Inner-outer Iterative Methods for Eigenvalue Problems - Convergence and Preconditioning

submitted by

Melina Annerose Freitag

for the degree of Doctor of Philosophy

of the

University of Bath

Department of Mathematical Sciences

September 2007

SUMMARY

Many methods for computing eigenvalues of a large sparse matrix involve shift-invert transformations which require the solution of a shifted linear system at each step. This thesis deals with shift-invert iterative techniques for solving eigenvalue problems where the arising linear systems are solved inexactly using a second iterative technique. This approach leads to an inner-outer type algorithm. We provide convergence results for the outer iterative eigenvalue computation as well as techniques for efficient inner solves. In particular eigenvalue computations using inexact inverse iteration, the Jacobi-Davidson method without subspace expansion and the shift-invert Arnoldi method as a subspace method are investigated in detail.

A general convergence result for inexact inverse iteration for the non-Hermitian generalised eigenvalue problem is given, using only minimal assumptions. This convergence result is obtained in two different ways; on the one hand, we use an equivalence result between inexact inverse iteration applied to the generalised eigenproblem and modified Newton's method; on the other hand, a splitting method is used which generalises the idea of orthogonal decomposition. Both approaches also include an analysis for the convergence theory of a version of inexact Jacobi-Davidson method, where equivalences between Newton's method, inverse iteration and the Jacobi-Davidson method are exploited.

To improve the efficiency of the inner iterative solves we introduce a new tuning strategy which can be applied to any standard preconditioner. We give a detailed analysis on this new preconditioning idea and show how the number of iterations for the inner iterative method and hence the total number of iterations can be reduced significantly by the application of this tuning strategy. The analysis of the tuned preconditioner is carried out for both Hermitian and non-Hermitian eigenproblems. We show how the preconditioner can be implemented efficiently and illustrate its performance using various numerical examples. An equivalence result between the preconditioned simplified Jacobi-Davidson method and inexact inverse iteration with the tuned preconditioner is given.

Finally, we discuss the shift-invert Arnoldi method both in the standard and restarted fashion. First, existing relaxation strategies for the outer iterative solves are extended to implicitly restarted Arnoldi's method. Second, we apply the idea of tuning the preconditioner to the inner iterative solve. As for inexact inverse iteration the tuned preconditioner for inexact Arnoldi's method is shown to provide significant savings in the number of inner solves.

The theory in this thesis is supported by many numerical examples.

ACKNOWLEDGEMENTS

It is a pleasure to thank the many people who have helped me during my time in Bath and made this thesis possible.

First of all I am very grateful to my supervisor, Alastair Spence, for assistance and valuable discussions over the past three years and for all the time and efforts he has put into this work. I thank him for his guidance and encouragements throughout my time as a PhD student.

I also wish to thank the following people: Mark Embree for a very enjoyable short visit to Rice University, Houston, last summer, for his enthusiasm and interest in my work, for introducing me into the world of pseudospectra and to American life style; Howard Elman for financial support which enabled me to visit his Department at the University of Maryland, College Park, last autumn, for his invaluable comments on my research during our discussions and the very pleasant time I had in Maryland, including a wonderful dinner with his family; Valeria Simoncini for her interest in my work and more than very helpful discussions during her visits to Bath and Manchester.

I am grateful to many more people for various discussions and fruitful comments on my work, those include Ivan Graham, Dianne O'Leary, Rich Lehoucq, Beresford Parlett, Miloud Sadkane, Hubert Schwetlick, Pete Stewart, Daniel Szyld and Andy Wathen. I thank both Valeria Simoncini and Ivan Graham for reading this thesis with great care. Their suggestions and remarks have certainly improved this work.

I further thank the entire Numerical Analysis group at Bath University, including Bill Morton, for creating a nice work environment. Special thanks go to the Badminton crowd, especially Darrel, Ann, Mark, Damien, Jason and Fran for several very enjoyable matches (and hopefully many more to come).

I thank Simone for bearing my complaints in my first year and him and his family for being such wonderful hosts in Italy. I also thank my other office mates and friends Andy, Dave, Laura, Patrick, Stefano and Zhivko and all other PG students, especially Matt, Nathan and Richard for creating an enjoyable working atmosphere and for sometimes delaying this thesis. I thank André, Martin, Soumyadip and many more friends, including my housemates over the years for their support and friendship, Bob and his friends for company in Maryland and Karolin and Robin and many others in Germany for keeping me up to date.

I thank EPSRC and the Department of Mathematical Sciences for financial support of the research that led to this thesis and for funding several visits to national and international conferences, including the USA, Germany and the Czech Republic. I also thank the always friendly staff and computer support team at the Department for keeping systems running smoothly.

Schließlich möchte ich mich bei meiner Familie, besonders bei meinen Eltern, für ihre Unterstützung während meiner Studienzeit bedanken.

“Alles Wissen und alles Vermehren unseres Wissens endet nicht mit einem Schlusspunkt, sondern mit einem Fragezeichen.”

Hermann Hesse (1877 - 1962), German author.

Für meine Eltern.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

NOTATION

The list below gives an overview of the notation frequently used in this thesis accompanied by a brief explanation and a page number of its definition or major occurence. In general, we write scalars (real or complex numbers) with greek letters or italic type and vectors as well as matrices in boldface roman type.

Symbol	**Definition**	**Page**
$\mathbf{A}$:	$n \times n$ matrix	1
$\mathbf{A}^H$:	Conjugate transpose of matrix $\mathbf{A}$	1
λ_i:	Eigenvalues of a matrix, $i = 1, \ldots, n$	1
$\mathbf{x}_i$:	Eigenvectors of a matrix, $i = 1, \ldots, n$	1
$\mathbb{C}$:	The set of complex numbers	1
$\mathbb{C}^n$:	The set of complex n dimensional vectors	1
$\mathbb{C}^{n \times n}$:	The set of complex $n \times n$ matrices over $\mathbb{C}$	1
rank():	The rank of a matrix	1
im():	The image of a matrix	1
$\mathcal{X}$:	Subspace	2
$\rho(\mathbf{x}^{(i)})$:	Rayleigh quotient for a vector $\mathbf{x}^{(i)}$	6
σ:	scalar value for a shift in a shift-invert method	6
$\mathcal{K}_i(\mathbf{A}, \mathbf{x})$:	Krylov subspace of dimension i	7
$\mathbf{x}^{(i)}$:	ith eigenvector iterate obtained from an iterative method	14
$\lambda^{(i)}$:	ith eigenvalue iterate obtained from an iterative method	15
$\mathbf{J}(\mathbf{z}^{(i)})$:	Jacobian for Newton's method i	15
$\mathbf{r}^{(i)}$:	eigenvalue residual at step i	17
$\lambda^{(0)}$:	starting guess for an eigenvalue	18
$\mathbf{x}^{(0)}$:	starting guess for an eigenvector	18
$\mathbf{d}^{(i)}$:	linear system residual at outer step i	18
$\mathcal{B}(\mathbf{z}^*, r)$:	Ball around $\mathbf{z}^*$ with radius r	19
Lip_γ:	Lipschitz continuous with Lipschitz constant γ	19

$\tau^{(i)}$:	solve tolerance for inexact solves	19
$\cos(\mathbf{x}_1, \mathbf{x}_2)$:	cosine of the angle between the vectors $\mathbf{x}_1$ and $\mathbf{x}_2$	22
$\mathbf{P}$:	$n \times n$ matrix preconditioner	27
$\mathbb{P}_i$:	$n \times n$ practical tuned matrix preconditioner at step i	33
$\phi(\mathbf{y}^{(i)})$:	scalar normalisation function	44
$\operatorname{sep}(\mathbf{A}, \mathbf{B})$:	separation function	49
$\mathbf{Q}^{(i)}$:	projection matrix	62
$\mathcal{P}^{\perp}$:	orthogonal projection	72
$\kappa^1(\mathbf{A})$:	reduced condition number of $\mathbf{A}$	72
$k^{(i)}$:	number of inner iterations k per outer iteration i	74
$\mathbb{P}$:	$n \times n$ ideal tuned matrix preconditioner	75
$\mathbb{L}$:	$n \times n$ Cholesky factor of ideal positive definite preconditioner $\mathbb{P}$	74
$\mathbb{L}_i$:	$n \times n$ Cholesky factor of tuned positive definite preconditioner $\mathbb{P}_i$	77
$\Pi_{1/2}$:	projection used in Jacobi-Davidson method	120
$\mathcal{P}$:	oblique projector	131
$\mathcal{R}(\mathbf{A})$:	range of a matrix $\mathbf{A}$	131
$\mathcal{N}(\mathbf{A})$:	nullspace of a matrix $\mathbf{A}$	131
$\mathbb{T}$:	ideal tuning matrix	136
$\mathbb{T}_i$:	approximate tuning operator	137
$\mathbf{H}_k$:	upper Hessenberg matrix of size k	162
$\mathbf{Q}_k$:	orthogonal matrix spanning Krylov subspace basis of dimension k	162
$\kappa(\mathbf{A})$:	condition number of $\mathbf{A}$	214
$\mathcal{F}(\mathbf{A})$:	field of values (numerical range) of a matrix $\mathbf{A}$	214
$\nu(\mathbf{A})$:	numerical radius of a matrix $\mathbf{A}$	214
$\Lambda_\varepsilon(\mathbf{A})$:	pseudospectra of a matrix $\mathbf{A}$	215
F_k:	Faber polynomials of degree k	216
$D(z_0, r)$:	disk with center z_0 and radius r	218
$E(z_0, d, a)$:	ellipse with center z_0, focal distance d and major semi axis a	219

PUBLICATIONS

Most of the work of this thesis has been published or is submitted for publication. Chapter 2 of this thesis has appeared in

> M. A. Freitag and A. Spence, *Convergence rates for inexact inverse iteration with application to preconditioned iterative solves*, BIT, 47 (2007), pp. 27-44.

and Chapter 3 has appeared in

> M. A. Freitag and A. Spence, *Convergence theory for inexact inverse iteration applied to the generalised nonsymmetric eigenproblem*, Electron. Trans. Numer. Anal., 28 (2007), pp. 40-64.

Parts of Chapter 4 are going to appear in

> M. A. Freitag and A. Spence, *A tuned preconditioner for inexact inverse iteration applied to Hermitian eigenvalue problems*, to appear in IMA J. Numer. Anal., `doi:10.1093/imanum/drm036`.

and Chapter 5 is going to appear in:

> M. A. Freitag and A. Spence, *Rayleigh quotient iteration and simplified Jacobi-Davidson method with preconditioned iterative solves*, to appear in Linear Algebra Appl.

Shorter versions of Chapters 6 and 7 are in preparation for submission:

> M. A. Freitag, A. Spence and E. Vainikko, *Tuning for Rayleigh quotient iteration applied to the nonsymmetric eigenproblem and comparison to Jacobi-Davidson*, in preparation.

> M. A. Freitag and A. Spence, *Inexact preconditioned Arnoldi's method and implicit restarts for eigenvalue computations*, in preparation.

CHAPTER 1

Introduction

1.1 The large sparse eigenproblem

Eigenvalues and eigenvectors of a given linear operator $\mathcal{A}$ arise in many areas of applied mathematics and the ability to approximate these quantities numerically is important in a wide variety of applications such as structural dynamics, quantum chemistry, electrical networks, control theory and material science. Furthermore eigenvalues and eigenvectors arise in the stability analysis of linear and nonlinear systems. A recent application is the search engine Google [52], which uses the eigenvector corresponding to the eigenvalue one for an extremely large sparse stochastic matrix. The increasing number of applications has fueled the development of new methods and software for the numerical solution of large scale algebraic eigenvalue problems. These include the widely used Arnoldi package ARPACK [80] and the Jacobi-Davidson package JDQR/JDQZ [39]. For a survey on software we refer to [58].

Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be an n by n matrix, the finite representation of the operator $\mathcal{A}$, $\mathbf{x} \in \mathbb{C}^n$ a column vector and $\lambda \in \mathbb{C}$ a scalar, such that

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x} \quad \text{with} \quad \mathbf{x} \neq \mathbf{0}. \tag{1.1}$$

Definition 1.1. *If (1.1) holds, λ is called an eigenvalue of $\mathbf{A}$ and $\mathbf{x}$ is called a (right) eigenvector. If $\mathbf{w}^H\mathbf{A} = \lambda\mathbf{w}^H$ then we call $\mathbf{w}$ a left eigenvector. The full set of eigenvalues of a matrix $\mathbf{A}$ is called the spectrum of $\mathbf{A}$ and denoted by $\Lambda(\mathbf{A})$.*

We call (1.1) a Hermitian eigenproblem if $\mathbf{A} = \mathbf{A}^H$. Furthermore we speak of a generalised eigenproblem if $\mathbf{A}\mathbf{x} = \lambda\mathbf{M}\mathbf{x}$, where $\mathbf{M}$ is an n by n matrix. For further classifications we refer to [4].

The concept of an eigenvector may be generalised to invariant subspaces.

Definition 1.2. *A subspace $\mathcal{S}$ is called an invariant subspace of $\mathbf{A}$ if $\mathbf{A}\mathcal{S} \subset \mathcal{S}$.*

Hence, if $\mathbf{x}$ is an eigenvector of $\mathbf{A}$ then $\operatorname{span}\{\mathbf{x}\}$ is a one-dimensional invariant subspace of $\mathbf{A}$. Note that if there exist $\mathbf{X} \in \mathbb{C}^{n\times k}$, $\mathbf{B} \in \mathbb{C}^{k\times k}$ with $\operatorname{rank}(\mathbf{X}) = k$ and $\mathbf{A}\mathbf{X} = \mathbf{X}\mathbf{B}$, then $\mathcal{S} = \operatorname{im}(\mathbf{X})$ is an invariant subspace of $\mathbf{A}$.

This introductory chapter is organised as follows. Section 1.2 gives a general overview of numerical methods for eigenproblems. We state the difference between direct and iterative methods and show why direct methods are infeasible for large, sparse eigenproblems. In Section 1.3 we introduce the shift-invert transformation and the concept of inner-outer iterations. Section 1.4 gives an overview of some iterative methods fore eigencomputations and Section 1.5 introduces preconditioned iterative solves for linear systems. Finally, in Section 1.6 we give an outline of the thesis.

1.2 A survey of numerical methods for eigenproblems

The problem (1.1) corresponds to finding the zeroes of the characteristic polynomial $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$ of $\mathbf{A}$. It is well-known that for $n \geq 5$ there is no expression for the roots of this polynomial for a general $\mathbf{A}$; therefore, determining the exact eigenvalues is generally not possible. However, there are many numerical methods that give very good approximations to eigenvectors and hence eigenvalues of a given matrix. Eigenvalue problems of moderate size, which means that a full $n \times n$ matrix can be stored conveniently, are often solved by direct methods, by which we mean methods where similarity transformations are applied until the eigenvalue estimates can be easily found.[1] The best known algorithm is the QR-Algorithm, based on the QR decomposition of a matrix (see, for example, [4] or Demmel [23] and Golub and Van Loan [48]), which is also implemented in the MATLAB function `eig`. The QR algorithm approximates the whole spectrum and the number of iterations needed is of order $\mathcal{O}(n^3)$, where n is the size of the matrix, which becomes very large for large problems.

There is another disadvantage of the QR method. If matrices are sparse, that is, the number of non-zero elements is small compared to the number of zero entries, and the matrix is structured, then the QR method generates matrices in which the sparse structure of the original matrix disappears. This leads to fill-in and an increasing storage requirement as the algorithm proceeds.

In many applications it is not necessary to calculate the complete eigenvalue decomposition of a matrix. Often only a few eigenvalues are of interest, which gives rise to faster, iterative methods. By iterative methods we mean methods based on matrix-vector multiplications using the original sparse matrix so that the sparse matrix storage and structure can be used to advantage. Hence, subspace algorithms are suitable for large sparse matrices. All subspace algorithms have the following structure in common:

1. Generate a sequence of subspaces $\mathcal{S}_1, \mathcal{S}_2, \ldots$
2. For each subspace $\mathcal{S}_i$ of dimension i construct a matrix $\mathbf{H}_i \in \mathbb{C}^{i\times i}$ which is the restriction and projection of $\mathbf{A}$ onto the subspace $\mathcal{S}_i$.

The matrices $\mathbf{H}_i$ are usually constructed with the Rayleigh-Ritz procedure, which can be described as projecting and restricting the full matrix $\mathbf{A}$ onto the subspace. Then the eigenvalues of the projected matrix are called Ritz values which are approximations to the wanted part of the spectrum. The corresponding eigenvectors of $\mathbf{A}$ are called Ritz vectors and they represent approximations to the exact eigenvectors of $\mathbf{A}$.

Different subspace methods are distinguished from the way the subspaces are generated. We can work with subspaces of both fixed and variable dimension. If the

[1] Note the slight abuse of terminology: For matrix size $n \geq 5$ all methods for eigenvalue computations are iterative. We distinguish between direct methods where similarity transforms are used and iterative methods, by which we mean methods based on matrix-vector multiplications using the original sparse matrix so that the sparse matrix storage and structure can be used to advantage.

dimension of the subspace is fixed to one then the most common methods obtained are the power method and Rayleigh quotient iteration (see for example Parlett [101] and Wilkinson [151] for details). The power method can be extended to subspaces with higher, but fixed dimension, where it is called subspace or simultaneous iteration which can be seen as a block power method. For details see Parlett [101], Demmel [23], Golub and Van Loan [48], Stewart [135] and Saad [110].

A further class of subspace methods is the one that uses nested subspaces of increasing dimension. Usually one starts with a subspace of dimension one and increases this dimension by one at each iteration step. Among the most popular of these methods are the Lanczos method (see Lanczos [76]) for symmetric matrices and the Arnoldi method (see Arnoldi [3]) for nonsymmetric matrices. These methods are Krylov subspace methods. More details on Arnoldi and Lanczos methods can be found in Demmel [23], Golub and Van Loan [48], Saad [110], [16], Bai et al. [4] and Trefethen and Bau [144]. The methods of Lanczos and Arnoldi have lead to the development of many other algorithms. For example, both methods can be generalised to block Lanczos and block Arnoldi algorithms, by working with p-dimensional subspaces instead of vectors. The iteration starts with a p-dimensional subspace and the dimension is increased by p at each step.

Since Lanczos and Arnoldi use subspaces of increasing dimension, storage and computing time is increased during the algorithm. In order to overcome this disadvantage, restarted Lanczos and Arnoldi methods have been developed, where a new starting vector is used at some stage of the iteration. A significant improvement over the standard Arnoldi method is the implicitly restarted Arnoldi or Lanczos method with exact shifts (see Sorensen [130], [131] and Watkins [149]). This polynomial filtering method is executed in the following way: After m Arnoldi steps m Ritz values are available, k of those are chosen to be wanted and $m-k$ as unwanted. A polynomial is constructed with the unwanted Ritz values as zeros, and then this polynomial in $\mathbf{A}$, applied to the previous starting vector, will have no components in the direction of the unwanted Ritz vectors.

There exist further subspace algorithms with increasing subspace dimension, where the subspace is expanded without using Krylov subspaces. A Newton iteration step, or an approximate Newton iteration step can be applied to obtain a new direction. Examples for this approach are the Davidson method and the Jacobi-Davidson method, see [124] and [63] for details.

A short discussion of basic iterative methods for eigenvalue problems together with numerical examples is provided in Appendix A.

1.3 The shift-invert transformation and inner-outer iterations

In this section we introduce the concept of inner-outer iterative methods which arise when the iterative methods discussed in the previous section are applied to the shift-invert transformation.

Subspace algorithms for eigenvalue problems usually approximate a small number of extreme eigenvalues of a matrix $\mathbf{A}$ only. Therefore, the computation of interior eigenvalues usually requires a so-called shift-invert strategy in conjunction with a subspace method.

Suppose we would like to compute the eigenvalues in a certain region near some target value σ. If we shift by σ and take the inverse we get a new matrix $(\mathbf{A}-\sigma\mathbf{I})^{-1}$ that has the same eigenvectors and invariant subspaces as $\mathbf{A}$, but different eigenvalues. Each eigenvalue λ of $\mathbf{A}$ corresponds to an eigenvalue $(\lambda-\sigma)^{-1}$ of $(\mathbf{A}-\sigma\mathbf{I})^{-1}$ and so the largest eigenvalues of $(\mathbf{A}-\sigma\mathbf{I})^{-1}$ correspond to the eigenvalues of $\mathbf{A}$ that are closest to σ. Hence, if we apply any of the above subspace algorithms to $(\mathbf{A}-\sigma\mathbf{I})^{-1}$ instead of $\mathbf{A}$ we approximate invariant subspaces that belong to eigenvalues close to σ.

Another spectral transformation that is often used to accelerate the convergence of eigencomputations is the Cayley transform [4], where two scalars μ and ν are chosen and the transformed eigenvalue problem $\mathbf{A}_C\mathbf{x}=\gamma\mathbf{x}$ with $\mathbf{A}_C=(\mathbf{A}-\mu\mathbf{I})^{-1}(\mathbf{A}-\nu\mathbf{I})$ and $\gamma=(\lambda-\mu)^{-1}(\lambda-\nu)$ is considered.

Both the shift-invert strategy and the Cayley transform involve a solve

$$(\mathbf{A}-\sigma\mathbf{I})\mathbf{p}=\mathbf{q}, \tag{1.2}$$

for a vector $\mathbf{p}\in\mathbb{C}^n$ at each iteration step of any of the subspace algorithms considered above, where $\mathbf{q}\in\mathbb{C}^n$ is given. If the shift σ is constant and the matrices are not large one can solve the arising system via an LU-factorisation. However, in this thesis we consider methods appropriate for large, sparse matrices where a direct solve of (1.2) via some factorisation might be very expensive. Also we want to exploit the structure of the matrices and therefore seek to solve the systems (1.2) iteratively using only matrix-vector products, since they can exploit the sparse structure of the matrix $\mathbf{A}$. Typical iterative methods for solving linear systems are the conjugate gradient method (CG) for symmetric positive definite systems, the minimum residual method (MINRES) for symmetric indefinite systems and the generalised minimum residual method (GMRES) for nonsymmetric matrices, although many other iterative methods are available [5, 111]. For a brief discussion on iterative methods for solving linear systems we refer to Appendix A. The convergence of Krylov subspace methods for linear systems depends on many factors. For symmetric solvers like CG and MINRES the error (or the residual) can be bounded in terms of the condition number of the system matrix whereas for GMRES applied to nonsymmetric matrices non-normality of the system matrix plays a role. The GMRES convergence bound depends both on the non-normality and the eigenvalue distribution of the system matrix. If the matrix is close to normal and has an eigenvalue distribution such that a polynomial of moderate degree and value one at the origin can be made small at all the eigenvalues the method converges fast. In order to improve the convergence of MINRES/CG or GMRES a preconditioner $\mathbf{P}$ is applied such that $\mathbf{A}\mathbf{P}^{-1}\approx\mathbf{I}$ and hence a faster convergence rate is achieved. For more details we refer to Appendix B.s

Using iterative methods the system (1.2) will only be solved to a certain tolerance, leading to a so-called "inexact" solve. Hence, if a shift-invert transformation is used within an iterative method for eigencomputations, then the matrix-vector multiplication arising at each step of this method will be carried out inexactly due to an iterative solve of the linear system.

