory for MINRES that is used to understand the inner iteration behaviour for both unpreconditioned and preconditioned solves in Algorithms 1 and 2. Specifically we derive a particular bound on the residual which will be used to provide bounds on the number of iterations needed for the inner solves.

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

$$\mathbf{B}\mathbf{z} = \mathbf{b}, \tag{47}$$

where $\mathbf{B}$ is a real symmetric $n \times n$ matrix. Define the Krylov space $\mathcal{K}_k(\mathbf{B}, \mathbf{b})$ by

$$\mathcal{K}_k(\mathbf{B}, \mathbf{b}) \quad = \quad \operatorname{span}\{\mathbf{b}, \mathbf{B}\mathbf{b}, \ldots, \mathbf{B}^{k-1}\mathbf{b}\}.$$

Throughout this paper we take an initial guess $\mathbf{z}_0 = \mathbf{0}$, though other choices are possible. MINRES seeks a solution $\mathbf{z}_k \in \mathcal{K}_k(\mathbf{B}, \mathbf{b})$ characterised by the property $\|\mathbf{b}-\mathbf{B}\mathbf{z}_k\|_2 = \min_{\mathbf{z}\in\mathcal{K}_k} \|\mathbf{b}-\mathbf{B}\mathbf{z}\|_2$.

Assume $\mathbf{B}$ has an eigenvalue decomposition of the form $\mathbf{B} = \mathbf{W}\mathbf{\Lambda}\mathbf{W}^T$, where $\mathbf{W}$ is orthogonal and $\mathbf{\Lambda} = \operatorname{diag}(\mu_1, \ldots, \mu_n)$. Then

$$\|\mathbf{b}-\mathbf{B}\mathbf{z}_k\|_2 \quad = \quad \min_{q\in P_k} \|q(\mathbf{B})\mathbf{b}\|_2$$

where P_k denotes the space of polynomials of degree k with $q(0) = 1$, and

$$\|\mathbf{b}-\mathbf{B}\mathbf{z}_k\|_2 \quad = \quad \min_{q\in P_k} \|q(\mathbf{\Lambda})\mathbf{W}^T\mathbf{b}\|_2, \tag{48}$$

from which a straightforward analysis (see, for example, p.54 of [7]) shows that

$$\|\mathbf{b} - \mathbf{B}\mathbf{z}_k\|_2 \quad \leq \quad 2\left(\sqrt{\frac{\kappa - 1}{\kappa + 1}}\right)^{k-1} \quad \|\mathbf{b}\|_2 \tag{49}$$

where $\kappa = \frac{\max_i |\mu_i|}{\min_i |\mu_i|}$. 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 (49).

Assume that the spectrum of $\mathbf{B}$ contains a small number of successive eigenvalues which are in some way distinguished from the remaining eigenvalues of $\mathbf{B}$. If $J \subset \mathbb{N}_n := \{1, 2, \dots n\}$ then $\{\mu_j\}_{j \in J}$ is the distinguished set and $|J|$ denotes the number of elements in it. Set $J^c := N_n - J$ and $\mathbf{Q}_J := \text{diag}\{\delta_1, \dots, \delta_n\}$ 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

$$p_J(t) \quad := \quad \prod_{j \in J} \frac{\mu_j - t}{\mu_j} \tag{50}$$

which vanishes for $t \in \{\mu_j\}_{j \in J}$. Clearly $q\, p_J \in \mathcal{P}_k$ for any $q \in \mathcal{P}_{k-|J|}$ and using the fact that $p_J(\mathbf{\Lambda}) = p_J(\mathbf{\Lambda})\mathbf{Q}_J$, (48) implies

$$\begin{aligned}
\|\mathbf{b} - \mathbf{B}\mathbf{z}_k\|_2 \quad &\leq \quad \min_{q \in \mathcal{P}_{k-|J|}} \|q(\mathbf{\Lambda})p_J(\mathbf{\Lambda})\mathbf{W}^T\mathbf{b}\|_2 \\
&= \quad \min_{q \in \mathcal{P}_{k-|J|}} \|q(\mathbf{\Lambda})p_J(\mathbf{\Lambda})\mathbf{Q}_J\mathbf{W}^T\mathbf{b}\|_2 \\
&\leq \quad \min_{q \in \mathcal{P}_{k-|J|}} \|q(\mathbf{\Lambda})p_J(\mathbf{\Lambda})\| \; \|\mathbf{Q}_J\mathbf{W}^T\mathbf{b}\|_2 \\
&= \quad \min_{q \in \mathcal{P}_{k-|J|}} \max_{j \in J^c} |q(\mu_j)p_J(\mu_j)| \|\mathbf{Q}_J\mathbf{W}^T\mathbf{b}\|_2 \\
&\leq \quad \left\{ \min_{q \in \mathcal{P}_{k-|J|}} \max_{j \in J^c} |q(\mu_j)| \right\} \max_{j \in J^c} |p_J(\mu_j)| \; \|\mathbf{Q}_J\mathbf{W}^T\mathbf{b}\|_2 \; .
\end{aligned} \tag{51}$$

Using (51) and standard results on Chebyshev polynomials (see, for example, [8, §3.1] or [9, §7.3.4]) we have the following theorem.

Theorem 3.1 *Suppose that the symmetric matrix* $\mathbf{B}$ *has eigenvalues* $\mu_1, \dots, \mu_n$ *with corresponding orthonormal eigenvectors* $\mathbf{w}_1, \dots, \mathbf{w}_n$. *Let* $\{\mu_j\}_{j \in J}$ *be* $|J|$ *successive eigenvalues of* $\mathbf{B}$ *and introduce the reduced condition number* $\kappa_J(\mathbf{B}) := \max_{j \in J^c} |\mu_j| / \min_{j \in J^c} |\mu_j|$. *With* $p_J(t)$ *and* $\mathbf{Q}_J$ *defined as above then*

$$\|\mathbf{b} - \mathbf{B}\mathbf{z}_k\|_2 \quad \leq \quad 2(\max_{j \in J^c} |p_J(\mu_j)|) \left\{ \frac{\sqrt{\kappa_J(\mathbf{B})} - 1}{\sqrt{\kappa_J(\mathbf{B})} + 1} \right\}^{k-|J|} \quad \|\mathbf{Q}_J\mathbf{W}^T\mathbf{b}\|_2$$

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

$$\|\mathbf{b} - \mathbf{B}\mathbf{z}_k\|_2 \quad \leq \quad 2(\max_{j \in J^c} |p_J(\mu_j)|) \left\{ \sqrt{\frac{\kappa_J(\mathbf{B}) - 1}{\kappa_J(\mathbf{B}) + 1}} \right\}^{k-|J|-1} \quad \|\mathbf{Q}_J\mathbf{W}^T\mathbf{b}\|_2$$

otherwise.

Note that the use of the matrix $\mathbf{Q}_J$ is nonstandard. It will play an important role in the analysis of the inner solves in Algorithm 2 later.

