Inexact Inverse Iteration and GMRES*

Jörg Berns–Müller† Alastair Spence‡

November 4, 2005

Abstract

In this paper we consider inexact inverse iteration applied to the nonsymmetric generalised eigenvalue problem $\mathbf{Ax} = \lambda\mathbf{Mx}$, where $\mathbf{M}$ is nonsingular and the problem is diagonalisable. This paper discusses the efficiency of inexact inverse iteration when GMRES is used to solve shifted linear systems with coefficient matrix $\mathbf{A} - \sigma\mathbf{M}$, where $\mathbf{A}$ and $\mathbf{M}$ are large and sparse. We consider three different formulations of inexact inverse iteration and show how they affect the performance of GMRES. Further we show that using a shift converging towards the sought eigenvalue is both practical and efficient. Several numerical experiments are presented that illustrate the theory.

1 Introduction

Consider the generalized eigenvalue problem

$$\mathbf{Ax} \;=\; \lambda\mathbf{Mx}, \quad \mathbf{x} \neq \mathbf{0}, \tag{1}$$

where $\mathbf{A}$ is a $n \times n$ non-symmetric matrix, $\mathbf{M}$ is a $n \times n$ nonsingular matrix, such that the standard eigenvalue problem $\mathbf{M}^{-1}\mathbf{A}$ is diagonalisable, with $\mathbf{x} \in \mathbb{C}^n$, $\lambda \in \mathbb{C}$. We have in mind the situation where n is large and $\mathbf{A}$ and $\mathbf{M}$ are sparse, as might arise from discretised partial differential equations or networks, and so methods like the QZ algorithm are not computationally feasible.

Inexact inverse iteration, the simplest inexact iterative method, involves the repeated approximate solution of shifted systems like

$$(\mathbf{A} - \sigma\mathbf{M})\mathbf{y} \;=\; \mathbf{Mx}. \tag{2}$$

*The work in this paper was sponsored by the Engineering and Physical Sciences Research Council, UK, grant number GR/M59075.

†Fachbereich Mathematik, JWG–Universität Frankfurt, Postfach 11 19 32, D-60054 Frankfurt, Germany. (`berns@math.uni-frankfurt.de`).

‡Department of Mathematical Sciences, University of Bath, United Kingdom. (`as@maths.bath.ac.uk`).

This iterative solve step is usually called the inner iteration and it is important to keep the costs of this step under control. The term outer iteration is used to describe the standard iterative process in inverse iteration where the approximate eigenvector and possibly the eigenvalue are updated.

In [1] we concentrated on the convergence of inexact inverse iteration with no particular linear solver in mind, though numerical results were presented using GMRES as a solver for the shifted linear systems. There we observed a growth in the number of iterations needed by GMRES to solve the shifted linear systems to a given tolerance. This growth has been remarked on before, see for example [17], though there was no analysis of this aspect. In this paper we give a detailed and careful analysis of the performance of unpreconditioned and preconditioned GMRES applied to shifted linear systems for three different formulations of inexact inverse iteration. This analysis allows us to understand the reasons for the growth in the number of inner iterations and explain why certain methods do not show this growth. Our analysis explains why a method based on shifts which converge to the desired eigenvalue (hence producing nearly singular systems) may still exhibit slow or even no growth in GMRES iterations. Additionally, our theory allows us to find a bound for the total number of GMRES iterations used by inexact inverse iteration and deduce that it is beneficial to choose a method that uses a shift converging towards the sought eigenvalue.

It was shown in [1] that, rather than solving $(\mathbf{A} - \sigma\mathbf{M})\mathbf{y} = \mathbf{M}\mathbf{x}$, it is useful to consider the more general linear system

$$(\mathbf{A} - \sigma\mathbf{M})\mathbf{y} = \mathbf{Z}\mathbf{x} \tag{3}$$

for some matrix $\mathbf{Z}$. If $\mathbf{Z} = \mathbf{M}$, then we recover the usual system, (2), but other choices are possible. Of particular interest is the choice $\mathbf{Z} = \mathbf{P}$, where $\mathbf{P}$ is the preconditioner used in GMRES. The idea of altering the right-hand side of the usual system when preconditioned iterative solves are used was first introduced for the symmetric eigenvalue problem in [20], but a related idea was also utilised in [19] when implementing the Lanczos method for a symmetric generalized eigenvalue problem. Our theory will also explain why this is a good idea and when it should be implemented.

Inexact inverse iteration is an 'inner-outer' type method with at least one free parameter, depending on the actual formulation. Previous work on inner-outer type methods has discussed the choice of one free parameter, usually the stopping threshold for the inner iteration, for example, see [17, 12, 22, 6, 20, 13] for inverse iteration type methods, or see [21] for a more general setting. While the stopping threshold for the inner iteration is an important quantity in our analysis, we shall show that the choice of shift and the form of the right-hand side also play crucial roles. We consider these choices in the context of three different ways to formulate inexact inverse iteration.

The outline of this paper is as follows. In Section 2 we describe three different formulations of inexact inverse iteration, which have the same iteration matrix but have different right-hand sides. Then in Section 3 we review the convergence theory for inexact inverse iteration given in [1] and [6], and give a brief discussion on the convergence of a few methods that might be used in practice. For nonsymmetric problems one might choose complex shifts so we discuss the convergence of GMRES applied to complex valued linear systems in Section 4. The convergence result presented there is standard however the form highlights the role of the right-hand side and is tailored to our later analysis. Then in Section 5 we state bounds on the number of inner iterations per outer iteration. These bounds lead to several observations. For example, we show that a strategy of shifting towards the desired eigenvalue is both a practical and efficient stragety when GMRES is the inexact solver. A bound on the total number of inner iterations is given in Section 6 from which we deduce that a strategy based on reducing the number of outer iterations may help to improve the efficiency of a given method. To avoid a cumbersome analysis in this paper we mainly consider left preconditioning. However in Section 7 we briefly comment on the right preconditioning case. In Section 8 we provide several numerical examples to illustrate the theory in this paper. Finally in Section 9 we present an eigenvalue perturbation result that is used to provide bounds needed in Sections 5 and 6.

The analysis of this paper is geared to the treatment of large sparse eigenvalue problems where preconditioning of shifted linear systems is essential. A fundamental assumption of this paper is that the preconditioner is of such a good quality that the number of inner iterations is reasonably small, and so other costs, for example, the cost of orthonormalising the Krylov basis in GMRES, may be ignored. Hence we shall take the number of inner iterations needed by the iterative solver as our measure of cost when comparing two different methods.

2 Three ways to formulate inexact inverse iteration

Here we describe three different ways of formulating inexact inverse iteration. These are mainly of interest when using (preconditioned) iterative methods for the solution of the shifted linear systems. We shall illustrate these formulations assuming a left preconditioner is available, though we could equally well have discussed right preconditioning, see Section 7. The important point to note is that all formulations have the same matrix but different right-hand sides and this will be seen to be important throughout this paper.

First, for reference we state a version of inverse iteration with exact solves:

Algorithm 1 (Inverse Iteration)
Given $\mathbf{x}^{(0)}$ *then iterate*
(1) *Choose* $\sigma^{(i)}$
(2) *Solve* $(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i+1)} = \mathbf{M}\mathbf{x}^{(i)}$
(3) *Set* $\mathbf{x}^{(i+1)} = \mathbf{y}^{(i+1)}/\varphi(\mathbf{y}^{(i+1)})$.

Here $\varphi(\mathbf{y}^{(i+1)})$ denotes a scalar normalising function. Common choices for φ are $\varphi(\mathbf{y}^{(i+1)}) = \|\mathbf{y}^{(i+1)}\|_{\mathbf{M}}$ and $\varphi(\mathbf{y}^{(i+1)}) = \mathbf{z}^H\mathbf{y}^{(i+1)}$ for some fixed vector $\mathbf{z}$. Often the choice $\mathbf{z} = \mathbf{e}_k$ (where $\mathbf{e}_k$ is the kth canonical unit vector) is made for some appropriate k, see [16]. One can keep $\sigma^{(i)}$ fixed, so that $\sigma^{(i)} = \sigma^{(0)}$, to obtain a fixed shift method. Alternatively, one can obtain a variable shift method by updating $\sigma^{(i)}$, say using the Rayleigh quotient or $\sigma^{(i+1)} = \sigma^{(i)} + 1/\varphi(\mathbf{y}^{(i)})$ if $\varphi(\mathbf{y}^{(i+1)}) = \mathbf{z}^H\mathbf{M}\mathbf{y}^{(i)}$.

The *first formulation* is merely the use of preconditioned GMRES to solve step 2 of Algorithm 1. That is, with $\mathbf{P}$ being a suitable preconditioner, apply GMRES to

$$\mathbf{P}^{-1}(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i+1)} \quad = \quad \mathbf{P}^{-1}\mathbf{M}\mathbf{x}^{(i)}, \tag{4}$$

or, equivalently

$$\mathbf{B}^{(i)}\mathbf{z}^{(i+1)} \quad = \quad \mathbf{b}^{(i)} \tag{5}$$

where

$$\mathbf{B}^{(i)} := \mathbf{P}^{-1}(\mathbf{A} - \sigma^{(i)}\mathbf{M}), \quad \mathbf{z}^{(i)} := \mathbf{y}^{(i)}, \quad \mathbf{b}^{(i)} := \mathbf{P}^{-1}\mathbf{M}\mathbf{x}^{(i)}. \tag{6}$$

Later, at the end of Section 3, we shall consider three practical versions of this formulation, depending on different choices of shift or solver tolerance. We shall see that there are some disadvantages to this formulation when we consider the performance of GMRES, and hence we are motivated to consider the next two formulations. Note that we have assumed that the solver is GMRES but that is not essential in this and the following section.

The *second formulation*, is a combination of the ideas in [19] and [20] for symmetric problems. The idea is to alter the right-hand side of (4) to hopefully improve the performance of GMRES. Specifically instead of (4) we apply GMRES to

$$\mathbf{P}^{-1}(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i+1)} \quad = \quad \mathbf{x}^{(i)}, \tag{7}$$

or, equivalently

$$\mathbf{B}^{(i)}\mathbf{z}^{(i+1)} \quad = \quad \mathbf{b}^{(i)} \tag{8}$$

where

$$\mathbf{B}^{(i)} := \mathbf{P}^{-1}(\mathbf{A} - \sigma^{(i)}\mathbf{M}), \quad \mathbf{z}^{(i)} := \mathbf{y}^{(i)}, \quad \mathbf{b}^{(i)} := \mathbf{x}^{(i)}. \tag{9}$$

As described in [1, Section 5] and summarised in the following section this change in the right-hand side alters the the basic structure of the inverse iteration algorithm, and hence may alter the outer convergence rate of inexact inverse iteration.

Before stating the third formulation we state an algorithm which incorporates the two formulations discussed above.

Algorithm 2 (Inexact Inverse Iteration)
Given $\mathbf{x}^{(0)}$ *and for an appropriately chosen* $\mathbf{Z}$, *iterate*
(1) *Choose* $\sigma^{(i)}$ *and* $\tau^{(i)}$
(2) *Find* $\mathbf{y}^{(i+1)}$ *such that* $\|(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i+1)} - \mathbf{Z}\mathbf{x}^{(i)}\| \leq \tau^{(i)}$
(3) *Set* $\mathbf{x}^{(i+1)} = \mathbf{y}^{(i+1)}/\varphi(\mathbf{y}^{(i+1)})$.

So the first formulation uses $\mathbf{Z} = \mathbf{M}$ and the second formulation uses $\mathbf{Z} = \mathbf{P}$.

The *third formulation* of inexact inverse iteration is sometimes called *inverse correction*, see for example [17]. This is simply the application to (4) of one step of classical iterative refinement for linear systems. There are different derivations, here we follow the version discussed in [6]. Specifically for a fixed shift, that is $\sigma^{(i)} = \sigma$ for all i, we reformulate (4) as

$$\begin{aligned} (\mathbf{A} - \sigma\mathbf{M})\mathbf{d}^{(i)} &= \mathbf{M}\mathbf{x}^{(i)} - (\mathbf{A} - \sigma\mathbf{M})\mathbf{y}^{(i)} =: \mathbf{r}_{IC}^{(i)} \\ \mathbf{y}^{(i+1)} &= \mathbf{y}^{(i)} + \mathbf{d}^{(i)} \\ \mathbf{x}^{(i+1)} &= \mathbf{y}^{(i)}/\varphi(\mathbf{y}^{(i+1)}). \end{aligned} \tag{10}$$

The application of left preconditioned GMRES to (10) produces

$$\mathbf{B}^{(i)}\mathbf{z}^{(i+1)} = \mathbf{b}^{(i)}$$

where

$$\mathbf{B}^{(i)} := \mathbf{P}^{-1}(\mathbf{A} - \sigma^{(i)}\mathbf{M}), \quad \mathbf{z}^{(i)} := \mathbf{d}^{(i)}, \quad \mathbf{b}^{(i)} := \mathbf{P}^{-1}\mathbf{r}_{IC}^{(i)}. \tag{11}$$

Solving the system defined by (11) inexactly we obtain a residual $\mathbf{res}_{IC}^{(i)} := \mathbf{r}_{IC}^{(i)} - (\mathbf{A} - \sigma\mathbf{M})\mathbf{d}^{(i)}$. Now as

$$\begin{aligned} (\mathbf{A} - \sigma\mathbf{M})\mathbf{y}^{(i+1)} &= (\mathbf{A} - \sigma\mathbf{M})\mathbf{y}^{(i)} + (\mathbf{A} - \sigma\mathbf{M})\mathbf{d}^{(i)} \\ &= (\mathbf{A} - \sigma\mathbf{M})\mathbf{y}^{(i)} + \mathbf{r}_{IC}^{(i)} - \mathbf{res}_{IC}^{(i+1)} \\ &= \mathbf{M}\mathbf{x}^{(i)} - \mathbf{res}_{IC}^{(i+1)}, \end{aligned}$$

we have $\mathbf{res}_{IC}^{(i)} = (\mathbf{A} - \sigma\mathbf{M})\mathbf{y}^{(i+1)} - \mathbf{M}\mathbf{x}^{(i)}$. Hence the third formulation, (11), has the same properties with respect to the outer convergence as the first formulation, however the performance of the linear solve might differ. So Algorithm 2 combines all three formulations, with $\mathbf{Z}$ for inverse correction depending on $\mathbf{y}^{(i)}$.