This discussion gives rise to the term inner-outer iterative method, since we have to distinguish between two iterative methods. By the outer iterative method we mean the subspace method for the eigenvalue computation. The inner part is then the (inexact) preconditioned iterative solution of the linear system (1.2). In the following chapters we typically use the variable i to denote the iteration number for the outer iteration and the letter k to count the inner iteration though we change this usage in Chapter 7.

In order to investigate inner-outer iterative methods for large sparse matrix problems, two main issues are to be considered. These are

1. The convergence rate of the outer iterative subspace method if the inner solves are carried out inexactly. Typical questions are:
   - How does the choice of the tolerance for the solution of the inner linear systems affect the rate of convergence of the outer method to the invariant subspace?
   - Is it possible to fix or even relax the tolerance to which the inner linear system is solved and still obtain accuracy for the computed eigenspace within some prescribed tolerance?
2. The efficiency of the inner solves. In this case important questions are:
   - What is the convergence rate for the inner iterations?
   - Can the number of inner iterations be reduced without significantly affecting the number of outer iterations?
   - How does one choose the preconditioner for the inner iteration?

Overall, we try to optimise an inner-outer iterative method by maximising the convergence rate of the outer iteration and minimising the total number of iterations.

This thesis focuses on both the above questions.

In the following sections we give short introductions and reviews of the iterative methods that we consider in this thesis and of preconditioners for linear systems. Inexact inverse iteration (presented in Section 1.4.1) computes the solution to (1.1) where the dimension of the subspace $\mathcal{S}_i$ is one. The behaviour of this vector iteration is the foundation for any other subspace methods, such as the Jacobi-Davidson method (see Section 1.4.2) and shift-invert Arnoldi's method (presented in Section 1.4.3). Section 1.5 gives an introduction to preconditioned iterative solves.

1.4 Iterative methods for computing eigenvalue

1.4.1 Inexact inverse iteration

Inexact methods for large sparse eigenvalue problems have received considerable attention in the recent literature. The early papers were concerned with one-dimensional subspace methods, such as inexact inverse iteration and inexact Rayleigh quotient iteration.

Consider the computation of one specific simple eigenpair $(\lambda_1, \mathbf{x}_1)$ of the eigenproblem $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ using inverse iteration, which obtains a new eigendirection by applying $(\mathbf{A} - \sigma\mathbf{I})^{-1}$ to $\mathbf{x}^{(i-1)}$, an approximate eigenvector, namely

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

and then rescaling the vector to obtain $\mathbf{x}^{(i)} = \mathbf{y}^{(i)}/\|\mathbf{y}^{(i)}\|$. The scalar $\sigma \in \mathbb{C}$ is a shift which is close to the sought eigenvalue. Details of this Algorithm can be found in Appendix A, Section A.1.1. This algorithm is merely the application of the power method to $(\mathbf{A} - \sigma\mathbf{I})^{-1}$ and involves repeated solution of the linear system

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

for some $\sigma \in \mathbb{C}$. For exact solves of this system convergence of inverse iteration has been proved to be linear with a convergence rate proportional to $|\lambda_1 - \sigma|/|\lambda_2 - \sigma|$, assuming that the fixed shift σ is chosen such that $|\lambda_1 - \sigma| < |\lambda_2 - \sigma| \leq \ldots |\lambda_n - \sigma|$ (see Parlett [100] and [101] for the symmetric problem). For variable Rayleigh quotient shifts $\sigma^{(i)} = \mathbf{x}^{(i)^H} \mathbf{A}\mathbf{x}^{(i)}$ (with the method being the so-called Rayleigh quotient iteration) the convergence is generally quadratic and, for Hermitian $\mathbf{A}$, it is even locally cubic (see Ostrowski [97]). However, if the arising linear system is solved inexactly, extra conditions have to be employed to ensure convergence. In this case we have

$$(\mathbf{A} - \sigma\mathbf{I})\mathbf{y} - \mathbf{x} = \mathbf{d}, \quad \text{where} \quad \|\mathbf{d}\| \leq \tau,$$

where τ is the solve tolerance of the inner iteration and the method is called inexact inverse iteration. One of the earliest papers in this area is [106], which contains results on inexact inverse iteration with the conjugate gradient method as an inner solver. More recently inexact inverse iteration has been considered in [50], [75], [128] and [129]. Those papers investigate the convergence theory of inexact inverse iteration and inexact Rayleigh quotient iteration, [128] and [129] consider the symmetric eigenproblem only, the results in [50] and [75] assume the problem is diagonalisable. Furthermore, very detailed investigations and generalisations of those earlier results were provided in [57], [119], [9], [10] and [11]. These works also give the first insight into the inner iteration, i.e. the iterative solve of the inner linear systems for the inner-outer iterative method. Other related papers on the topic of inexact eigenvalue solvers include [72], , [91], [92], [73] and [95]. An extension to inexact subspace iteration, that, is a block form of inexact inverse iteration has recently been given in [104], where both the outer convergence theory and an insight to the inner iteration was given. All results have one theme in common, namely that in order to achieve convergence in inexact inverse iteration, either the solve tolerance of the inner solver has to be reduced (if a fixed shift is chosen), or, for a small enough fixed solve tolerance, a variable shift approaching the eigenvalue gives convergence. This can mainly be explained by the fact that inverse iteration is merely a form of Newton method and it is well known that inexact Newton method achieves linear (or quadratic) convergence, depending on the accuracy of the inexact solve (see, for example [71]). We explore this connection in Chapter 2. The thesis contains a convergence theory for inexact inverse iteration in the most general case in Chapter 3 and also gives an extensive analysis of the iterative solves for the arising inner system (see Chapter 4 for the Hermitian problem and Chapter 6 for the non-Hermitian problem). In practice, subspace methods like shift-invert (restarted) Arnoldi (see Chapter 7) and Jacobi-Davidson are more likely to be used in eigenvalue computations, though inexact inverse iteration has proved to be a useful tool in improving estimates obtained from inexact shift-invert Arnoldi's method with very coarse tolerances, see [54]. Also, the behaviour of inexact inverse iteration is the foundation for other subspace methods, such as the Jacobi-Davidson method (see Section 1.4.2) and shift-invert Arnoldi's method (see Section 1.4.3)

1.4.2 The Jacobi-Davidson method

As we shall see, inverse iteration is closely related to (a simplified version of) the Jacobi-Davidson method, an iterative method proposed by Sleijpen and van der Vorst [124] which we now explain. Let $(\lambda_1, \mathbf{x}_1)$ be an eigenpair of $\mathbf{A}$ and let $(\theta, \mathbf{x})$ be an approximate eigenpair with unit vector $\mathbf{x}$. Further let the residual be given by $\mathbf{r} = \mathbf{A}\mathbf{x} - \theta\mathbf{x}$. Then the correction pair $(\delta, \mathbf{y})$ with the correction $\mathbf{y}$ orthogonal to $\mathbf{x}$ should ideally satisfy

$$\mathbf{A}(\mathbf{x} + \mathbf{y}) = (\theta + \delta)(\mathbf{x} + \mathbf{y}), \quad \text{and} \quad \mathbf{y} \perp \mathbf{x},$$

which is equivalent to

$$(\mathbf{A} - \theta\mathbf{I})\mathbf{y} = -\mathbf{r} + \mathbf{x}\delta + \mathbf{y}\delta \quad \text{and} \quad \mathbf{y} \perp \mathbf{x}.$$

Consider the projection of this equation onto the orthogonal complement of the current approximate eigenvector $\mathbf{x}$ by multiplying the equation by $(\mathbf{I} - \mathbf{x}\mathbf{x}^H)$. Then, using $\mathbf{y}^H\mathbf{x} = 0$ we get

$$(\mathbf{I} - \mathbf{x}\mathbf{x}^H)(\mathbf{A} - \theta\mathbf{I})(\mathbf{I} - \mathbf{x}\mathbf{x}^H)\mathbf{y} = -\mathbf{r} + \mathbf{y}\delta.$$

Neglecting the second order term $\mathbf{y}\delta$ we obtain the so-called correction equation

$$(\mathbf{I} - \mathbf{x}\mathbf{x}^H)(\mathbf{A} - \theta\mathbf{I})(\mathbf{I} - \mathbf{x}\mathbf{x}^H)\mathbf{y} = -\mathbf{r} \tag{1.3}$$

for the Jacobi-Davidson method. The correction $\mathbf{y}$ is then used to expand the search space. Starting with a subspace of dimension one, which contains only the eigenvector approximation $\mathbf{x}$, the dimension of the search space is increased by one at each step using the modified Gram-Schmidt algorithm applied to the current search space and the correction $\mathbf{y}$. This process is called subspace expansion.

If no subspace expansion is applied then we obtain a simplified version of the Jacobi-Davidson method. Further, if in that case the correction equation is solved exactly then the simplified Jacobi-Davidson method is equivalent to both inverse iteration and Newton's method applied to the eigenproblem (subject to some normalisation).

In practice the Jacobi-Davidson method is designed to use preconditioned iterative solves in order to solve the correction equation (1.3). Convergence analysis is usually carried out by equivalence results with Rayleigh quotient iteration or Newton's method. Convergence theory for Jacobi-Davidson applied to the Hermitian eigenproblem has been given in [147] and, for a special inner solver, in [93]. Furthermore, equivalence results between a simplified version of Jacobi-Davidson method and Newton's method for exact solves have been pointed out in [94, 124–126]. Jacobi-Davidson methods for generalised eigenproblems have been investigated in [123]. For further details on the Jacobi-Davidson method we refer to Appendix A and the original paper by Sleijpen and van der Vorst [124].

Relations between the convergence of inexact inverse iteration and the simplified Jacobi-Davidson method are analysed in Chapter 3. An equivalence result between preconditioned versions of inexact Rayleigh quotient iteration and the inexact simplified Jacobi-Davidson method is shown in Chapter 5.

1.4.3 The inexact Arnoldi method and implicitly restarted Arnoldi method

The Arnoldi process [3] or, in the special case of Hermitian matrices, Lanczos' process [76], is a method that constructs an orthonormal basis for the k-dimensional Krylov subspace

$$\mathcal{K}_k(\mathbf{A}, \mathbf{q}) = \text{span}\{\mathbf{q}, \mathbf{A}\mathbf{q}, \mathbf{A}^2\mathbf{q}, \ldots, \mathbf{A}^{k-1}\mathbf{q}\}. \tag{1.4}$$

The resulting Arnoldi relation is given by

$$\mathbf{A}\mathbf{Q}_k = \mathbf{Q}_k\mathbf{H}_k + \mathbf{q}_{k+1}\mathbf{h}_{k+1,k}\mathbf{e}_k^H,$$

where the columns of $\mathbf{Q}_k$ form an orthonormal basis for the Krylov subspace (1.4) and $\mathbf{H}_k = \mathbf{Q}_k{}^H\mathbf{A}\mathbf{Q}_k$, the Rayleigh-Ritz projection of $\mathbf{A}$ onto $\mathcal{K}_k(\mathbf{A},\mathbf{q})$, is an upper Hessenberg matrix of size k. If the Arnoldi process breaks down, that is $\mathbf{h}_{k+1,k} = 0$, we have found an invariant subspace and since $\mathbf{A}\mathbf{Q}_k = \mathbf{Q}_k\mathbf{H}_k$ the eigenvalues of $\mathbf{H}_k$ are eigenvalues of $\mathbf{A}$. If $\mathbf{h}_{k+1,k} \neq 0$ we get at least the following error estimate for the Ritz values: If $(\theta, \mathbf{u})$ ($\|\mathbf{u}\|_2 = 1$) is an eigenpair of $\mathbf{H}_k$ obtained by the Rayleigh-Ritz procedure applied to $\mathbf{A}$ and $\mathbf{Q}_k$ then, with $\mathbf{z} = \mathbf{Q}_k\mathbf{u}$, we have

$$\|\mathbf{A}\mathbf{z} - \theta\mathbf{z}\|_2 = |\mathbf{h}_{k+1,k}||\mathbf{u}(k,1)|,$$

where $\mathbf{u}(k,1)$ denotes the last component of $\mathbf{y}$. For a bound on the rate of convergence of Arnoldi's method we refer to [110] (see also [107] for the special case of Lanczos method). Since one does not know how large the Krylov subspace has to be in order to obtain good convergence of the eigenvalues of $\mathbf{H}_k$ to the eigenvalues of $\mathbf{A}$, we consider two possible ways of accelerating the method: spectral transformation and implicit restarts.

If one replaces $\mathbf{A}$ by $(\mathbf{A}-\sigma\mathbf{I})^{-1}$ then convergence to the eigenvalues close to σ is more rapid. However, for large sparse $\mathbf{A}$ this shift-invert strategy results in inexact methods since the multiplication by $(\mathbf{A} - \sigma\mathbf{I})^{-1}$ is replaced by a preconditioned iterative solve. Therefore the basis vectors of the Krylov subspace are obtained inexactly: they are still orthonormal but they do not built a Krylov subspace any more. Recent insights into inexact Arnoldi iterations have been given by several numerical experiments carried out in [14], where it has been observed that one can relax the tolerance for inexact solves and still achieve convergence. Further investigations into these relaxation strategies have then been carried out by Simoncini [118] and Golub et al. [51]. The approach in [118] is motivated by ideas for inexact Krylov subspace methods for linear systems (see, for example [120, 121, 148]) and theoretically justifies the results obtained in [14]. A different approach in terms of inexact solves for the shifted linear system has been taken in [142], where the spectral transformation is approximated by a fixed-polynomial operator which is computed prior to the Arnoldi iteration. Finally, we note that rational Krylov sequence methods are methods where the Krylov subspace (1.4) is constructed using $(\mathbf{A} - \sigma^{(i)}\mathbf{I})^{-1}$ instead of $\mathbf{A}$ and where the shift $\sigma^{(i)}$ is varied at every step. Ruhe [105] shows how to build an orthogonal basis for the rational Krylov sequence subspace, that is a subspace of the form

$$\mathrm{span}\{\mathbf{q}, (\mathbf{A} - \sigma^{(1)}\mathbf{I})^{-1}\mathbf{q}, \ldots, (\mathbf{A} - \sigma^{(k-1)})^{-1}\ldots(\mathbf{A} - \sigma^{(1)}\mathbf{I})^{-1}\mathbf{q}\}.$$

Inexact solves within the rational Krylov method were considered by Lehoucq and Meerbergen [78].

For Arnoldi's method restarting is generally needed to reduce storage requirements and orthogonalisation costs. Arnoldi's method is successively restarted with a modified starting vector $\mathbf{q}_1$ using eigenvector information from the Krylov subspace previously obtained. With the better starting vector the convergence of Arnoldi's method is hoped to be improved. Polynomial restart methods update the starting vector with $\tilde{\mathbf{q}}_1 = \Psi(\mathbf{A})\mathbf{q}_1$, where $\Psi(\mathbf{A})$ is a filter polynomial which filters out the unwanted part of the spectrum. One of the easiest methods uses a linear combination of the previous Ritz vectors [108] or Chebychev polynomials on ellipses, which minimise the polynomial in the unwanted part of the spectrum [109]. A straightforward approach which is also called the exact shift approach, is to select the unwanted eigenvalues of the current $\mathbf{H}_k$ as roots of Ψ, suggested by Sorensen [130]. Also in this paper an implicit restarting strategy (IRA) was proposed, which implements polynomial restarting by applying a sequence of p implicit updates to an m step Arnoldi factorisation, where $m = k + p$. The factorisation is reduced back to order k via a QR factorisation with p implicit shifts. Details on implicitly restarted Arnoldi's method are given in Chapter 7. In that chapter we consider inexact solves for both Arnoldi's method and implicitly restarted Arnoldi's method. Inexact solves with IRA have also been considered in [142].

1.5 Iterative solves for linear systems and preconditioning

Inner-outer iterative methods require the solution of a linear system

$$\mathbf{B}\mathbf{z} = \mathbf{b}, \tag{1.5}$$

where for our problems $\mathbf{B} := \mathbf{A} - \sigma\mathbf{I}$ and $\mathbf{b}$ is a given right-hand side. The most common iterative methods compute an approximate solution $\mathbf{z}_k$ in a subspace $\mathbf{z}_0 + \mathcal{K}_k$ of dimension k by imposing the Petrov-Galerkin condition

$$\mathbf{b} - \mathbf{B}\mathbf{z}_k \perp \mathcal{L}_k,$$

where $\mathcal{L}_k$ is another subspace of dimension k. If $\mathcal{K}_k := \mathcal{K}_k(\mathbf{B}, \mathbf{b})$ is a Krylov subspace (see (1.4)), then these methods are called Krylov subspace methods. The choice of $\mathcal{L}_k$ then determines the methods used. For $\mathcal{L}_k = \mathcal{K}_k$, the Full Orthogonalisation Method (FOM) is obtained (which becomes the Conjugate Gradient method (CG) for Hermitian problems). For the projection based on taking $\mathcal{L}_k = \mathbf{B}\mathcal{K}_k$ we obtain the Generalised Minimum Residual Method (GMRES) and the Minimum Residual Method (MINRES) for Hermitian systems. Many more variations of Krylov subspace methods exist, we refer the reader to [55] and [111].

If the coefficient matrix $\mathbf{B}$ is close to the identity or close to normal with eigenvalues tightly clustered around some point away from the origin, which usually does not hold in practice, all these algorithms converge fast (see Appendix A for an explanation). Hence, if we apply a preconditioner $\mathbf{P}$ so that $\mathbf{B}\mathbf{P}^{-1}$ or $\mathbf{P}^{-1}\mathbf{B}$ is close to the identity we obtain a modified system

$$\mathbf{B}\mathbf{P}^{-1}\tilde{\mathbf{z}} = \mathbf{b}, \quad \mathbf{P}^{-1}\tilde{\mathbf{z}} = \mathbf{z} \quad \text{or} \quad \mathbf{P}^{-1}\mathbf{B}\mathbf{z} = \mathbf{P}^{-1}\mathbf{b}.$$

Since the convergence of the iterative method depends on the condition number and eigenvalue clustering of the system matrix (see Appendix B for details), $\mathbf{P}$ should be chosen such that $\mathbf{B}\mathbf{P}^{-1}$ (and $\mathbf{P}^{-1}\mathbf{B}$) is some approximation of the identity so that Krylov solvers will converge quickly. For symmetric matrices $\mathbf{P}$ should be chosen such that the condition number of $\mathbf{B}\mathbf{P}^{-1}$ decreases and for nonsymmetric matrices $\mathbf{B}\mathbf{P}^{-1}$ should be close to normal and have eigenvalues clustered away from the origin. Note that often $\mathbf{P}$ is not constructed explicitly and only the application of $\mathbf{P}^{-1}$ to $\mathbf{A}$ is known.

In this thesis we consider preconditioners where (1.5) arises from an eigenvalue problem, hence $\mathbf{B}$ can be almost singular and $\mathbf{b}$ is some approximation to the eigenvector of the system. It turns out that preconditioners $\mathbf{P}$ with a small rank change of the usual preconditioner are advantageous usually due to the special right hand side of the system. We remark that dependence on the right hand side for an iterative solver, namely GMRES, has also been considered in [81], where a convection-diffusion model problem was investigated with different right hand sides arising from a change in boundary conditions. Preconditioners with a small rank change have previously been considered by Vuik, Nabben and others [37, 40, 88, 89]. In [40] so-called deflation based preconditioners for symmetric linear systems have been examined which project out an unwanted subspace corresponding to small eigenvalues and therefore reducing the condition number. [88, 89] give a comparison of the deflation based preconditioner to the balancing preconditioner, a similar preconditioner with a rank change but which projects unwanted parts of the spectrum of $\mathbf{B}$ onto one and [37] gives an extension of those results to the nonsymmetric problem using GMRES. Note that our motivation for preconditioners with a rank one change is different. We consider the special right-hand side of the system and furthermore the preconditioner we use is nonsingular, whilst the work discussed by Vuik et al. uses projections which are singular.

1.6 Structure of this thesis

This thesis gives new results on both the convergence theory of certain inexact methods for eigenvalue problems and the efficiency of the inner iterative solves.

In particular we improve the existing convergence theory for inexact inverse iteration in two different ways using, first a modified Newton method in Chapter 2 and, second, a splitting approach which generalises the orthogonal decomposition introduced in [101] (Chapter 3). Furthermore we show how the convergence theory of Jacobi-Davidson method can be interpreted as an inexact Newton method and as inexact inverse iteration, which extends the existing results and fills some gaps in the present theory. In addition we extend the available convergence results for inexact shift-invert Arnoldi's method for finding an eigenvector to invariant subspaces and to an inexact version of implicitly restarted Arnoldi's method (Chapter 7).

Concerning the second question in Section 1.3 of efficient inner solves, we introduce a new tuned preconditioner which gives significant improvement on the standard preconditioner. Using convergence analysis of Krylov methods for linear systems, we show that both for Hermitian and for non-Hermitian eigenproblems the tuned preconditioner reduces the total number of inner iterations in comparison to the standard preconditioner (Chapters 4 and 6). We also show how this tuning strategy for the preconditioner is comparable to a simplified preconditioned Jacobi-Davidson method (Chapter 5). Finally, we show how the tuned preconditioner also gives a considerable saving in the number of total iterations for Arnoldi's method and implicitly restarted Arnoldi method (Chapter 7).

In particular, we have the following structure:

First, in *Chapter 2* we give an analysis of inexact inverse iteration applied to nonsymmetric generalised eigenproblems $\mathbf{Ax} = \lambda\mathbf{Mx}$ using modified Newton's method. Convergence rates for inexact inverse iteration with variable shift for the calculation of an algebraically simple eigenvalue are obtained. In particular, it is shown that if the inexact solves are carried out with a tolerance chosen proportional to the eigenvalue residual then quadratic convergence is achieved. Furthermore a simple version of inexact Jacobi-Davidson is shown to be equivalent to inexact Newton method. We also show how modifying the right hand side in inverse iteration still provides a convergent method, but the rate of convergence will be quadratic only under certain conditions on the right hand side. We discuss the implications of this for the preconditioned iterative solution of the linear systems. Finally we introduce a new preconditioner which is a simple modification to the usual preconditioner, but which has advantages both for the standard form of inverse iteration and for the version with a modified right hand side.

Chapter 3 then provides a different account of the convergence theory of inexact inverse iteration for a generalised eigenproblem using a splitting method, which generalises the orthogonal decomposition in Parlett [101]. This theory is very general and requires few assumptions on $\mathbf{A}$ and $\mathbf{M}$ and extends the theory currently in the literature. In particular, there is no need for $\mathbf{A}$ to be diagonalisable or for $\mathbf{M}$ to be symmetric positive definite or even nonsingular, as was required in earlier approaches in the literature. The theory includes both fixed and variable shift strategies, and the bounds obtained are improvements on those currently in the literature. In addition, the analysis developed here is used to provide a convergence theory for a version of inexact simplified Jacobi-Davidson's method.