From now on we shall assume that the distinguished set of eigenvalues of $\mathbf{B}$ consists of a simple eigenvalue μ_1 so that $J = \{1\}$. This is the simplest case but is all that we need in the remainder of this paper. In this case we write $\mathbf{Q}_J = \mathbf{Q}_1 = \mathrm{diag}\{0, 1, \ldots, 1\}$, $\kappa_J(\mathbf{B}) = \kappa_1(\mathbf{B}) = \max_{j=2,\ldots,n} |\mu_j| \,/ \min_{j=2,\ldots,n} |\mu_j|$, and $p_1(t) = (\mu_1 - t)/\mu_1$. Also we define the two quantities

$$q_e \quad := \quad \frac{\sqrt{\kappa_1(\mathbf{B})} - 1}{\sqrt{\kappa_1(\mathbf{B})} + 1}, \tag{52}$$

$$\text{and} \qquad q_i \quad := \quad \sqrt{\frac{\kappa_1(\mathbf{B}) - 1}{\kappa_1(\mathbf{B}) + 1}}, \tag{53}$$

where q_e refers to the case where μ_1 is an extreme eigenvalue and thus $\mu_2, \ldots, \mu_n$ are of the same sign, and q_i covers all other situations. For this choice of J we have the following key result.

Corollary 3.1 *Let the assumptions of Theorem 3.1 hold with $J = \{1\}$. Then*

$$\|\mathbf{b} - \mathbf{B}\mathbf{z}_k\|_2 \quad \leq \quad 2\left(\max_{j=2,\ldots,n} \frac{|\mu_1 - \mu_j|}{|\mu_1|}\right)(q)^{k-\delta}\ \|\mathbf{Q}_1\mathbf{W}^T\mathbf{b}\|_2, \tag{54}$$

where $q = q_e, \delta = 1$ if μ_1 is an extreme eigenvalue and $\mu_2, \ldots, \mu_n$ have all the same sign, and $q = q_i, \delta = 2$ otherwise. In addition, if

$$k \geq \delta + \left(\log 2(\max_{j=2,\ldots,n} |\mu_1 - \mu_j|) + \log \frac{\|\mathbf{Q}_1\mathbf{W}^T\mathbf{b}\|}{|\mu_1|\tau}\right) / \log q^{-1} \tag{55}$$

then $\|\mathbf{b} - \mathbf{B}\mathbf{z}_k\| \leq \tau$.

Proof: Equation (54) follows by setting $J = \{1\}$ in Theorem 3.1 and using $\max_{j\in J^c} |p_J(\mu_j)| = \max_{j=2,\ldots,n} \dfrac{|\mu_1 - \mu_j|}{|\mu_1|}$. From (54) $\| \mathbf{b} - \mathbf{B}\mathbf{z}_k \| \leq \tau$ will be satisfied provided k is such that

$$2\left(\max_{j=2,\ldots,n} |\mu_1 - \mu_j|\right) \frac{\|\mathbf{Q}_1\mathbf{W}^T\mathbf{b}\|}{|\mu_1|\ \tau}(q)^{k-\delta} \leq 1,$$

and the last result follows by taking logs. □

Note that if $q = q_e$ then the residual reduction indicated by (54) is the same as that achieved by the Conjugate Gradient method applied to a positive definite symmetric matrix.

We shall use bound (55) in the following subsections to help understand the behaviour of MINRES inner iterations in inexact inverse iteration. To do this we require bounds on $\|\mathbf{Q}_1\mathbf{W}^T\mathbf{b}\|$ for the three main choices for $\mathbf{B}$ and $\mathbf{b}$. This is done in the following Lemma.

Lemma 3.1 *Let* $\mathbf{W}$ *be the orthogonal matrix whose columns are orthonormal eigenvectors of* $\mathbf{B}$ *and let* $\mathbf{x}^{(i)}$ *be as in (6).*

(a) (No Preconditioning) With $\mathbf{B} = (\mathbf{A} - \sigma^{(i)}\mathbf{I}),\ \mathbf{b} = \mathbf{x}^{(i)}$ *then*

$$\|\mathbf{Q}_1\mathbf{W}^T\mathbf{b}\| = \|\mathbf{Q}_1\mathbf{W}^T\mathbf{x}^{(i)}\| = |s^{(i)}|. \tag{56}$$

(b) (Standard Preconditioning) With $\mathbf{B} = \mathbf{P}_1^{-1}(\mathbf{A} - \sigma^{(i)}\mathbf{I})\mathbf{P}_1^{-T},\ \mathbf{b} = \mathbf{P}_1^{-1}\mathbf{x}^{(i)}$ *then*

$$\|\mathbf{Q}_1\mathbf{W}^T\mathbf{b}\| = \|\mathbf{Q}_1\mathbf{W}^T\mathbf{P}_1^{-1}\mathbf{x}^{(i)}\| \leq \|\mathbf{P}_1^{-1}\|\ . \tag{57}$$

(c) (Simoncini-Elden Preconditioning) With $\mathbf{B} = \mathbf{P}_1^{-1}(\mathbf{A} - \sigma^{(i)}\mathbf{I})\mathbf{P}_1^{-T},\ \mathbf{b} = \mathbf{P}_1^T\mathbf{x}^{(i)}$ *and* $\sigma^{(i)}$ *chosen to be the Rayleigh quotient then*

$$\|\mathbf{Q}_1\mathbf{W}^T\mathbf{b}\| = \|\mathbf{Q}_1\mathbf{W}^T\mathbf{P}_1^T\mathbf{x}^{(i)}\| \leq\ C'|s^{(i)}| \tag{58}$$

where C' *is a positive constant independent of* i.

Proof: For (a), note that $\mathbf{W}$ is merely the matrix of eigenvectors of $\mathbf{A}$ and so

$$\mathbf{Q}_1\mathbf{W}^T\mathbf{x}^{(i)} = s^{(i)}(0, \mathbf{v}_2^T\mathbf{u}^{(i)}, \ldots, \mathbf{v}_n^T\mathbf{u}^{(i)})$$

which gives that $\|\mathbf{Q}_1\mathbf{W}^T\mathbf{x}^{(i)}\| = |s^{(i)}|$, since $\|\mathbf{u}^{(i)}\| = 1$. The proof of part (b) is straightforward. To prove (c) write $\mathbf{B}$ in the form

$$\mathbf{B} = \mathbf{P}_1^{-1}(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{P}_1^{-T} + (\lambda_1 - \sigma^{(i)})\mathbf{P}_1^{-1}\mathbf{P}_1^{-T}$$

where we note that $\mathbf{P}_1^{-1}(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{P}_1^{-T}(\mathbf{P}_1^T\mathbf{v}_1) = \mathbf{0}$. Standard perturbation theory for simple eigenvalues of symmetric matrices shows that $\mathbf{B}$ has a simple eigenvalue μ_1 say near 0 with corresponding eigenvector $\mathbf{w}_1$ near $\mathbf{P}_1^T\mathbf{v}_1$. In fact

$$\|\mathbf{P}_1^T\mathbf{v}_1 - \mathbf{w}_1\| \leq\ C|\lambda_1 - \sigma^{(i)}|, \tag{59}$$

for some C independent of i. Thus,

$$\begin{aligned}\mathbf{Q}_1\mathbf{W}^T\mathbf{b} &= \mathbf{Q}_1\mathbf{W}^T\mathbf{P}_1^T\mathbf{x}^{(i)}\ =\ c^{(i)}\mathbf{Q}_1\mathbf{W}^T\mathbf{P}_1^T\mathbf{v}_1 + s^{(i)}\mathbf{Q}_1\mathbf{W}^T\mathbf{P}_1\mathbf{u}^{(i)} \\ &= c^{(i)}\mathbf{Q}_1\mathbf{W}^T(\mathbf{w}_1 + (\mathbf{P}_1^T\mathbf{v}_1 - \mathbf{w}_1)) + s^{(i)}\mathbf{Q}_1\mathbf{W}^T\mathbf{P}_1\mathbf{u}^{(i)}.\end{aligned}$$