Each of the formulations given by (6), (9) and (11) requires the inexact solution up to a specified tolerance of a system $\mathbf{B}^{(i)}\mathbf{z}^{(i)} = \mathbf{b}^{(i)}$, with $\mathbf{B}^{(i)}$ having the same shifted form, but with a different right-hand side $\mathbf{b}^{(i)}$, each of the form $\mathbf{b}^{(i)} = \mathbf{Z}\mathbf{x}^{(i)}$. This allows us to provide (in Section 5) a unified description of the performance of GMRES for each of the three formulations of inverse iteration. In particular recognising the special form of each of the different right-hand sides $\mathbf{b}^{(i)}$ in (6), (9) and (11) is crucial in understanding the performance of GMRES at each outer iteration. The analysis in Section 5 is specially tuned to pick up this distinction.

In the following section we summarise and extend slightly the convergence theory needed for this paper.

3 Inexact Inverse Iteration

Here we review and slightly extend the treatment of inexact inverse iteration given in [1]. This analysis is also valid for residual based approaches as used in [17] and [6].

We consider the case where $\mathbf{M}$ is invertible and the eigenvalue problem $\mathbf{M}^{-1}\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ is diagonalisable, that is, there exists an invertible matrix $\mathbf{V}$, a diagonal matrix $\mathbf{\Lambda}$ and $\mathbf{U} = (\mathbf{M}\mathbf{V})^{-1}$ (all possibly complex) such that

$$\mathbf{A}\mathbf{V} \;=\; \mathbf{M}\mathbf{V}\mathbf{\Lambda}, \quad \text{and} \quad \mathbf{U}\mathbf{A} \;=\; \mathbf{\Lambda}\mathbf{U}\mathbf{M}. \tag{12}$$

Thus the columns of $\mathbf{V}$ are the right eigenvectors, that is, $\mathbf{A}\mathbf{v}_j = \lambda_j\mathbf{M}\mathbf{v}_j$, $j = 1,\dots,n$ with $\mathbf{\Lambda} = \operatorname{diag}\{\lambda_j\}$. Similarly the jth row of $\mathbf{U}$ provides the corresponding left eigenvector $\mathbf{u}_j^T\mathbf{A} = \lambda_j\mathbf{u}_j^T\mathbf{M}$, $j = 1,\dots,n$. Note that for the theory we leave the scaling of the eigenvectors open, but we could ask $\|\mathbf{v}_j\| = 1$ or $\|\mathbf{v}_j\|_{\mathbf{M}} = 1$. In either case $\mathbf{U}\mathbf{M}\mathbf{V} = \mathbf{I}$ provides the corresponding scaling for $\mathbf{u}_j$.

Assume we seek a vector in $\operatorname{span}\{\mathbf{v}_1\}$. We write the approximation $\mathbf{x}^{(i)}$ in the following form

$$\mathbf{x}^{(i)} \;=\; \alpha^{(i)}(c^{(i)}\mathbf{v}_1 + s^{(i)}\widetilde{\mathbf{w}}^{(i)}), \tag{13}$$

where $\widetilde{\mathbf{w}}^{(i)} \in \operatorname{span}(\mathbf{v}_2,\dots,\mathbf{v}_n)$ and $\|\mathbf{U}\mathbf{M}\widetilde{\mathbf{w}}^{(i)}\| = 1$. Thus $\mathbf{V}^{-1}\widetilde{\mathbf{w}}^{(i)} \in \operatorname{span}(\mathbf{e}_2,\dots,\mathbf{e}_n)$ and $\|\mathbf{V}^{-1}\widetilde{\mathbf{w}}^{(i)}\| = 1$. Further we define

$$\alpha^{(i)} \;:=\; \|\mathbf{U}\mathbf{M}\mathbf{x}^{(i)}\|. \tag{14}$$

Equation (13) is the analogue of the orthogonal splitting used in [14] for the symmetric eigenvalue problem. To see this multiply (13) by $\mathbf{U}\mathbf{M}$ and divide by $\alpha^{(i)}$ to obtain

$$\frac{\mathbf{U}\mathbf{M}\mathbf{x}^{(i)}}{\|\mathbf{U}\mathbf{M}\mathbf{x}^{(i)}\|} \;=\; c^{(i)}\mathbf{e}_1 + s^{(i)}\mathbf{U}\mathbf{M}\widetilde{\mathbf{w}}^{(i)}, \tag{15}$$

where $\mathbf{U}\mathbf{M}\widetilde{\mathbf{w}}^{(i)}$ is a unit vector orthogonal to $\mathbf{e}_1$. Geometrically (15) describes a right angled triangle with hypotenuse of length 1 and perpendicular sides of length $c^{(i)}$ and $s^{(i)}$. Thus $s^{(i)}$ and $c^{(i)}$ can be interpreted as a generalized sine and generalized cosine respectively. Hence, it is natural to use the generalized tangent

$$t^{(i)} \;:=\; \frac{|s^{(i)}|}{|c^{(i)}|} \;=\; \frac{\|(\mathbf{I} - \mathbf{e}_1\mathbf{e}_1^T)\mathbf{U}\mathbf{M}\mathbf{x}^{(i)}|}{|\mathbf{e}_1^T\mathbf{U}\mathbf{M}\mathbf{x}^{(i)}|}, \tag{16}$$

as a measure of the convergence of $\mathbf{x}^{(i)}$ to $\operatorname{span}(\mathbf{v}_1)$ since $\|\mathbf{x}^{(i)}/(\alpha^{(i)}c^{(i)}) - \mathbf{v}_1\| = t^{(i)}\|\widetilde{\mathbf{w}}^{(i)}\|$ (cf. [6, Section 2]). Note $t^{(i)}$ is independent of the scaling of $\mathbf{x}^{(i)}$.

For future reference we recall that for $\mathbf{x} \in \mathbb{C}^n$ the Rayleigh quotient for (1) is defined by

$$\varrho(\mathbf{x}) \;:=\; \mathbf{x}^H\mathbf{A}\mathbf{x}/\mathbf{x}^H\mathbf{M}\mathbf{x} \tag{17}$$

and it is readily shown that

$$|\lambda_1 - \varrho(\mathbf{x}^{(i)})| \;\le\; C|s^{(i)}| \tag{18}$$

for some positive constant C. Also, for $\mathbf{x}^{(i)}$ given by (13), the eigenvalue residual $\mathbf{r}^{(i)}$ defined by

$$\mathbf{r}^{(i)} := (\mathbf{A} - \varrho(\mathbf{x}^{(i)})\mathbf{M})\mathbf{x}^{(i)} \tag{19}$$

satisfies

$$\|\mathbf{r}^{(i)}\| \leq C|s^{(i)}| \tag{20}$$

for some positive constant C.

Let us assume that the sought eigenvalue, say λ_1, is simple and well separated, in the sense that, for any shift σ

$$0 < |\lambda_1 - \sigma| < |\lambda_2 - \sigma| \leq \ldots \leq |\lambda_n - \sigma|. \tag{21}$$

Further let the residual of the generalized system (3) be defined by

$$\mathbf{res}^{(i)} := \mathbf{Z}\mathbf{x}^{(i)} - (\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)}, \tag{22}$$

and so step 2 of Algorithm 2 asks that $\|\mathbf{res}^{(i)}| \leq \tau^{(i)}$. For our later analysis we require

$$|\mathbf{u}_1^T\mathbf{res}^{(i)}| < \beta|\mathbf{u}_1^T\mathbf{Z}\mathbf{x}^{(i)}| \tag{23}$$

for some $\beta \in (0,1)$. When $\mathbf{Z} = \mathbf{I}$ and $\mathbf{res}^{(i)} = \mathbf{0}$ and $i = 0$ this condition reduces to the standard assumption that the starting vector is not deficient in the sought eigendirection. Using condition (23) we can introduce the quantity $T^{(i)}$ defined by

$$T^{(i)} := \frac{\|(\mathbf{I} - \mathbf{e}_1\mathbf{e}_1^T)\mathbf{U}\mathbf{Z}\mathbf{x}^{(i)}\| + \|(\mathbf{I} - \mathbf{e}_1\mathbf{e}_1^T)\mathbf{U}\mathbf{res}^{(i)}\|}{|\mathbf{u}_1^T\mathbf{Z}\mathbf{x}^{(i)}| - |\mathbf{u}_1^T\mathbf{res}^{(i)}|} \tag{24}$$

which will measure the effect of the inexact solve on the outer convergence rate of Algorithm 2. For example, in the case of $\mathbf{Z} = \mathbf{M}$ and exact solves (that is $\mathbf{res}^{(i)} = \mathbf{0}$) then $T^{(i)} = t^{(i)}$ defined by (16). Note that $T^{(i)}$ is independent of the scaling of $\mathbf{x}^{(i)}$. The following Lemma, which is a combination of Lemmas 3.1 and 4.1 in [1], shows how $T^{(i)}$ enters in the convergence theory.

Lemma 3.1 *Assume the shifts $\sigma^{(i)}$ satisfy (21), and that (23) holds then*

$$t^{(i+1)} \leq \frac{|\lambda_1 - \sigma^{(i)}|}{|\lambda_2 - \sigma^{(i)}|}\, T^{(i)}, \tag{25}$$

where $T^{(i)}$ is defined by (24). Further, with $\mathbf{Z} = \mathbf{M}$ and $\|\mathbf{res}^{(i)}\| \leq \tau^{(i)} \leq Ct^{(i)}$ for some positive constant C then $T^{(i)} \leq C't^{(i)}$ for some positive constant C' and the same asymptotic rate of convergence as with exact solves is attained. If $\angle(\mathbf{Z}\mathbf{v}_1, \mathbf{M}\mathbf{v}_1) \neq 0$ or $0 < \widetilde{C} \leq \|\mathbf{res}^{(i)}\| \leq C$ then $T^{(i)} \leq C'$ for some positive constant C' and the asymptotic rate of convergence is one order less than using exact solves.

Proof. The proof is given in Lemma 4.1 of [1]. ∎

We now present five possible practical versions of Algorithm 2 starting with the choice $\mathbf{Z} = \mathbf{M}$. In this case $|\mathbf{e}_1^T\mathbf{U}\mathbf{Z}\mathbf{x}^{(i)}| = |c^{(i)}|\alpha^{(i)}$ and $\|(\mathbf{I}-\mathbf{e}_1\mathbf{e}_1^T)\mathbf{U}\mathbf{Z}\mathbf{x}^{(i)}\| = |s^{(i)}|\alpha^{(i)}$ hold, hence

$$
\begin{aligned}
T^{(i)} &\leq \frac{|s^{(i)}| + \|\mathbf{U}\mathbf{res}^{(i)}\|/\alpha^{(i)}}{|c^{(i)}| - |\mathbf{u}_1^T\mathbf{res}^{(i)}|/\alpha^{(i)}} \\
&\leq \frac{|s^{(i)}| + \|\mathbf{U}\|\tau^{(i)}/\alpha^{(i)}}{|c^{(i)}|(1-\beta)}
\end{aligned}
\tag{26}
$$

where $\tau^{(i)}$ is the tolerance on the residual in Algorithm 2. From this one can deduce convergence for a variety of practical methods, which we now discuss. Here we consider three versions using the first formulation of inexact inverse iteration, (6), so $\mathbf{Z} = \mathbf{M}$.

FSd Consider the use of a fixed shift $\sigma^{(i)} = \sigma^{(0)}$ and a decreasing tolerance $\tau^{(i)} = \min\{\tau_0, \tau_1\|\mathbf{r}^{(i)}\|/\sigma^{(0)}\}\|\mathbf{M}\mathbf{x}^{(i)}\|$, where τ_0 is chosen such that condition (23) holds while (20) ensures that $\tau^{(i)} \leq C|s^{(i)}|$. Hence, as the method proceeds the linear systems are solved more and more accurately. Lemma 3.1 indicates that **FSd** should exhibit linear convergence.

RQIf Consider the use of the Rayleigh quotient as shift, that is $\sigma^{(i)} = \varrho(\mathbf{x}^{(i)})$, then we have according to (18) $|\lambda_1 - \sigma^{(i)}| \leq C|s^{(i)}|$. Further let the tolerance be constant $\tau^{(i)} = \tau_0$ with τ_0 being small enough such that (23) holds. Lemma 3.1 indicates that **RQIf** should exhibit linear convergence.