Chapter 4 investigates the computation of an eigenvalue and corresponding eigenvector of a large sparse Hermitian positive definite matrix using either inexact inverse iteration with a fixed shift or inexact Rayleigh quotient iteration. The large sparse linear systems arising at each iteration are solved approximately by means of symmetrically preconditioned MINRES. We consider preconditioners based on the incomplete Cholesky factorisation and derive a new tuned Cholesky preconditioner which shows considerable improvement over the standard preconditioner. This improvement is analysed using the convergence theory for MINRES. We also compare the spectral properties of the tuned preconditioned matrix with those of the standard preconditioned matrix. In particular, we provide both a perturbation result and an interlacing result, and these results show that the spectral properties of the tuned preconditioner are similar to those of the standard preconditioner. For Rayleigh quotient shifts, comparison is also made with a technique introduced by Simoncini and Eldén [119] which involves changing the right hand side of the inverse iteration step.

In *Chapter 5* we show that, for the non-Hermitian eigenvalue problem, simplified Jacobi-Davidson with preconditioned iterative solves is equivalent to inexact inverse iteration where the preconditioner is altered by a simple rank one change. This extends existing equivalence results to the case of preconditioned iterative solves.

Chapter 6 considers the case of preconditioning the non-Hermitian generalised eigenproblem. We discuss inexact inverse iteration with a fixed shift and Rayleigh quotient shifts. The convergence theory of GMRES as the iterative method used for the solution of the inner system is provided in Appendix B. The performance of the method is measured in terms of the number of inner iterations needed at each outer solve. For both unpreconditioned and preconditioned GMRES it is shown that the number of inner iterations increases as the outer iteration proceeds. We derive a tuning strategy for the generalised eigenproblem, and show how a rank one change to the preconditioner produces savings in overall costs while the solve tolerances for the inner solver are reduced or while the shift converges to the sought eigenvalue.

In *Chapter 7* we consider the computation of a few eigenvectors and corresponding eigenvalues in an isolated cluster around a given shift using shift-and-invert Arnoldi's method with and without implicit restarts. For the inner iterations we use GMRES as the iterative solver. The costs of the inexact solves are measured by the number of inner iterations needed by the iterative solver at each outer step of the algorithm. We first extend the relaxation strategy developed by Simoncini [118] to implicitly restarted Arnoldi's method which yields an improvement in the overall costs of the method. Secondly, we apply a new preconditioning strategy to the inner solver. We show that small rank changes of the preconditioner can produce significant savings in the total number of iterations.

Chapter 8 summarises the contributions of the thesis and suggests some areas of further work.

Finally, Appendix A gives some background on iterative methods for eigenvalue computations and iterative methods for linear systems. Appendix B deals with a general iterative method, which is applied to most of the nonsymmetric linear systems arising in the inner iteration of most of the considered shift-invert eigenvalue subspace methods. For the inner solution the most general method we use is GMRES, the generalised minimum residual method for nonsymmetric linear systems. This short appendix summarises partly well-known convergence bounds for GMRES, and also derives convergence theory for MINRES and CG, derived as special cases. Appendix C contains some basic theory for the eigenvalues and eigenvectors of perturbed matrices. The results in Appendix C summarise some well-known perturbation theory for simple eigenvectors.

We remark that the chapters have been designed to be read independently as well as sequentially. Please note that, in general, constants arising within one chapter are independent from the ones in other chapters. All chapters contain various numerical examples to illustrate the theoretical results and computations have been carried out in MATLAB.

CHAPTER 2

Convergence of inexact inverse iteration using Newton's method with application to preconditioned iterative solves

2.1 Introduction

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

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{M}\mathbf{x}, \quad \mathbf{x} \neq \mathbf{0},$$

using inverse iteration with iterative solves of the resulting linear systems

$$(\mathbf{A} - \sigma\mathbf{M})\mathbf{y} = \mathbf{M}\mathbf{x}. \tag{2.1}$$

Here σ is a complex shift chosen to be close to the desired eigenvalue. We call this method "inexact inverse iteration", since the linear system is solved to some prescribed tolerance only. It is well known that, using exact solves, inverse iteration achieves linear convergence with a fixed shift and quadratic convergence for a Rayleigh quotient shift (see [101] and [100]). For more information about inverse iteration we refer to the classic articles [49] and [103], and the more recent survey [66].

In [119] it was noted that Rayleigh quotient iteration can be related to Newton's method on a Grassmann manifold (see [29]). Wu et al. [152] considered several inexact Newton preconditioning techniques for large eigenproblems. For inexact inverse iteration applied to the nonsymmetric eigenvalue problem we refer to [75] and [50] for a fixed shift, and [12] for a variable shift strategy. In many of these papers the convergence analysis is based on eigenvector expansions and convergence is determined by looking at a (generalised) tangent of the error in the desired eigendirection. Often, in such accounts the norm of a matrix of all the eigenvectors arises in the convergence analysis and in error bounds, which is a drawback to the approach.

In this chapter a completely different and novel approach to the analysis for variable shifts is used which provides a much simpler interpretation, and also suggests a way of analysing preconditioned iterative solves when the right hand side is modified as in [119]. We show that inexact inverse iteration is a modified Newton's method and hence obtain a convergence analysis for inexact inverse iteration applied to the calculation of an algebraically simple eigenvalue. In addition, the approach here suggests a "tuning" strategy for the preconditioner that works well in numerical examples.

The plan of this chapter is as follows. Section 2.2 contains a review of some known results about Newton's method and inverse iteration. The main theory of this chapter is contained in Section 2.3 where the convergence results for inexact inverse iteration applied to the generalised eigenvalue problem are obtained. In Section 2.4 a comparison of inexact inverse iteration to a simplified Jacobi-Davidson algorithm is carried out using an inexact Newton approach. In Section 2.5 we discuss how to maintain quadratic convergence for a version of inexact inverse iteration where the right-hand side is modified to improve the performance of a preconditioned iterative solver. We illustrate this theory by introducing a "tuned" ILU preconditioner which is a simple rank one modification of the standard preconditioner. This tuned preconditioner turns out to have a significantly improved performance over the standard ILU preconditioner in several different numerical examples. In the later Chapters 4 and 6 this new tuning strategy for inexact inverse iteration will be discussed in detail.

Throughout this chapter we use $\|\mathbf{z}\|=\|\mathbf{z}\|_\infty$ unless otherwise stated.

2.2 Inverse iteration and Newton's method

It is well-known that inverse iteration can be formulated as a Newton method. This was first done in [146] but was then rediscovered in [103] and [138] (see also [19]). Tapia et al [141] give an extension of the results in [103] for the symmetric eigenproblem, where convergence of order $1+\sqrt{2}$ is derived for the projected Newton method (which coincides with Newton's method if a special normalisation is used as we will see in this section). A more recent paper [152] gives a summary of different Newton methods for eigenvalue problems, amongst them the use of the augmented system which we discuss in this section. We revise the convergence theory briefly for a generalised eigenvalue problem.

Let $\mathbf{A}$ and $\mathbf{M}$ be real or complex $n \times n$ matrices, and consider the generalised eigenvalue problem

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{M}\mathbf{x}, \quad \lambda \in \mathbb{C},\, \mathbf{x} \in \mathbb{C}^n. \tag{2.2}$$

Assume that $(\mathbf{x}_1, \lambda_1)$ is an algebraically simple finite eigenpair of (2.2) with $\mathbf{u}_1^H$ the corresponding left eigenvector, so that,

$$\mathbf{u}_1^H \mathbf{M}\mathbf{x}_1 \neq 0. \tag{2.3}$$

Also, for some non-zero constant vector $\mathbf{c} \in \mathbb{C}^n$ assume the normalisation

$$\mathbf{c}^H \mathbf{x}_1 = 1. \tag{2.4}$$

One version of inverse iteration is given by Algorithm 1.
Note that from steps (2) and (3) of Algorithm 1, $\mathbf{c}^H\mathbf{x}^{(i+1)} = 1\ \forall i$ and hence $\mathbf{c}^H\Delta\mathbf{x}^{(i)} = 0$, where $\Delta\mathbf{x}^{(i)} = \mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}$.

Now, let us introduce the nonlinear system $\mathbf{F}(\mathbf{z}) = \mathbf{0}$ where, for $\mathbf{z} := (\mathbf{x}^T, \lambda)^T$,

$$\mathbf{F}(\mathbf{z}) = \begin{bmatrix} (\mathbf{A} - \lambda\mathbf{M})\mathbf{x} \\ \mathbf{c}^H\mathbf{x} - 1 \end{bmatrix}. \tag{2.5}$$

Algorithm 1 Inverse Iteration as Newton's Method

Input: $\lambda^{(0)}$ and $\mathbf{x}^{(0)}$ with $\mathbf{c}^H\mathbf{x}^{(0)} = 1$, i_{max}.
  for $i = 1, \ldots, i_{max}$ **do**
    Solve $(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{M}\mathbf{x}^{(i)}$,
    Set $\Delta\lambda^{(i)} = \dfrac{1}{\mathbf{c}^H\mathbf{y}^{(i)}}$; $\quad \lambda^{(i+1)} = \lambda^{(i)} + \Delta\lambda^{(i)}$,
    Update $\mathbf{x}^{(i+1)} = \Delta\lambda^{(i)}\mathbf{y}^{(i)}$,
    Evaluate $\mathbf{r}^{(i+1)} = (\mathbf{A} - \lambda^{(i+1)}\mathbf{M})\mathbf{x}^{(i+1)}$,
    Test for convergence
  end for
Output: $\lambda^{(i_{max})}$, $\mathbf{x}^{(i_{max})}$.

Then the steps in Algorithm 1 may be rewritten in the following block matrix form

$$\begin{bmatrix} (\mathbf{A} - \lambda^{(i)}\mathbf{M}) & -\mathbf{M}\mathbf{x}^{(i)} \\ \mathbf{c}^H & 0 \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x}^{(i)} \\ \Delta\lambda^{(i)} \end{bmatrix} = \begin{bmatrix} -(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{x}^{(i)} \\ 0 \end{bmatrix}, \tag{2.6}$$

with

$$\begin{bmatrix} \mathbf{x}^{(i+1)} \\ \lambda^{(i+1)} \end{bmatrix} = \begin{bmatrix} \mathbf{x}^{(i)} + \Delta\mathbf{x}^{(i)} \\ \lambda^{(i)} + \Delta\lambda^{(i)} \end{bmatrix}. \tag{2.7}$$

Equations (2.6) and (2.7) are merely Newton's method applied to (2.5), namely,

$$\mathbf{J}(\mathbf{z}^{(i)})\Delta\mathbf{z}^{(i)} = -\mathbf{F}(\mathbf{z}^{(i)}), \quad \mathbf{z}^{(i+1)} = \mathbf{z}^{(i)} + \Delta\mathbf{z}^{(i)}, \tag{2.8}$$

where $\mathbf{J}(\mathbf{z}^{(i)})$ denotes the Jacobian

$$\mathbf{J}(\mathbf{z}^{(i)}) = \begin{bmatrix} (\mathbf{A} - \lambda^{(i)}\mathbf{M}) & -\mathbf{M}\mathbf{x}^{(i)} \\ \mathbf{c}^H & 0 \end{bmatrix}. \tag{2.9}$$

Lemma 2.1. *If $\mathbf{z}_1 = (\mathbf{x}_1^T, \lambda_1)^T$ is an algebraically simple finite eigenpair of (2.2) then under (2.4), $\mathbf{J}(\mathbf{z}_1)$ is nonsingular.*

Proof. Lemma 2.8 in [70] (see also [53, Lemma 3.1]) shows that if $\operatorname{rank}(\mathbf{A} - \lambda_1\mathbf{M}) = n - 1$, and if (2.3) and (2.4) hold, then $\mathbf{J}(\mathbf{z}_1)$ is nonsingular. □

Note that one can obtain explicit bounds on the norm of $\mathbf{J}(\mathbf{z}_1)^{-1}$ as discussed in [1], [114] and [115, page 6] where the bounds depend on $\xi^{-1} = |\mathbf{u}_1^H\mathbf{M}\mathbf{x}_1|^{-1}$. Lemma 2.2 characterises the norm of the inverse of the bordered matrix $\mathbf{J}(\mathbf{z}_1)$ in terms of the vectors $\mathbf{c}$ and $\mathbf{M}\mathbf{x}_1$. Note that $\|\cdot\|_2$ and $\|\cdot\|_\infty$ are equivalent [48] and hence it is sufficient to describe $\|\mathbf{J}(\mathbf{z}_1)^{-1}\|_2$.

Lemma 2.2. *Let $\mathbf{J}(\mathbf{z}_1)$ be given by*

$$\mathbf{J}(\mathbf{z}_1) := \begin{bmatrix} \mathbf{A} - \lambda_1\mathbf{M} & -\mathbf{M}\mathbf{x}_1 \\ \mathbf{c}^H & 0 \end{bmatrix},$$

where $\mathbf{c}^H\mathbf{x}_1 = 1$ and $\mathbf{u}_1^H\mathbf{M}\mathbf{x}_1 = \xi \neq 0$ and let $\|\cdot\| := \|\cdot\|_2$. Then

$$\|\mathbf{J}(\mathbf{z}_1)^{-1}\| \leq \frac{\|\mathbf{u}_1\|\|\mathbf{x}_1\|}{|\xi|}\sqrt{\|(\mathbf{A} - \lambda_1\mathbf{M})^\dagger\|^2\|\mathbf{M}\mathbf{x}_1\|^2\|\mathbf{c}\|^2 + \|\mathbf{M}\mathbf{x}_1\|^2 + \|\mathbf{c}\|^2}, \tag{2.10}$$

where $(\mathbf{A} - \lambda_1\mathbf{M})^\dagger$ denotes the Moore-Penrose pseudo-inverse of $\mathbf{A} - \lambda_1\mathbf{M}$.

Proof. The proof is similar to the one given in [115, page 6]. Let the singular value decomposition of $\mathbf{A} - \lambda_1 \mathbf{M}$ be given by

$$\mathbf{A} - \lambda_1 \mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{X}^H = [\mathbf{U}_1 \, c_1 \mathbf{u}_1] \begin{bmatrix} \mathbf{\Sigma}_1 & \mathbf{0} \\ \mathbf{0}^H & 0 \end{bmatrix} [\mathbf{X}_1 \, c_2 \mathbf{x}_1]^H,$$

where $\mathbf{\Sigma}_1 = \operatorname{diag}(\sigma_1, \ldots, \sigma_{n-1}) \in \mathbb{C}^{(n-1)\times(n-1)}$ is nonsingular (since λ_1 is a simple eigenvalue) and $\mathbf{U} \in \mathbb{C}^{n\times n}$, $\mathbf{X} \in \mathbb{C}^{n\times n}$ are unitary matrices. The constants c_1 and c_2 were introduced in order to assure the normalisation $\|c_1 \mathbf{u}_1\| = 1$ and $\|c_2 \mathbf{x}_1\| = 1$. Clearly, $c_1 = 1/\|\mathbf{u}_1\|$ and $c_2 = 1/\|\mathbf{x}_1\|$. Using this decomposition we may write

$$\widehat{\mathbf{J}(\mathbf{z}_1)} = \begin{bmatrix} \mathbf{U}^H & \mathbf{0} \\ \mathbf{0}^H & 1 \end{bmatrix} \mathbf{J}(\mathbf{z}_1) \begin{bmatrix} \mathbf{X}^H & \mathbf{0} \\ \mathbf{0}^H & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{\Sigma}_1 & \mathbf{0} & -\mathbf{U}_1^H \mathbf{M}\mathbf{x}_1 \\ \mathbf{0}^H & 0 & -c_1 \mathbf{u}_1^H \mathbf{M}\mathbf{x}_1 \\ \mathbf{c}^H \mathbf{X}_1 & c_2 \underbrace{\mathbf{c}^H \mathbf{x}_1}_{=1} & 0 \end{bmatrix}.$$

We know from Lemma 2.1, that $\mathbf{J}(\mathbf{z}_1)$ and hence $\widehat{\mathbf{J}(\mathbf{z}_1)}$ is nonsingular. Block Gaussian elimination gives

$$\widehat{\mathbf{J}(\mathbf{z}_1)}^{-1} = \begin{bmatrix} \mathbf{\Sigma}_1^{-1} & -\mathbf{\Sigma}_1^{-1} \dfrac{\mathbf{U}_1^H \mathbf{M}\mathbf{x}_1}{c_1 \mathbf{u}_1^H \mathbf{M}\mathbf{x}_1} & \mathbf{0} \\ -\dfrac{\mathbf{c}^H \mathbf{X}_1}{c_2} \mathbf{\Sigma}_1^{-1} & \dfrac{\mathbf{c}^H \mathbf{X}_1}{c_2} \mathbf{\Sigma}_1^{-1} \dfrac{\mathbf{U}_1^H \mathbf{M}\mathbf{x}_1}{c_1 \mathbf{u}_1^H \mathbf{M}\mathbf{x}_1} & \dfrac{1}{c_2} \\ \mathbf{0}^H & -\dfrac{1}{c_1 \mathbf{u}_1^H \mathbf{M}\mathbf{x}_1} & 0 \end{bmatrix} =: \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & 0 \end{bmatrix},$$

where

$$\mathbf{K}_{11} = \begin{bmatrix} \mathbf{I} \\ -\dfrac{\mathbf{c}^H \mathbf{X}_1}{c_2} \end{bmatrix} \mathbf{\Sigma}_1^{-1} \begin{bmatrix} \mathbf{I} & -\dfrac{\mathbf{U}_1^H \mathbf{M}\mathbf{x}_1}{c_1 \mathbf{u}_1^H \mathbf{M}\mathbf{x}_1} \end{bmatrix}, \quad \mathbf{K}_{12} = \begin{bmatrix} \mathbf{0} \\ \dfrac{1}{c_2} \end{bmatrix}$$

$$\text{and} \quad \mathbf{K}_{21} = \begin{bmatrix} \mathbf{0}^H & -\dfrac{1}{c_1 \mathbf{u}_1^H \mathbf{M}\mathbf{x}_1} \end{bmatrix}.$$

We then have

$$\|\mathbf{J}(\mathbf{z}_1)^{-1}\| = \|\widehat{\mathbf{J}(\mathbf{z}_1)}^{-1}\| \leq \sqrt{\|\mathbf{K}_{11}\|^2 + \|\mathbf{K}_{12}\|^2 + \|\mathbf{K}_{21}\|^2}. \tag{2.11}$$

Clearly $\|\mathbf{K}_{12}\| = \dfrac{1}{c_2}$ and $\|\mathbf{K}_{21}\| = \dfrac{1}{c_1 |\mathbf{u}_1^H \mathbf{M}\mathbf{x}_1|}$. Furthermore we may apply a unitary decomposition to $\mathbf{c}$ and $\mathbf{M}\mathbf{x}_1$, since $\mathbf{U} \in \mathbb{C}^{n\times n}$, $\mathbf{X} \in \mathbb{C}^{n\times n}$ are unitary matrices:

$$\begin{aligned} \mathbf{c} &= \mathbf{X}\mathbf{X}^H \mathbf{c} = \mathbf{X}_1 \mathbf{X}_1^H \mathbf{c} + c_2 \mathbf{x}_1 (c_2 \mathbf{x}_1^H \mathbf{c}) = \mathbf{X}_1 \mathbf{X}_1^H \mathbf{c} + c_2 \mathbf{x}_1 c_2 \\ \mathbf{M}\mathbf{x}_1 &= \mathbf{U}\mathbf{U}^H \mathbf{M}\mathbf{x}_1 = \mathbf{U}_1 \mathbf{U}_1^H \mathbf{M}\mathbf{x}_1 + c_1 \mathbf{u}_1 (c_1 \mathbf{u}_1^H \mathbf{M}\mathbf{x}_1) \end{aligned}$$

These unitary decompositions show that

$$\left\| \begin{bmatrix} \mathbf{I} \\ -\dfrac{\mathbf{c}^H \mathbf{X}_1}{c_2} \end{bmatrix} \right\| = \sqrt{1 + \frac{\|\mathbf{c}^H \mathbf{X}_1\|^2}{|c_2|^2}} = \frac{\|\mathbf{c}\|}{c_2},$$

and

$$\left\| \begin{bmatrix} \mathbf{I} & -\dfrac{\mathbf{U}_1^H \mathbf{M}\mathbf{x}_1}{c_1 \mathbf{u}_1^H \mathbf{M}\mathbf{x}_1} \end{bmatrix} \right\| = \sqrt{1 + \frac{\|\mathbf{U}_1^H \mathbf{M}\mathbf{x}_1\|^2}{\|c_1 \mathbf{u}_1^H \mathbf{M}\mathbf{x}_1\|^2}} = \frac{\|\mathbf{M}\mathbf{x}_1\|}{c_1 |\mathbf{u}_1^H \mathbf{M}\mathbf{x}_1|},$$

and hence

$$\|\mathbf{K}_{11}\| \leq \frac{\|\mathbf{\Sigma}_1^{-1}\|\|\mathbf{c}\|\|\mathbf{M}\mathbf{x}_1\|}{c_1c_2|\mathbf{u}_1^H\mathbf{M}\mathbf{x}_1|} = \frac{\|(\mathbf{A}-\lambda_1\mathbf{M})^\dagger\|\|\mathbf{c}\|\|\mathbf{M}\mathbf{x}_1\|}{c_1c_2|\mathbf{u}_1^H\mathbf{M}\mathbf{x}_1|}.$$

Putting these results together and using (2.11) as well as the definitions of c_1 and c_2 and bounds resulting from $\mathbf{c}^H\mathbf{x}_1 = 1$ and $\mathbf{u}_1^H\mathbf{M}\mathbf{x}_1 = \xi$ we obtain (2.10). □

Remark 2.3. *Lemma 2.2 shows that $\|\mathbf{J}(\mathbf{z}_1)^{-1}\|$ can be bounded above where the bound depends on $|\mathbf{u}_1^H\mathbf{M}\mathbf{x}_1|^{-1}$ and on $\|(\mathbf{A}-\lambda_1\mathbf{M})^\dagger\|$. Note the quantity $\|(\mathbf{A}-\lambda_1\mathbf{M})^\dagger\|$ is often called the reduced resolvent norm [16].*

(a) *We see the importance of condition (2.3). If $|\mathbf{u}_1^H\mathbf{M}\mathbf{x}_1|$ is close to zero then the bounds on $\|\mathbf{J}(\mathbf{z}_1)^{-1}\|$ will be large and the radius of the ball of convergence in the Newton theory in Theorem 2.6 will be correspondingly small.*

(b) *Also, if the eigenvalue λ_1 is close to its neighbouring eigenvalue then $\|(\mathbf{A}-\lambda_1\mathbf{M})^\dagger\|$ will be large and again, the radius of the ball of convergence in the Newton theory in Theorem 2.6 will be correspondingly small. We will see in Chapter 3 (Theorem 3.1) that, using the generalised Schur decomposition there exist unitary matrices $\mathbf{Q}$ and $\mathbf{Z}$ such that*

$$\mathbf{Q}^H(\mathbf{A}-\lambda_1\mathbf{M})\mathbf{Z} = \begin{bmatrix} t_{11} & \mathbf{t}_{12}^H \\ \mathbf{0} & \mathbf{T}_{22}\end{bmatrix} - \lambda_1 \begin{bmatrix} s_{11} & \mathbf{s}_{12}^H \\ \mathbf{0} & \mathbf{S}_{22}\end{bmatrix}, \quad \textit{where} \quad \lambda_1 = \frac{t_{11}}{s_{11}}.$$

With σ_{n-1} denoting the second smallest singular value well-known linear algebra results (see [48]) give

$$\begin{aligned}\|(\mathbf{A}-\lambda_1\mathbf{M})^\dagger\| &= \frac{1}{\sigma_{n-1}(\mathbf{A}-\lambda_1\mathbf{M})} = \frac{1}{\sigma_{n-1}(\mathbf{T}_{22}-\lambda_1\mathbf{S}_{22})} \\ &= \frac{1}{\min_{\|\mathbf{a}\|=1}\|(\mathbf{T}_{22}-\lambda_1\mathbf{S}_{22})\mathbf{a}} =: \frac{1}{sep(\lambda_1,(\mathbf{T}_{22},\mathbf{S}_{22}))},\end{aligned}$$

where $sep(\lambda_1,(\mathbf{T}_{22},\mathbf{S}_{22}))$ denotes the separation between the eigenvalue λ_1 and the rest of the spectrum. Hence, if λ_1 and the rest of the spectrum is not well separated then $sep(\lambda_1,(\mathbf{T}_{22},\mathbf{S}_{22}))$ is small leading to a large bound on $\|\mathbf{J}(\mathbf{z}_1)^{-1}\|$ and the radius of the ball of convergence in the Newton theory in Theorem 2.6 will be correspondingly small.