Now we see the importance of the $\mathbf{Q}_1$ matrix, which in this case is $\mathrm{diag}\{0, 1, \ldots, 1\}$. Since $\mathbf{W}^T\mathbf{w}_1 = \mathbf{e}_1$ we have that $\mathbf{Q}_1\mathbf{W}^T\mathbf{w}_1 = \mathbf{0}$, and so we immediately obtain

$$\|\mathbf{Q}_1\mathbf{W}^T\mathbf{b}\| \quad\leq\quad |\lambda_1 - \sigma^{(i)}|C + s^{(i)}\ \|\mathbf{P}_1\| \leq C's^{(i)}, \tag{60}$$

for some positive constants C and C' independent of i, since $\sigma^{(i)}$ is the Rayleigh quotient. □

3.1 Eigenvalue bounds

In the analysis of §3.3 we shall assume bounds for $|\lambda_1 - \sigma^{(i)}| \,/\, |\mu_1|$, where μ_1 is the smallest eigenvalue (in modulus) of $\mathbf{B} = \mathbf{P}_1^{-1}(\mathbf{A} - \sigma^{(i)}\mathbf{I})\mathbf{P}_1^{-T}$. 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) *If* $\mathbf{A}$ *is a symmetric positive definite matrix arising from discretization of an elliptic PDE and a symmetric positive definite preconditioner* $\mathbf{P}^{-1}$ *is constructed using domain decomposition methods then one typically has a bound on the condition number of* $\mathbf{P}^{-1}\mathbf{A}$*. Thus if we denote the eigenvalues of any matrix* $\mathbf{M}$ *by* $\lambda_j(\mathbf{M})$ *we assume*

$$\gamma_L \;\leq\; \lambda_j(\mathbf{P}^{-1}\mathbf{A}) \;\leq\; \gamma_U$$

for some positive constants γ_L, γ_U*. Now with* $\mathbf{B} = \mathbf{P}_1^{-1}(\mathbf{A} - \sigma^{(i)}\mathbf{I})\mathbf{P}_1^{-T}$ *we have*

$$\lambda_j(\mathbf{B}) \;\;=\;\; \lambda_j(\mathbf{P}_1^{-1}\mathbf{A}^{\frac{1}{2}}(\mathbf{I} - \sigma^{(i)}\mathbf{A}^{-1})\mathbf{A}^{\frac{1}{2}}\mathbf{P}_1^{-T})$$

and Sylvester's Inertia Theorem can be used to provide bounds on $\lambda_j(\mathbf{B})$*. For example, if* $\lambda_1 < \sigma^{(i)} < \lambda_2$ *then with* $\mu_1 = \lambda_1(\mathbf{B})$ *we have* $\mu_1 < 0$ *and*

$$(1 - \frac{\sigma^{(i)}}{\lambda_1})\lambda_n(\mathbf{P}^{-1}\mathbf{A}) \;\leq\; \mu_1 \;\leq\; (1 - \frac{\sigma^{(i)}}{\lambda_1})\lambda_1(\mathbf{P}^{-1}\mathbf{A}),$$

using Sylvester's Inertia Theorem, so that

$$\frac{\lambda_1}{\gamma_U} \;\leq\; \frac{\lambda_1}{\lambda_n(\mathbf{P}^{-1}\mathbf{A})} \;\leq\; \frac{|\lambda_1 - \sigma^{(i)}|}{|\mu_1|} \;\leq\; \frac{\lambda_1}{\lambda_1(\mathbf{P}^{-1}\mathbf{A})} \;\leq\; \frac{\lambda_1}{\gamma_L}. \tag{61}$$

Example 3.2 (Cholesky Preconditioners) *If* $\mathbf{A}$ *is symmetric positive definite, and an incomplete Cholesky factorisation of* $\mathbf{A}$ *is used to find* $\mathbf{P}$*, i.e.* $\mathbf{A} = \mathbf{P}_1\mathbf{P}_1^T + \mathbf{E}$ *with* $\mathbf{E}$ *'small', say* $\|\mathbf{E}\| < \lambda_1$*. Then using ideas in [20] we write*

$$\begin{aligned} \mathbf{P}_1^{-1}(\mathbf{A} - \sigma^{(i)}\mathbf{I})\mathbf{P}_1^{-T}\mathbf{w}_1 \;&=\; \mu_1\mathbf{w}_1, \\ \textit{as} \qquad (\mathbf{A} - \sigma^{(i)}\mathbf{I})\widetilde{\mathbf{w}} = \mu_1\mathbf{P}_1\mathbf{P}_1^T\widetilde{\mathbf{w}} \;&=\; \mu_1(\mathbf{A} - \mathbf{E})\widetilde{\mathbf{w}}, \end{aligned}$$

where $\widetilde{\mathbf{w}} = \mathbf{P}_1^{-T}\mathbf{w}_1$*. So*

$$(\mathbf{A} + \frac{\mu_1}{1 - \mu_1}\mathbf{E})\widetilde{\mathbf{w}} \;\;=\;\; \frac{\sigma^{(i)}}{1 - \mu_1}\widetilde{\mathbf{w}}.$$

Now comparing with $\mathbf{A}\mathbf{v}_1 = \lambda_1\mathbf{v}_1$ *the Bauer-Fike Theorem gives*

$$\left|\frac{\sigma^{(i)}}{1 - \mu_1} - \lambda_1\right| \;\;\leq\;\; \left|\frac{\mu_1}{1 - \mu_1}\right| \|\mathbf{E}\|$$

and hence, since $\lambda_1 > 0$ *and* $\|\mathbf{E}\| < \lambda_1$ *by the assumptions above,*

$$\lambda_1 - \|\mathbf{E}\| \leq \frac{|\lambda_1 - \sigma^{(i)}|}{|\mu_1|} \leq \lambda_1 + \|\mathbf{E}\| \, . \tag{62}$$

Thus in both examples considered here we can say that, under suitable assumptions,

$$|\lambda_1| \, C' \leq \frac{|\lambda_1 - \sigma^{(i)}|}{|\mu_1|} \leq |\lambda_1| \, C \tag{63}$$

for some positive constants C' and C independent of i.