RQId Here the shift is chosen to be the Rayleigh quotient, that is $\sigma^{(i)} = \varrho(\mathbf{x}^{(i)})$, so that, again, $|\lambda_1 - \sigma^{(i)}| \leq C|s^{(i)}|$. A decreasing tolerance as in **FSd** is chosen, that is $\tau^{(i)} = \min\{\tau_0,\tau_1\|\mathbf{r}^{(i)}\|/\sigma^{(0)}\}\|\mathbf{M}\mathbf{x}^{(i)}\|$, so that $\tau^{(i)} \leq C|s^{(i)}|$. Here τ_0 is chosen such that condition (23) holds. Lemma 3.1 indicates that **RQIf** should exhibit quadratic convergence.

The second formulation of inexact inverse iteration discussed in Section 2 alters the right-hand side of the standard formulation in that we take $\mathbf{Z} = \mathbf{P}$ instead of $\mathbf{Z} = \mathbf{M}$ in Algorithm 2 where $\mathbf{P}$ is the preconditioner for $\mathbf{A} - \sigma^{(i)}\mathbf{M}$ used in GMRES. Now, independent of $\|\mathbf{res}^{(i)}\|$, $T^{(i)}$ as defined in (24) does not tend to zero unless $\angle(\mathbf{P}\mathbf{v}_1, \mathbf{M}\mathbf{v}_1) = 0$, and so we use a fixed absolute residual $\tau^{(i)} = \tau^{(0)}$. An example for such a method is the approach suggested by Simoncini and Eldén in [20].

ModRhs In Algorithm 2 let $\mathbf{Z} = \mathbf{P}$, where $\mathbf{P}$ is the preconditioner used in GMRES. Further let the shift be chosen as the Rayleigh quotient $\sigma^{(i)} = \varrho(\mathbf{x}^{(i)})$ and so $|\lambda_1 - \sigma^{(i)}| \leq C|s^{(i)}|$. Let the tolerance be given by $\tau^{(i)} = \tau_0\|\mathbf{M}\mathbf{x}^{(i)}\|$ and so, this is essentially a fixed tolerance. Lemma 3.1 indicates that **ModRhs** should converge linearly for an accurate enough preconditioner.

The third formulation of inexact inverse iteration discussed in Section 2 is the Inverse Correction method suggested in [17, 6]. Though it has the form of Algorithm 2 and hence is covered by Lemma 3.1 when implementing the method, it is more convenient to write out the steps in detail as Algorithm 3 (following [6]).

Algorithm 3 (Inverse Correction)
Given $\mathbf{x}^{(0)}$, σ. *Set* $\mathbf{y}^{(0)} = \mathbf{0}$ *and iterate*
(1) *Set* $\mathbf{r}_{IC}^{(i)} = \mathbf{M}\mathbf{x}^{(i)} - (\mathbf{A} - \sigma\mathbf{M})\mathbf{y}^{(i)}$
(2) *Choose* $\tau_{IC}^{(i)}$,
(3) *Find* $\mathbf{d}^{(i)}$ *such that* $\|\mathbf{r}_{IC}^{(i)} - (\mathbf{A} - \sigma\mathbf{M})\mathbf{d}^{(i)}\| \leq \tau_{IC}^{(i)}$
(4) *Set* $\mathbf{y}^{(i+1)} = \mathbf{y}^{(i)} + \mathbf{d}^{(i)}$
(5) *Update* $\mathbf{x}^{(i+1)} = \mathbf{y}^{(i+1)}/\varphi(\mathbf{y}^{(i+1)})$.

For later reference we define the following method.

IC Consider the use of Algorithm 3 with $\tau_{IC}^{(i)} = \min\{\tau_0, \tau_1\|\mathbf{r}_{IC}^{(i)}\|/\sigma^{(0)}\}$. **IC** has linear convergence.

Since the shift is fixed and the tolerance is decreasing, Lemma 3.1 indicates that this version should have linear convergence, just like **FSd**. However, in **IC** the system $(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{d}^{(i)} = \mathbf{r}_{IC}^{(i)}$ is solved to a tolerance proportional to $\|\mathbf{r}_{IC}^{(i)}\|$. Hence, though the tolerance is decreasing, it is in fact a *fixed* relative tolerance, and as a consequence we shall see that, **IC** requires fewer inner iterations than **FSd** even though both methods have the same outer convergence rate.

While the above results are valid independent of the linear solver we will now assume that GMRES is the solver applied to the shifted linear systems.

Numerical results illustrating the convergence rate achieved by the methods described above and the performance of GMRES are given in Section 8.

4 GMRES

In many problems $\mathbf{A}$ and $\mathbf{M}$ will be real but non-symmetric and so the shifts in inverse iteration may be complex. Hence, in this section we provide a convergence result for GMRES applied to the complex linear system

$$\mathbf{B}\mathbf{z} \;=\; \mathbf{b} \tag{27}$$

which later on we shall apply to equations (6), (9) and (11) to help understand the performance of the inexact solves in inexact inverse iteration. The results here on GMRES for complex systems are known but we have not found them in a convenient form in the literature.

Denote the eigenvalues of $\mathbf{B}$ by $\mu_1, \ldots, \mu_n$ and the eigendecomposition of $\mathbf{B}$ by $\mathbf{B} = \mathbf{W}\mathbf{D}\mathbf{W}^{-1}$. Now assume the initial approximation $\mathbf{z}_0 = \mathbf{0}$, and denote the approximate solution after k steps of GMRES by $\mathbf{z}_k$ then $\mathbf{z}_k \in \mathcal{K}_k(\mathbf{B}, \mathbf{b})$ where $\mathcal{K}_k(\mathbf{B}, \mathbf{b})$ is the Krylov space spanned by $\{\mathbf{b}, \mathbf{B}\mathbf{b}, \mathbf{B}^2\mathbf{b}, \ldots, \mathbf{B}^{k-1}\mathbf{b}\}$. It is a standard result that

$$\|\mathbf{b} - \mathbf{B}\mathbf{z}_k\| \;=\; \min_{p \in \Pi_k^1} \|p(\mathbf{B})\mathbf{b}\| \;\leq\; \kappa(\mathbf{W}) \min_{p \in \Pi_k^1} \max_{\xi \in \Omega} |p(\xi)| \; \|\mathbf{b}\|, \tag{28}$$

where $\Omega \subset \mathbb{C}$ contains all eigenvalues of $\mathbf{B}$ and $\mathbf{W}$ denotes the matrix of eigenvectors of $\mathbf{B}$ while $\kappa(\mathbf{W})$ denotes the condition number of $\mathbf{W}$. Here Π_k^1 is the set of polynomials with degree $\leq k$ which equal 1 at the origin. Convergence is obtained when $\min\max|p(\xi)|$ tends to zero as k tends to infinity. To bound this min max problem one usually uses Chebyshev polynomials however they are not necessarily optimal in cases where $\Omega \not\subset \mathbb{R}$, see [4].

In the case where $\mathbf{B}$ is real symmetric and has an eigenvalue well separated from the rest of the spectrum, it is also standard, see for example [10], to take advantage of this property. For example, if the eigenvalues of $\mathbf{B}$ satisfy $\mu_2, \mu_3 \ldots, \mu_n \in \Omega$ and $\mu_1 \notin \Omega$ with $\mu_1 \neq 0$ then, following [10], one has the bound

$$\|\mathbf{b} - \mathbf{B}\mathbf{z}_k\| \leq \kappa_2(\mathbf{W}) \max_{\xi\in\Omega} \left|\frac{\mu_1 - \xi}{\mu_1}\right| \min_{p\in\Pi^1_{k-1}} \max_{\xi\in\Omega} |p(\xi)| \, \|\mathbf{b}\|. \tag{29}$$

When $\mathbf{B}$ is symmetric or symmetric positive definite it is standard to further bound $\min\max|p(\xi)|$ in terms of a reduced condition number of $\mathbf{B}$, see, for example, [10, 7]. By this way of doing the analysis highlights the role well separated eigenvalues play with respect to the rate of convergence, and is advantageous when in (29) the minimum over Π^1_{p-1} tends to zero considerably quicker than the minimum over Π^1_p in (28).

Here we utilise the approach of splitting the spectrum into two to analyse the convergence of (27). First, let us assume that the eigenvalues of $\mathbf{B}$ fall into two sets, a critical set $\mathcal{J} := \{\mu_1, \ldots, \mu_J\}$ with $J \ll n$, and $\Omega := \{\mu_{J+1}, \ldots, \mu_n\}$. We then use the approach in [10], that is we separate the effect of the critical set of eigenvalues from the rest in the expression for $\|\mathbf{b} - \mathbf{B}\mathbf{z}_k\|$. To accomplish this we introduce the orthogonal projection $\mathbf{Q}_J$ defined by

$$\mathbf{Q}_J := \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{I}_{n-J} \end{pmatrix}. \tag{30}$$

The use of $\mathbf{Q}_J$ is nonstandard and we will use $\mathbf{Q}_J$ to project out the eigenvectors of $\mathbf{B}$ corresponding to the set of critical eigenvalues, $\mathcal{J}$. This will prove important in our later analysis on the effect the choice of right hand side has on the performance of GMRES.

While in the symmetric case the minimum over the constrained polynomials, as in (29), tends to zero for $k \to \infty$ this is not necessarily the case in the complex world. Therefore, we need an assumption on the min max term in the bounds of $\|\mathbf{b} - \mathbf{B}\mathbf{z}_k\|$ to ensure that they converge to zero as k tends to infinity.

Assumption 1 *Let $\mathbf{B}$ be diagonalisable with eigenvalues $\{\mu_1, \mu_2, \ldots, \mu_n\}$, split into the sets $\mathcal{J} = \{\mu_1, \ldots, \mu_J\}$ and $\Omega = \{\mu_{J+1}, \ldots, \mu_n\}$. We assume that there exists constants $C > 0$ and $q \in (0,1)$ such that for all $k \in \mathcal{N}$*

$$\min_{p\in\Pi^1_k} \max_{\mu_j\in\Omega} |p(\mu_j)| \leq C\, q^k. \tag{31}$$

Later we apply the GMRES theory to preconditioned systems $\mathbf{B}^{(i)}$, where, for a good preconditioner, most of the eigenvalues will be clustered and lie in the set Ω which can be enclosed by a certain ellipse not enclosing the origin. In such cases it is known that a bound like (31) exists with q determined by the radius of this ellipse divided by the radius of a second ellipse with the same foci but with the origin on the boundary, for details, see [4]. (For more on constrained minimal polynomials over complex sets see, for example, [15, 18, 5].) The set $\mathcal{J}$ contains the eigenvalues outside the ellipse. Often when considering inverse iteration applied to a well separated eigenvalue we shall have $\mathcal{J} = \{\mu_1\}$, so there is only one critical eigenvalue.

The following result is standard as it uses the same technique as that used to obtain (29) with the exception that the use of the projection $\mathbf{Q}_J$ on the right-hand side is nonstandard. In our later analysis the term $\|\mathbf{Q}_J\mathbf{W}^{-1}\mathbf{b}\|$ which appears in (32) plays a prominent part.

Lemma 4.1 *Consider GMRES applied to $\mathbf{B}\mathbf{z} = \mathbf{b}$ where $\mathbf{B} \in \mathbb{C}^{n\times n}$ and $\mathbf{b} \in \mathbb{C}^n$. Let Assumption 1 hold and let $\mathbf{Q}_J$ be defined by (30). Then for all $k \geq J$*

$$\|\mathbf{b} - \mathbf{B}\mathbf{z}_k\|_2 \quad \leq \quad p\, q^{k-J}|\mu_1|^{-1}\|\mathbf{Q}_J\mathbf{W}^{-1}\mathbf{b}\|_2, \tag{32}$$

where

$$p \quad := \quad \|\mathbf{W}\mathbf{Q}_J|_2 C \max_{\xi\in\Omega}\left(|\mu_1 - \xi|\prod_{j=2}^{J}\frac{|\mu_j - \xi|}{|\mu_j|}\right), \tag{33}$$

and where C is as in Assumption 1.