Standard convergence theory for Newton's method (see, for example, [25]) applied to (2.8) provides the following well-known convergence result.

Corollary 2.4. *If $\mathbf{z}_1 = (\mathbf{x}_1^T, \lambda_1)^T$ is an algebraically simple eigenpair of (2.2) and if (2.4) holds, then Algorithm 1 converges quadratically for a close enough starting guess.*

This quadratic rate of convergence is observed in practice (see, for example, the numerical results given by the solid line in Figure 2-1).

For $(\mathbf{x}^{(i)^T}, \lambda^{(i)})^T$ the eigenvalue residual

$$\mathbf{r}^{(i)} = (\mathbf{A}-\lambda^{(i)}\mathbf{M})\mathbf{x}^{(i)}, \tag{2.12}$$

is calculated in step (4) of Algorithm 1. Since $\mathbf{c}^H\mathbf{x}^{(i)} = 1$, $\forall i$, we have $\|\mathbf{r}^{(i)}\| = \|\mathbf{F}(\mathbf{z}^{(i)})\|$. Now with $\mathbf{z}_1 = (\mathbf{x}_1^T, \lambda_1)^T$ denoting the root of $\mathbf{F}(\mathbf{z}) = \mathbf{0}$,

$$\|\mathbf{F}(\mathbf{z}^{(i)})\| = \|\mathbf{F}(\mathbf{z}^{(i)}) - \mathbf{F}(\mathbf{z}_1)\| \leq C_1\, \|\mathbf{z}^{(i)} - \mathbf{z}_1\|,$$

holds, where C_1 is a bound on the norm of $\mathbf{J}(\mathbf{z})$ in some ball centered on $\mathbf{z}_1$. Hence

$$\|\mathbf{r}^{(i)}\| \leq C_1 \, \|\mathbf{z}^{(i)} - \mathbf{z}_1\| \,. \tag{2.13}$$

It is important to note that, in practice, to stop a Newton iteration one would typically use a relative stopping condition, see, for example, [71, Section 5.2] or [26, Chapter 2] where a clear scaling-invariant account of Newton's method is given.

In this section we have shown how exact inverse iteration can be regarded as a Newton method. Note that a straightforward generalisation applies to subspace iteration, where the single eigenvector $\mathbf{x}^{(i)}$ is replaced by an invariant subspace. So-called block Newton iterations have for approximating invariant subspaces have been considered by Lösche et al. [84] and proved to be at least quadratic under appropriate conditions. In the next section we describe how inexact inverse iteration can be interpreted as a modified Newton method and hence derive corresponding convergence results.

2.3 Inexact inverse iteration & modified Newton's method

Let us now consider a version of inexact inverse iteration that introduces two changes from Algorithm 1. First, as the name implies, we solve the linear systems iteratively to a given residual tolerance (and hence the linear systems are solved "inexactly"). Second, instead of (2.1) we consider the linear system

$$(\mathbf{A} - \sigma\mathbf{M})\mathbf{y} = \mathbf{Z}(\lambda)\mathbf{x}, \tag{2.14}$$

where $\mathbf{Z}(\lambda)$ is a complex $n \times n$ matrix depending on λ. If $\mathbf{Z}(\lambda) = \mathbf{M}$, then (2.14) reduces to (2.1). However, in Section 2.5 we consider the system $(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \frac{1}{\lambda^{(i)}}\mathbf{P}\mathbf{x}^{(i)}$, where $\mathbf{P}$ is a preconditioner for $\mathbf{A} - \lambda^{(i)}\mathbf{M}$, and so we consider the convergence theory for the more general form given by (2.14). Thus we discuss the following extension of Algorithm 1.

Algorithm 2 Inexact Inverse Iteration as modified Newton method

Input: $\lambda^{(0)}$ and $\mathbf{x}^{(0)}$ with $\mathbf{c}^H\mathbf{x}^{(0)} = 1$, i_{max}.
  for $i = 1, \dots, i_{max}$ **do**
    Choose $\tau^{(i)}$,
    Solve $(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)}$ inexactly, that is,

$$\| \, (\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{y}^{(i)} - \mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)} \, \| \leq \tau^{(i)} \, \|\mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)}\|,$$

    Set $\Delta\lambda^{(i)} = \frac{1}{\mathbf{c}^H\mathbf{y}^{(i)}}$; $\quad \lambda^{(i+1)} = \lambda^{(i)} + \Delta\lambda^{(i)}$,
    Update $\mathbf{x}^{(i+1)} = \Delta\lambda^{(i)}\mathbf{y}^{(i)}$,
    Evaluate $\mathbf{r}^{(i+1)} = (\mathbf{A} - \lambda^{(i+1)}\mathbf{M})\mathbf{x}^{(i+1)}$,
    Test for convergence.
  end for
Output: $\lambda^{(i_{max})}$, $\mathbf{x}^{(i_{max})}$.

To analyse this algorithm let us introduce the linear system residual

$$\mathbf{d}^{(i)} := (\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{y}^{(i)} - \mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)} \tag{2.15}$$

which should not be confused with the eigenvalue residual $\mathbf{r}^{(i)}$ defined by (2.12). From step (2) in Algorithm 2 we know

$$\|\mathbf{d}^{(i)}\| \leq \tau^{(i)} \, \|\mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)}\| \,. \tag{2.16}$$

Now, using $\mathbf{x}^{(i+1)} = \Delta\lambda^{(i)}\mathbf{y}^{(i)}$ from step (4) of Algorithm 2, we may write (2.15) as

$$(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{x}^{(i+1)} = \Delta\lambda^{(i)}(\mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)} + \mathbf{d}^{(i)}),$$

or, equivalently,

$$(\mathbf{A} - \lambda^{(i)}\mathbf{M})\Delta\mathbf{x}^{(i)} - \Delta\lambda^{(i)}(\mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)} + \mathbf{d}^{(i)}) = -(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{x}^{(i)}.$$

This equation along with $\mathbf{c}^H\Delta\mathbf{x}^{(i)} = 0$ gives

$$\begin{bmatrix} \mathbf{A} - \lambda^{(i)}\mathbf{M} & -(\mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)} + \mathbf{d}^{(i)}) \\ \mathbf{c}^H & 0 \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x}^{(i)} \\ \Delta\lambda^{(i)} \end{bmatrix} = \begin{bmatrix} -(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{x}^{(i)} \\ 0 \end{bmatrix}, \tag{2.17}$$

which with (2.7) we can write as

$$\widetilde{\mathbf{J}}(\mathbf{z}^{(i)})\Delta\mathbf{z}^{(i)} = -\mathbf{F}(\mathbf{z}^{(i)}), \quad \mathbf{z}^{(i+1)} = \mathbf{z}^{(i)} + \Delta\mathbf{z}^{(i)}, \tag{2.18}$$

where $\mathbf{z}^{(i)} = (\mathbf{x}^{(i)^T}, \lambda^{(i)})^T$ and $\widetilde{\mathbf{J}}$ is defined by

$$\widetilde{\mathbf{J}}(\mathbf{z}^{(i)}) = \begin{bmatrix} \mathbf{A} - \lambda^{(i)}\mathbf{M} & -(\mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)} + \mathbf{d}^{(i)}) \\ \mathbf{c}^H & 0 \end{bmatrix}. \tag{2.19}$$

Clearly (2.18) is a modified Newton method for $\mathbf{F}(\mathbf{z}) = \mathbf{0}$ with $\widetilde{\mathbf{J}}(\mathbf{z}^{(i)})$ being an approximation to the exact Jacobian $\mathbf{J}(\mathbf{z}^{(i)})$ given by (2.9). In fact,

$$\widetilde{\mathbf{J}}(\mathbf{z}^{(i)}) - \mathbf{J}(\mathbf{z}^{(i)}) = \begin{bmatrix} \mathbf{O} & (\mathbf{M}\mathbf{x}^{(i)} - \mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)} - \mathbf{d}^{(i)}) \\ \mathbf{0}^H & 0 \end{bmatrix}. \tag{2.20}$$

where $\mathbf{O}$ denotes the $n \times n$ zero matrix.

Hence the convergence of the inexact inverse iteration method given by Algorithm 2 can be proved using the convergence theory of modified Newton's method. We state a convergence theorem for modified Newton's method (see for example [25] and [24, Theorem 3.4]) that is used to prove Theorem 2.6.

Theorem 2.5. *Assume* $\mathbf{F} : \mathbb{C}^n \longrightarrow \mathbb{C}^n$ *and let* $\mathbf{F}(\mathbf{z}^*) = \mathbf{0}$. *For some* $r > 0$, *define* $\mathcal{B} := \mathcal{B}(\mathbf{z}^*, r)$ *and assume* $\mathbf{J}(\mathbf{z}) \in Lip_\gamma\mathcal{B}$, *where* $\mathbf{J}(\mathbf{z})$ *is the Jacobian of* $\mathbf{F}(\mathbf{z})$. *Further, assume* $\|\mathbf{J}(\mathbf{z}^*)^{-1}\| \leq \beta$. *For each* $\mathbf{z}$ *let* $\widetilde{\mathbf{J}}(\mathbf{z})$ *be a complex* $n \times n$ *matrix satisfying, for some* δ, $\quad 0 \leq \delta < 1$,

$$\|\mathbf{J}(\mathbf{z}^*)^{-1}(\widetilde{\mathbf{J}}(\mathbf{z}) - \mathbf{J}(\mathbf{z}^*))\| \leq \delta. \tag{2.21}$$

Then $\widetilde{\mathbf{J}}(\mathbf{z})^{-1}$ *exists in* $\mathcal{B}$ *and* $\|\widetilde{\mathbf{J}}(\mathbf{z})^{-1}\| \leq \dfrac{\beta}{1-\delta}$. *Next consider the solution of*

$$\mathbf{F}(\mathbf{z}) = \mathbf{0}, \quad \mathbf{z} \in \mathbb{C}^n \tag{2.22}$$

using modified *Newton's method:*

$$\mathbf{z}^{(i+1)} = \mathbf{z}^{(i)} - \widetilde{\mathbf{J}}(\mathbf{z}^{(i)})^{-1}\mathbf{F}(\mathbf{z}^{(i)}), \quad \mathbf{z}^{(0)} \in \mathcal{B}. \tag{2.23}$$

If

$$\left\{ \frac{\beta\gamma r}{2(1-\delta)} + \delta \right\} =: \alpha < 1, \quad \forall \mathbf{z} \in \mathcal{B} \tag{2.24}$$

then, with $\mathbf{e}^{(i)} := \mathbf{z}^{(i)} - \mathbf{z}^*$,

1. *modified Newton's method converges* linearly *to* $\mathbf{z}^*$,

2. *we have*

$$\|\mathbf{e}^{(i+1)}\| \leq \frac{\beta}{(1-\delta)} \left\{ \frac{\gamma}{2} \|\mathbf{e}^{(i)}\| + \|\mathbf{J}(\mathbf{z}^{(i)}) - \widetilde{\mathbf{J}}(\mathbf{z}^{(i)})\| \right\} \|\mathbf{e}^{(i)}\|,$$

and

3. *if*

$$\|(\mathbf{J}(\mathbf{z}^{(i)}) - \widetilde{\mathbf{J}}(\mathbf{z}^{(i)}))\mathbf{e}^{(i)}\| \leq C \|\mathbf{e}^{(i)}\|^2$$

for some constant C *independent of* i*, then modified Newton's method converges* quadratically*.*

Proof. See [24, Theorem 3.4]). □

Hence, with this result we can state and prove Theorem 2.6.

Theorem 2.6 (Convergence of Inexact Inverse Iteration). *Let* $\mathbf{z}_1 = (\mathbf{x}_1^T, \lambda_1)^T$ *be an algebraically simple eigenpair of (2.2) satisfying (2.4). Since* $\mathbf{J}(\mathbf{z}_1)$ *defined by (2.9) is nonsingular we assume* $\|\mathbf{J}(\mathbf{z}_1)^{-1}\| \leq \beta$ *(see Lemma 2.1). For some* $\tau_{max}, r > 0$*, consider the use of Algorithm 2 with* $\tau^{(i)} \leq \tau_{max}$*,* $\forall i$*, with starting value* $\mathbf{z}^{(0)} = (\mathbf{x}^{(0)^T}, \lambda^{(0)})^T \in \mathcal{B} = \mathcal{B}(\mathbf{z}_1, r)$*. If* r, τ_{max} *and* $\mathbf{Z}(\lambda)$ *are such that*

$$\beta\{|\lambda_1 - \lambda| \, \|\mathbf{M}\| + \|\mathbf{Z}(\lambda)\mathbf{x} - \mathbf{M}\mathbf{x}_1\| + \tau_{max} \|\mathbf{Z}(\lambda)\mathbf{x}\|\} =: \delta < 1 \tag{2.25}$$

for $\mathbf{z} = (\mathbf{x}^T, \lambda)^T \in \mathcal{B}$*, and if*

$$\left\{ \frac{\beta \, \|\mathbf{M}\| \, r}{1-\delta} + \delta \right\} =: \alpha < 1, \tag{2.26}$$

then with $\mathbf{e}^{(i)} := \mathbf{z}^{(i)} - \mathbf{z}_1$*,*

1. *Algorithm 2 converges linearly to* $\mathbf{z}_1 = (\mathbf{x}_1^T, \lambda_1)^T$ *with*

$$\|\mathbf{z}^{(i+1)} - \mathbf{z}_1\| \leq \alpha \, \|\mathbf{z}^{(i)} - \mathbf{z}_1\|,$$

2. $\mathbf{e}^{(i+1)}$ *satisfies*

$$\|\mathbf{e}^{(i+1)}\| \leq \frac{\beta}{1-\delta} \left(\|\mathbf{M}\| \|\mathbf{e}^{(i)}\| + \|\mathbf{M}\mathbf{x}^{(i)} - \mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)}\| + \|\mathbf{d}^{(i)}\| \right) \|\mathbf{e}^{(i)}\|, \tag{2.27}$$

3. *if, in addition,* $\tau^{(i)}$ *in Algorithm 2 satisfies*

$$\tau^{(i)} = C_2 \, \|\mathbf{r}^{(i)}\|, \tag{2.28}$$

for some constant C_2 *independent of* i *with* $\mathbf{r}^{(i)}$ *given by (2.12), and*

$$\|\mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)} - \mathbf{M}\mathbf{x}^{(i)}\| \leq C_3 \, \|\mathbf{e}^{(i)}\|, \tag{2.29}$$

for some constant C_3 *independent of* i*, then Algorithm 2 converges quadratically.*

Proof. The proof consists of verifying the conditions of Theorem 2.5 on the convergence of modified Newton's method for $\mathbf{F}$ defined by (2.5). First note that γ, the Lipschitz constant of $\mathbf{J}$, can be taken as $2 \parallel \mathbf{M} \parallel$. Next by reducing r and τ_{max} and taking $\mathbf{Z}(\lambda)$ close enough to $\mathbf{M}$, conditions (2.25) and (2.26) can always be made to hold. Thus conditions (2.21) and (2.24) of Theorem 2.5 hold and so the linear convergence of Algorithm 2 (part (a)) is proved. Part (b) of Theorem 2.6 follows immediately from part (b) in Theorem 2.5. Under (2.28) and (2.29) (and recalling (2.13)),

$$\|\mathbf{J}(\mathbf{z}^{(i)}) - \widetilde{\mathbf{J}}(\mathbf{z}^{(i)})\| \leq \|\mathbf{M}\mathbf{x}^{(i)} - \mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)}\| + \|\mathbf{d}^{(i)}\| \leq C_4 \|\mathbf{e}^{(i)}\|, \tag{2.30}$$

for some constant C_4 independent of i. The quadratic convergence follows from case (c) in Theorem 2.5. □

We see from (2.27) that the possibility of achieving quadratic convergence in Algorithm 2 is determined by the size of $\|\mathbf{M}\mathbf{x}^{(i)} - \mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)}\|$ and how $\|\mathbf{d}^{(i)}\|$ is controlled. However if $\tau^{(i)}$ is held fixed, or if (2.29) does not hold, then linear convergence is all that can be expected. We discuss the natural case $\mathbf{Z}(\lambda^{(i)}) = \mathbf{M}$ in the following subsection, but we end this theoretical section with a corollary.

Corollary 2.7. *Assume that the conditions of Theorem 2.6 are satisfied. Let $\tau^{(i)}$ be chosen as in (2.28) and assume $\lambda^{(i)} \neq 0$ and*

$$|\lambda^{(i)}| \geq C_L, \quad \forall i, \tag{2.31}$$

with $C_L > 0$. If $\mathbf{Z}(\lambda^{(i)})$ satisfies

$$\|\mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)} - \frac{1}{\lambda^{(i)}}\mathbf{A}\mathbf{x}^{(i)}\| \leq C_5 \|\mathbf{e}^{(i)}\|, \tag{2.32}$$

then Algorithm 2 exhibits quadratic convergence.

Proof. We have

$$\begin{aligned} \|\mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)} - \mathbf{M}\mathbf{x}^{(i)}\| \leq& \|\mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)} - \frac{1}{\lambda^{(i)}}\mathbf{A}\mathbf{x}^{(i)}\| \\ &+ \|\frac{1}{\lambda^{(i)}}(\mathbf{A}\mathbf{x}^{(i)} - \lambda^{(i)}\mathbf{M}\mathbf{x}^{(i)})\| . \end{aligned} \tag{2.33}$$

Using (2.32) and the fact that $(\mathbf{x}^{(i)}, \lambda^{(i)})^T$ is an approximate eigenpair we can apply the properties of the eigenvalue residual (2.12) and (2.13) to get

$$\|\mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)} - \mathbf{M}\mathbf{x}^{(i)}\| \leq C_5 \|\mathbf{e}^{(i)}\| + \frac{1}{|\lambda^{(i)}|} C_1 \|\mathbf{e}^{(i)}\| \leq \tilde{C}_3 \|\mathbf{e}^{(i)}\|,$$

where $\tilde{C}_3 := C_5 + \dfrac{C_1}{C_L}$. Hence (2.28) and (2.29) in Theorem 2.6 hold, with $C_3 := \tilde{C}_3$, proving that the convergence is quadratic. □

Thus we see from (2.32) that if the right hand side of (2.14) can be made to approximate $\dfrac{1}{\lambda^{(i)}}\mathbf{A}\mathbf{x}^{(i)}$ then there is the possibility of superlinear or even quadratic convergence.

2.3.1 Standard inexact inverse iteration

In this subsection we consider the standard form of inexact inverse iteration by making the choice $\mathbf{Z}(\lambda^{(i)}) = \mathbf{M}$. In this context we discuss two choices of $\tau^{(i)}$ in Algorithm 2, namely, either $\tau^{(i)}$ is chosen to decrease or $\tau^{(i)}$ is held fixed (cases (a) and (b) respectively in the following Corollary).

Corollary 2.8. *Assume $\mathbf{Z}(\lambda^{(i)}) = \mathbf{M}$ in Algorithm 2 and let the conditions of Theorem 2.6 hold. Then we obtain the following rates of convergence, depending on the tolerance $\tau^{(i)}$.*

(a) Decreasing tolerance. *If $\tau^{(i)} = C_2 \, \|\mathbf{r}^{(i)}\|$ in step (1) of Algorithm 2 then, for a close enough starting guess, Algorithm 2 achieves quadratic convergence, which is equal to the rate achieved by Algorithm 1 for exact inverse iteration.*

(b) Fixed tolerance. *If $\tau^{(i)} = \tau$ in step (1) of Algorithm 2, where τ is fixed but small enough, then, for a close enough starting guess, Algorithm 2 converges linearly.*

Proof. For $\mathbf{Z}(\lambda^{(i)}) = \mathbf{M}$ condition (2.29) in Theorem 2.6 is obviously satisfied with $C_3 = 0$. In the case of a decreasing tolerance (a), condition (2.28) of Theorem 2.6 is assumed and therefore quadratic convergence follows immediately from Theorem 2.6. In the case of a fixed tolerance (b), the bound in (2.16) becomes

$$\|\mathbf{d}^{(i)}\| \leq \tau \, \|\mathbf{M}\mathbf{x}^{(i)}\|, \tag{2.34}$$

where τ is fixed. Then by considering (2.27) we obtain

$$\|\mathbf{e}^{(i+1)}\| \leq \frac{\beta}{1-\delta} \left(\|\mathbf{M}\| \|\mathbf{e}^{(i)}\|^2 + \tau \|\mathbf{M}\mathbf{x}^{(i)}\| \, \|\mathbf{e}^{(i)}\| \right),$$

and hence only linear convergence can be proved. □

We now present some numerical results to illustrate the theoretical results from Corollary 2.8, and also provide a comparison with Algorithm 1 (exact solves).

2.3.2 Numerical example

Here we present numerical results to illustrate the convergence behaviour of inexact inverse iteration for two different choices of the solve tolerance in step (2) of Algorithm 2.