3.2 Inner Iterations for Unpreconditioned MINRES

Consider now the use of (55) to help understand the behaviour of MINRES in inexact inverse iteration. Let us consider in detail the case where λ_1 is an extremal, well-separated eigenvalue of $\mathbf{A}$, so that λ_1 satisfies $\lambda_1 < \lambda_2 \leq \ldots \leq \lambda_n$, or $\lambda_n \leq \lambda_{n-1} \leq \ldots \leq \lambda_2 < \lambda_1$. Assume 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 take $\mathbf{B} = (\mathbf{A} - \sigma^{(i)}\mathbf{I}),\ \mathbf{b} = \mathbf{x}^{(i)}$, and

$$\mathbf{res}^{(i)}_{k^{(i)}} := \mathbf{x}^{(i)} - (\mathbf{A} - \sigma^{(i)}\mathbf{I})\mathbf{y}^{(i)}_{k^{(i)}}, \tag{64}$$

where $\mathbf{y}^{(i)}_{k^{(i)}}$ denotes the $k^{(i)}$th iterate of MINRES. Using case (a) in Lemma 3.1, and $\delta = 1$ in (55) we find that to achieve (12) it is sufficient for $k^{(i)}$ to satisfy

$$k^{(i)} \quad \geq \quad 1 + \left\{ \log 2|\lambda_1 - \lambda_n| + \log \frac{|s^{(i)}|}{|\lambda_1 - \sigma^{(i)}|\tau^{(i)}} \right\} / \log q_e^{-1}. \tag{65}$$

The theory in Section 2.3 assumed the upper bounds $\tau^{(i)} \leq C_2|s^{(i)}|^\beta$ and $|\lambda_1 - \sigma^{(i)}| \leq C_1|s^{(i)}|^\alpha$. Now, in addition, assume the two sided bounds

$$C_1'|s^{(i)}|^\alpha \leq |\lambda_1 - \sigma^{(i)}| \leq C_1|s^{(i)}|^\alpha, \tag{66}$$

and

$$C_2'|s^{(i)}|^\beta \leq \tau^{(i)} \leq C_2|s^{(i)}|^\beta, \tag{67}$$

where the constants C_1', C_1, C_2' and C_2 are positive and independent of i. Note that assumptions (66) and (67) are reasonable. First, the bounds on $\tau^{(i)}$ are seen to be satisfied for a constant tolerance ($\beta = 0$) or a decreasing tolerance ($\beta = 1$) with $\tau^{(i)}$ proportional to the norm of the eigenvalue residual (9) and assuming (4) (see also, Lemma 2.6 in [22]). Second, the lower bound in (66) is satisfied for a Rayleigh quotient shift $\sigma^{(i)} = \varrho(\mathbf{x}^{(i)})$, since (8) then shows that $|\lambda_1 - \sigma^{(i)}| = (s^{(i)})^2|\lambda_2 - \varrho(\mathbf{u}^{(i)})|$, and the lower bound is given by Lemma 2.7 of [22].

Using the lower bounds for $\tau^{(i)}$ and $|\lambda_1 - \sigma^{(i)}|$ in (66) and (67) we see that (12) is satisfied if

$$k^{(i)} \quad \geq \quad 1 + \left\{ \log \frac{2|\lambda_1 - \lambda_n|}{C_1' C_2'} + (\alpha + \beta - 1) \log(|s^{(i)}|^{-1}) \right\} / \log q_e^{-1}. \quad (68)$$

We have thus proved the following Lemma.

Lemma 3.2 *(a) Assume Algorithm 1 is used to compute an extreme eigenvalue of* $\mathbf{A}$ *using unpreconditioned MINRES for the inexact solves. Assume in addition that the lower bounds given by (66) and (67) hold. If*

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

then $\mathbf{res}^{(i)}$ *defined by (64) satisfies* $\| \mathbf{res}^{(i)} \| \leq \tau^{(i)}$ *and Algorithm 1 converges with a rate predicted by Theorem 2.2.*

(b) In particular, for convergence to occur in the inexact Rayleigh quotient methods of Theorem 2.1 we see that for a decreasing tolerance (case (a)*) it is sufficient that*

$$k^{(i)} \quad \geq \quad 1 + \left\{ \log C + 2 \log(|s^{(i)}|^{-1}) \right\} / \log(q_e^{-1}) \qquad (69)$$

for some C *independent of* i*. (Here* $\alpha = 2, \beta = 1$*). Moreover for a fixed tolerance (case* (b)*) it is sufficient that*

$$k^{(i)} \quad \geq \quad 1 + \left\{ \log C + \log(|s^{(i)}|^{-1}) \right\} / \log(q_e^{-1}) \qquad (70)$$

for some C *independent of* i*. (Here* $\alpha = 2, \beta = 0$*).*

We note that if $\alpha + \beta = 1$ (i.e. linear convergence in Theorem 2.2) then the bound on $k^{(i)}$ does not grow. This is confirmed by numerical results in Example 3.3. However, for **RQId** and **RQIf**, Lemma 3.2, case (b), predicts an increase in the number of inner iterations required for each outer iteration, with **RQId** being more expensive than **RQIf**. This behaviour is indeed observed in Table 1. Significantly, we see that the effect of letting $\sigma^{(i)}$ converge quadratically to λ_1 produces only logarithmic growth in the number of inner iterations needed to achieve convergence, so that we need not be concerned that the Krylov solver is applied to a matrix which is becoming more and more singular. The explanation lies in the interplay between the choices of shift, tolerance and right hand side. For unpreconditioned solves the term $\log \dfrac{\|\mathbf{Q}_1 \mathbf{W}^T \mathbf{b}\|}{|\mu_1| \tau}$ in the bound for k given by (55) produces the $\log \dfrac{|s^{(i)}|}{|\lambda_1 - \sigma^{(i)}| \tau^{(i)}}$ term in (65), which provides nothing worse than logarithmic growth in $|s^{(i)}|^{-1}$.

It is well known that the bound given by (54) is, at best, descriptive and is unlikely to be sharp. As a consequence (68) is unlikely to provide a realistic estimate for $k^{(i)}$ and should only be used in a 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, especially when applied to discretizations of PDEs.

For any well-separated interior eigenvalue λ_1 the bound (65) still holds provided $\sigma^{(i)} \neq \lambda_1$. The bounds on $\tau^{(i)}$ given by (67) also hold. However for Rayleigh quotient shifts approximating an interior eigenvalue, the discussion above Lemma 2.7 in [22] shows that for a special choice of $\mathbf{u}^{(i)}$ it is possible to make $\varrho(\mathbf{u}^{(i)}) = \lambda_1$, and in this instance $C_1' = 0$ in (66). Nonetheless, the likelihood of this happening is extremely small and this was never observed in any of our experiments. Under the assumption that the left hand bound of (66) holds in practice, then for a well-separated interior eigenvalue a bound similar to (68) is obtained with the main difference that the factor q_e is replaced by q_i given by (53), and comments about the likely increase in the number of inner iterations remain the same as for the extreme eigenvalue case.

3.3 Inner Iterations for Preconditioned MINRES

In this subsection we give an analysis of the number of inner iterations for the two preconditioned methods using preconditioned MINRES with Rayleigh quotient shifts and a fixed tolerance. (The two methods are covered by Theorems 2.1 and 2.3). The main result is given in the following Lemma.