Proof. We use the auxiliary polynomial $g(\xi) := \prod_{j\leq J}(\mu_j - \xi)/\mu_j$ then for any $f \in \Pi^1_{k-J}$ the product fg is in Π^1_k. Note that $\mathbf{B}$ is diagonalisable with $\mathbf{B} = \mathbf{W}\mathbf{D}\mathbf{W}^{-1}$ and that $\mathbf{Q}_J$ and $\mathbf{D}$ comute, while ${\mathbf{Q}_J}^2 = \mathbf{Q}_J$. Further we have $g(\mathbf{D}) = g(\mathbf{D}\mathbf{Q}_J)$ and therefore $f(\mathbf{D})g(\mathbf{D}) = f(\mathbf{D}\mathbf{Q}_J)g(\mathbf{D}\mathbf{Q}_J)\mathbf{Q}_J$. Hence

$$\begin{aligned}
\|\mathbf{b} - \mathbf{B}\mathbf{z}_k\| &= \min_{f\in\Pi^1_k}\|\mathbf{W}f(\mathbf{D})\mathbf{W}^{-1}\mathbf{b}\| \\
&\leq \min_{f\in\Pi^1_{k-J}}\|\mathbf{W}f(\mathbf{D}\mathbf{Q}_J)g(\mathbf{D}\mathbf{Q}_J)\mathbf{Q}_J\mathbf{W}^{-1}\mathbf{b}\|, \\
&\leq \min_{f\in\Pi^1_{k-J}}\|\mathbf{W}\mathbf{Q}_J\|\,\|f(\mathbf{D}\mathbf{Q}_J)\|\,\|g(\mathbf{D}\mathbf{Q}_J)\|\,\|\mathbf{Q}_J\mathbf{W}^{-1}\mathbf{b}\|,
\end{aligned}$$

and applying Assumption 1 to $\min_{f\in\Pi^1_{k-J}}\|f(\mathbf{D})\|$ we obtain the claimed result. ☐

Now we are in the position to obtain a bound on the number of iterations needed by GMRES that we will use later. Assume we seek a solution of (27) such that $\|\mathbf{b} - \mathbf{B}\mathbf{z}_k\| \leq \tau_G$ for a given tolerance τ_G. It is common to present a condition on k such that for k greater than a certain quantity the bound on the residual is achieved. Here we present this type of result in a slightly different form, namely, we provide an upper bound for k_{actual}, the actual number of iterations used by GMRES to satisfy the residual condition, that is,

$$\|\mathbf{b} - \mathbf{B}\mathbf{z}_k\| > \tau_G, \;\; \forall k < k_{actual} \quad \text{and} \quad \|\mathbf{b} - \mathbf{B}\mathbf{z}_{k_{actual}}\| \leq \tau_G. \tag{34}$$

Corollary 4.2 *Assume the conditions of Lemma 4.1 hold and define the real number K by*

$$K \; := \; J + \log\left(\frac{p\|\mathbf{Q}_J\mathbf{W}^{-1}\mathbf{b}\|}{|\mu_1|\,\tau_G}\right) / |\log q|. \tag{35}$$

Hence, if $k \geq K$ then $\|\mathbf{b} - \mathbf{B}\mathbf{z}_k\| \leq \tau_G$ and thus

$$k_{actual} \;\; \leq \;\; K + 1. \tag{36}$$

Proof. Using (32) we have $\|\mathbf{b} - \mathbf{B}\mathbf{z}_k\| \leq \tau_G$ if

$$\left(\frac{1}{q}\right)^k \;\; \geq \;\; \left(\frac{1}{q}\right)^J \frac{p\|\mathbf{Q}_J\mathbf{W}^{-1}\mathbf{b}\|}{|\mu_1|\,\tau_G},$$

then by taking logs and dividing by $\log(1/q)$ we obtain the result. $\square$

As pointed out by [8] the convergence of GMRES is not entirely determined by the eigenvalues of $\mathbf{B}$, therefore the bounds presented here are best used as qualitative than quantitative information. For other descriptive bounds on the convergence of GMRES see [3].

5 Inner Iteration Analysis for Preconditioned GMRES

In this section we combine the convergence of inexact inverse iteration with the theory for GMRES in order to understand better the behaviour of GMRES when applied to the shifted system $\mathbf{B}^{(i)} = \mathbf{P}^{-1}(\mathbf{A} - \sigma^{(i)}\mathbf{M})$. For the remainder of this paper we assume $\alpha^{(i)} = 1$. (The term $\alpha^{(i)}$ has no affect on the analysis or theoretical results and provides no further inside.) Let us first review the problem.

Assume that (1) has an algebraically simple eigenpair $(\mathbf{v}_1, \lambda_1)$ with λ_1 well separated from the other eigenvalues (that is, (21) holds). We seek to compute λ_1 using inexact inverse iteration which involves the repeated solution of preconditioned linear systems with the matrix $\mathbf{B}^{(i)} = \mathbf{P}^{-1}(\mathbf{A} - \sigma^{(i)}\mathbf{M})$. Let us denote the spectrum of $\mathbf{B}^{(i)}$ by $\{\mu_1, \ldots, \mu_n\}$. Theorem 9.1 in Section 9 shows that under fairly general conditions $\mathbf{B}^{(i)}$ has an algebraic simple eigenvalue μ_1 satisfying

$$C_1|\lambda_1 - \sigma^{(i)}| \; \leq \; |\mu_1| \; < \; |\mu_j|, \quad j = 2, \ldots, n \tag{37}$$

for some constant C_1. Thus we may split the spectrum of $\mathbf{B}^{(i)}$ into two sets $\mathcal{J} = \{\mu_1\}$ and $\Omega = \{\mu_2, \ldots, \mu_n\}$ and apply Corollary 4.2 with $J = 1$ to obtain a bound on $k_{actual}^{(i)}$, needed by GMRES. Before doing this, note that the theory in Section 3 defines $\mathbf{res}^{(i)}$ as the residual of the original system (22), and asks that $\|\mathbf{res}^{(i)}\| \leq \tau^{(i)}$, whereas the theory in Section 4 uses the residual of the preconditioned system (27), and asks that $\|\mathbf{b}^{(i)} - \mathbf{B}^{(i)}\mathbf{z}_k^{(i)}\| \leq \tau_G^{(i)}$. To match these conditions we take

$$\tau_G^{(i)} \;\; = \;\; \tau^{(i)}/\|\mathbf{P}\| \tag{38}$$

so that if $\|\mathbf{b}^{(i)} - \mathbf{B}^{(i)}\mathbf{z}_k^{(i)}\| \leq \tau_G^{(i)}$ we have

$$\|\mathbf{Z}\mathbf{x}^{(i)} - (\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}_k^{(i)}\| \; = \; \|\mathbf{P}(\mathbf{b}^{(i)} - \mathbf{B}^{(i)}\mathbf{z}_k^{(i)})\| \; \leq \; \|\mathbf{P}\|\tau_G^{(i)} \; = \; \tau^{(i)},$$

(using (6), (9) or (11) as appropriate). Thus, given a shift $\sigma^{(i)}$, a tolerance $\tau^{(i)}$ and a right hand side $\mathbf{b}^{(i)}$, then a typical question is how many iterations are required to solve the linear system.

Theorem 5.1 *Let $\sigma^{(i)}$ satisfy* (21) *and assume that Assumption 1 holds for* $\mathbf{B}^{(i)} = \mathbf{P}^{-1}(\mathbf{A} - \sigma^{(i)}\mathbf{M})$. *Assume that* (1) *has an algebraically simple eigenvalue satisfying* (21). *If* $\mathbf{u}_1^T\mathbf{P}\mathbf{v}_1 \neq 0$ *then* $k_{actual}^{(i)}$, *defined by* (34), *satisfies*

$$k_{actual}^{(i)} \;\; \leq \;\; 2 + \left(\log \frac{p\|\mathbf{P}\|}{C_1} + \log \frac{\|\mathcal{R}\mathbf{b}^{(i)}\|}{|\lambda_1 - \sigma^{(i)}|\tau^{(i)}} \right) / \log(1/q), \tag{39}$$

where

$$\mathcal{R}\mathbf{b}^{(i)} \;\; := \;\; \mathbf{Q}_J(\mathbf{W}^{(i)})^{-1}\mathbf{b}^{(i)} \tag{40}$$

and C_1 is the constant in (37).

Proof. We combine Corollary 4.2 with (37) and with $\tau_G^{(i)} = \tau^{(i)}s//\|\mathbf{P}\|$. ☐

The key term in (39) is

$$\log \left(\frac{\|\mathcal{R}\mathbf{b}^{(i)}\|}{|\lambda_1 - \sigma^{(i)}|\tau^{(i)}} \right). \tag{41}$$

It shows that $k_{actual}^{(i)}$ depends on the interplay of the choice of right-hand side $\mathbf{b}^{(i)}$, the choice of shift $\sigma^{(i)}$ and the choice of $\tau^{(i)}$. We shall discuss the term $\|\mathcal{R}\mathbf{b}^{(i)}\|$ in the following subsection. But first we comment on the choices of $\sigma^{(i)}$ and $\tau^{(i)}$.

From the convergence theory of inexact inverse iteration we know that reducing $\tau^{(i)}$ and letting the shift $\sigma^{(i)}$ tend towards the desired eigenvalue helps to speed up the outer convergence. As we see from (41) both of these actions will increase the cost of the linear solve.

Inequality (39) shows how expensive each linear solve is, however it does not relate the cost of a linear solve with the gain for the outer method. To fully understand the interaction of GMRES with inexact inverse iteration such a *cost-gain* relation is essential. This cost-gain relation is best expressed as an a posteriori bound.

Theorem 5.2 *Assume the conditions of Theorem 5.1 hold. If* $\exists C_2 > 0$ *such that*

$$T^{(i)} \;\; \leq \;\; C_2\tau^{(i)} \tag{42}$$

where $T^{(i)}$ is defined by (24) *then*

$$k_{actual}^{(i)} \;\; \leq \;\; C_3 + \log \frac{\|\mathcal{R}\mathbf{b}^{(i)}\|}{t^{(i+1)}} / \log(1/q), \tag{43}$$

for some constant C_3 independent of i. *Proof* Using Lemma 3.1 we have

$$t^{(i+1)} \;\leq\; \frac{|\lambda_1 - \sigma^{(i)}|}{|\lambda_2 - \sigma^{(i)}|} T^{(i)} \;\leq\; 2\,\frac{|\lambda_1 - \sigma^{(i)}|}{|\lambda_2 - \lambda_1|}\, C_2\, \tau^{(i)}$$

which is equivalent to

$$\frac{1}{|\lambda_1 - \sigma^{(i)}|\tau^{(i)}} \;\leq\; \frac{2C_2}{|\lambda_2 - \lambda_1|}\,\frac{1}{t^{(i+1)}}.$$

Applying this inequality to Theorem 5.1 we obtain the claimed result with

$$C_3 = 2 + \log\left(2\, C_2\, \|\mathbf{P}\|\, p/(C_1\, |\lambda_2 - \lambda_1|\right) / \log(1/q). \qquad \square$$

Note that the bound on $k^{(i)}_{actual}$ presented in Theorem 5.2 does not explicitly depend on $\lambda_1 - \sigma^{(i)}$ or $\tau^{(i)}$, however both these terms influence the bound through the term $t^{(i+1)}$. We can deduce from (43) that it is of equal merit to either shift towards the sought eigenvalue and ask for a crude approximation of the linear system or to keep the shift away from the eigenvalue and solve the linear system rather accurately. This leads to the following remark.

Remark 5.3. Given a right-hand side $\mathbf{b}^{(i)}$ then Theorem 5.1 shows that a strategy of shifting towards the desired eigenvalue with a fixed tolerance should have the same effect on the number of GMRES iterations as a strategy based on keeping the shift fixed and decreasing the tolerance.

From Lemma 3.1 we know that $T^{(i)}$ is at most linear in $t^{(i)}$ for methods with $\mathbf{Z} = \mathbf{M}$ and constant for methods with $\mathbf{Z} = \mathbf{P}$, hence residual tolerances which are constant or decreasing like $t^{(i)}$ are consistent with Theorem 5.2, due to the condition (42). This condition ensures that solving the linear system up to $\tau^{(i)}$ still improves the outer convergence. In contrast solving the linear systems more accurately, so in effect oversolving the system, provides little or no improvement for the outer iteration.

5.1 Analysis of the right-hand side term $\mathcal{R}\mathbf{b}^{(i)}$

In this section we discuss the behaviour of the term $\mathcal{R}\mathbf{b}^{(i)}$ defined in (40).

Recall that $\mathcal{J} = \{\mu_1\}$, and

$$\mathbf{Q}_{J} \;=\; \begin{pmatrix} 0 & \\ & \mathbf{I}_{n-1} \end{pmatrix},$$

and $\mathbf{W}^{(i)}$ is the matrix of eigenvectors $\mathbf{w}_j^{(i)}$, $j = 1, \ldots, n$, of $\mathbf{B}^{(i)}$. Thus if $\mathbf{b}^{(i)}$ were expanded as a linear combination of the $\mathbf{w}_j^{(i)}$, say $\mathbf{b}^{(i)} = \sum \beta_j^{(i)} \mathbf{w}_j^{(i)}$, then

$$\mathcal{R}\mathbf{b}^{(i)} \;=\; \mathbf{Q}_{J}(\mathbf{W}^{(i)})^{-1}\mathbf{b}^{(i)} \;=\; \sum_{j=2}^{n} \beta_j^{(i)} \mathbf{w}_j^{(i)},$$

that is the eigendirection corresponding to μ_1 has been projected out. It is this construction that allows a proper understanding of the effect of the different right-hand sides $\mathbf{b}^{(i)}$ (in (6), (9) and (11)) on the performance of GMRES as we now explain.

Consider first a simple example, namely, the standard eigenvalue problem, $\mathbf{Ax} = \lambda\mathbf{x}$ and the use of unpreconditioned GMRES to solve $(\mathbf{A} - \sigma^{(i)}\mathbf{I})\mathbf{y}^{(i)} = \mathbf{x}^{(i)}$. So, in the notation of Sections 2 and 3, $\mathbf{M} = \mathbf{I}$, $\mathbf{P} = \mathbf{Z} = \mathbf{I}$, $\mathbf{b}^{(i)} = \mathbf{x}^{(i)}$ and $(\mathbf{W}^{(i)})^{-1} = \mathbf{U}$, the matrix of left eigenvectors. Therefore

$$\|\mathcal{R}\mathbf{b}^{(i)}\| = \|\mathbf{Q}_J(\mathbf{W}^{(i)})^{-1}\mathbf{b}^{(i)}\| = \|\mathbf{Q}_J\mathbf{U}\mathbf{x}^{(i)}\| = \|(\mathbf{I} - \mathbf{e}_1\mathbf{e}_1^T)\mathbf{U}\mathbf{x}^{(i)}\| = s^{(i)}$$

using (13) and assuming $\alpha^{(i)} = 1$. Consider now the term (41), when method **FSd** in Section 3 is used. In this case, with a fixed shift and with $\tau^{(i)} \leq C|s^{(i)}|$ we see that $\|\mathcal{R}\mathbf{b}^{(i)}\|/(|\lambda_1 - \sigma^{(i)}|\tau^{(i)}) \leq C$, for some C. Hence we expect no growth in the number of inner iterations per outer iteration even though the system are being solved with a decreasing tolerance. This is indeed observed in Table 1, Example 8.1 and was previously noted in [12]. As we will show, this effect is lost when preconditioned linear solves are applied to a standard eigenvalue problem or when a generalised eigenvalue problem is considered. However the approach of modifying the right-hand side as in the second formulation in Section 2 recovers this property as is shown in part (iii) of the following lemma. This lemma provides bounds on $\|\mathcal{R}\mathbf{b}^{(i)}\|$ for each of the formulations discussed in Section 2.