Example 2.9. *Consider the standard eigenvalue problem $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ where $\mathbf{A}$ is the finite difference discretisation (central differences) on a 32×32 grid of the following eigenvalue problem of the convection-diffusion equation*

$$-\Delta u + 5u_x + 5u_y = \lambda u \quad on \quad (0,1)^2, \tag{2.35}$$

with homogeneous Dirichlet boundary conditions. Take $\mathbf{Z}(\lambda^{(i)}) = \mathbf{M} = \mathbf{I}$. This eigenvalue problem is also discussed in [50]. Consider finding the smallest eigenvalue ($\lambda_1 \approx 32.18560954$) by Algorithm 1 and by Algorithm 2 with both decreasing and fixed tolerances. We take an initial vector $\mathbf{x}^{(0)}$ with $\cos(\mathbf{x}_1, \mathbf{x}^{(0)}) \approx 0.84$ and an initial eigenvalue $\lambda^{(0)} = 20$, knowing that the smallest eigenvalue of the Laplacian $-\Delta$ is equal to

$2\pi^2 \approx 20$. *We use GMRES as the inexact solver and for the inexact solves with fixed tolerance we take* $\tau^{(i)} = \tau = 0.3$ *(case (b) in Corollary 2.8) and for the inexact solves with decreasing tolerance (case (a) in Corollary 2.8) we take*

$$\tau^{(i)} = \min\{\tau, \|\mathbf{r}^{(i)}\|\}, \quad \textit{with} \quad \tau = 0.3, \tag{2.36}$$

where the eigenvalue residual $\|\mathbf{r}^{(i)}\|$ *is given by (2.12). As an overall stopping condition we use the norm of the relative eigenvalue residual, so that, once*

$$\left\| \frac{\mathbf{r}^{(i)}}{\lambda^{(i)}} \right\| < 10^{-10}$$

is satisfied, the computation stops. Note that the computations use $\|\cdot\| = \|\cdot\|_2$*, since GMRES minimises the* 2*-norm of the linear system residual.*

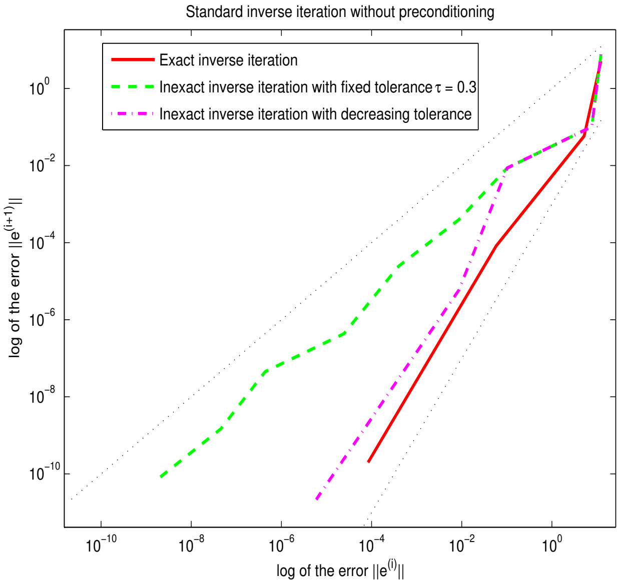


Figure 2-1: *Numerical results for Example 2.9. The slopes of the solid, dashed and dashed-dotted lines indicate the rates of convergence achieved. The dotted lines indicate the slopes expected for linear and quadratic convergence.*

The results are illustrated in Figure 2-1 which gives logarithmic plots for the norm of the error at step $i + 1$ against the norm of the error at step i. The dotted outer lines indicate the slopes expected for linear and quadratic convergence. Clearly, exact inverse iteration specified by the solid line yields quadratic convergence as expected, since it corresponds to Newton's method (see Section 2.2). Also inexact inverse iteration with decreasing tolerance indicated by the dash-dotted line gives quadratic convergence as expected from Corollary 2.8 part (a). For inexact inverse iteration with fixed tolerance plotted in the dashed line we get only linear convergence as predicted in Corollary 2.8 part (b).

If the convection term in the problem is increased in (2.35), a closer starting guess is required since the spectrum of the convection-diffusion operator becomes more bunched, with complex eigenvalues moving close to the desired eigenvalue λ_1.

We remark that numerical results using $\mathbf{Z}(\lambda^{(i)}) = \dfrac{1}{\lambda^{(i)}}\mathbf{A}$ (so that $C_5 = 0$ in Corollary 2.7) using both exact solves and inexact solves with decreasing tolerance chosen as in (2.36) produce the expected quadratic convergence. We do not reproduce these results here.

2.4 Simplified Jacobi-Davidson method as an inexact Newton method

This section contains an analysis of simplified Jacobi-Davidson method using Newton's method. We show how inverse iteration with a variable shift, a version of simplified Jacobi-Davidson method and Newton's method are equivalent to each other. Equivalences between the Jacobi-Davidson method and an accelerated Newton scheme have first been shown in [124] and [125]. We also show how inexact versions of all three methods correspond to each other.

The Jacobi-Davidson method was introduced by Sleijpen and van der Vorst for the linear eigenproblem (see [124] and [126]) and it has been applied to the generalised eigenproblem and matrix pencils (see [39] and [123]). Here we consider a simplified version of it for the generalised eigenproblem, where the dimension of the subspace is not increased at each step.

Assume $(\lambda^{(i)}, \mathbf{x}^{(i)})$ approximates the exact eigenpair $(\lambda_1, \mathbf{x}_1)$. Introduce the following orthogonal projections

$$\mathbf{P}^{(i)} = \mathbf{I} - \frac{\mathbf{M}\mathbf{x}^{(i)}{\mathbf{x}^{(i)}}^H\mathbf{M}^H}{{\mathbf{x}^{(i)}}^H\mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}} \quad \text{and} \quad \mathbf{Q}^{(i)} = \mathbf{I} - \frac{\mathbf{x}^{(i)}{\mathbf{x}^{(i)}}^H\mathbf{M}^H\mathbf{M}}{{\mathbf{x}^{(i)}}^H\mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}}. \tag{2.37}$$

With $\mathbf{r}^{(i)}$ defined by (2.12) we then solve the correction equation

$$\mathbf{P}^{(i)}(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{Q}^{(i)}\mathbf{s}^{(i)} = -\mathbf{r}^{(i)}, \quad \text{where} \quad \mathbf{s}^{(i)} \perp \mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}, \tag{2.38}$$

for $\mathbf{s}^{(i)}$. This is the Jacobi-Davidson correction equation and the matrix $\mathbf{P}^{(i)}(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{Q}^{(i)}$ maps $(\text{span}\{\mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}\})^\perp$ onto $(\text{span}\{\mathbf{M}\mathbf{x}^{(i)}\})^\perp$, where $(\text{span}\{\mathbf{M}\mathbf{x}^{(i)}\})^\perp$ denotes the orthogonal complement of $\text{span}\{\mathbf{M}\mathbf{x}^{(i)}\}$. Then, an improved guess for the eigenvector is given by a suitably normalised $\mathbf{x}^{(i)} + \mathbf{s}^{(i)}$. Other choices of the projections are possible: for further discussion on the correction equation (2.38) we refer to [123]. If (2.38) is solved inexactly via an iterative method, then we can describe the inexact solve by

$$\mathbf{P}^{(i)}(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{Q}^{(i)}\mathbf{s}^{(i)} = -\mathbf{r}^{(i)} + \mathbf{d}_{JD}^{(i)} \quad \text{where} \quad \mathbf{s}^{(i)} \perp \mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}, \tag{2.39}$$

with

$$\|\mathbf{d}_{JD}^{(i)}\| \leq \tau_{JD}^{(i)}\|\mathbf{r}^{(i)}\|, \quad \text{and} \quad \tau_{JD}^{(i)} < 1.$$

Algorithm 3 gives a description of the method investigated in this section.
With $\mathbf{s}^{(i)} \perp \mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}$, $\mathbf{Q}^{(i)}\mathbf{s}^{(i)} = \mathbf{s}^{(i)}$, and so we can rewrite (2.39) as

$$\mathbf{P}^{(i)}(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{s}^{(i)} = -\mathbf{r}^{(i)} + \mathbf{d}_{JD}^{(i)},$$

and hence, with the definition of $\mathbf{P}^{(i)}$ from (2.37) we have

$$(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{s}^{(i)} - \frac{{\mathbf{x}^{(i)}}^H\mathbf{M}^H(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{s}^{(i)}}{{\mathbf{x}^{(i)}}^H\mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}}\mathbf{M}\mathbf{x}^{(i)} = -\mathbf{r}^{(i)} + \mathbf{d}_{JD}^{(i)},$$

Algorithm 3 Simplified Jacobi-Davidson

Input: $\mathbf{x}^{(0)}$, $i_{max}, \lambda^{(0)}$.
  for $i = 0, \dots, i_{max}$ **do**
    Choose $\tau^{(i)}$,
    $\mathbf{r}^{(i)} = (\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{x}^{(i)}$,
    Find $\mathbf{s}^{(i)}$ such that

$$\|\mathbf{P}^{(i)}(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{Q}^{(i)}\mathbf{s}^{(i)} + \mathbf{r}^{(i)}\| \leq \tau^{(i)}\|\mathbf{r}^{(i)}\| \quad \text{for} \quad \mathbf{s}^{(i)} \perp \mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)},$$

    Set $\mathbf{x}^{(i+1)} = \dfrac{\mathbf{x}^{(i)} + \mathbf{s}^{(i)}}{\|\mathbf{M}(\mathbf{x}^{(i)} + \mathbf{s}^{(i)})\|}$,
    Update $\lambda^{(i)}$.
    Test for convergence.
  end for
Output: $\mathbf{x}^{(i_{max})}$.

or

$$(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{s}^{(i)} - \alpha^{(i)}\mathbf{M}\mathbf{x}^{(i)} = -\mathbf{r}^{(i)} + \mathbf{d}_{JD}^{(i)}, \tag{2.40}$$

where $\alpha^{(i)} = \dfrac{{\mathbf{x}^{(i)}}^H\mathbf{M}^H(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{s}^{(i)}}{{\mathbf{x}^{(i)}}^H\mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}}$. Then the solution of the correction equation (2.39) may be written in the block matrix form

$$\begin{bmatrix} \mathbf{A} - \lambda^{(i)}\mathbf{M} & -\mathbf{M}\mathbf{x}^{(i)} \\ {\mathbf{x}^{(i)}}^H\mathbf{M}^H\mathbf{M} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{s}^{(i)} \\ \alpha^{(i)} \end{bmatrix} = -\begin{bmatrix} \mathbf{r}^{(i)} \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{d}_{JD}^{(i)} \\ \mathbf{0} \end{bmatrix}. \tag{2.41}$$

Now with Algorithm 3, let

$$\begin{bmatrix} \mathbf{x}^{(i+1)} \\ \lambda^{(i+1)} \end{bmatrix} = \begin{bmatrix} \mathbf{x}^{(i)} \\ \lambda^{(i)} \end{bmatrix} + \begin{bmatrix} \mathbf{s}^{(i)} \\ \alpha^{(i)} \end{bmatrix}, \tag{2.42}$$

where the update in $\lambda^{(i)}$ is only of theoretical nature. Introducing the nonlinear system $\mathbf{G}(\mathbf{z}) = \mathbf{0}$, where $\mathbf{z} := (\mathbf{x}^T, \lambda)^T$ and

$$\mathbf{G}(\mathbf{z}) = \begin{bmatrix} (\mathbf{A} - \lambda\mathbf{M})\mathbf{x} \\ \dfrac{\mathbf{x}^H\mathbf{M}^H\mathbf{M}\mathbf{x}}{2} - \dfrac{1}{2} \end{bmatrix}, \tag{2.43}$$

we observe that with $\mathbf{d}_{JD}^{(i)} = \mathbf{0}$ equations (2.41) and (2.42) are equivalent to the application of one step of Newton's method applied to (2.43), that is

$$\mathbf{J}(\mathbf{z}^{(i)})\Delta\mathbf{z}^{(i)} = -\mathbf{G}(\mathbf{z}^{(i)}), \quad \mathbf{z}^{(i+1)} = \mathbf{z}^{(i)} + \Delta\mathbf{z}^{(i)},$$

where $\mathbf{J}(\mathbf{z}^{(i)})$ denotes the Jacobian

$$\mathbf{J}(\mathbf{z}^{(i)}) = \begin{bmatrix} \mathbf{A} - \lambda^{(i)}\mathbf{M} & -\mathbf{M}\mathbf{x}^{(i)} \\ {\mathbf{x}^{(i)}}^H\mathbf{M}^H\mathbf{M} & \mathbf{0} \end{bmatrix}.$$

Lemmas 2.1 and 2.2 as well as Corollary 2.4 apply. For inexact solves, that is $\mathbf{d}_{JD}^{(i)} \neq \mathbf{0}$, we have with (2.41) and

$$\mathbf{G}(\mathbf{z}^{(i)}) = \begin{bmatrix} \mathbf{r}^{(i)} \\ 0 \end{bmatrix}, \quad \|\mathbf{G}(\mathbf{z}^{(i)})\| = \|\mathbf{r}^{(i)}\|,$$

as well as $\|\mathbf{d}_{JD}^{(i)}\| \leq \tau_{JD}^{(i)}\|\mathbf{r}^{(i)}\|$

$$\|\mathbf{J}(\mathbf{z}^{(i)})\Delta\mathbf{z}^{(i)} + \mathbf{G}(\mathbf{z}^{(i)})\| \leq \tau_{JD}^{(i)}\|\mathbf{G}(\mathbf{z}^{(i)})\|, \quad \mathbf{z}^{(i+1)} = \mathbf{z}^{(i)} + \Delta\mathbf{z}^{(i)}, \tag{2.44}$$

with $\tau_{JD}^{(i)} < 1$. Clearly (2.44) is an inexact Newton method for $\mathbf{G}(\mathbf{z}) = 0$ and hence one step of inexact Newton method corresponds to one step of simplified Jacobi-Davidson as stated in Algorithm 3. We use the following theorem on inexact Newton's method.

Theorem 2.10. *Assume $\mathbf{G} : \mathbb{C}^n \to \mathbb{C}^n$, and let $\mathbf{G}(\mathbf{z}^*) = 0$. Assume $\mathbf{J}(\mathbf{z}) \in Lip_\gamma\mathcal{B}$ where $\mathbf{J}(\mathbf{z})$ is the Jacobian of $\mathbf{G}(\mathbf{z})$ and $\mathcal{B} = \mathcal{B}(\mathbf{z}^*, r)$ for some $r > 0$. Furthermore assume that $\mathbf{J}(\mathbf{z}^*)$ is nonsingular. Then there are $\delta \leq r$ and $\eta_{\max}$ such that if $\mathbf{z}_0 \in \mathcal{B}(\mathbf{z}^*, \delta)$ and $\{\eta^{(i)}\} \subset [0, \eta_{\max}]$, then the inexact Newton iteration*

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

where

$$\|\mathbf{J}(\mathbf{z}^{(i)})\Delta\mathbf{z}^{(i)} + \mathbf{G}(\mathbf{z}^{(i)})\| \leq \eta^{(i)}\|\mathbf{G}(\mathbf{z}^{(i)})\|$$

converges linearly *to $\mathbf{z}^*$. Moreover, if*

$$\eta^{(i)} \leq K_\eta\|\mathbf{G}(\mathbf{z}^{(i)})\| \tag{2.45}$$

for some $K_\eta > 0$ the convergence is quadratic.

Proof. See [71, Theorem 6.1.2] and [22, Corollary 3.5]. □

Hence, we have the following consequence for the convergence of simplified Jacobi-Davidson.

Theorem 2.11 (Convergence of Algorithm 3)**.** *Let $\mathbf{z}_1 = (\mathbf{x}_1, \lambda_1)^T$ be an algebraically simple eigenpair of (2.2) satisfying $\mathbf{x}_1^H\mathbf{M}^H\mathbf{M}\mathbf{x}_1 = 1$. For some $\tau_{\max}, r > 0$, consider the use of Algorithm 3 with $\tau^{(i)} \leq \tau_{\max}$, $\forall i$, with starting value $\mathbf{z}^{(0)} = (\mathbf{x}^{(0)}, \lambda^{(0)})^T \in \mathcal{B} = \mathcal{B}(\mathbf{z}_1, r)$. Then, Algorithm 3 converges* linearly. *If, in addition $\tau^{(i)} \leq K_\tau\|\mathbf{r}^{(i)}\|$, for some $K_\tau > 0$ and $\forall i$, then Algorithm 3 converges* quadratically.

Proof. We check the conditions in Theorem 2.10. Since $\mathbf{z}_1 = (\mathbf{x}_1, \lambda_1)^T$ is an algebraically simple eigenpair $\mathbf{J}(\mathbf{z}^*)$ is nonsingular and clearly $\mathbf{J}(\mathbf{z}) \in Lip_\gamma\mathcal{B}$ from (2.41). Then, for fixed $\tau \leq \tau_{\max}$ we obtain linear convergence. With $\mathbf{G}(\mathbf{z}^{(i)}) = \mathbf{r}^{(i)}$ we obtain from Theorem 2.10 that $\tau^{(i)} \leq K_\tau\|\mathbf{r}^{(i)}\|$ leads to quadratic convergence. □

The conclusions of this theorem are similar to the ones obtained in Chapter 3, Section 3.6 where we compare simplified Jacobi-Davidson method to inexact inverse iteration. In that section we will also show that for a fixed and small enough tolerance $\tau^{(i)}$, linear convergence of simplified Jacobi-Davidson is achieved whilst for a decreasing tolerance $\tau^{(i)}$ quadratic convergence of the simplified Jacobi-Davidson method can be obtained. In section 3.6 different techniques are used, namely an equivalence result to inexact Rayleigh quotient iteration.

Note that the approach here only gives the equivalence of the inexact simplified Jacobi-Davidson to the inexact Newton method for the eigenvector $\mathbf{x}^{(i)}$, since we do not make use of the update in $\lambda^{(i)}$ from (2.42), but rather use the Rayleigh quotient for an update of $\lambda^{(i)}$.

We also refer to the numerical examples in Chapter 3, Section 3.6.

2.5 Preconditioned iterative solves

In this section we consider the preconditioned iterative solution of the shifted linear systems in inverse iteration. First we show how a modified right hand side as in (2.14) can arise, and then we show how quadratic convergence in Algorithm 2 can be maintained by a simple rank one update to the standard preconditioner. Our motivation for choosing a different right hand side arises in the consideration of the performance of the iterative solver used in inexact inverse iteration. It was noted in [119] for the standard symmetric eigenvalue problem that it was advantageous to alter the right hand side in inverse iteration to reduce the number of iterations used by a Krylov solver. We consider this idea applied to the nonsymmetric, generalised case.

The obvious way to implement the (left) preconditioned solution of the shifted system $(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{M}\mathbf{x}^{(i)}$, with $\mathbf{P}$ a suitable preconditioner, is

$$\mathbf{P}^{-1}(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{P}^{-1}\mathbf{M}\mathbf{x}^{(i)}, \tag{2.46}$$

However, the idea in [119] is to alter the right hand side of (2.46) to produce a linear system whose solution requires fewer steps of GMRES. For the nonsymmetric eigenvalue problem we argue heuristically as follows. If $(\mathbf{x}^{(i)}, \lambda^{(i)})$ is close enough to $(\mathbf{x}_1, \lambda_1)$, then $\mathbf{A}\mathbf{x}^{(i)} \approx \lambda^{(i)}\mathbf{M}\mathbf{x}^{(i)}$ and so the right hand side of (2.46), namely $\mathbf{P}^{-1}\mathbf{M}\mathbf{x}^{(i)}$, can be approximated by (assuming $\lambda^{(i)} \neq 0$) $\mathbf{P}^{-1}\mathbf{M}\mathbf{x}^{(i)} \approx \frac{1}{\lambda^{(i)}}\mathbf{P}^{-1}\mathbf{A}\mathbf{x}^{(i)}$, and if, in addition, $\mathbf{P}^{-1}\mathbf{A} \approx \mathbf{I}$ then $\mathbf{P}^{-1}\mathbf{M}\mathbf{x}^{(i)} \approx \frac{1}{\lambda^{(i)}}\mathbf{x}^{(i)}$ and so

$$\mathbf{P}^{-1}(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{y}^{(i)} \approx \frac{1}{\lambda^{(i)}}\mathbf{x}^{(i)}.$$

Thus, if the preconditioner for $\mathbf{A} - \lambda^{(i)}\mathbf{M}$ is chosen to approximate $\mathbf{A}$, then it is reasonable to replace (2.46) by

$$\mathbf{P}^{-1}(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \frac{1}{\lambda^{(i)}}\mathbf{x}^{(i)}. \tag{2.47}$$

Hence, the right hand side vector is roughly in the direction of the approximate null vector of the iteration matrix $\mathbf{P}^{-1}(\mathbf{A} - \lambda^{(i)}\mathbf{M})$. A detailed account of the costs of Krylov solvers applied to shifted linear systems in inverse iteration is further discussed in [10] (for symmetric problems) and in [11] (for nonsymmetric problems). For more discussions on the cost of the inner iterations we refer to Chapters 4 and 6. From the viewpoint of outer convergence theory (2.47) reduces to the equation

$$(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \frac{1}{\lambda^{(i)}}\mathbf{P}\mathbf{x}^{(i)}, \tag{2.48}$$

that is, the equation analysed in Section 2.3 with $\mathbf{Z}(\lambda^{(i)}) = \frac{1}{\lambda^{(i)}}\mathbf{P}$. Due to the factor $\frac{1}{\lambda^{(i)}}$ we introduce an additional assumption that for some C_L, with $C_L > 0$ and independent of i, (2.31) holds. The following corollary provides the key theoretical result, where we allow the preconditioner to depend on i.

Corollary 2.12. *Let $\mathbf{P}_i$ be a preconditioner for $\mathbf{A} - \lambda^{(i)}\mathbf{M}$, where $\mathbf{P}_i$ depends on i. Assume that the conditions of Theorem 2.6 are satisfied and that (2.31) holds. Let $\tau^{(i)}$ be chosen as in (2.28). If $\mathbf{Z}(\lambda^{(i)})$ is chosen as*

$$\mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)} = \frac{1}{\lambda^{(i)}}\mathbf{P}_i\mathbf{x}^{(i)}, \tag{2.49}$$

and $\mathbf{P}_i$ satisfies

$$\mathbf{P}_i\mathbf{x}^{(i)} = \mathbf{A}\mathbf{x}^{(i)} \tag{2.50}$$

then Algorithm 2 exhibits quadratic convergence.