Lemma 3.3 *(a) (Standard Preconditioning) Assume Algorithm 1 with Rayleigh quotient shifts and a fixed tolerance $\tau^{(i)} = \tau^{(0)}$ is used to compute an extreme eigenvalue of $\mathbf{A}$ using preconditioned MINRES for the inexact solves. Additionally, assume $|(\lambda_1 - \sigma^{(i)})/\mu_1|$ satisfies (63) and that the bounds (66) and (67) hold. If*

$$k^{(i)} \quad \geq \quad 1 + \{\log C + 2\log(|s^{(i)}|^{-1}\})/\log q_e^{-1} \tag{71}$$

where C is a known constant independent of i, then the outer iteration converges quadratically.

(b) (Simoncini-Eldén Preconditioning) Assume Algorithm 2 with Rayleigh quotient shifts and a fixed tolerance $\tau^{(i)} = \tau^{(0)}$ is used to compute an extreme eigenvalue of $\mathbf{A}$ using preconditioned MINRES for the inexact solves. Additionally, assume $|(\lambda_1 - \sigma^{(i)})/\mu_1|$ satisfies (63) and that the bounds (66) and (67) hold. If

$$k^{(i)} \quad \geq \quad 1 + \{\log C' + \log(|s^{(i)}|^{-1}\})/\log q_e^{-1} \tag{72}$$

where C' is a known constant independent of i then the outer iteration converges quadratically.

Proof: The standard preconditioned case is given by (in the notation of (54) and (55))

$$\mathbf{z}_k = \mathbf{P}_1^T\mathbf{y}^{(i)}, \quad \mathbf{B} = \mathbf{P}_1^{-1}(\mathbf{A} - \sigma^{(i)}\mathbf{I})\mathbf{P}_1^{-T}, \quad \mathbf{b} = \mathbf{P}_1^{-1}\mathbf{x}^{(i)}, \tag{73}$$

with, from (57), $\|\mathbf{Q}_1\mathbf{W}^T\mathbf{b}\|\leq\|\mathbf{P}_1^{-1}\|$, and in the Simoncini-Eldén preconditioned form given in Algorithm 2 we have

$$\mathbf{z}_k = \mathbf{P}_1^T\mathbf{y}^{(i)}, \quad \mathbf{B} = \mathbf{P}_1^{-1}(\mathbf{A}-\sigma^{(i)}\mathbf{I})\mathbf{P}_1^{-T}, \quad \mathbf{b} = \mathbf{P}_1^T\mathbf{x}^{(i)},$$

with, from (58), $\|\,\mathbf{Q}_1\mathbf{W}^T\mathbf{b}\,\|\leq C|s^{(i)}|$, for some constant C independent of i. Here we see that the Simoncini-Elden preconditioning retains the $|s^{(i)}|$ factor in the bound on $\|\mathbf{Q}_1\mathbf{W}^T\mathbf{b}\|$. This turns out to be the main difference between the methods and will explain the improved inner iteration performance of the Simoncini-Elden preconditioning over standard preconditioning.

Proof of (a). For the standard preconditioned case we stop Algorithm 1 using the residual of the unpreconditioned system not the residual of $\mathbf{B}\mathbf{z}_{k^{(i)}} = \mathbf{b}$. Now

$$\|(\mathbf{A}-\sigma^{(i)}\mathbf{I})\mathbf{y}^{(i)}-\mathbf{x}^{(i)}\|=\|\mathbf{P}_1(\mathbf{B}\mathbf{z}_{k^{(i)}}-\mathbf{b})\|\leq\|\mathbf{P}_1\|\|\mathbf{B}\mathbf{z}_{k^{(i)}}-\mathbf{b}\|, \tag{74}$$

and so, if $k^{(i)}$ is such that

$$\|\mathbf{B}\mathbf{z}_{k^{(i)}}-\mathbf{b}\|\leq\tau^{(i)}\ \|\mathbf{P}_1\|^{-1}, \tag{75}$$

then we satisfy

$$\|(\mathbf{A}-\sigma^{(i)}\mathbf{I})\mathbf{y}^{(i)}-\mathbf{x}^{(i)}\|\leq\tau^{(i)}. \tag{76}$$

Using (54), (55) and (57), condition (75) holds if $k^{(i)}$ satisfies,

$$k^{(i)} \geq 1 + \left\{\log\left(2\max_{j=2,\ldots,n}|\mu_1-\mu_j|\kappa(\mathbf{P}_1)\right)+\log\frac{1}{|\mu_1|\ \tau^{(i)}}\right\}/\log q_e^{-1}. \tag{77}$$

or, equivalently,

$$k^{(i)} \geq 1 + \left\{\log\left(2\max_{j=2,\ldots,n}|\mu_1-\mu_j|\kappa(\mathbf{P}_1)\frac{|\lambda_1-\sigma^{(i)}|}{|\mu_1|}\right)+\log\frac{1}{|\lambda_1-\sigma^{(i)}|\ \tau^{(i)}}\right\}/\log q_e^{-1}. \tag{78}$$

Here q_e is given by (52). The quantity $|\lambda_1-\sigma^{(i)}|/|\mu_1|$ relates the eigenvalue nearest zero of $(\mathbf{A}-\sigma^{(i)}\mathbf{I})$ to the value of the corresponding eigenvalue of $\mathbf{P}_1^{-1}(\mathbf{A}-\sigma^{(i)}\mathbf{I})\mathbf{P}_1^{-T}$. We shall assume the bound (63) for this ratio as is discussed in detail in §3.1. Note that $\tau^{(i)}=\tau^{(0)}$, a constant, and we no longer have a factor of $|s^{(i)}|$ in the numerator of the second log term in (78) compared to (65). Using (66) this log term produces the $2\log|s^{(i)}|^{-1}$ term in (71).

Proof of (b). For the Simoncini-Eldén preconditioned approach in Algorithm 2 the analysis proceeds similarly. To achieve

$$\|(\mathbf{A}-\sigma^{(i)}\mathbf{I})\mathbf{y}^{(i)}-\mathbf{P}\mathbf{x}^{(i)}\|\leq\tau^{(i)}\ \|\mathbf{P}\mathbf{x}^{(i)}\| \tag{79}$$

it is sufficient that $\|\mathbf{P}_1(\mathbf{P}_1^{-1}(\mathbf{A}-\sigma^{(i)}\mathbf{I})\mathbf{P}_1^T\widetilde{\mathbf{y}}^{(i)}-\mathbf{P}_1^T\mathbf{x}^{(i)})\|\leq\tau^{(i)}\ \|\mathbf{P}\mathbf{x}^{(i)}\|$, where $\widetilde{\mathbf{y}}^{(i)}=\mathbf{P}_1^{-T}\mathbf{y}^{(i)}$. In the notation of Lemma 3.1, (c), condition (79) will hold if $\|\mathbf{P}_1\|\|\mathbf{B}\mathbf{z}_k-\mathbf{b}\|\leq\tau^{(i)}\ \|\mathbf{P}\mathbf{x}^{(i)}\|$ and this is achieved provided

$$k^{(i)} \geq 1 + \left\{\log\left(2\max_{j=2,\ldots,n}|\mu_1-\mu_j|\frac{|\lambda_1-\sigma^{(i)}|\ \|\mathbf{P}_1\|}{|\mu_1|\ \|\mathbf{P}\mathbf{x}^{(i)}\|}\right)+\log\frac{|s^{(i)}|}{|\lambda_1-\sigma^{(i)}|\ \tau^{(i)}}\right\}/\log q_e^{-1}. \tag{80}$$

where the second log term now has $|s^{(i)}|$ in the numerator and has the same form as the second term in the unpreconditioned inequality (65). Using (66) this term produces the $\log |s^{(i)}|^{-1}$ term in (72). $\square$