Lemma 5.4 *Let $\mathcal{R}\mathbf{b}^{(i)}$ be defined by* (40). *Then we have the following bounds on $\|\mathcal{R}\mathbf{b}^{(i)}\|$ for several different $\mathbf{b}^{(i)}$.*

(i) *For the first formulation in Section 2 leading to equation* (6) *in case of a standard eigenvalue problem $\mathbf{M} = \mathbf{I}$ and without preconditioning, so $\mathbf{P} = \mathbf{I}$, then*

$$\|\mathcal{R}\mathbf{b}^{(i)}\| \leq Ct^{(i)}, \tag{44}$$

for some positive constant C independent of i.

(ii) *For the first formulation in Section 2 leading to equation* (6) *with preconditioning,*

$$\|\mathcal{R}\mathbf{b}^{(i)}\| \leq C + C't^{(i)}, \tag{45}$$

for some positive constant C' independent of i, and with

$$C \leq \|(\mathbf{W}^{(i)})^{-1}(\mathbf{P}^{-1}\mathbf{M}\mathbf{v}_1 - \mathbf{v}_1)\|.$$

(iii) *For the second formulation in Section 2 leading to equation* (9) *on which version* **ModRhs** *is based we have*

$$\|\mathcal{R}\mathbf{b}^{(i)}\| \leq Ct^{(i)}, \tag{46}$$

for some positive constant C independent of i.

(iv) *For the third formulation in Section 2 leading to equation* (11) *on which version* **IC** *is based we have*

$$\|\mathcal{R}\mathbf{b}^{(i)}\| \;\leq\; Ct^{(i)}, \tag{47}$$

for some positive constant C independent of i.

Proof. We have proved case (i) above while the result for case (iv) is obvious.

For case (iii) we observe that $\mathbf{S} := \mathbf{P}^{-1}(\mathbf{A} - \lambda_1\mathbf{M})$ has the simple eigenpair $(0, \mathbf{v}_1)$. From Theorem 9.1, we know that $\mathbf{B}^{(i)}$ has an eigenpair $(\mu_1^{(i)}, \mathbf{w}_1^{(i)})$ with $\|\mathbf{w}_1^{(i)} - \mathbf{v}_1\| \leq \widetilde{C}|\lambda_1 - \sigma^{(i)}|$ for some constant $\widetilde{C} > 0$. Now with $\mathbf{b}^{(i)} = \mathbf{x}^{(i)}$

$$\begin{aligned} \mathcal{R}\mathbf{b}^{(i)} &= \mathbf{Q}_J(\mathbf{W}^{(i)})^{-1}\mathbf{x}^{(i)} \\ &= \mathbf{Q}_J(\mathbf{W}^{(i)})^{-1}(\mathbf{x}^{(i)} - c^{(i)}\mathbf{w}_1^{(i)}), \end{aligned} \tag{48}$$

since $\mathbf{Q}_J(\mathbf{W}^{(i)})^{-1}\mathbf{w}_1^{(i)} = \mathbf{0}$. Then

$$\mathcal{R}\mathbf{b}^{(i)} \;=\; \mathbf{Q}_J(\mathbf{W}^{(i)})^{-1}(\mathbf{x}^{(i)} - c^{(i)}\mathbf{v}_1^{(i)}) + c^{(i)}\mathbf{Q}_J(\mathbf{W}^{(i)})^{-1}(\mathbf{v}_1^{(i)} - \mathbf{w}_1^{(i)}). \tag{49}$$

Now, using (13) and the fact that $\|\mathbf{v}_1 - \mathbf{w}_1^{(i)}\| \leq \widetilde{C}|\lambda_1 - \sigma^{(i)}| \leq C|s^{(i)}|$, since $\sigma^{(i)}$ is the Rayleigh quotient, we obtain the bound $\|\mathcal{R}\mathbf{b}^{(i)}\| \leq Ct^{(i)}$, for some C.

Finally for case (ii) we have $\mathbf{b}^{(i)} = \mathbf{P}^{-1}\mathbf{M}\mathbf{x}^{(i)}$, and therefore

$$\mathcal{R}\mathbf{b}^{(i)} \;=\; \mathbf{Q}_J(\mathbf{W}^{(i)})^{-1}(\mathbf{P}^{-1}\mathbf{M}\mathbf{x}^{(i)} - c^{(i)}\mathbf{w}_1^{(i)}), \tag{50}$$

since $\mathbf{Q}_J(\mathbf{W}^{(i)})^{-1}\mathbf{w}_1^{(i)} = \mathbf{0}$. Then

$$\begin{aligned} \mathcal{R}\mathbf{b}^{(i)} \;=\; & \Big(\mathbf{Q}_J(\mathbf{W}^{(i)})^{-1}\mathbf{P}^{-1}\mathbf{M}\mathbf{x}^{(i)} - \mathbf{Q}_J(\mathbf{W}^{(i)})^{-1}\mathbf{P}^{-1}\mathbf{M}c^{(i)}\mathbf{v}_1^{(i)}\Big) \\ & + c^{(i)}\mathbf{Q}_J(\mathbf{W}^{(i)})^{-1}(\mathbf{P}^{-1}\mathbf{M}\mathbf{v}_1 - \mathbf{v}_1) + c^{(i)}\mathbf{Q}_J(\mathbf{W}^{(i)})^{-1}(\mathbf{v}_1 - \mathbf{w}_1^{(i)}). \end{aligned} \tag{51}$$

Now the first and third terms are bounded by $C|s^{(i)}|$ as in case (i), but the second term is bounded by $\|(\mathbf{W}^{(i)})^{-1}(\mathbf{P}^{-1}\mathbf{M}\mathbf{v}_1 - \mathbf{v}_1)\|$. Hence

$$\mathcal{R}\mathbf{b}^{(i)} \;\leq\; C + C't^{(i)} \tag{52}$$

with $C = \|(\mathbf{W}^{(i)})^{-1}(\mathbf{P}^{-1}\mathbf{M}\mathbf{v}_1 - \mathbf{v}_1)\|$. □

This result provides the reason to chose $\mathbf{Z} = \mathbf{M}$ or $\mathbf{Z} = \mathbf{P}$, in Algorithm 2, where $\mathbf{P}$ is the preconditioner in GMRES. While the choice $\mathbf{Z} = \mathbf{M}$ is beneficial for the outer convergence the choice $\mathbf{Z} = \mathbf{P}$ is beneficial for the inner convergence. The reason for the latter statement is seen by noting that for $\mathbf{Z} = \mathbf{P}$, $\|\mathcal{R}\mathbf{b}^{(i)}\| \leq Ct^{(i)}$, and looking again at (41) we gain a $t^{(i)}$ term on the numerator of the log term, which gives a consequent decrease in the bound on k_{actual} in (39). For preconditioned solves of the standard right hand side with $\mathbf{Z} = \mathbf{M}$ we do not have this helpful feature in the numerator of (41).

A heuristic explanation of why versions **ModRhs** in Lemma 5.4 is better than versions based on the standard formulation (6) is that the Krylov subspace in GMRES is generated with a $\mathbf{x}^{(i)}$. Therefore, the angle between the initial subspace and the final solution in each outer iteration is of order $O(t^{(i)})$.

As case (ii) in Lemma 5.4 indicates, the size of $\|\mathcal{R}\mathbf{b}^{(i)}\|$ depends on the angle $\angle(\mathbf{v}_1, \mathbf{P}^{-1}\mathbf{M}\mathbf{v}_1)$. We illustrate this later in Example 8.3.

In the following we present some Remarks combining the results of Theorem 5.2 and Lemma 5.4.

Remark 5.5. For cases in Lemma 5.4 where $\|\mathcal{R}\mathbf{b}^{(i)}\| \leq Ct^{(i)}$, that is formulations based on (9) and (11), then from (43) we see that $k^{(i)}_{actual}$ is proportional to $\log(t^{(i)}/t^{(i+1)})$. Hence the number of inner iterations needed by preconditioned GMRES is related to the improvement made in one step of the outer iteration. Numerical examples in Section 8 (Example 8.2) illustrate this point. The same situation arises when unpreconditioned GMRES is used for the standard eigenvalue problem $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$.

Remark 5.6. For preconditioned solves of the formulation given by (6) the number of inner iterations, $k^{(i)}_{actual}$, is independent of $t^{(i)}$. Therefore, it does not matter how good the current approximation is, the cost of the next solve depends only on the resulting tangent as now $k^{(i)}_{actual}$ grows like $\log(1/t^{(i+1)})$.

One unfortunate effect of the situation discussed in Remark 5.6 is that a small improvement in the final iteration of inexact inverse iteration using the standard formulation (6) is only obtained with the high cost of a very accurate solve of an ill conditioned system. Whereas Remark 5.5 indicates that when versions **ModRhs** and **IC** are used a small improvement in the final iteration may be achieved with little effort.

6 Total number of inner iterations

In Section 5 we provided bounds on the actual number of GMRES iterations per outer iteration, and were therefore able to understand differences in the numerical performance of various formulations of inexact inverse iteration. In this section we shall take this analysis one further and consider the total number of inner iterations for a given version of inexact inverse iteration. (Recall from the remarks at the end of the introduction that we take our measure of cost as the number of inner iterations.)

Let $\mathcal{N}$ denote the total number of outer iterations performed and define k_{total} by

$$k_{total} = \sum_{i=0}^{\mathcal{N}-1} k^{(i)}_{actual} \tag{53}$$

where $k^{(i)}_{actual}$ is defined by (34).

Theorem 6.1 *Assume the conditions of Lemma 3.1 hold and that for some constant* $C \in (0,1)$

$$\frac{|\lambda_1 - \sigma^{(i)}|}{|\lambda_2 - \sigma^{(i)}|} T^{(i)} \leq Ct^{(i)} \tag{54}$$

holds. Further assume the conditions of Theorem 5.2 hold. Now let $\mathcal{N}$ be the number of outer iterations used by inexact inverse iteration to achieve the tangent t^* where $t^* = 10^{-\gamma} t^{(0)}$ with $\gamma > 0$.

(i) *In case of the first formulation with $\|(\mathbf{W}^{(i)})^{-1}(\mathbf{P}^{-1}\mathbf{M}\mathbf{v}_1 - \mathbf{v}_1)\| = 0$ and in case of the second and third formulation there exists a constant C' such that*

$$k_{total} \;\leq\; \mathcal{N}C' + \gamma \frac{\log 10}{\log(1/q)}. \tag{55}$$

(ii) *In case of the first formulation with $\|(\mathbf{W}^{(i)})^{-1}(\mathbf{P}^{-1}\mathbf{M}\mathbf{v}_1 - \mathbf{v}_1)\| \neq 0$ there exists a constant C' such that*

$$k_{total} \;\leq\; \mathcal{N}C' + \gamma \frac{\log 10}{\log(1/q)} + \sum_{i=0}^{\mathcal{N}-1} \frac{\log(1/t^{(i)})}{\log(1/q)}. \tag{56}$$

Proof. As (54) holds with $C < 1$ we conclude using Lemma 3.1 that inexact inverse iteration converges. Applying Theorem 5.2 and Lemma 5.4 we have for some positive constant C_3

$$\sum_{i=0}^{\mathcal{N}-1} k_{actual}^{(i)} \;\leq\; \sum_{i=0}^{\mathcal{N}-1} \left(C_3 + \log\left(\frac{\|\mathcal{R}\mathbf{b}^{(i)}\|}{t^{(i+1)}} \right) \log(1/q) \right). \tag{57}$$

Now, in case of the first formulation and $\|(\mathbf{W}^{(i)})^{-1}(\mathbf{P}^{-1}\mathbf{M}\mathbf{v}_1 - \mathbf{v}_1)\| = 0$ we can apply (i) of Lemma 5.4 to gain $\|\mathcal{R}\mathbf{b}^{(i)}\| \leq \widetilde{C} t^{(i)}$ for some constant $\widetilde{C}$. Similarly we obtain in case of the second and third formulation $\|\mathcal{R}\mathbf{b}^{(i)}\| \leq \widetilde{C} t^{(i)}$ for some constant $\widetilde{C}$ using Lemma 5.4. Hence we obtain for (i)

$$k_{total} \;\leq\; \mathcal{N}\left(C_3 + \frac{\log \widetilde{C}}{\log(1/q)} \right) + \frac{1}{\log(1/q)} \sum_{i=0}^{\mathcal{N}-1} \log \frac{t^{(i)}}{t^{(i+1)}} \tag{58}$$

from which (55) follows with $C' = C_3 + \log \widetilde{C} / \log(1/q)$ and $\sum \log(t^{(i)}/t^{(i+1)}) = \gamma \log 10$.