Proof. Using (2.50) we can write (2.48) as

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

Theorem 2.6 can now be applied with $\mathbf{Z}(\lambda^{(i)}) = \frac{1}{\lambda^{(i)}}\mathbf{A}$. We have that

$$\|\mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)} - \mathbf{M}\mathbf{x}^{(i)}\| = \|\frac{1}{\lambda^{(i)}}(\mathbf{A}\mathbf{x}^{(i)} - \lambda^{(i)}\mathbf{M}\mathbf{x}^{(i)})\|, \tag{2.51}$$

and so, using (2.12), (2.13) and (2.31),

$$\|\mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)} - \mathbf{M}\mathbf{x}^{(i)}\| \le \frac{1}{|\lambda^{(i)}|}C_1 \ \|\mathbf{e}^{(i)}\| \le C_3 \ \|\mathbf{e}^{(i)}\|,$$

where $C_3 := \frac{C_1}{C_L}$. Hence (2.28) and (2.29) in Theorem 2.6 hold, proving that the convergence is quadratic. ☐

Thus we see from (2.49) and (2.50) that if the right hand side of (2.48) can be made to approximate $\frac{1}{\lambda^{(i)}}\mathbf{A}\mathbf{x}^{(i)}$ then there is the possibility of quadratic outer convergence, with the added advantage of an efficient solution procedure for the shifted linear systems. In the following section we explain how it is possible to satisfy (2.50) and hence to achieve this quadratic convergence rate using an ILU preconditioner.

2.5.1 Incomplete LU preconditioning

In this subsection we consider the use of an incomplete LU factorisation of $\mathbf{A}$ as a preconditioner. Assume

$$\mathbf{A} = \tilde{\mathbf{L}}\tilde{\mathbf{U}} + \mathbf{E}, \tag{2.52}$$

where $\tilde{\mathbf{L}}$ is a lower triangular matrix and $\tilde{\mathbf{U}}$ is an upper triangular matrix approximating the matrices $\mathbf{L}$ and $\mathbf{U}$ from the complete LU decomposition. $\mathbf{E}$ is the error matrix associated with the incomplete factorisation. We take as preconditioner the matrix

$$\tilde{\mathbf{P}} = \tilde{\mathbf{L}}\tilde{\mathbf{U}}.$$

The previous discussion suggests two options for preconditioning. First, we may simply apply preconditioning in the obvious way to the standard inverse iteration method, that is,

$$\tilde{\mathbf{P}}^{-1}(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \tilde{\mathbf{P}}^{-1}\mathbf{M}\mathbf{x}^{(i)}, \tag{2.53}$$

which does not change the analysis of Section 2.3.1 for the outer rate of convergence of Algorithm 2. Alternatively, we may modify the right hand side and use

$$\tilde{\mathbf{P}}^{-1}(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \frac{1}{\lambda^{(i)}}\mathbf{x}^{(i)}, \tag{2.54}$$

which, from the point of view of the outer convergence rate, is equivalent to $(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \frac{1}{\lambda^{(i)}}\tilde{\mathbf{P}}\mathbf{x}^{(i)}$ (cf. (2.48) but with the right hand side scaled by $\frac{1}{\lambda^{(i)}}$). This is motivated by the similar scaling given in (2.47). So, in the notation of Section 2.3, (2.54) is written as $(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)}$ with $\mathbf{Z}(\lambda^{(i)}) = \frac{1}{\lambda^{(i)}}\tilde{\mathbf{P}}$. For this approach quadratic convergence is lost, but we do achieve a linear rate of convergence as is shown in the following corollary.

Corollary 2.13. *Assume*

$$\mathbf{Z}(\lambda^{(i)}) = \frac{1}{\lambda^{(i)}}\tilde{\mathbf{L}}\tilde{\mathbf{U}} \tag{2.55}$$

in Algorithm 2 and let the conditions of Theorem 2.6 hold. Then we obtain the following linear rates of convergence, depending on the tolerance $\tau^{(i)}$.

1. Decreasing tolerance. *If $\tau^{(i)} \leq C_2\|\mathbf{r}^{(i)}\|$ in step (1) of Algorithm 2 then, for a close enough starting guess, Algorithm 2 achieves linear convergence with*

$$\begin{aligned}\|\mathbf{e}^{(i+1)}\| \leq & \frac{\beta}{1-\delta}\frac{\|\mathbf{E}\mathbf{x}^{(i)}\|}{|\lambda^{(i)}|}\|\mathbf{e}^{(i)}\| \\ & + \frac{\beta}{1-\delta}\left\{\|\mathbf{M}\| + C_1C_2\|\mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)}\| + \frac{C_1}{|\lambda^{(i)}|}\right\}\|\mathbf{e}^{(i)}\|^2,\end{aligned} \tag{2.56}$$

   where $\mathbf{e}^{(i)} = \mathbf{z}^{(i)} - \mathbf{z}_1$.

2. Fixed tolerance. *If $\tau^{(i)} = \tau$ in step (1) of Algorithm 2, where τ is fixed but small enough, then, for a close enough starting guess, Algorithm 2 converges linearly with*

$$\begin{aligned}\|\mathbf{e}^{(i+1)}\| \leq & \frac{\beta}{1-\delta}\left\{\frac{1}{|\lambda^{(i)}|}\|\mathbf{E}\mathbf{x}^{(i)}\| + \tau\|\mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)}\|\right\}\|\mathbf{e}^{(i)}\| \\ & + \frac{\beta}{1-\delta}\left\{\|\mathbf{M}\| + \frac{C_1}{|\lambda^{(i)}|}\right\}\|\mathbf{e}^{(i)}\|^2.\end{aligned} \tag{2.57}$$

Proof. With $\mathbf{Z}(\lambda^{(i)})$ defined by (2.55) and using (2.52) we may write

$$\begin{aligned}\mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)} - \mathbf{M}\mathbf{x}^{(i)} =& \frac{1}{\lambda^{(i)}}(\mathbf{A} - \mathbf{E})\mathbf{x}^{(i)} - \mathbf{M}\mathbf{x}^{(i)} \\ =& \frac{1}{\lambda^{(i)}}(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{x}^{(i)} - \frac{1}{\lambda^{(i)}}\mathbf{E}\mathbf{x}^{(i)}.\end{aligned} \tag{2.58}$$

Hence, with (2.12) we get

$$\|\mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)} - \mathbf{M}\mathbf{x}^{(i)}\| = \frac{1}{|\lambda^{(i)}|}\|\mathbf{r}^{(i)} - \mathbf{E}\mathbf{x}^{(i)}\|, \tag{2.59}$$

and using (2.13) we obtain

$$\|\mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)} - \mathbf{M}\mathbf{x}^{(i)}\| \leq \frac{C_1}{|\lambda^{(i)}|}\|\mathbf{e}^{(i)}\| + \frac{1}{|\lambda^{(i)}|}\|\mathbf{E}\mathbf{x}^{(i)}\|. \tag{2.60}$$

Now using (2.27), and with $\tau^{(i)}$ satisfying (2.28), we obtain the convergence result (2.56) which shows linear convergence. Similarly, if $\tau^{(i)}$ satisfies a fixed tolerance (i.e. $\tau^{(i)} = \tau$) then (2.57) is derived from (2.27). □

This proof illustrates the importance of the scaling factor $\frac{1}{\lambda^{(i)}}$ on the right hand side of (2.54). Inequalities (2.56) and (2.57) show the dominant terms that influence the linear rate of convergence for a decreasing and a fixed tolerance respectively. The following example illustrates that (2.56) does indeed describe what is observed in practice.

Example 2.14. *We use the same matrix* $\mathbf{A}$ *as in Example 2.9 but here* $\mathbf{M}$ *is a symmetric tridiagonal matrix with* $2/3$ *as diagonal and* $1/6$ *as sub- and superdiagonals. Again, we seek the smallest eigenvalue, which in this case is given by* $\lambda_1 \approx 32.17511440$.

We apply inexact inverse iteration to the problem with modified right hand side (2.54) and we use the algorithm with decreasing tolerance (2.28). Again, the initial vector $\mathbf{x}^{(0)}$ *is chosen to be sufficiently close to the eigenvector* $\mathbf{x}_1$*, the decreasing tolerance* $\tau^{(i)}$ *and the stopping condition are chosen as in Example 2.9. Hence (2.56) holds.*

We apply GMRES to (2.54) where $\tilde{\mathbf{P}} = \tilde{\mathbf{L}}\tilde{\mathbf{U}}$*, with* $\tilde{\mathbf{L}}$ *and* $\tilde{\mathbf{U}}$ *chosen as in (2.52). Note, the preconditioner is only applied to the left hand side of the equation and the right hand side is scaled by the computed eigenvalue. Furthermore we perform the incomplete LU factorization of* $\mathbf{A}$ *with different drop tolerances ranging from* 10^{-2} *to* 10^{-4} *giving an increasingly better preconditioner.*

Table 2.1: *Results for Example 2.14. The table gives values for* $\|\mathbf{E}\|_\infty$ *and the numerical values of the slopes of the corresponding lines in Figure 2-2 for different drop tolerances of the preconditioner*

Drop Tolerance	$\|\mathbf{E}\|_\infty$	Slope in Figure 2-2
1.0e-2	1.672e+2	0.769
1.0e-3	2.395e+1	0.287
1.0e-4	2.846e+0	0.035

In Figure 2-2 the results for the convergence rate of the outer iteration with decreasing tolerance are given. As predicted in (2.56) we observe linear convergence in each experiment. Table 2.1 gives values of $\|\mathbf{E}\|_\infty$ and the numerical values of the slopes of the corresponding lines in Figure 2-2. We see that as the drop tolerance decreases by a factor of 10 the slope reduces approximately by a factor of 10 as predicted by (2.56).

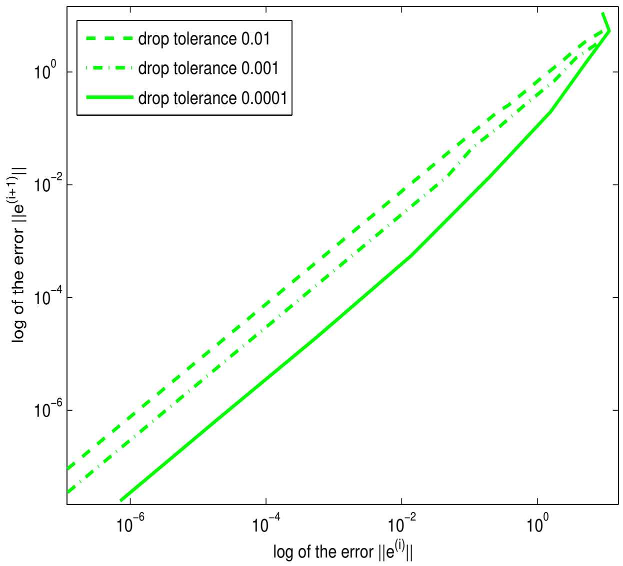


Figure 2-2: *Numerical results for Example 2.14. Outer convergence rate are shown for inexact inverse iteration using ILU preconditioned GMRES with different drop tolerances and scaled modified right hand side*

We would like to note that the preconditioner improves the number of inner iterations whereas the scaling of the right hand side reduces the number of outer iterations and therefore the convergence rate.
We now compute larger eigenvalues to investigate the influence of the term $\frac{1}{|\lambda^{(i)}|}\|\mathbf{E}\mathbf{x}^{(i)}\|$ from (2.56) on the linear rate of convergence.

Example 2.15. *In this example we use the same setup as in Example 2.14, but here we seek the first and the fourth smallest eigenvalues which are given by $\lambda_1 \approx 32.1751$ and $\lambda_9 \approx 177.8825$. We keep the drop tolerance to 10^{-4} and choose the initial vectors sufficiently close to the eigenvectors $\mathbf{x}_1$ and $\mathbf{x}_9$.*

Figure 2-3 shows the results for this test. Note that $\frac{1}{|\lambda_1|}\|\mathbf{E}\mathbf{x}_1\| \approx 0.0111$ and $\frac{1}{|\lambda_9|}\|\mathbf{E}\mathbf{x}_9\| \approx 0.0036$. We see that we obtain a better linear convergence rate for the larger eigenvalue, as predicted by formula (2.56).

The next experiment compares the costs of solving (2.53) with the costs of solving (2.54), where $\tilde{\mathbf{P}} = \tilde{\mathbf{L}}\tilde{\mathbf{U}}$. The costs are measured in terms of the number of inner iterations at each outer iteration, which in this case equals the number of matrix-vector products used.

Example 2.16. *Consider Example 2.14 with a drop tolerance of 10^{-4}. We use (2.53) with a decreasing tolerance, so as to retain quadratic convergence by Corollary 2.8. For (2.54) we use a fixed tolerance, since, by Corollary 2.13 we can only achieve linear convergence no matter the choice of $\tau^{(i)}$. For the fixed tolerance we use $\tau = 0.1$, and for the decreasing tolerance, $\tau^{(i)} = \min\{0.1, 0.1\|\mathbf{r}^{(i)}\|\}$. As a stopping condition we use $\|\mathbf{r}^{(i)}\| \leq 10^{-10}$.*

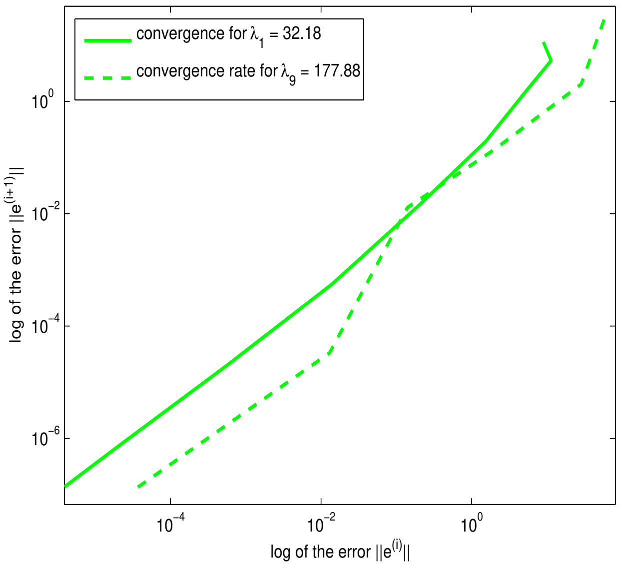


Figure 2-3: *Numerical results for Example 2.15. Outer convergence rates for inexact inverse iteration using ILU preconditioning for the solves by GMRES for different eigenvalues*

Table 2.2: *Iteration numbers for Example 2.16. Total number of iterations and number of inner iterations for inexact inverse iteration using either solves of (2.53) with decreasing tolerance or (2.54) with fixed tolerance*

Outer Iterations	Example (2.54)	Example (2.53)
1	2	2
2	2	2
3	2	3
4	2	5
5	2	10
6	3	16
7	3	
8	4	
9	3	
10	3	
11	4	
total	30	38

The number of inner solves per outer iteration are listed in Table 2.2. We observe that the method based on (2.54) needs more outer iterations, since the convergence is only linear whereas the method based on (2.53) and decreasing tolerance achieves quadratic convergence. However for (2.54) the number of inner iterations is approximately constant at each outer iteration whereas the number of inner iterations per outer iteration increases for (2.53). This phenomenon is discussed for symmetric problems in [10] and for nonsymmetric problems in [11]. For this example with this drop tolerance the linearly convergent method based on using (2.54) is seen to be more efficient than the quadratically convergent method based on (2.53). In other experiments on this example the quadratically convergent method was more efficient, so it is not clear which approach should be preferred in other applications.

Note that further experiments show that we can increase the fixed tolerance for (2.54) to about $\tau = 0.7$ without getting any significant change in the total number of iterations. This can be explained by formula (2.57), where the two terms in $\frac{1}{|\lambda^{(i)}|}\|\mathbf{E}\mathbf{x}^{(i)}\| + \tau\|\mathbf{Z}(\lambda^{(i)})\mathbf{x}^{(i)}\|$ have similar orders of magnitude up to around $\tau = 0.7$.

We have seen that exact preconditioning with a scaling in the right hand side gives quadratic convergence and cheap inner solves, whereas preconditioning with an ILU factorisation together with a scaling of the right hand side gives only linear convergence. In the next section we show how both advantages can be combined, that is, how Corollary 2.12 can be used to achieve both quadratic convergence and cheap inner solves.

2.5.2 Incomplete LU preconditioning and tuning

In this subsection we again consider the use of an incomplete LU factorisation of $\mathbf{A}$ as a preconditioner for $\mathbf{A} - \lambda^{(i)}\mathbf{M}$. This is common in applications involving discretised PDEs, where there is a well-established technology for obtaining a good preconditioner for $\mathbf{A}$ and where $\mathbf{M}$ usually represents a discretised lower order operator. We shall explain how condition (2.50) in Corollary 2.12 may be achieved and implemented and then present two numerical examples.

Assume the incomplete factorisation

$$\mathbf{A} = \mathbf{L}\mathbf{U} + \mathbf{E},$$

where $\mathbf{L}$ is a lower triangular matrix and $\mathbf{U}$ is an upper triangular matrix. The matrix $\mathbf{E}$ is the error matrix. We take

$$\mathbf{P}_S := \mathbf{L}\mathbf{U}, \tag{2.61}$$

as the preconditioner for $\mathbf{A} - \lambda^{(i)}\mathbf{M}$. We shall call this the "standard" preconditioner. However, there is no reason why $\mathbf{P}_S$ should satisfy (2.50), but we now show how a simple modification of $\mathbf{P}_S$ can ensure that (2.50) is achieved. We define

$$\mathbf{f}^{(i)} := \mathbf{A}\mathbf{x}^{(i)} - \mathbf{P}_S\mathbf{x}^{(i)} \tag{2.62}$$

for a given $\mathbf{x}^{(i)}$ and introduce the preconditioner

$$\mathbb{P}_i := \mathbf{P}_S + \mathbf{f}^{(i)}\mathbf{c}^H, \tag{2.63}$$

where $\mathbf{c}^H\mathbf{x}^{(i)} = 1$, with $\mathbf{c}$ being the normalisation vector in Algorithm 2. Clearly, by construction,

$$\mathbb{P}_i\mathbf{x}^{(i)} = \mathbf{A}\mathbf{x}^{(i)} \tag{2.64}$$

and (2.50) holds for $\mathbf{P}_i := \mathbb{P}_i$. We say that $\mathbb{P}_i$ is "tuned" in the sense that, as well as being a preconditioner in the usual sense, $\mathbb{P}_i$ agrees with $\mathbf{A}$ in the direction $\mathbf{x}^{(i)}$, the current estimate for $\mathbf{x}_1$. Note that $\mathbb{P}_i$ is a rank-one change of $\mathbf{P}_S$, and so, using the Sherman-Morrison formula (see, for example, [23, p. 95]), the additional costs of calculating the action of $\mathbb{P}_i^{-1}$ compared with the action of $\mathbf{P}_S^{-1}$ for a given i, is merely one forward and one back substitution. Therefore, in Algorithm 2, step (2), we solve

$$(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \frac{1}{\lambda^{(i)}}\mathbb{P}_i\mathbf{x}^{(i)} \tag{2.65}$$

so that $\mathbf{Z}(\lambda^{(i)}) = \frac{1}{\lambda^{(i)}}\mathbb{P}_i$, with $\mathbb{P}_i$ given by (2.63), and Algorithm 2 should achieve quadratic convergence. Note that (2.65) is implemented as

$$\mathbb{P}_i^{-1}(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \frac{1}{\lambda^{(i)}}\mathbf{x}^{(i)}. \tag{2.66}$$

We now give two numerical examples to illustrate this Corollary and also to compare the performance of $\mathbb{P}_i$ and $\mathbf{P}_S$ as preconditioners for the standard shifted system $(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{M}\mathbf{x}^{(i)}$.[1]

Example 2.17. *We consider the same convection-diffusion operator with Dirichlet boundary conditions as in Example 2.9 but a generalised eigenvalue problem $\mathbf{A}\mathbf{x} = \lambda\mathbf{M}\mathbf{x}$ is derived by discretising (2.35) using a Galerkin-FEM on regular triangular elements with piecewise linear functions. We use a 32×32 grid leading to 961 degrees of freedom. Again, we seek the smallest eigenvalue, which in this case is given by $\lambda_1 \approx 32.15825765$. As initial guess we take a vector with $\cos(\mathbf{x}_1, \mathbf{x}^{(0)}) \approx 0.79$ and as an initial eigenvalue we take $\lambda^{(0)} = 20$. As solver we take preconditioned GMRES with either the usual ILU preconditioner, $\mathbf{P}_S$ given by (2.61), or the tuned preconditioner, $\mathbb{P}_i$ given by (2.63). We compare the costs of the following three methods.*

(a) "$\mathbb{P}_i$/modified-rhs": the tuned preconditioner is applied to the inverse iteration system with a modified right hand side, namely,

$$\mathbb{P}_i^{-1}(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \frac{1}{\lambda^{(i)}}\mathbf{x}^{(i)}, \quad \mathbb{P}_i = \mathbf{LU} + \mathbf{f}^{(i)}\mathbf{c}^H, \tag{2.67}$$

where $\mathbf{U}$ and $\mathbf{L}$ are given by the ILU decomposition, and with $\mathbf{f}^{(i)}$ given by $\mathbf{f}^{(i)} = \mathbf{A}\mathbf{x}^{(i)} - \mathbf{LU}\mathbf{x}^{(i)}$.

(b) "$\mathbb{P}_i$/standard-rhs": the tuned preconditioner is applied to the standard inverse iteration system, namely,

$$\mathbb{P}_i^{-1}(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbb{P}_i^{-1}\mathbf{M}\mathbf{x}^{(i)}, \quad \mathbb{P}_i = \mathbf{LU} + \mathbf{f}^{(i)}\mathbf{c}^H, \tag{2.68}$$

and $\mathbf{f}^{(i)} = \mathbf{A}\mathbf{x}^{(i)} - \mathbf{LU}\mathbf{x}^{(i)}$.

(c) "$\mathbf{P}_S$/standard-rhs": the usual ILU preconditioner is applied to the standard inverse iteration system, namely,

$$\mathbf{P}_S^{-1}(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{P}_S^{-1}\mathbf{M}\mathbf{x}^{(i)}, \quad \mathbf{P}_S = \mathbf{LU}. \tag{2.69}$$

In each case we use the decreasing tolerance $\tau^{(i)} = \min\{\tau, \|\mathbf{r}^{(i)}\|\}$ with $\tau = 0.5$. So all three methods have quadratic convergence using Corollaries 2.8 and 2.12. The iteration stops once the relative residual satisfies $\left\|\frac{\mathbf{r}^{(\mathbf{i})}}{\lambda^{(i)}}\right\| < 10^{-14}$. As in Example 2.9 $\|\cdot\| = \|\cdot\|_2$.

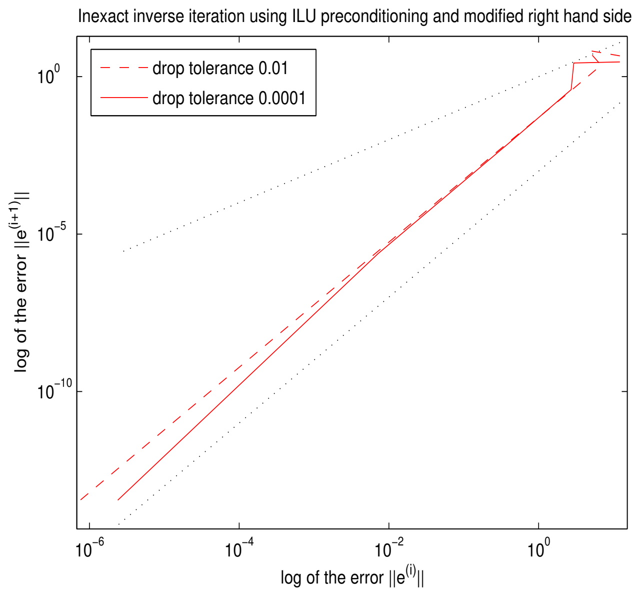


Figure 2-4: *Numerical results for Example 2.17. The quadratic outer convergence rate for method "$\mathbb{P}_i$/modified-rhs" with different drop tolerances is readily observed.*

Table 2.3: *Iteration numbers for Example 2.17. Total number of iterations and number of inner iterations for the three methods using (2.67), (2.68) or (2.69) with decreasing tolerance. In each method the drop tolerances were* 10^{-2} *and* 10^{-4}.