Because of the different factors multiplying the $\log |s^{(i)}|^{-1}$ terms in (71) and (72) we expect that Algorithm 2, which uses the modified right hand side, may be less expensive in terms of the total number of inner iterations compared with preconditioned Algorithm 1. This is indeed observed in Table 2, where we compare the number of inner iterations needed by methods **RQIf** (Lemma 3.3, case (a)) and **SEf** (Lemma 3.3, case (b)) applied to the matrix in Example 2.3 to compute the interior eigenvalue. We see that both methods exhibit an increase in the number of inner iterations as the outer iteration proceeds as predicted by (71) and (72), but **RQIf** is more expensive than **SEf** with regard to the total number of inner iterations to achieve comparable accuracy which can be explained by the multiplier 2 in front of the second log term in (71), compared with the multiplier 1 in (72).

We have presented the theory of preconditioned solves for extremal values, but remarks similar to those in the last paragraph of the previous section also hold. For example, (78), with q_e replaced by q_i, describes the behaviour of MINRES as experienced in practice for any well separated eigenvalue.

3.4 'A posteriori' bounds for preconditioned MINRES

The discussion in the previous subsection is a natural consequence of the inequality (55) on the number of inner iterations in MINRES. In this subsection we look at an alternative approach that extracts additional information available from (78) and (80) by combining the MINRES theory with the inexact inverse iteration convergence theory of §2. Note that the theory in this section does not rely on $\sigma^{(i)}$ being the Rayleigh quotient.

In the previous sections we considered 'a priori' information about the cost of an inner solve. Here we will use 'a posteriori' information to obtain an upper bound on the overall cost of the inner solves of a convergent method. To this end we define $k_M^{(i)}$ to be the actual number of inner iterations used by MINRES at step i, that is,

$$\|\mathbf{res}^{(i)}_{k_M^{(i)}}\| \leq \tau^{(i)}, \quad \|\mathbf{res}^{(i)}_k\| > \tau^{(i)}, \ \forall k < k_M^{(i)}. \tag{81}$$

Now the inequalities in §3.2 and §3.3 produce sufficient conditions on the number of inner steps needed to ensure a residual tolerance is satisfied. Equally well, however, these sufficient conditions provide *upper* bounds on the actual number of inner iterations needed. This is seen in the following consequence of Corollary 3.1.

Corollary 3.2 *Let the assumptions of Corollary 3.1 hold and assume μ_1 is an extreme eigenvalue. Define $k_*^{(i)}$ by*

$$k_*^{(i)} := 1 + \left(\log(2 \max_{j=2,\dots,n} |\mu_1 - \mu_j|) + \log \frac{\|\mathbf{Q}_1 \mathbf{W}^T \mathbf{b}^{(i)}\|}{|\mu_1|\tau} \right) / \log q_e^{-1}, \tag{82}$$

and let $k_M^{(i)}$ denote the actual number of inner iterations used by MINRES to solve $\mathbf{Bz} = \mathbf{b}^{(i)}$ to a tolerance τ. Then

$$k_M^{(i)} \leq k_*^{(i)} + 1. \tag{83}$$

Proof: The quantity $k_*^{(i)}$ may not be an integer and so one needs to add 1 to ensure an upper bound on the integer $k_M^{(i)}$. □

The idea in this subsection is to use the right hand side of (14) to link $k_M^{(i)}$ with $t^{(i+1)}$ as follows. Assume the inexact inverse iteration algorithm under consideration converges with the second equation in (66) with $\beta = 1$ satisfied or with $\tau^{(i)}$ fixed. Clearly $\tau^{(i)} \neq 0$ from (66) so we can say

$$\frac{|s^{(i)}| + \tau^{(i)}}{|c^{(i)}| - \tau^{(i)}} \leq C_5 \tau^{(i)} \tag{84}$$

for some C_5 independent of i. Thus (14) gives

$$t^{(i+1)} \leq \frac{|\lambda_1 - \sigma^{(i)}|}{|\lambda_2 - \sigma^{(i)}|}\, C_5 \tau^{(i)}, \tag{85}$$

and hence

$$\frac{1}{|\lambda_1 - \sigma^{(i)}|\tau^{(i)}} \leq \frac{C_5}{|\lambda_2 - \sigma^{(i)}|\; t^{(i+1)}}. \tag{86}$$

Hence, taking logs we have

$$\log \frac{1}{|\lambda_1 - \sigma^{(i)}|\tau^{(i)}} \leq \log \frac{C_5}{|\lambda_2 - \sigma^{(i)}|} + \log \frac{1}{t^{(i+1)}}. \tag{87}$$

Now for standard preconditioned solves $\|\mathbf{Q}_1\mathbf{W}^T\mathbf{b}\| \leq \|\mathbf{P}_1^{-1}\|$ and $\mathbf{Bz} = \mathbf{b}^{(i)}$ is solved to a tolerance $\tau = \tau^{(i)}\ \|\mathbf{P}_1\|^{-1}$ (cf. the proof of Lemma 3.3(a)). Thus we can bound the second log term in (82) as follows.

$$\begin{aligned} \log \frac{\|\mathbf{Q}_1\mathbf{W}^T\mathbf{b}^{(i)}\|}{|\mu_1|\tau} &= \log \frac{\|\mathbf{Q}_1\mathbf{W}^T\mathbf{b}^{(i)}\|}{|\mu_1|\tau^{(i)}\ \|\mathbf{P}_1\|^{-1}} \\ &\leq \log \|\mathbf{P}_1^{-1}\|\|\mathbf{P}_1\| + \log \frac{1}{|\mu_1|\tau^{(i)}}. \end{aligned}$$

Thus using (82) and (83) we have

$$k_M^{(i)} \leq 2 + \left(\log(2 \max_{j=2,\ldots,n} |\mu_1 - \mu_j|\ \kappa(\mathbf{P}_1)\frac{|\lambda_1 - \sigma^{(i)}|}{|\mu_1|}) + \log \frac{1}{|\lambda_1 - \sigma^{(i)}|\tau^{(i)}}\right) / \log q_e^{-1} \tag{88}$$

and hence, using (86),

$$k_M^{(i)} \leq 2 + (\log C + \log \frac{1}{t^{(i+1)}})/\log q_e^{-1} \tag{89}$$

where

$$C = 2 \max_{j=2,\dots,n} |\mu_1 - \mu_j| \; \kappa(\mathbf{P}_1) \; \frac{|\lambda_1 - \sigma^{(i)}|}{\mu_1} \; \frac{C_5}{|\lambda_2 - \sigma^{(i)}|}. \tag{90}$$

Under assumptions (4) and (63), C is bounded independent of i.