Finally in case of the first formulation and $\|(\mathbf{W}^{(i)})^{-1}(\mathbf{P}^{-1}\mathbf{M}\mathbf{v}_1 - \mathbf{v}_1)\| \neq 0$ there exists $\widetilde{C}$ such that $\|\mathcal{R}\mathbf{b}^{(i)}\| \leq \widetilde{C}$ and therefore

$$k_{total} \;\leq\; \mathcal{N}\left(C_3 + \frac{\log \widetilde{C}}{\log(1/q)} \right) + \frac{1}{\log(1/q)} \sum_{i=0}^{\mathcal{N}-1} \log \frac{1}{t^{(i+1)}},$$

from which the claimed results follows by adding and subtracting $\gamma \frac{\log 10}{\log(1/q)}$. ∎

In case (i) the bound on k_{total} consists of two terms the first is linear in the number of outer iterations $\mathcal{N}$ and therefore can be reduced by allowing a faster convergent method for example letting the shift tend towards the sought eigenvalue. The other term in the bound on the total inner iterations k_{total}, depends only on the quality of the eigenvalue approximation. This last contribution reflects the difficulty of the task and is independent of the applied method.

In case (ii) k_{total} the bound consists of three parts, the first two are the same as in case (i). Now the third term in (56) not only depends on the quality of the eigenvalue approximation after each outer iteration it also grows with the number of outer iteration. Therefore, we make the following remark.

Remark 6.2. In order to reduce the total number of inner iterations k_{total}, it is important to reduce the number of outer iterations.

As a converging shift helps to reduce the number of outer iterations but does not lead to unnecessary costs, see Remark 5.3, we advocate the use of variable shift methods like **RQIf**, **RQId**, and **ModRhs**. Which of these methods is most efficient will depend on the applied preconditioner, we illustrate this point in Example 8.3.

7 Right preconditioned GMRES

Up till now we have assumed that left preconditioning was used with GMRES. In this section we briefly discuss the case of right preconditioning applied to the **ModRhs** version of inexact inverse iteration, since this is the most interesting case. What happens when right preconditioning is used for the other versions discussed in this paper is easily deduced to be the same as for left preconditioning and hence is omitted.

The right preconditioned version of **ModRhs** uses $\sigma^{(i)}$ as the Rayleigh quotient and has a fixed tolerance $\tau^{(i)}$. At each outer step i the system

$$(\mathbf{A}-\sigma\mathbf{M})\mathbf{P}^{-1}\widetilde{\mathbf{y}}^{(i)} \quad = \quad \mathbf{P}\mathbf{x}^{(i)} \tag{59}$$

is solved and $\mathbf{y}^{(i)}$ is retrieved from $\mathbf{y}^{(i)}=\mathbf{P}^{-1}\widetilde{\mathbf{y}}^{(i)}$. Here the right-hand side has the form $\mathbf{P}\mathbf{x}^{(i)}$ instead of the standard right-hand side $\mathbf{M}\mathbf{x}^{(i)}$. In cases when only the action of $\mathbf{P}^{-1}$ is known it is possible to find $\mathbf{P}\mathbf{x}^{(i)}$ since, using the standard scaling in step 3 of Algorithm 2,

$$\mathbf{P}\mathbf{x}^{(i)} \;=\; \frac{1}{\varphi(\mathbf{y}^{(i-1)})}\mathbf{P}\mathbf{y}^{(i-1)} \;=\; \frac{1}{\varphi(\mathbf{y}^{(i-1)})}\widetilde{\mathbf{y}}^{(i-1)} \tag{60}$$

with $\widetilde{\mathbf{y}}^{(i)}$ known from (59). Often right preconditioning is prefered over left preconditioning on the basis that in standard linear systems solves the right-hand side is not changed by the preconditioner. However, as one can see from (59) the right preconditioned **ModRhs** does change the right-hand side.

Now, for right preconditioning, the perturbation theorem in Section 9 shows that $(\mathbf{A}-\sigma^{(i)}\mathbf{M})\mathbf{P}^{-1}$ has a simple eigenpair $(\mu_1,\mathbf{w}_1^{(i)})$ with

$$\|\mathbf{w}_1^{(i)}-\mathbf{P}\mathbf{v}_1\| \;\leq\; C|\lambda_1-\sigma^{(i)}| \tag{61}$$

for some constant C, see (76).

We know from the theory in Section 5 and, in particular, from the form of (41) that it is desirable to have $\|\mathcal{R}\mathbf{b}^{(i)}\|\leq Ct^{(i)}$, where $\mathbf{b}^{(i)}$ is the right-hand side of (59). The following lemma shows that this property indeed holds for right preconditioned **ModRhs**.

Lemma 7.1 *With $\mathcal{R}$ defined by* (40) *then*

$$\|\mathcal{R}\mathbf{b}^{(i)}\| \;\leq\; Ct^{(i)} \tag{62}$$

for some positive constant C independent of i.

Proof. From (13) we have

$$\mathcal{R}\mathbf{b}^{(i)} \;=\; c^{(i)}\mathcal{R}\mathbf{P}\mathbf{v}_1 + O(t^{(i)}), \tag{63}$$

as $\alpha^{(i)} = 1$ and so, using (61) and since $\sigma^{(i)}$ is the Rayleigh quotient with $|\lambda_1 - \sigma^{(i)}| \leq O(t^{(i)})$,

$$\mathcal{R}\mathbf{P}\mathbf{b}^{(i)} \;=\; \mathcal{R}\mathbf{w}_1^{(i)} + O(t^{(i)}). \tag{64}$$

Now $\mathcal{R}\mathbf{w}_1^{(i)} = \mathbf{Q}_J(\mathbf{W}^{(i)})^{-1}\mathbf{w}_1^{(i)} = \mathbf{0}$ (as was used in the proof of Lemma 5.4), and so

$$\|\mathcal{R}\mathbf{b}^{(i)}\| \;=\; O(t^{(i)}) \tag{65}$$

which is the required result. □

Hence we expect GMRES to exhibit the same behaviour in right preconditioned **ModRhs** as in the left preconditioned case. Numerical results for right preconditioning are presented in Table 6 in the next section.

8 Numerical Examples

We start by considering some eigenvalue problems constructed to illustrate various points made in previous sections. Later we apply inexact inverse iteration to a more realistic eigenvalue problem, arising from a discretised PDE.

First we consider unpreconditioned GMRES applied to solve the linear systems arising in inverse iteration applied to a standard eigenvalue problem (that is, $\mathbf{M} = \mathbf{I}$ in (1)). According to Remark 5.5 we expect that the number of inner iterations should be roughly constant for linearly converging methods like **FSd**, **RQIf** and **IC**. This is indeed observed as we now describe. In this first example we consider a modified Toeplitz matrix, where 5 entries violate the Toeplitz structure which results in three well separated eigenvalues.

Example 8.1. Let $\mathbf{T}$ be the 400×400 Toeplitz matrix with $\mathbf{T}_{i,i-1} = -0.2$, $\mathbf{T}_{i,i} = 1$, and $\mathbf{T}_{i,i+1} = \mathbf{T}_{i,i+2} = 0.1$ while all other entries are zero. Then we consider the eigenvalue problem $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ with $\mathbf{A}_{i,j} = \mathbf{T}_{i,j}$ for all i, j except $\mathbf{A}_{1,1} = \mathbf{A}_{2,2} = \mathbf{A}_{3,3} = 0.4$ and $\mathbf{A}_{1,2} = 0.2$ while $\mathbf{A}_{1,3} = 0$. The matrix $\mathbf{A}$ has three well separated eigenvalues at 0.311 and $0.425 \pm 0.14i$. The remainder of the spectrum is contained in the triangle with corners at 0.6 and $1.21 \pm 0.18i$.

In Table 1 we present numerical results obtained using unpreconditioned GMRES with algorithms **FSd**, **RQIf** and **IC** described at the end of Section 2, to calculate $\lambda_1 \approx 0.425 + 0.14i$. For all three methods we use the same starting vector $\mathbf{x}^{(0)} = c^{(0)}\mathbf{v}_1 + s^{(0)}\mathbf{V}\mathbf{u}$ where $\mathbf{u}$ is a vector of uniformly distributed random variables in $(0, 1)$ with $\mathbf{u} \perp \mathbf{e}_1$ and $\|\mathbf{u}\| = 1$.

	a) **FSd**		b) **RQIf**		c) **IC**	
	$t^{(i)}$	$k^{(i-1)}_{actual}$	$t^{(i)}$	$k^{(i-1)}_{actual}$	$t^{(i)}$	$k^{(i-1)}_{actual}$
0	1.0e-01		1.0e-01		1.0e-01	
1	1.4e-02	6	1.4e-02	6	6.1e-03	9
2	7.2e-04	8	1.9e-03	7	4.7e-04	12
3	5.7e-05	9	4.6e-05	7	4.0e-05	13
4	5.1e-06	8	3.5e-06	8	3.5e-06	13
5	4.7e-07	8	2.3e-07	7	3.1e-07	13
6	4.1e-08	8	7.6e-09	8	2.8e-08	12
7	2.9e-09	7	8.1e-10	7	2.5e-09	13
8	2.7e-10	8	8.5e-11	7	2.3e-10	12
9	2.4e-11	8	9.1e-12	7	2.1e-11	11
10	3.3e-12	4	1.1e-11	3	1.9e-12	11
$\sum k^{(i-1)}_{actual}$		74		67		119

Table 1: Generalized tangent $t^{(i)}$ and number of inner iterations $k^{(i)}_{actual}$ a) for **FSd** with $\tau_0 = 0.2$ and $\tau_1 = 1$, b) for **RQIf** with $\tau_0 = 0.2$ and c) for **IC** with $\tau_0 = 0.2$ and $\tau_1 = 1$.

In all examples we use GMRES with the inner iteration stopping conditions.

$$\|\mathbf{x}^{(i)} - (\mathbf{A} - \sigma^{(i)}\mathbf{I})\mathbf{y}_k^{(i)}\| \quad \leq \quad \tau^{(i)}. \tag{66}$$

In addition (following Appendix A in [20]) after each inner iteration step we monitor the relative eigenvalue residual and hence stop the whole Algorithm when

$$\|\mathbf{r}^{(i)}\|/|\varrho^{(i)}| \quad \leq \quad tol, \tag{67}$$

where $\mathbf{r}^{(i)}$ is given by (19), and tol is the sought eigenvalue tolerance (in Example 8.1 $tol = 10^{-12}$). This later stopping condition adds to the robustness of inexact inverse iteration and avoids unnecessary work by stopping when the sought eigenvalue accuracy is achieved. Due to the second stopping condition, (67), the linear solve residual will probably not be satisfied in the last outer iteration. As a result the final value of the tangent may be larger than the previous value, see for example $t^{(9)}$ and $t^{(10)}$ for **RQIf** in Table 1.

As predicted by our convergence results all three methods converge linearly. More importantly note that the number of inner iterations is about constant as predicted by Remark 5.5. While this feature for **IC** was observed in [6] and for **FSd** in [12] it has not been previously noted for **RQIf**.

Further we observe that there is only a minor difference in performance between **FSd** and **RQIf**. While the outer rate of convergence for **FSd** cannot be made any faster than $|\lambda_1 - \sigma^{(i)}|/|\lambda_2 - \sigma^{(i)}|$, one can improve the convergence for **RQIf** by reducing τ_0 (see definition of **RQIf** in Section 3) and thereby reduce the total number of inner iterations $k_{total} = \sum k^{(i)}_{actual}$.

Finally, comparing **IC** with **FSd**, in our experience method **IC** always required more inner iterations than method **FSd** for standard eigenvalue problems. However

	a) **FSd**		b) **RQIf**		c) **RQIf**		d) **RQId**	
	$t^{(i)}$	$k_{actual}^{(i-1)}$	$t^{(i)}$	$k_{actual}^{(i-1)}$	$t^{(i)}$	$k_{actual}^{(i-1)}$	$t^{(i)}$	$k_{actual}^{(i-1)}$
0	5.0e-02		5.0e-02		5.0e-02		5.0e-02	
1	1.8e-01	10	2.2e-01	9	8.6e-02	11	3.8e-02	12
2	3.6e-02	11	3.7e-02	11	1.1e-03	16	4.5e-06	24
3	3.2e-03	15	1.9e-03	15	3.3e-06	24	8.9e-11	39
4	4.2e-04	18	1.2e-04	19	6.0e-09	33		
5	3.5e-05	22	6.6e-06	23	9.0e-11	39		
6	4.5e-06	25	4.2e-07	27				
7	6.1e-07	28	2.5e-08	31				
8	8.4e-08	31	1.5e-09	35				
9	1.2e-08	34	9.0e-11	39				
10	9.7e-10	38						
11	1.3e-10	41						
12	5.7e-11	42						
$\sum k_{actual}^{(i-1)}$		315		209		123		75

Table 2: Generalized tangent $t^{(i)}$ and number of inner iterations $k_{actual}^{(i)}$ a) for **FSd** with $\tau_0 = 0.1$ and $\tau_1 = 2.5$, b) for **RQIf** with $\tau_0 = 0.2$ and c) for **RQIf** with $\tau_0 = 0.05$ and d) for **RQId** with $\tau_0 = 0.1$ and $\tau_1 = 0.1$.

this advantage is reversed for generalised eigenvalue problems (see Example 8.2) and when preconditioned GMRES is used as solver for standard eigenvalue problems.

Our next example is a generalized eigenvalue problem. This illustrates that $k_{actual}^{(i)}$ depends on $\log(1/t^{(i+1)})$ for methods using the standard form (6) but that $k_{actual}^{(i)}$ depends on $\log(t^{(i)}/t^{(i+1)})$ for **ModRhs** and **IC** as discussed in Lemma 5.4.