	"$\mathbb{P}_i$/modified-rhs"		"$\mathbb{P}_i$/standard-rhs"		"$\mathbf{P}_S$/standard-rhs"	
Outer It.	10^{-2}	10^{-4}	10^{-2}	10^{-4}	10^{-2}	10^{-4}
1	1	1	1	1	1	1
2	3	3	7	5	9	5
3	3	3	11	7	13	8
4	6	5	13	8	18	13
5	8	7	16	8	28	18
6	13	13				
7	18					
total	52	32	48	29	69	45

In Figure 2-4 we give logarithmic plots of the errors obtained from method "$\mathbb{P}_i$/ modified-rhs" for two different drop tolerances. The dotted lines indicate the slopes expected for linear and quadratic convergence. As predicted in Corollary 2.12 we achieve quadratic convergence. Table 2.3 shows the number of inner iterations for the inexact solves of the three methods. We see that for both drop tolerances the tuned preconditioner applied to the standard inverse iteration formulation, method

[1]Note that the *costs* of the methods are measured in terms of the number of inner iterations per outer iterations used. For the tuned preconditioner the cost of applying preconditioner is slightly higher, since one extra solve with $\mathbf{P}$ and a matrix-vector product per outer iteration and some inner products are needed. We assume here that these costs are negligible, since the modification of $\mathbf{P}$ is just done with vectors and only one extra daxpy is needed per outer iteration.

"$\mathbb{P}_i$/standard-rhs", requires fewer iterations than the other two methods. In particular, comparing the results for "$\mathbb{P}_i$/standard-rhs" with "$\mathbf{P}_S$/standard-rhs" we see that the tuned preconditioner is significantly better than the usual ILU preconditioner. Method "$\mathbb{P}_i$/standard-rhs" requires fewer outer iterations than "$\mathbb{P}_i$/modified-rhs", which may be explained by considering the constants in the convergence theory.

In particular, method (a) is sensitive to the starting guess whereas methods (b) and (c) are more robust with respect to the starting vector. For example if we choose a starting vectors with $\cos(\mathbf{x}_1, \mathbf{x}^{(0)}) \approx 0.47$ method (a) fails to work, whereas methods (b) and (c) work fine, with (b) again proving superior to (c).

Hence, if the preconditioner is tuned as explained in this section, it appears to be best to apply it on the standard system $(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{M}\mathbf{x}^{(i)}$ rather than consider modifying the right hand side. This gives the best result in terms of the total number of iterations and the convergence rate. A simple, heuristic, explanation of this is as follows. We see from (2.68) that near convergence

$$\mathbb{P}_i^{-1}(\mathbf{A} - \lambda^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbb{P}_i^{-1}\mathbf{M}\mathbf{x}^{(i)} \approx \frac{1}{\lambda^{(i)}}\mathbb{P}_i^{-1}\mathbf{A}\mathbf{x}^{(i)} = \frac{1}{\lambda^{(i)}}\mathbf{x}^{(i)},$$

and so the right hand side $\mathbb{P}_i^{-1}\mathbf{M}\mathbf{x}^{(i)}$ is roughly in the desired direction, thus keeping the costs of the inner solves of GMRES low.

Next we present an example arising in reactor design (see [113] for details).

Example 2.18. *The standard model to describe the neutron balance in a 2D model of a nuclear reactor is given by the two-group neutron equations*

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

where u_1 and u_2 are defined on $[0,1] \times [0,1]$ and represent the density distributions of fast and thermic neutrons respectively. K_1 and K_2 are diffusion coefficients and $\Sigma_{a,1}, \Sigma_{a,2}, \Sigma_s, \Sigma_{f,1}$ and $\Sigma_{f,2}$ measure interaction probabilities and take different piece-wise constant values in different regions of the reactor, which for this example are given in Figure 2-5 and Table 2.4. The largest μ_1 such that $1/\mu_1$ is an eigenvalue of the

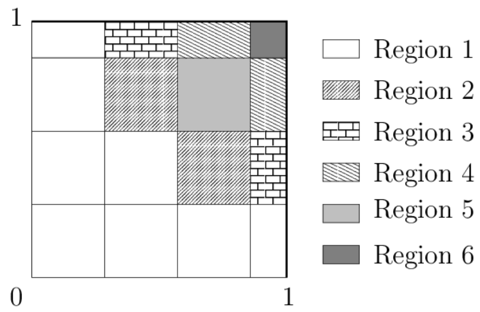


Figure 2-5: *Nuclear reactor problem geometry.*

Table 2.4: *Data for the nuclear reactor problem.*

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

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

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

Discretising the problem using a finite difference approximation on a $h \times h$ grid, where $h = 1/m$ we obtain a $2m^2 \times 2m^2$ discrete eigenproblem $\mathbf{Au} = \lambda \mathbf{Mu}$, where $\mathbf{A}$ and $\mathbf{M}$ are both nonsymmetric and $\mathbf{M}$ is singular. We seek the smallest eigenvalue $\lambda_1 (= 1/\mu_1)$, which determines the criticality of the reactor. We choose $m = 32$, which leads to a system of size $n = 2048$. For initial conditions, we take $\lambda^{(0)} = 1$, since, as discussed earlier, for $\mu_1 = \frac{1}{lambda_1} = 1$ the reactor is in the critical phase. Furthermore we take $\mathbf{u}^{(0)} = [1, \ldots, 1]^T / \sqrt{n}$. We use the decreasing tolerance $\tau^{(i)} = \min\{\tau, \| \mathbf{r}^{(i)} \|\}$ with $\tau = 0.3$. The iteration stops once the relative residual satisfies $\left\| \frac{\mathbf{r^{(i)}}}{\lambda^{(i)}} \right\| < 10^{-11}$. As in Example 2.9 $\| \cdot \| = \| \cdot \|_2$. In fact, the exact eigenvalue is given by $\lambda_1 = 0.9707$ and $\cos(\mathbf{u}_1, \mathbf{u}^{(0)}) \approx 0.44$. We compare methods (a), (b) and (c) as in Example 2.17.

Table 2.5: *Iteration numbers for Example 2.18. Total number of iterations and number of inner iterations for the three methods using (2.67), (2.68) or (2.69) with decreasing tolerance. In each method the drop tolerances were 10^{-1} and 10^{-2}.*

	"$\mathbb{P}_i$/modified-rhs"		"$\mathbb{P}_i$/standard-rhs"		"$\mathbf{P}_S$/standard-rhs"	
Outer It.	10^{-1}	10^{-2}	10^{-1}	10^{-2}	10^{-1}	10^{-2}
1	6	7	4	4	1	4
2	4	10	2	11	8	11
3	12	21	5	22	3	27
4	22	29	27	27	5	37
5	43		38		26	45
6	65		59		38	
7					55	
8					76	
total	152	67	135	64	212	124

Table 2.5 shows the results obtained by methods (a), (b) and (c). Again, we use an ILU preconditioner with, in this case, drop tolerances of 0.1 and 0.01. We observe that the use of the tuned preconditioner applied to the standard formulation (see the middle columns in Table 2.5) provides the best results with respect to overall costs. Also, the standard preconditioner applied to the standard formulation (see the right hand columns) performs least well. These numerical results are consistent with those obtained in the previous example and confirm the usefulness and applicability of the tuned preconditioner.

2.6 Conclusions

We have analysed inexact inverse iteration for a generalised eigenvalue problem and we have shown that for an algebraically simple eigenvalue it is a modified Newton method. Using the convergence theory of the modified Newton method we obtained convergence rates for inexact solves with either fixed or decreasing tolerances. Furthermore, we have analysed a simplified version of inexact Jacobi-Davidson's method using inexact Newton theory. This approach is much simpler than previous approaches involving eigenvector expansions. Using the same tool, we also analysed preconditioned iterative solves. In situations where the right hand side in inexact inverse iteration is modified we have shown how an ILU preconditioner may be tuned to recover quadratic convergence. Additionally, we have given two examples which indicate that the application of the tuned preconditioner may be advantageous when applied to the standard inverse iteration formulation.

This chapter has shown that it is advantageous to consider special preconditioners for the solution of eigenvalue problems, due to the structure of the problem.

CHAPTER 3

Convergence theory for inexact inverse iteration applied to the generalised nonsymmetric eigenproblem

3.1 Introduction

As in the previous chapter, we consider the computation of a simple, finite eigenvalue and corresponding eigenvector of the generalised eigenvalue problem

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{M}\mathbf{x}, \quad \mathbf{x} \neq \mathbf{0}, \tag{3.1}$$

where $\mathbf{A} \in \mathbb{C}^{n\times n}$ and $\mathbf{M} \in \mathbb{C}^{n\times n}$ are large and sparse. We shall explore, under minimal assumptions, convergence rates attained by inexact inverse iteration, illustrate the theory with reference to some physical examples, and obtain a convergence result for a version of the inexact Jacobi-Davidson method. Unlike in the previous chapter we use a splitting approach to obtain this convergence result and, in doing so, obtain a more flexible treatment that can be used to analyse a variety of shift strategies.

The paper by Golub and Ye [50] provided a convergence theory of inexact inverse iteration for a fixed shift strategy for nonsingular $\mathbf{M}$ with $\mathbf{M}^{-1}\mathbf{A}$ diagonalisable. Linear convergence is proved if a suitable solve tolerance is chosen to decrease linearly. An early paper, which also considers inexact inverse iteration applied to a diagonalisable problem is the one by Lai et al. [75]. They provide a theory for the standard eigenproblem with a fixed shift strategy and obtain linear convergence for both the eigenvalue and the eigenvector if the solve tolerance decreases depending on a quantity containing unknown parameters. They also give numerical results on a transformed generalised eigenvalue problem. In [12] a convergence theory is given for Rayleigh quotient shifts assuming $\mathbf{M}$ is symmetric positive definite. Following [50], the convergence theory in [12] used a decomposition in terms of the right eigenvectors. One result in [12] is that for a variable shift strategy, the linear systems need not be solved accurately to obtain a convergent method.

In this chapter we consider a quite general setting, where $\mathbf{A}$ and $\mathbf{M}$ are nonsymmetric matrices with both $\mathbf{A}$ and $\mathbf{M}$ allowed to be singular, but without a common null vector. We only assume that the sought eigenpair $(\lambda_1, \mathbf{x}_1)$ is simple, well-separated and finite. We provide a convergence theory for inexact inverse iteration applied to this generalised eigenproblem for both fixed and variable shifts. This theory extends the results of the previous chapter (see also [43]), since this new theory holds for any shift, not just the shift that gives the equivalence of Newton's method to inverse iteration. Also, the convergence rate is seen to depend on how close the sought eigenvalue is to the rest of the spectrum, a natural result that is somewhat hidden in the theory in Chapter 2 (see also [43]). We use a decomposition that allows us to consider nondiagonalisable problems where $\mathbf{M}$ may be singular. To be precise, we use a splitting of the approximate right eigenvector in terms of the exact right eigenvector and a basis of a right invariant subspace. This is an approach used by Stewart [136] to provide a perturbation theory of invariant subspaces, and allows us to overcome the theoretical dependence of the allowed solve tolerance on the basis of eigenvectors, which appeared in [50] and [12]. If a decreasing solve tolerance is required then we take it to be proportional to the eigenvalue residual, as was done in [12].

It is well-known that there is a close connection between inverse iteration and the Jacobi-Davidson method, see [123, 124, 126] and Section 2.4. We shall use the convergence theory developed here for inexact inverse iteration applied to (3.1) to provide a convergence theory for a version of inexact simplified Jacobi-Davidson.

The chapter is organised as follows. In Section 3.2 standard results on the generalised eigenproblem are summarised and a generalised Rayleigh quotient is discussed. Section 3.3 provides the main result of the chapter; a new convergence measure is introduced and the main convergence result for inexact inverse iteration applied to the generalised non-Hermitian eigenproblem is stated and proved. Section 3.4 contains some additional convergence results. In Section 3.5 we give numerical tests on examples arising from modeling of a nuclear reactor and the linearised incompressible Navier-Stokes equations. Section 3.6 presents a convergence analysis for the inexact simplified Jacobi-Davidson method and provides some numerical results to illustrate the theory.

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

3.2 Standard results on the generalised eigenproblem

In order to state convergence results for Algorithm 4 stated on page 44 we need some results about the generalised eigenproblem. First recall that the eigenvalues of (3.1) are given by $\lambda(\mathbf{A},\mathbf{M}) := \{z \in \mathbb{C} : \det(\mathbf{A} - z\mathbf{M}) = 0\}$.

Note, that $\lambda(\mathbf{A},\mathbf{M})$ could be finite, empty or infinite, since either $\mathbf{A}$ or $\mathbf{M}$ or even both $\mathbf{A}$ and $\mathbf{M}$ could be singular. In particular $\det(\mathbf{A}-z\mathbf{M}) = 0$ for all $z \in \mathbb{C}$ whenever $\mathbf{A}$ and $\mathbf{M}$ have a common null space, which means that in this case $\lambda(\mathbf{A},\mathbf{M}) = \mathbb{C}$.

We use the following theorem for a canonical form of (3.1), which is a generalisation of the Schur Decomposition of the standard eigenproblem.

Theorem 3.1 (Generalised Schur Decomposition). *If $\mathbf{A} \in \mathbb{C}^{n\times n}$ and $\mathbf{M} \in \mathbb{C}^{n\times n}$, then there exist unitary matrices $\mathbf{Q}$ and $\mathbf{Z}$ such that $\mathbf{Q}^H\mathbf{A}\mathbf{Z} = \mathbf{T}$ and $\mathbf{Q}^H\mathbf{M}\mathbf{Z} = \mathbf{S}$ are upper triangular. If for some j, t_{jj} and s_{jj} are both zero, then $\lambda(\mathbf{A},\mathbf{M}) = \mathbb{C}$. If $s_{jj} \neq 0$ then $\lambda(\mathbf{A},\mathbf{M}) = \{t_{jj}/s_{jj}\}$, otherwise, the jth eigenvalue of problem (3.1) is an infinite eigenvalue.*

Proof. See [48, page 377]. □

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

$$\mathbf{Q}^H\mathbf{A}\mathbf{Z} = \begin{bmatrix} t_{11} & \mathbf{t}_{12}^H \\ \mathbf{0} & \mathbf{T}_{22} \end{bmatrix} \quad \text{and} \quad \mathbf{Q}^H\mathbf{M}\mathbf{Z} = \begin{bmatrix} s_{11} & \mathbf{s}_{12}^H \\ \mathbf{0} & \mathbf{S}_{22} \end{bmatrix}, \tag{3.2}$$

where $\mathbf{T}_{22} \in \mathbb{C}^{(n-1)\times(n-1)}$ and $\mathbf{S}_{22} \in \mathbb{C}^{(n-1)\times(n-1)}$. If λ_1, the desired eigenvalue, is finite, then $s_{11} \neq 0$ and $\lambda_1 = t_{11}/s_{11}$. The factorisation (3.2) provides a orthogonal similarity transform, but in order to decompose the problem for the convergence analysis into span$\{\mathbf{x}_1\}$ and the invariant subspace containing the other eigenvectors we make a further transformation to block diagonalise the problem. To this end we define the linear transformation $\Phi : \mathbb{C}^{(n-1)\times 2} \to \mathbb{C}^{(n-1)\times 2}$ by

$$\Phi(\mathbf{h}, \mathbf{g}) := (t_{11}\mathbf{h} - \mathbf{T}_{22}^H\mathbf{g}, s_{11}\mathbf{h} - \mathbf{S}_{22}^H\mathbf{g}), \tag{3.3}$$

where $\mathbf{g} \in \mathbb{C}^{(n-1)\times 1}$ and $\mathbf{h} \in \mathbb{C}^{(n-1)\times 1}$. (This transformation is a simplification of that suggested by Stewart in [133].) The following lemma states conditions under which Φ is nonsingular. A generalisation for the case $\mathbf{g} \in \mathbb{C}^{(n-p)\times p}$, with $p > 1$ was proved in [133].

Lemma 3.2. *The operator Φ from (3.3) is nonsingular if and only if $\lambda_1 = \dfrac{t_{11}}{s_{11}} \notin \lambda(\mathbf{T}_{22}, \mathbf{S}_{22})$.*

Proof. In order to show that Φ is nonsingular we need to show that the system of equations

$$\begin{aligned} t_{11}\mathbf{h} - \mathbf{T}_{22}^H\mathbf{g} &= \mathbf{a}, & (3.4)\\ s_{11}\mathbf{h} - \mathbf{S}_{22}^H\mathbf{g} &= \mathbf{b}, & (3.5) \end{aligned}$$

has a unique solution for any $\mathbf{a}$, $\mathbf{b} \in \mathbb{C}^{n-1}$. This holds if and only if

$$\det \begin{bmatrix} t_{11}\mathbf{I} & -\mathbf{T}_{22}^H \\ s_{11}\mathbf{I} & -\mathbf{S}_{22}^H \end{bmatrix} \neq 0. \tag{3.6}$$

If $t_{11} \neq 0$ we have

$$\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\dfrac{t_{11}}{s_{11}} & \mathbf{I} \end{bmatrix} \begin{bmatrix} t_{11}\mathbf{I} & -\mathbf{T}_{22}^H \\ s_{11}\mathbf{I} & -\mathbf{S}_{22}^H \end{bmatrix} = \begin{bmatrix} t_{11}\mathbf{I} & -\mathbf{T}_{22}^H \\ \mathbf{0} & \dfrac{t_{11}}{s_{11}}\mathbf{T}_{22}^H - \mathbf{S}_{22}^H \end{bmatrix},$$

and hence condition (3.6) is satisfied if and only if

$$\det \begin{bmatrix} t_{11}\mathbf{I} & -\mathbf{T}_{22}^H \\ \mathbf{0} & \dfrac{t_{11}}{s_{11}}\mathbf{T}_{22}^H - \mathbf{S}_{22}^H \end{bmatrix} \neq 0,$$

which holds if and only if $\det\left(\dfrac{t_{11}}{s_{11}}\mathbf{T}_{22}^H - \mathbf{S}_{22}^H\right) \neq 0$ which again holds if and only if $\lambda_1 = \dfrac{t_{11}}{s_{11}} \notin \lambda(\mathbf{T}_{22}, \mathbf{S}_{22})$, proving the statement.

For $t_{11} = 0$ condition (3.6) becomes

$$\det \begin{bmatrix} \mathbf{0} & -\mathbf{T}_{22}^H \\ s_{11}\mathbf{I} & -\mathbf{S}_{22}^H \end{bmatrix} \neq 0,$$

which is satisfied if and only if

$$\det \begin{bmatrix} s_{11}\mathbf{I} & -\mathbf{S}_{22}^H \\ \mathbf{0} & -\mathbf{T}_{22}^H \end{bmatrix} \neq 0,$$

or, equivalently if and only if $\det(\mathbf{T}_{22}) \neq 0$. Hence, system (3.4), (3.5) has a unique solution if and only if $\dfrac{t_{11}}{s_{11}} \notin \lambda(\mathbf{T}_{22}, \mathbf{S}_{22})$, which concludes the proof. □

Hence Φ is nonsingular if and only if λ_1 is a simple eigenvalue of (3.1). With Lemma 3.2 we can prove the following result.