Now consider Algorithm 2 with a fixed tolerance so that (42) holds with $\|\mathbf{res}^{(i)}\| \leq \tau^{(0)} \; \|\mathbf{P}\mathbf{x}^{(i)}\|$. Assuming (41) and (43) also hold and that the method is convergent, then we may write

$$t^{(i+1)} \quad \leq \quad \frac{|\lambda_1 - \sigma^{(i)}|}{|\lambda_2 - \sigma^{(i)}|} C$$

for some C independent of i using (42). Reasoning as in the standard preconditioned case we obtain a similar bound for $k_M^{(i)}$ but with the key difference that the $\|\mathbf{Q}_1 \mathbf{W}^T \mathbf{b}^{(i)}\|$ term leads to a $\log \frac{t^{(i)}}{t^{(i+1)}}$ term rather than the $\log \frac{1}{t^{(i+1)}}$ term in (89), and this provides a simplification as is shown in the following theorem.

Theorem 3.2 *(a) (Standard Preconditioning) Assume Algorithm 1 converges, where the system* $(\mathbf{A} - \sigma^{(i)}\mathbf{I})\mathbf{y}^{(i)} = \mathbf{x}^{(i)}$ *is solved by preconditioned MINRES. Let* $k_M^{(i)}$ *denote the actual number of inner iterations used by MINRES at step* i*. Then*

$$k_M^{(i)} \quad \leq \quad 2 + (\log C + \log \frac{1}{t^{(i+1)}}) / \log(q_e^{-1}) \tag{91}$$

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

$$\sum_{i=0}^{\mathcal{N}-1} k_M^{(i)} \quad \leq \quad 2\mathcal{N} + \Big[\mathcal{N} \log C + \sum_{i=0}^{\mathcal{N}-1} \log \frac{1}{t^{(i+1)}}\Big] / \log(q_e^{-1}). \tag{92}$$

.

(b) (Simoncini-Elden Preconditioning) Assume Algorithm 2 with a fixed tolerance satisfying (43) converges, where the system $(\mathbf{A} - \sigma^{(i)}\mathbf{I})\mathbf{y}^{(i)} = \mathbf{P}\mathbf{x}^{(i)}$ *is solved by preconditioned MINRES with Rayleigh quotient shifts. Let* $k_M^{(i)}$ *denote the actual number of inner iterations used by MINRES at step* i *so that* $\|(\mathbf{A} - \sigma^{(i)}\mathbf{I})\mathbf{y}^{(i)} - \mathbf{P}\mathbf{x}^{(i)}\| \leq \tau^{(i)} \; \|\mathbf{P}\mathbf{x}^{(i)}\|$ *holds. Then*

$$k_M^{(i)} \quad \leq \quad 1 + (\log C + \log \frac{t^{(i)}}{t^{(i+1)}}) / \log(q_e^{-1}) \tag{93}$$

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

$$\sum_{i=0}^{\mathcal{N}-1} k_M^{(i)} \quad \leq \quad 2\mathcal{N} + \Big[\mathcal{N} \log C + \gamma \log 10\Big] / \log(q_e^{-1}). \tag{94}$$

	unpreconditioned **FSd**				preconditioned **FSd**			
	$\tau_1 = 0.05$				$\tau_1 = 0.05$			
	$\|\mathbf{r}^{(i)}\| \,/\, \|\rho^{(i)}\|$	$\tan\theta^{(i)}$	$k^{(i-1)}$	$k_*^{(i-1)}$	$\|\mathbf{r}^{(i)}\| \,/\, \|\rho^{(i)}\|$	$\tan\theta^{(i)}$	$k^{(i-1)}$	$k_*^{(i-1)}$
0	3.0e-01	1.0e-02			3.0e-01	1.0e-02		
1	1.9e-04	5.6e-04	47	786	1.4e-04	1.3e-04	25	49
2	7.8e-06	4.4e-05	123	799	2.4e-06	2.4e-05	30	54
3	6.0e-07	5.3e-06	143	808	4.0e-07	5.8e-06	37	63
4	8.1e-08	1.1e-06	139	815	9.7e-08	1.5e-06	40	67
5	1.8e-08	2.7e-07	138	821	2.4e-08	3.6e-07	43	70
6	4.4e-09	6.6e-08	109	823	6.1e-09	9.1e-08	43	73
7	1.1e-09	1.6e-08	117	823	1.5e-09	2.3e-08	44	76
8	2.7e-10	4.1e-09	118	823	3.8e-10	5.7e-09	45	79
9	6.8e-11	1.0e-09	116	823	9.5e-11	1.4e-09	47	82
10	1.7e-11	2.6e-10	118	823	2.4e-11	3.6e-10	48	85
11	4.3e-12	6.3e-11	112	823	5.9e-12	8.9e-11	50	88
12	1.1e-12	1.6e-11	114	823	1.5e-12	2.2e-11	49	91
13	9.4e-13	1.3e-11	4	824	8.1e-13	5.6e-12	48	94
$\sum k^{i-1}$			1398				549	

Table 3: Numerical results showing the behaviour of $k^{(i)}$ and $k_*^{(i)}$ (as defined in (82)) for unpreconditioned and preconditioned **FSd**.

Comparing (94) with (92) we anticipate that the total number of inner iterations in case (b) would be significantly lower than in case (a) since $\gamma \log 10$ in (94) will be significantly smaller than $\sum_{i=0}^{\mathcal{N}-1} \log \frac{1}{t^{(i+1)}}$, the corresponding term in (92). Note also that if we take $\mathbf{P} = \mathbf{I}$ in case (b) above then the behaviour for an unpreconditioned solver is described.

The following example helps to illustrate the results of Theorem 3.2.

Example 3.3 *Consider again Example 2.1 but discretised using a* 31×31 *regular grid. We consider the approximation of the 10th smallest eigenvalue. The methods* **RQIf** *and* **SEf** *are described in Example 2.3. Let us also consider the following linearly convergent method.*

FSd*: Algorithm 1 with fixed shift* $\sigma^{(i)} = \sigma^{(0)}$ *and decreasing tolerance* $\tau^{(i)} = \min\{\tau_0, \tau_1 \|\mathbf{r}^{(i)}\|\}$.

In Table 3 we present results obtained using **FSd** *and preconditioned* **FSd** *in the calculation of the 10th smallest eigenvalue to a residual accuracy of* 10^{-12}. *Further, in Table 4 we present corresponding results obtained by using* **RQIf** *and* **SEf**.

First consider the results obtained by unpreconditioned **FSd** on the left hand side of Table 3. The outer rate of convergence is linear, and so Lemma 3.2 (part (a) with $\alpha = 0, \beta = 1$) predicts no growth in inner iterations as the outer iteration proceeds. This is indeed observed.