Example 8.2. Let $\mathbf{A}$ be as in Example 8.1 and $\mathbf{M}$ be the Toeplitz matrix with nonzero entries $\mathbf{M}_{i,i} = 1$ and $\mathbf{M}_{i-1,i} = \mathbf{M}_{i+1,i} = 0.1$. The generalized eigenvalue problem $\mathbf{Ax} = \lambda\mathbf{Mx}$ has three well separated eigenvalues at 0.262 and $0.456 \pm 0.123i$ while the remainder of the spectrum is contained in the triangle with corners at 0.5 and $1.5 \pm 0.22i$. In Table 2 we present results for **FSd**, **RQIf** and **RQId** applied to calculate $\lambda_1 = 0.456 + 0.123i$ using unpreconditioned GMRES. Similarly we present in Table 3 results for **ModRhs** and **IC** applied to the same eigenvalue problem. For all results we use the same starting vector and same stopping condition as in Example 8.1.

Table 2 contains numerical results for methods derived from (6), and so Remark 5.6 is relevant. We see that in all cases in Table 2 the number of inner iterations increases with the progress of the outer method. Further we observe that the actual number of inner iterations $k_{actual}^{(i)}$ is related to $\log(1/t^{(i+1)})$. For example, observe that for $t^{(i)} \approx 4.5 \cdot 10^{-6}$ we have $k_{actual}^{(i-1)} \approx 24$ independent of the method and the progress achieved. Also the increase in $k_{actual}^{(i)}$ is linear for the linear converging methods **FSd** and **RQIf**. Finally we note that the method with the fewest outer iterations, namely **RQId** is significantly faster than the other methods, and this is consistent with Remark 6.2.

	a) **ModRhs**		b) **IC**	
	$t^{(i)}$	$k_{actual}^{(i-1)}$	$t^{(i)}$	$k_{actual}^{(i-1)}$
0	5.0e-02		5.0e-02	
1	2.4e-01	6	8.6e-02	11
2	3.0e-02	8	4.2e-03	15
3	1.6e-03	8	2.1e-04	17
4	3.4e-05	11	2.2e-05	16
5	2.0e-06	9	2.0e-06	16
6	1.4e-07	9	3.0e-07	15
7	5.3e-09	9	4.5e-08	15
8	7.1e-10	8	5.9e-09	15
9	1.2e-10	7	1.0e-09	14
10			1.7e-10	14
11			4.1e-11	14
12			1.1e-11	14
$\sum k_{actual}^{(i-1)}$		75		176

Table 3: Generalized tangent $t^{(i)}$ and number of inner iterations $k_{actual}^{(i)}$ a) for **ModRhs** with $\tau_0 = 0.2$ and b) for **IC** with $\tau_0 = 0.1$ and $\tau_1 = 5$

Table 3 contains numerical results for **ModRhs** and **IC**, see equations (9) and (11) respectively, and for which Remark 5.5 applies. In these cases we expect the number of iterations $k_{actual}^{(i)}$ to remain roughly constant and this is indeed observed in Table 3. In contrast to the standard approach the number of inner iterations $k_{actual}^{(i)}$ does not increase when applying **ModRhs** or **IC** as we observe from Table 3. Despite needing 9 outer iterations **ModRhs** needs only the same total number inner iterations as **RQId** which converges in just 3 outer iterations. Comparing **RQIf** and **ModRhs** we clearly observe the advantage of using the modified right hand side with respect to the number of inner iterations needed for each outer iteration. Further for this generalised eigenvalue problem we observe that **IC** is much more efficient than **FSd** and can compete with **RQIf**.

Results similar to Table 3 above can easily be produced using preconditioned GMRES. We do not reproduce these here.

In our next example we highlight the effect the preconditioner has on the right hand side as described in Lemma 5.4 case (ii). There we showed that $\|\mathcal{R}\mathbf{b}^{(i)}\| \leq C + C't^{(i)}$ with $C \leq \|(\mathbf{W}^{(i)})^{-1}(\mathbf{P}^{-1}\mathbf{M}\mathbf{v}_1 - \mathbf{v}_1)\|$ for **RQIf**. We will now consider a family of preconditioners of the form

$$\mathbf{P} \;=\; \mathbf{A} + \delta\mathbf{F} + \varepsilon\mathbf{G}, \tag{68}$$

where

$$\mathbf{F} \;=\; \mathbf{T}\left(\mathbf{I} - \frac{1}{\mathbf{u}_1^T\mathbf{M}\mathbf{v}_1}\mathbf{v}_1\mathbf{u}_1^T\mathbf{M}\right) \quad \text{and} \tag{69}$$

$$\mathbf{G} \;=\; \mathbf{V}(\mathbf{1} - \mathbf{e}_1)\mathbf{u}_1^T\mathbf{M}, \tag{70}$$

where $\mathbf{1}$ denotes the vector with all entries one and $\mathbf{T}$ is the skew symmetric matrix with all entries zeros except $\mathbf{T}_{i,i+1} = -1$ and $\mathbf{T}_{i,i-1} = 1$. The purpose of the $\delta\mathbf{F}$

	RQIf $\varepsilon = 10^{-1}$		**RQIf** $\varepsilon = 10^{-5}$		**RQIf** $\varepsilon = 10^{-9}$	
	$t^{(i)}$	$k^{(i-1)}_{actual}$	$t^{(i)}$	$k^{(i-1)}_{actual}$	$t^{(i)}$	$k^{(i-1)}_{actual}$
0	5.0e-02		5.0e-02		5.0e-02	
1	1.4e-01	11	1.3e-01	6	1.3e-01	6
2	1.4e-02	14	1.4e-03	8	1.4e-03	8
3	5.0e-04	19	6.7e-05	8	1.4e-04	6
4	2.3e-05	24	5.9e-06	12	1.6e-06	9
5	2.4e-07	31	9.3e-08	18	2.4e-08	9
6	2.5e-09	38	2.3e-09	24	5.6e-10	12
7	9.5e-11	43	3.4e-10	27	7.1e-11	15
$\sum k^{(i-1)}_{actual}$		180		103		65

Table 4: Generalised tangent $t^{(i)}$ and number of inner iterations $k^{(i)}_{actual}$ for **RQIf** using preconditioned GMRES with a preconditioner given by (68). In all cases $\delta = 0.2$ and $\tau_0 = 0.2$ while in a) $\varepsilon = 0.1$, b) $\varepsilon = 10^{-5}$, and in c) $\varepsilon = 10^{-9}$.

term is to control the general quality of the preconditioner without perturbing the sought eigenvector, so a larger value of δ makes the preconditioner less effective. **G** is the term of interest as in this case $\|(\mathbf{W}^{(i)})^{-1}(\mathbf{P}^{-1}\mathbf{M}\mathbf{v}_1 - \mathbf{v}_1)\|$ is proportional to ε.

Example 8.3. We consider the eigenvalue problem of Example 8.2 and apply **RQIf** and **ModRhs** using left preconditioned GMRES with preconditioner **P** as given by (68). In Table 4 we present results for **RQIf** with $\tau_0 = 0.2$ using different values ϵ for the preconditioner **P**, while $\delta = 0.2$. Further in Table 5 we present results for **ModRhs** with $\tau_0 = 0.1$ using also different values for ϵ while $\delta = 0.2$.

Table 4 shows that for **RQIf** $k^{(i)}_{actual}$ is about constant as long as $t^{(i)} \geq \varepsilon$. This agrees with the fact that $\|\mathcal{R}\mathbf{b}^{(i)}\|/t^{(i+1)}$ is about constant in this case (see part (ii) in Lemma 5.4), and hence the bound on $k^{(i)}_{actual}$ is constant (see Theorem 5.2). If $t^{(i)} \ll \varepsilon$ then the number of inner iterations grows again like $\log(1/t^{(i+1)})$ as shown in Example 8.2. So a preconditioner with $\mathbf{P}\mathbf{v}_1 \approx \mathbf{M}\mathbf{v}_1$ leads to lower costs for methods using the standard approach.

In Table 5 we consider the same matrix and preconditioner but use method **ModRhs**, now $k^{(i)}_{actual}$ does not depend on the value of ε. However the rate of convergence of **ModRhs** depends on ε, and for values of ε larger than 0.1 the convergences deteriorates. **ModRhs** performs robustly and effectively for small values of ε but it can fail for large values of ε. So, we find that in practice a reasonable preconditioner is required when using **ModRhs**.

As a last example we consider a more realistic eigenvalue problem.

Example 8.4. Consider **A** and **M** derived by discretising

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

using the Galerkin-FEM on regular triangular elements with piecewise linear func-

	ModRhs $\varepsilon = 0.025$		ModRhs $\varepsilon = 0.005$		ModRhs $\varepsilon = 0.001$	
	$t^{(i)}$	$k_{actual}^{(i-1)}$	$t^{(i)}$	$k_{actual}^{(i-1)}$	$t^{(i)}$	$k_{actual}^{(i-1)}$
0	5.0e-02		5.0e-02		5.0e-02	
1	2.1e-01	7	1.5e-01	8	4.0e-02	7
2	3.3e-02	7	9.0e-03	8	1.4e-03	8
3	4.5e-03	8	1.3e-04	10	1.5e-05	9
4	6.1e-04	8	6.5e-06	8	2.5e-07	8
5	7.7e-05	8	1.6e-07	9	9.4e-09	7
6	9.3e-06	8	4.1e-09	9	1.2e-10	10
7	1.1e-06	8	1.9e-10	8		
8	1.3e-07	8	4.5e-11	3		
9	1.6e-08	8				
10	1.9e-09	8				
11	2.3e-10	8				
12	2.7e-11	8				
13	9.8e-12	6				
$\sum k_{actual}^{(i-1)}$		100		63		49

Table 5: Generalised tangent $t^{(i)}$ and number of inner iterations $k_{actual}^{(i)}$ for **ModRhs** using preconditioned GMRES with a preconditioner according to (68). In all cases $\delta = 0.2$ and $\tau_0 = 0.1$ while in a) $\varepsilon = 0.025$, b) $\varepsilon = 0.005$, and in c) $\varepsilon = 0.001$.

tions. This eigenvalue problem is also discussed in [6, 1]. Here we use a 64 by 64 grid which leads to 3969 degrees of freedom. For the discrete eigenvalue problem we have $\lambda_1 \approx 32.2$ and $\lambda_2 \approx 61.8$ with all other eigenvalues satisfying $Re(\lambda_j) > 61.8$. In Table 6 we present numerical results for **RQIf**, **RQId** and **ModRhs** using right preconditioned GMRES.

Using multigrid as right preconditioner $\mathbf{P}$ we have only the action of $\mathbf{P}^{-1}$ on a vector available, but as discussed at the beginning of Section 7 we can readily obtain the action of $\mathbf{P}$ on $\mathbf{x}^{(i)}$ which is needed for right preconditioned **ModRhs**. In the first iteration of **ModRhs** we use the standard right hand side $\mathbf{Mx}$ and ask the linear solver, GMRES, to return the solution $\mathbf{y}$ and the vector $\mathbf{Py}$. Now we can use this additional vector $\mathbf{Py}$ to construct the next right hand side $\mathbf{Px}^{(i+1)} = \mathbf{Py}^{(i)}/\varphi(\mathbf{y}^{(i)})$.

As we expect there are no significant differences between the results for right and for left preconditioned GMRES and so we present results here for right preconditioning. In Table 6 columns a) and b), we see that the number of inner iteration $k_{actual}^{(i)}$ increases again like $\log(1/t^{(i+1)})$ for **RQIf** and **RQId**. From several tests we observe that the parameter $\tau_1 = 1$ in **RQId** leads for this problem to oversolving the linear systems, hence we obtain a slightly higher value for the actual number of inner iterations in **RQId** as compared to **RQIf** to achieve roughly the same tangent.

	RQIf		RQId		ModRhs	
	$\|\|\mathbf{r}^{(i)}\|\|$	$k_{actual}^{(i-1)}$	$\|\|\mathbf{r}^{(i)}\|\|$	$k_{actual}^{(i-1)}$	$\|\|\mathbf{r}^{(i)}\|\|$	$k_{actual}^{(i-1)}$
0	8.1e-03		8.1e-03		8.1e-03	
1	3.0e-05	4	3.1e-05	6	3.0e-05	4
2	4.8e-06	3	4.2e-06	5	9.4e-06	3
3	1.7e-07	4	1.2e-07	7	2.3e-06	4
4	3.5e-09	6	1.0e-10	11	5.1e-08	4
5	9.0e-11	8	5.0e-15	13	8.1e-09	3
6	4.4e-14	12			6.8e-10	4
7	7.1e-15	13			6.8e-11	4
8					6.5e-12	4
9					6.2e-13	4
10					5.9e-14	4
11					5.9e-15	3
$\sum k_{actual}^{i-1}$		50		42		41

Table 6: Norm of the eigenvalue residual $\mathbf{r}^{(i)}$ and number of inner iterations $k_{actual}^{(i)}$ for a) **RQIf** with $\tau_0 = 0.1$, b) **RQId** with $\tau_0 = 0.2$ and $\tau_1 = 1$, and c) **ModRhs** with $\tau_0 = 0.1$, using right-preconditioned GMRES with multigrid as preconditioner. Here we use multigrid as preconditioner with 3 Jacobi-iterations for pre and post smoothing

9 Some Eigenvalue Bounds

In our analysis in Section 5 (specially in equation (37)) we used an upper bound on $|(\lambda_1 - \sigma^{(i)})/\mu_1|$ where μ_1 is the eigenvalue of smallest modulus of the preconditioned shifted linear system. Similarly in the proofs of Lemmas 5.4 and 7.1 we used a perturbation type result for the corresponding eigenvector. In this section we provide proofs of the results we need using a variation of the usual perturbation theory for simple eigenvalues.