Using preconditioned linear solves we expect that the number of inner iterations at each outer iteration will be significantly decreased compared with the

	preconditioned **RQIf**				preconditioned **SEf**			
	$\tau_0 = 0.5$				$\tau_0 = 0.5$			
	$\|\mathbf{r}^{(i)}\| \,/\, \|\rho^{(i)}\|$	$\tan\theta^{(i)}$	$k^{(i-1)}$	$k_*^{(i-1)}$	$\|\mathbf{r}^{(i)}\| \,/\, \|\rho^{(i)}\|$	$\tan\theta^{(i)}$	$k^{(i-1)}$	$k_*^{(i-1)}$
0	3.0e-01	1.0e-02			3.0e-01	1.0e-02		
1	5.6e-03	1.7e-02	6	41	5.6e-03	1.7e-02	6	39
2	9.4e-06	1.1e-05	28	51	3.4e-06	8.2e-06	26	50
3	3.4e-12	2.1e-12	45	73	3.9e-12	3.2e-12	36	62
4	9.4e-13	2.7e-12	49	93	6.1e-13	2.0e-12	5	52
$\sum k^{i-1}$			128				73	

Table 4: Numerical results showing the behaviour of $k^{(i)}$ and $k_*^{(i)}$ (as defined in (82)), and the total number of inner iterations for **RQIf** and **SEf**.

unpreconditioned case. This is indeed observed in the right hand columns of Table 3. However we note that in the preconditioned case the number of inner iterations increases slowly with the progress of the outer iteration as predicted by (91). Note that in Table 3 we have provided values for the bound $k_*^{(i)}$ as defined by (82). As was mentioned previously, it is readily seen that these values are considerably larger than the $k^{(i)}$ values. However they show the same trend as $k^{(i)}$ as i increases by increasing or remaining constant according to the method used, and as such provide the same qualitative information.

Next, (91) indicates that, whatever the outer rate of convergence of the standard preconditioned method the bound for the number of inner iterations $k^{(i)}$ to produce an error angle $t^{(i+1)}$ depends on $\log(1/t^{(i+1)})$ and is independent of the previous error angle. This is confirmed in Table 3 where for $i = 2$ preconditioned **FSd** needed 30 iterations to produce an error angle of 2.4×10^{-05}, whereas in Table 4 at $i = 2$ preconditioned **RQIf** achieved an error angle of 1.1×10^{-05} after 28 iterations, essentially the same cost, even though the previous error angles were considerably different. This shows the effect of $\log \frac{1}{t^{(i+1)}}$ in (91). Also, from Table 4 we observe an important difference between **RQIf** and **SEf**. After $i = 3$ both methods have almost achieved the desired accuracy of 10^{-12}. To reach the desired accuracy **RQIf** needed a further 49 inner iterations as predicted by the $\log \dfrac{1}{t^{(i+1)}}$ term in (91), whereas **SEf** needed only 5 more inner iterations as suggested by the $\log \dfrac{t^{(i)}}{t^{(i+1)}}$ term in (93).

Finally, comparing preconditioned **RQIf** with preconditioned **SEf** in Table 4 we see the superiority of **SEf** in terms of overall costs to achieve a given accuracy. This is predicted in Theorem 3.2 where for **SEf** the total cost depends on a $\gamma \log 10$ term in (94), whereas the corresponding term in (92) contains a sum of $\log \frac{1}{t^{(i+1)}}$ terms.

References

[1] Jörg Berns-Müller. *Inexact Inverse Iteration using Galerkin Krylov solvers.* PhD thesis, University of Bath, 2003.

[2] Rafael Bru, E. Elsner, and M. Neumann. Convergence of infinite products of matrices and inner-outer iteration schemes. *ETNA*, 2:183–193, 1994.

[3] Eric de Sturler. Nested Krylov methods based on GCR. *Journal of Computational and Applied Mathematics*, 67:15–41, 1997.

[4] Jack J. Dongarra, C.B. Moler, and J.H. Wilkinson. Improving the accuracy of computed eigenvalues and eigenvectors. *SIAM Journal on Numerical Analysis*, 20(1):23–45, 1983.

[5] Gene H. Golub and Qiang Ye. Inexact preconditioned conjugate gradient method with inner-outer iteration. *SIAM Journal on Scientific Computing*, 21(4):1305–1320, 1999.

[6] Gene H. Golub and Qiang Ye. Inexact inverse iteration for generalized eigenvalue problems. *BIT*, 40(4):671–684, 2000.

[7] Anne Greenbaum. *Iterative Methods for Solving Linear Systems.* SIAM, 1997.

[8] Anne Greenbaum, M. Rozložník, and Zdeněk Strakoš. Numerical behavior of the modified Gram-Schmidt GMRES implementation. *BIT*, 37(3):706–719, 1997.

[9] Wolfgang Hackbusch. *Iterative solution of large sparse systems of equations.* Springer-Verlag, Berlin, 1994.

[10] Andrew V. Knyazev. Preconditioned eigensolvers - an oxymoron? *Electronic transaction on Numerical Analysis*, 7:104–123, 1998.

[11] Andrew V. Knyazev. Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. Technical Report 2, 2001.

[12] Andrew V. Knyazev and Klaus Neymeyr. A geometric theory for preconditioned inverse iteration. iii: a short and sharp convergence estimate for generalized eigenvalue problems. *Linear Algebra and its Applications*, 358:95–114, 2003.

[13] Yu-Ling Lai, Kun-Yi Lin, and Lin Wen-Wei. An inexact inverse iteration for large sparse eigenvalue problems. *Numerical Linear Algebra with Applications*, 1:1–13, 1997.

[14] Yven Notay. Convergence analysis of inexact Rayleigh quotient iterations. *SIAM Journal on Matrix Analysis and Applications*, 24:627–644, 2001.

[15] Alexander M. Ostrowski. On the convergence of the Rayleigh quotient iteration for the computation of the characteristic roots and vectors. I. *Archive for Rational Mechanics and Analysis*, 1:233–241, 1957.

[16] Christopher C. Paige and M.A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis*, 12:617–629, 1975.

[17] Beresford N. Parlett. *The Symmetric Eigenvalue Problem.* Prentice-Hall, second edition, 1998.

[18] Ulrich Rüde and W. Schmid. Inverse Multigrid Correction for Generalized Eigenvalue Computations. Technical report, Universität Augsburg, September 1995.

[19] Axel Ruhe and Torbjörn Wiberg. The method of conjugate gradients used in inverse iteration. *BIT*, 12:543–554, 1972.

[20] Valeria Simoncini and Lars Eldén. Inexact Rayleigh quotient-type methods for eigenvalue computations. *BIT*, 42(1):159–182, 2002.

[21] Gerald L. G. Sleijpen, Henk A. van der Vorst, and Ellen Meijerink. Efficient expansion of subspaces in the Jacobi-Davidson method for standard and generalized eigenproblems. *ETNA*, 7:75–89, 1998.

[22] Paul Smit and Michael H. C. Paardekooper. The effects of inexact solvers in algorithms for symmetric eigenvalue problems. *Linear Algebra and its Applications*, 287:337–357, 1999.