As these results are needed for both right and left preconditioned GMRES we consider here the system matrix $\mathbf{P}_L^{-1}(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{P}_R^{-1}$.

For a given shift σ close to λ_1 consider the eigenvalue problem

$$\mathbf{P}_L^{-1}(\mathbf{A} - \sigma\mathbf{M})\mathbf{P}_R^{-1}\mathbf{w} \quad = \quad \mu\mathbf{w}, \tag{71}$$

where μ is the eigenvalue of smallest modulus and $\mathbf{w}$ the corresponding eigenvector. We now rewrite (71) in the equivalent form

$$(\mathbf{A} - \lambda_1\mathbf{M} + (\lambda_1 - \sigma)\mathbf{M})\mathbf{x} \quad = \quad \mu\mathbf{R}\mathbf{x}, \tag{72}$$

where $\mathbf{R} = \mathbf{P}_L\mathbf{P}_R$ and $\mathbf{x} = \mathbf{P}_R^{-1}\mathbf{w}$. Setting $\varepsilon := \lambda_1 - \sigma$, (72) is equivalent to the generalised eigenvalue problem

$$(\mathbf{A} - \lambda_1\mathbf{M} - \mu\mathbf{R} + \varepsilon\mathbf{M})\mathbf{x} \quad = \quad \mathbf{0}. \tag{73}$$

Now we present a perturbation result which uses the implicit function theorem applied to (73) and which is proved in the Appendix.

Theorem 9.1 *Assume* $(\mathbf{v}_1, \lambda_1)$ *is an algebraically simple eigenvalue of* $\mathbf{A}\mathbf{v} = \lambda\mathbf{M}\mathbf{v}$ *with corresponding left eigenvector* $\mathbf{u}_1^H$. *Consider the matrix* $\mathbf{P}_L^{-1}(\mathbf{A} - \sigma\mathbf{M})\mathbf{P}_R^{-1}$ *with* $\mathbf{P}_L$ *and* $\mathbf{P}_R$ *nonsingular and*

$$\mathbf{u}_1^T \mathbf{P}_L \mathbf{P}_R \mathbf{v}_1 \neq 0. \tag{74}$$

Then for $\lambda_1 - \sigma$ *sufficiently small,* $\mathbf{P}_L^{-1}(\mathbf{A} - \sigma\mathbf{M})\mathbf{P}_R^{-1}$ *has an algebraically simple eigenvalue* μ_1 *with*

$$C_1|\lambda_1 - \sigma| \leq |\mu_1| \leq C_2|\lambda_1 - \sigma|, \tag{75}$$

and corresponding eigenvector $\mathbf{w}_1$ *with*

$$\|\mathbf{P}_R^{-1}\mathbf{w}_1 - \mathbf{v}_1\| \leq C_3|\lambda_1 - \sigma| \tag{76}$$

for some positive constants C_1, C_2 *and* C_3. *Additionally*

$$\mu_1 \;=\; (\lambda_1 - \sigma)\frac{\mathbf{u}_1^T \mathbf{M}\mathbf{v}_1}{\mathbf{u}_1^T \mathbf{P}_L \mathbf{P}_R \mathbf{v}_1} + O(\lambda_1 - \sigma)^2, \tag{77}$$

Note that Theorem 9.1 implies that μ_1 is the only eigenvalue of $\mathbf{P}_L^{-1}(\mathbf{A} - \sigma\mathbf{M})\mathbf{P}_R^{-1}$ near 0, and so if the other eigenvalues of $\mathbf{P}_L^{-1}(\mathbf{A} - \sigma\mathbf{M})\mathbf{P}_R^{-1}$ are denoted μ_j, $j = 2, \ldots, n$ we have

$$|\mu_1| \;<\; |\mu_j|, \quad j = 2, \ldots, n. \tag{78}$$

References

[1] J. Berns-Müller and A. Spence, *Inexact inverse iteration and the non-symmetric generalised eigenvalue problem*, SIAM Journal on Matrix Analysis and Application, (2005). submitted.

[2] B. Buffoni and J. Toland, *Analytic Theory of Global Bifurcation*, Princeton University Press, 2003.

[3] M. Embree, *How descriptive are GMRES convergence bounds?*, Tech. Report OUCL Numerical Analysis Group Technical Report 99/08, Oxford University Computing Laboratory, 1999.

[4] B. Fischer and R. W. Freund, *Chebyshev polynomials are not always optimal*, Journal of Approximation Theory, 65 (1991), pp. 261–272.

[5] B. Fischer and F. Peherstorfer, *Chebyshev approximation via polynomial mappings and the convergence behaviour of Krylov subspace methods*, ETNA, 12 (2001), pp. 205–2–15.

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

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

[8] A. Greenbaum, V. Pták, and Z. Strakoš, *Any noninctreasing convergence curve is possible for GMRES*, SIAM Journal on Matrix Analysis and Applications, 17 (1996), pp. 465–469.

[9] P. Griffiths and J. Harris, *Principles of Algebraic Geometry*, John Wiley, 1978.

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

[11] H. B. Keller, *Numerical solution of bifurcation and nonlinear eigenvalue problems*, Academic Press, NewYork, 1977.

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

[13] Y. Notay, *Convergence analysis of inexact Rayleigh quotient iterations*, SIAM Journal on Matrix Analysis and Applications, 24 (2003), pp. 627–644.

[14] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, 1980.

[15] F. Peherstorfer, *Minimal polynomials for compact sets of the complex plane*, Constructive Approximation, 12 (1996), pp. 481–488.

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

[17] U. Rüde and W. Schmid, *Inverse Multigrid Correction for Generalized Eigenvalue Computations*, tech. report, Universität Augsburg, September 1995.

[18] K. Schiefermayr, *Extremal polynomials with respect to the maximum norm on several intervals and arcs in the complex plane*, PhD thesis, Johannes-Kepler Universität Linz, 1997.

[19] D. S. Scott, *Solving sparse symmetric generalized eigenvalue problems without factorization*, SIAM Journal on Numerical Analysis, 18 (1981), pp. 102–110.

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

[21] V. Simoncini and D. B. Szyld, *Theory of inexact Krylov subspace methods and applications to scientific computing*, SIAM Journal on Scientific Computing, 25 (2003), pp. 454–477.

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

A A Perturbation Theorem

Theorem 9.1 in Section 9 follows immediately by putting $\varepsilon = \lambda_1 - \sigma$ and setting $\mu(\varepsilon) = \mu_1$ from the following theorem.

Theorem A.1 *Consider the $n \times n$ generalised eigenvalue problem*

$$(\mathbf{A} - \lambda_1 \mathbf{M})\mathbf{v} - \mu \mathbf{R}\mathbf{v} + \varepsilon \mathbf{M}\mathbf{v} \quad = \quad \mathbf{0} \tag{79}$$

where $\mathbf{A}$, $\mathbf{M}$ and $\mathbf{R}$ are complex matrices. Assume

- *$(\mathbf{v}_1, \lambda_1)$ is an algebraically simple eigenpair of $\mathbf{A}\mathbf{v} = \lambda \mathbf{M}\mathbf{v}$, with $\mathbf{u}_1^H$ the corresponding left eigenvector, that is, $\mathbf{u}_1^H(\mathbf{A} - \lambda_1 \mathbf{M}) = 0$.*
- *$\mathbf{c}$ is a constant normalising vector satisfying $\mathbf{c}^H \mathbf{v}_1 = 1$.*
- *that $\mathbf{u}_1^H \mathbf{R} \mathbf{v}_1 \neq 0$.*

Then, for sufficiently small ε, $(\mathbf{A} - (\lambda - \varepsilon)\mathbf{M})\mathbf{v} = \mu \mathbf{R}\mathbf{v}$ has an algebraically simple eigenpair $(\mathbf{v}(\varepsilon), \mu(\varepsilon))$ with $(\mathbf{v}(0), mu(0)) = (\mathbf{v}_1, 0)$ and

$$|\mu(\varepsilon)| \leq C_2 |\varepsilon|, \quad \|\mathbf{v}(\varepsilon) - \mathbf{v}_1\| \leq C_3 |\varepsilon| \tag{80}$$

for some positive constants C_2 and C_3. Also

$$\mu(\varepsilon) \quad = \quad \varepsilon \frac{\mathbf{u}_1^H \mathbf{M} \mathbf{v}_1}{\mathbf{u}_1^H \mathbf{R} \mathbf{v}_1} + O(\varepsilon^2). \tag{81}$$

and hence there exists a positive constant C_1 such that

$$C_1 |\varepsilon| \quad \leq \quad |\mu(\varepsilon)|. \tag{82}$$

Proof. Define $\mathbf{F} : \mathbb{C}^{n+2} \to \mathbb{C}^{n+1}$ by

$$\mathbf{F}(\mathbf{y}, \varepsilon) \quad := \quad \begin{pmatrix} (\mathbf{A} - \lambda_1 \mathbf{M})\mathbf{v} - \mu \mathbf{R}\mathbf{v} + \varepsilon \mathbf{M}\mathbf{v} \\ \mathbf{c}^H \mathbf{v} - 1 \end{pmatrix}, \qquad \mathbf{y} \; = \; \begin{pmatrix} \mathbf{v} \\ \mu \end{pmatrix}. \tag{83}$$

Clearly $\mathbf{F}$ is an analytic function of $\mathbf{y}$. Now (79) and the first two conditions imply

$$\mathbf{F}(\mathbf{y}_0, 0) = \mathbf{0}, \quad \mathbf{y}_0 = \begin{pmatrix} \mathbf{v}_1 \\ 0 \end{pmatrix}. \tag{84}$$

Also

$$\mathbf{F}_{\mathbf{y}}(\mathbf{y}, \varepsilon) = \begin{pmatrix} (\mathbf{A} - \lambda_1 \mathbf{M}) - \mu \mathbf{R} + \varepsilon \mathbf{M} & -\mathbf{R}\mathbf{v} \\ \mathbf{c}^H & 0 \end{pmatrix} \tag{85}$$

and

$$\mathbf{F}_{\mathbf{y}}(\mathbf{y}_0, 0) = \begin{pmatrix} (\mathbf{A} - \lambda_1 \mathbf{M}) - \mu \mathbf{R} & -\mathbf{R}\mathbf{v}_1 \\ \mathbf{c}^H & 0 \end{pmatrix}. \tag{86}$$

Using Lemma 2.8 (part (ii) with $\nu = 1$) in [11], $\mathbf{F}_{\mathbf{y}}(\mathbf{y}_0, 0)$ is nonsingular if

$$\mathbf{u}_1^T \mathbf{R}\mathbf{v}_1 \neq 0 \tag{87}$$

and

$$\mathbf{c}^H \mathbf{v}_1 \neq 0, \tag{88}$$

which hold by the last two assumptions.

The Implicit Function Theorem for analytic functions (see for example [2, p56] or [9, p19]) shows $\mathbf{F}(\mathbf{y}(\varepsilon), \varepsilon) = \mathbf{0}$ and $\mathbf{y}(\varepsilon)$ is an analytic function of ε. Also $\|\mathbf{y}(\varepsilon) - \mathbf{y}(0)\| \leq C|\varepsilon|$ for some constant C, and so $\|\mathbf{v}(\varepsilon) - \mathbf{v}_1\| \leq C|\varepsilon|$ and $|\mu(\varepsilon)| \leq C|\varepsilon|$ since $\mu(0) = 0$, which proves (80) in the statement of the theorem.

The first row of $\mathbf{F}(\mathbf{y}(\varepsilon), \varepsilon) = \mathbf{0}$ gives

$$(\mathbf{A} - \lambda_1 \mathbf{M})\mathbf{v}(\varepsilon) - \mu(\varepsilon)\mathbf{R}\mathbf{v}(\varepsilon) + \varepsilon \mathbf{M}\mathbf{v}(\varepsilon) = \mathbf{0}. \tag{89}$$

Differentiating with respect to ε gives

$$(-\mu'(\varepsilon)\mathbf{R} + \mathbf{M})\mathbf{v}(\varepsilon) + (\mathbf{A} - \lambda_1 \mathbf{M} - \mu(\varepsilon)\mathbf{R} + \varepsilon \mathbf{M})\mathbf{v}'(\varepsilon) = \mathbf{0}. \tag{90}$$

Setting $\varepsilon = 0$ and multiplying on the left by $\mathbf{u}_1^H$ we obtain, using the first assumption.

$$\mu'(0)\mathbf{u}_1^H \mathbf{R}\mathbf{v}_1 = \mathbf{u}_1^H \mathbf{M}\mathbf{v}_1 \tag{91}$$

So, since $\mathbf{u}_1^H \mathbf{R}\mathbf{v}_1 \neq 0$, and $\mu(0) = 0$,

$$\mu(\varepsilon) = \varepsilon \frac{\mathbf{u}_1^T \mathbf{M}\mathbf{v}_1}{\mathbf{u}_1^T \mathbf{R}\mathbf{v}_1} + O(\varepsilon^2). \tag{92}$$

Since $\mathbf{u}_1^T \mathbf{M}\mathbf{v}_1 \neq 0$, (92) implies that for small enough ε there is a positive constant, C_1 say, such that $C_1|\varepsilon| \leq |\mu(\varepsilon)|$. $\square$