Lemma 3.3. *If the operator Φ from (3.3) is nonsingular then the matrices* $\mathbf{G}$ *and* $\mathbf{H}$ *are given by*

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

where $\Phi(\mathbf{h}_{12}, \mathbf{g}_{12}) = (-\mathbf{t}_{12}, -\mathbf{s}_{12})$, and, with $\mathbf{T}$ *and* $\mathbf{S}$ *defined in Theorem 3.1,*

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

Furthermore,

$$\|\mathbf{H}\|^2 = \|\mathbf{H}^{-1}\|^2 = C_{\|\mathbf{h}_{12}\|}, \quad C_{\|\mathbf{h}_{12}\|} := (\|\mathbf{h}_{12}\|^2 + \sqrt{\|\mathbf{h}_{12}\|^4 + 4\,\|\mathbf{h}_{12}\|^2} + 2)/2, \tag{3.7}$$

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

Proof. Since Φ is nonsingular the vectors $\mathbf{g}_{12}$ and $\mathbf{h}_{12}$ exist and simple calculation gives $\mathbf{G}^{-1}\mathbf{T}\mathbf{H} = \mathrm{diag}(t_{11}, \mathbf{T}_{22})$ and $\mathbf{G}^{-1}\mathbf{S}\mathbf{H} = \mathrm{diag}(s_{11}, \mathbf{S}_{22})$. Result (3.7) follows by direct calculation of the spectral radius of $\mathbf{H}^H\mathbf{H}$. We may use the special structure of $\mathbf{H}$ to determine $\|\mathbf{H}\|$. Using $\|\mathbf{H}\|_2 = (\lambda_{\max}(\mathbf{H}^H\mathbf{H}))^{\frac{1}{2}}$ (see [60]) we have to determine the largest eigenvalue of

$$\mathbf{H}^H\mathbf{H} = \begin{bmatrix} 1 & \mathbf{h}_{12}^H \\ \mathbf{h}_{12} & \mathbf{I}_{n-1} + \mathbf{h}_{12}\mathbf{h}_{12}^H \end{bmatrix}.$$

A simple calculation shows that $(1, [0\,, \mathbf{h}_{12}^\perp]^H)$ is an eigenpair of $\mathbf{H}^H\mathbf{H}$, where $\mathbf{h}_{12}^\perp \in \mathbb{C}^{n-1}$ is a vector in the $n-2$-dimensional subspace orthogonal to $\mathbf{h}_{12}$. There are $n-2$ of these vectors, so only two more eigenvalues have to be found. Let $[\zeta\,, \mathbf{h}_{12}]^H$ be another eigenvector, which is orthogonal to $[0\,, \mathbf{h}_{12}^\perp]^H$, we can then calculate

$$\begin{bmatrix} 1 & \mathbf{h}_{12}^H \\ \mathbf{h}_{12} & \mathbf{I}_{n-1} + \mathbf{h}_{12}\mathbf{h}_{12}^H \end{bmatrix} \begin{bmatrix} \zeta \\ \mathbf{h}_{12} \end{bmatrix} = \begin{bmatrix} \left(1 + \dfrac{\mathbf{h}_{12}^H\mathbf{h}_{12}}{\zeta}\right)\zeta \\ (\zeta + 1 + \mathbf{h}_{12}^H\mathbf{h}_{12})\mathbf{h}_{12} \end{bmatrix}.$$

Solving the quadratic equation $1 + \dfrac{\mathbf{h}_{12}^H\mathbf{h}_{12}}{\zeta} = \zeta + 1 + \mathbf{h}_{12}^H\mathbf{h}_{12}$ for ζ gives the required further two eigenvalues $\dfrac{\|\mathbf{h}_{12}\|^2 \pm \sqrt{\|\mathbf{h}_{12}\|^4 + 4\|\mathbf{h}_{12}\|^2} + 2}{2}$ of which the largest one is given by

$$\lambda_{\max}(\mathbf{H}^H\mathbf{H}) = \frac{\|\mathbf{h}_{12}\|^2 + \sqrt{\|\mathbf{h}_{12}\|^4 + 4\|\mathbf{h}_{12}\|^2} + 2}{2}.$$

Therefore $\|\mathbf{H}\|_2 = \max\{1, \sqrt{C_{\|\mathbf{h}_{12}\|}}\}$ where $C_{\|\mathbf{h}_{12}\|}$ is given by (3.7). Furthermore, $\|\mathbf{H}\| = \|\mathbf{H}^{-1}\|$, because $\mathbf{H}^{-1} = \begin{bmatrix} 1 & -\mathbf{h}_{12}^H \\ \mathbf{0} & \mathbf{I}_{n-1} \end{bmatrix}$ and $\|\mathbf{H}\|$ does not depend on the sign of the upper right block of $\mathbf{H}$. □

Note that $C_{\|\mathbf{h}_{12}\|}$ and $C_{\|\mathbf{g}_{12}\|}$ measure the conditioning of the eigenvalue λ_1, with large values of $C_{\|\mathbf{h}_{12}\|}$ and $C_{\|\mathbf{g}_{12}\|}$ implying a poorly conditioned problem. We shall see in Section 3.3 that $\|\mathbf{g}_{12}\|$ and $\|\mathbf{h}_{12}\|$ appear in the bounds in the convergence theory. Combining Theorem 3.1 and Lemma 3.3 gives the following corollary:

Corollary 3.4. *Define*

$$\mathbf{U} = \mathbf{Q}\mathbf{G} \tag{3.8}$$

and

$$\mathbf{X} = \mathbf{Z}\mathbf{H}. \tag{3.9}$$

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

$$\mathbf{U}^{-1}(\mathbf{A} - \lambda\mathbf{M})\mathbf{X} = \begin{bmatrix} t_{11} & \mathbf{0}^H \\ \mathbf{0} & \mathbf{T}_{22} \end{bmatrix} - \lambda \begin{bmatrix} s_{11} & \mathbf{0}^H \\ \mathbf{0} & \mathbf{S}_{22} \end{bmatrix}. \tag{3.10}$$

For our purposes, decomposition (3.10) has advantages over the Schur factorisation (3.2), since (3.10) allows the eigenvalue problem $\mathbf{A}\mathbf{x} = \lambda\mathbf{M}\mathbf{x}$ to be split into two problems. The first problem is the trivial $\lambda\ t_{11} = s_{11}$. The second problem arising from the $(n-1)\times(n-1)$ block is that of finding $\lambda(\mathbf{T}_{22}, \mathbf{S}_{22})$ which contains the $(n-1)$ eigenvalues excluding λ_1. From (3.10) we have

$$(\mathbf{A} - \lambda_1\mathbf{M})\mathbf{x}_1 = \mathbf{0} \quad \text{and} \quad \mathbf{u}_1^H(\mathbf{A} - \lambda_1\mathbf{M}) = \mathbf{0}, \tag{3.11}$$

where $\lambda_1 = \dfrac{t_{11}}{s_{11}}$ is an eigenvalue of (3.1), with corresponding right and left eigenvectors, $\mathbf{x}_1 = \mathbf{X}\mathbf{e}_1$ and $\mathbf{u}_1 = \mathbf{U}^{-H}\mathbf{e}_1$, where $\mathbf{e}_1$ is the first canonical vector.

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

$$\mathbf{u}_1^H\mathbf{M}\mathbf{x}_1 \neq 0, \tag{3.12}$$

since, by (3.10) and the special structure of $\mathbf{G}$ and $\mathbf{H}$ in Lemma 3.3, we have

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

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

$$\frac{\mathbf{c}^H\mathbf{A}\mathbf{x}}{\mathbf{c}^H\mathbf{M}\mathbf{x}}, \tag{3.13}$$

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

$$\rho(\mathbf{x}) := \frac{\mathbf{x}^H\mathbf{M}^H\mathbf{A}\mathbf{x}}{\mathbf{x}^H\mathbf{M}^H\mathbf{M}\mathbf{x}}, \tag{3.14}$$

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

$$\|\mathbf{A}\mathbf{x} - \rho(\mathbf{x})\mathbf{M}\mathbf{x}\| = \min_{z\in\mathbb{C}} \|\mathbf{A}\mathbf{x} - z\mathbf{M}\mathbf{x}\|. \tag{3.15}$$

(This property can be verified using simple least-squares approximation as in [144, page 203].) If we normalise $\mathbf{x}$ such that $\mathbf{x}^H\mathbf{M}^H\mathbf{M}\mathbf{x} = 1$, then $\rho(\mathbf{x}) = \mathbf{x}^H\mathbf{M}^H\mathbf{A}\mathbf{x}$.

Note that the choice $\mathbf{c}^{(i)} = \mathbf{u}^{(i)}$ in (3.13), where $\mathbf{u}^{(i)}$ is an approximation to the left eigenvector provides the generalised two-sided Rayleigh quotient $\rho(\mathbf{u}^{(i)}, \mathbf{x}^{(i)}) := \frac{\mathbf{u}^{(i)H}\mathbf{A}\mathbf{x}^{(i)}}{\mathbf{u}^{(i)H}\mathbf{M}\mathbf{x}^{(i)}}$ as it was defined in [151, page 179] (see also [114]). Note that in these references the name generalised Rayleigh quotient was used for what we call the two-sided generalised Rayleigh quotient. Since we do not compute the left eigenvector here we will not use this approximation.

3.3 Inexact inverse iteration

We assume that the generalised nonsymmetric eigenproblem (3.1) has a simple, well-separated eigenvalue (λ_1 satisfying (3.11) and(3.12)). This section contains the convergence theory for inexact inverse iteration described by Algorithm 4.

Algorithm 4 Inexact Inverse Iteration for the generalised eigenproblem

Input: $\mathbf{x}^{(0)}$, i_{max}.
  for $i = 1, \ldots, i_{max}$ **do**
    Choose $\sigma^{(i)}$ and $\tau^{(i)}$,
    Find $\mathbf{y}^{(i)}$ such that $\|(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} - \mathbf{M}\mathbf{x}^{(i)}\| \leq \tau^{(i)}\|\mathbf{M}\mathbf{x}^{(i)}\|$,
    Set $\mathbf{x}^{(i+1)} = \mathbf{y}^{(i)}/\phi(\mathbf{y}^{(i)})$,
    Set $\lambda^{(i+1)} = \rho(\mathbf{x}^{(i+1)})$,
    Evaluate $\mathbf{r}^{(i+1)} = (\mathbf{A} - \lambda^{(i+1)}\mathbf{M})\mathbf{x}^{(i+1)}$,
    Test for convergence.
  end for
Output: $\mathbf{x}^{(i_{max})}$.

Note that we choose $\lambda^{(i+1)} = \rho(\mathbf{x}^{(i+1)})$ to make use of the minimisation property (3.15). Also, in Algorithm 4 the function $\phi(\mathbf{y}^{(i)})$ is a scalar normalisation. Common choices for this normalisation are $\phi(\mathbf{y}^{(i)}) = \mathbf{z}^{(i)H}\mathbf{y}^{(i)}$, for some $\mathbf{z}^{(i)} \in \mathbb{C}^n$, or a norm of $\mathbf{y}^{(i)}$, such as $\phi(\mathbf{y}^{(i)}) = \|\mathbf{y}^{(i)}\|_2$ or, if $\mathbf{M}$ is positive definite, $\phi(\mathbf{y}^{(i)}) = \|\mathbf{y}^{(i)}\|_{\mathbf{M}}$.

We introduce a new convergence measure in Section 3.3.1, provide a one step bound in Section 3.3.2 and finally give convergence results for both fixed and variable shifts in Section 3.3.3. In Section 3.4 we discuss some properties of the function $\phi(\mathbf{y})$.

3.3.1 The measure of convergence

In order to analyse the convergence of inexact inverse iteration we use a different approach to the one used in [12], [50] where the splitting was done in terms of the right eigenvectors of the problem. We split the approximate right eigenvector into two components: the first is in the direction of the exact right eigenvector, and the second lies in the right invariant subspace not containing the exact eigenvector. This decomposition is based on that used by [136] for the perturbation theory of invariant subspaces. However, we introduce a scaling, namely $\alpha^{(i)}$ as in [12], which turns out to be advantageous in the analysis. Let us decompose $\mathbf{x}^{(i)}$, the vector approximating $\mathbf{x}_1$, as

$$\frac{\mathbf{x}^{(i)}}{\phi(\mathbf{x}^{(i)})} = (\mathbf{x}_1 q^{(i)} + \mathbf{X}_2\mathbf{p}^{(i)}),$$

where $q^{(i)} \in \mathbb{C}$, $\mathbf{p}^{(i)} \in \mathbb{C}^{(n-1)\times 1}$ and $\mathbf{X}_2 = \mathbf{X}\bar{\mathbf{I}}_{n-1}$, where $\mathbf{X}$ is given by (3.9) and

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

with $\mathbf{I}_{n-1}$ being the identity matrix of size $(n-1)$ and $\phi(\mathbf{x}^{(i)})$ determines the normalisation of $\mathbf{x}^{(i)}$. Then with $\alpha^{(i)} := \phi(\mathbf{x}^{(i)})$ we have the decomposition

$$\mathbf{x}^{(i)} = \alpha^{(i)}(\mathbf{x}_1 q^{(i)} + \mathbf{X}_2\mathbf{p}^{(i)}). \tag{3.16}$$

for the approximate eigenvector $\mathbf{x}^{(i)}$. For the convergence theory we leave the scaling of the approximate eigenvector and exact right eigenvector $\mathbf{x}^{(i)}$ and $\mathbf{x}_1$ open, however, in Sections 3.4 and 3.5, we will use $\|\mathbf{M}\mathbf{x}^{(i)}\| = 1$.

Clearly $q^{(i)}$ and $\mathbf{p}^{(i)}$ measure how close the approximate eigenvector $\mathbf{x}^{(i)}$ is to the sought eigenvector $\mathbf{x}_1$. As we shall see in the following analysis the advantage of this splitting is that we need not be concerned about any highly non-normal behaviour in the matrix pair $(\mathbf{T}_{22}, \mathbf{S}_{22})$. This is in contrast to the approach in [12], where the splitting only existed for positive definite $\mathbf{M}$ and involved a bound on the condition number of the matrix of eigenvectors. However, our analysis uses the separation between λ_1 and the spectrum of the matrix pair $(\mathbf{T}_{22}, \mathbf{S}_{22})$ given by $\mathrm{sep}(\lambda_1, (\mathbf{T}_{22}, \mathbf{S}_2 2))$. This quantity can be arbitrarily smaller than the actual distance between the eigenvalue λ_1 and the eigenvalues of the matrix pair $(\mathbf{T}_{22}, \mathbf{S}_{22})$ (see [137, page 234]). Hence, our bound (see 3.30) might lead to unnecessarily slow theoretical convergence rates. Note that the splitting in (3.16) is a generalisation of orthogonal decomposition introduced in [101], where

$$\mathbf{x}^{(i)} = \cos\theta^{(i)}\mathbf{x}_1 + \sin\theta^{(i)}\mathbf{x}_\perp^{(i)}, \quad \mathbf{x}_\perp^{(i)} \perp \mathbf{x}_1, \tag{3.17}$$

with $\|\mathbf{x}_1\| = \|\mathbf{x}_\perp^{(i)}\| = 1$ and $\theta^{(i)} = \angle(\mathbf{x}^{(i)}, \mathbf{x}_1)$. The error can than be measured by $\tan\theta^{(i)}$. However, this decomposition is only possible if the matrix $\mathbf{A}$ is Hermitian, i.e. has a complete set of orthonormal eigenvectors. Now set

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

and multiply (3.16) from the left by $\mathbf{U}^{-1}\mathbf{M}$. Using

$$\mathbf{U}^{-1}\mathbf{M}\mathbf{x}_1 = s_{11}\mathbf{e}_1 \quad \text{and} \quad \mathbf{U}^{-1}\mathbf{M}\mathbf{X}_2 = \begin{bmatrix} \mathbf{e}_2 & \dots & \mathbf{e}_n \end{bmatrix}\mathbf{S}_{22} = \bar{\mathbf{I}}_{n-1}\mathbf{S}_{22}, \tag{3.18}$$

from (3.10), where $\mathbf{e}_i$ is the ith canonical vector, we have

$$\begin{aligned} 1 = \frac{\|\mathbf{U}^{-1}\mathbf{M}\mathbf{x}^{(i)}\|}{\alpha^{(i)}} &= \|s_{11}q^{(i)}\mathbf{e}_1 + \bar{\mathbf{I}}_{n-1}\mathbf{S}_{22}\mathbf{p}^{(i)}\| \\ &= ((s_{11}q^{(i)})^2 + \|\mathbf{S}_{22}\mathbf{p}^{(i)}\|^2)^{\frac{1}{2}}. \end{aligned} \tag{3.19}$$

Thus $|s_{11}q^{(i)}|$ and $\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|$ can be interpreted as generalisations of the cosine and sine functions as used in the orthogonal decomposition for the symmetric eigenproblem, [101]. Also, from (3.19), we have $|s_{11}q^{(i)}| \leq 1$ and $\|\mathbf{S}_{22}\mathbf{p}^{(i)}\| \leq 1$. Note that (3.19) also indicates why $\alpha^{(i)}$ was introduced in (3.16). This scaling is not used in [136] or [137]. It is now natural to introduce

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

as our measure for convergence. Comparing with the orthogonal splitting in (3.17) where the convergence is measured by $\tan\theta^{(i)} = \sin\theta^{(i)}/\cos\theta^{(i)}$, equation (3.19) shows that $T^{(i)}$ can be interpreted as a generalised tangent. Using (3.16) we have, for $\alpha^{(i)}q^{(i)} \neq 0$,

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

and also

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

Using the last two bounds together with (3.9) we obtain

$$\frac{1}{\|\mathbf{H}^{-1}\|}\frac{\|\mathbf{p}^{(i)}\|}{|q^{(i)}|} \leq \|\frac{\mathbf{x}^{(i)}}{\alpha^{(i)}q^{(i)}} - \mathbf{x}_1\| \leq \|\mathbf{H}\|\frac{\|\mathbf{p}^{(i)}\|}{|q^{(i)}|}, \tag{3.20}$$

with expressions on $\|\mathbf{H}\|$ and $\|\mathbf{H}^{-1}\|$ given by (3.7).

Hence (3.20) yields that $\dfrac{\|\mathbf{p}^{(i)}\|}{|q^{(i)}|} \to 0$ if and only if $\text{span}\{\mathbf{x}^{(i)}\} \to \text{span}\{\mathbf{x}_1\}$. Further we have

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

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

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

$$\mathbf{r}^{(i)} := (\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbf{x}^{(i)}. \tag{3.21}$$

Lemma 3.5. *The generalised Rayleigh quotient $\rho(\mathbf{x}^{(i)})$ given in (3.14) satisfies*

$$|\rho(\mathbf{x}^{(i)}) - \lambda_1| \leq C_{\|\mathbf{g}_{12}\|}\|\mathbf{T}_{22} - \lambda_1\mathbf{S}_{22}\|\|\mathbf{p}^{(i)}\|, \tag{3.22}$$

and the eigenvalue residual (3.21) satisfies

$$\|\mathbf{r}^{(i)}\| \leq C_{\|\mathbf{g}_{12}\|}\|\mathbf{T}_{22} - \lambda_1\mathbf{S}_{22}\|\|\mathbf{p}^{(i)}\|, \tag{3.23}$$

where $\mathbf{p}^{(i)}$ is given in (3.16) and $C_{\|\mathbf{g}_{12}\|}$ is given in (3.7).

Proof. Since $(\mathbf{A} - \lambda_1\mathbf{M})\mathbf{x}^{(i)} = \alpha^{(i)}(\mathbf{A} - \lambda_1\mathbf{M})\mathbf{X}_2\mathbf{p}^{(i)}$ using (3.16) we have

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

Hence, using (3.10) and the definition of $\alpha^{(i)}$ we get

$$\begin{aligned}
|\rho(\mathbf{x}^{(i)}) - \lambda_1| &= \frac{\|\mathbf{U}^{-1}\mathbf{M}\mathbf{x}^{(i)}\||{\mathbf{x}^{(i)}}^H\mathbf{M}^H\mathbf{U}\bar{\mathbf{I}}_{n-1}(\mathbf{T}_{22} - \lambda_1\mathbf{S}_{22})\mathbf{p}^{(i)}\|}{\|\mathbf{M}\mathbf{x}^{(i)}\|^2} \\
&\leq \|\mathbf{U}^{-1}\|\|\mathbf{U}\|\|(\mathbf{T}_{22} - \lambda_1\mathbf{S}_{22})\mathbf{p}^{(i)}\|. \qquad (3.24)
\end{aligned}$$

Now we have

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

since $\mathbf{Q}$ is unitary. Hence, from equation (3.24), we obtain

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

as required. The eigenvalue residual can be written as

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

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

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

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

Lemma 3.5 shows that the generalised Rayleigh quotient $\rho(\mathbf{x}^{(i)})$ defined by (3.14) converges linearly in $\|\mathbf{p}^{(i)}\|$ to λ_1 and the norm of the eigenvalue residual $\|\mathbf{r}^{(i)}\|$ converges linearly in $\|\mathbf{p}^{(i)}\|$ to zero. This observation leads to more practical measures of convergence than the generalised tangent $T^{(i)}$, which is only of theoretical nature. Nonetheless, one must recognise the limitations of this approach: if $C_{\|\mathbf{g}_{12}\|}$ is large then the error in the generalised Rayleigh quotient and the residual may be large, even if $\|\mathbf{p}^{(i)}\|$ is small. Clearly, $C_{\|\mathbf{g}_{12}\|}$ becomes large if $\|\mathbf{g}_{12}\|$ gets large. Let

$$\|(\mathbf{h}_{12}, \mathbf{g}_{12})\| = \max\{\|\mathbf{h}_{12}\|, \|\mathbf{g}_{12}\|\},$$

and introduce the operator dif (see [133]) given by

$$\text{dif}[(t_{11}, \mathbf{T}_{22}), (s_{11}, \mathbf{S}_{22})] = \inf_{\|(\mathbf{h}_{12}, \mathbf{g}_{12})\|=1} \|\Phi(\mathbf{h}_{12}, \mathbf{g}_{12})\|,$$

then $\text{dif}[(t_{11}, \mathbf{T}_{22}), (s_{11}, \mathbf{S}_{22})] = 0$ if and only if Φ is singular, and the operator dif (similar to sep) is a measure of separation between $\lambda_1 = t_{11}/s_{11}$ and $\lambda(\mathbf{T}_{22}, \mathbf{S}_{22})$. For a discussion on the relation between the functions sep and dif we refer to [133]. Now we have

$$\text{dif}[(t_{11}, \mathbf{T}_{22}), (s_{11}, \mathbf{S}_{22})] \leq \frac{\|(\mathbf{t}_{12}, \mathbf{s}_{12})\|}{\|(\mathbf{h}_{12}, \mathbf{g}_{12})\|},$$

and hence

$$\|\mathbf{g}_{12}\| \leq \frac{\|(\mathbf{t}_{12}, \mathbf{s}_{12})\|}{\text{dif}[(t_{11}, \mathbf{T}_{22}), (s_{11}, \mathbf{S}_{22})]}.$$

From this inequality we see that $\|\mathbf{g}_{12}\|$ (and hence $C_{\|\mathbf{g}_{12}\|}$) can become large if the generalised eigenproblem is highly non-normal or if the eigenvalue λ_1 is not well-separated from the rest of the spectrum of the pair $(\mathbf{A}, \mathbf{M})$.

The Lemma in the following subsection provides a bound on the generalised tangent $T^{(i)}$ after one step of inexact inverse iteration, and is a generalisation of Lemma 3.1 proved in [12] for a diagonalisable problem with symmetric positive definite $\mathbf{M}$.

3.3.2 A one step bound

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

Lemma 3.6. *Let the generalised eigenproblem $\mathbf{Ax} = \lambda\mathbf{Mx}$ have a simple finite eigenpair $(\lambda_1, \mathbf{x}_1)$ and let (3.16) hold for the approximate eigenpair. Assume the shift satisfies $\sigma^{(i)} \notin \lambda(\mathbf{T}_{22}, \mathbf{S}_{22})$. Further let*

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

with $\|\mathbf{d}^{(i)}\| \leq \tau^{(i)}\|\mathbf{Mx}^{(i)}\|$ in Algorithm 4 and

$$\tau^{(i)} < \beta\alpha^{(i)}\frac{|s_{11}q^{(i)}|}{\|\mathbf{u}_1\|\|\mathbf{Mx}^{(i)}\|} \tag{3.25}$$

with $\beta \in (0,1)$ then

$$T^{(i+1)} = \frac{\|\mathbf{S}_{22}\mathbf{p}^{(i+1)}\|}{|s_{11}q^{(i+1)}|} \leq \frac{|\lambda_1 - \sigma^{(i)}|\|\mathbf{S}_{22}\|}{\|(\mathbf{T}_{22} - \sigma^{(i)}\mathbf{S}_{22})^{-1}\|^{-1}}\frac{\left(\|\alpha^{(i)}\mathbf{S}_{22}\mathbf{p}^{(i)}\| + \|\mathbf{d}^{(i)}\|\right)}{(1-\beta)|\alpha^{(i)}s_{11}q^{(i)}|}. \tag{3.26}$$

Proof. Using

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

from Algorithm 4 together with the splitting (3.16) for $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(i+1)}$ we obtain

$$\phi(\mathbf{y}^{(i)})(\mathbf{A}-\sigma^{(i)}\mathbf{M})(\alpha^{(i+1)}\mathbf{x}_1q^{(i+1)}+\alpha^{(i+1)}\mathbf{X}_2\mathbf{p}^{(i+1)}) = \mathbf{M}(\alpha^{(i)}\mathbf{x}_1q^{(i)}+\alpha^{(i)}\mathbf{X}_2\mathbf{p}^{(i)})-\mathbf{d}^{(i)}. \tag{3.27}$$

Using (3.10) we get that

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

where $\bar{\mathbf{I}}_{n-1}$ is defined in (3.18). Thus, multiplying (3.27) by $\mathbf{U}^{-1}$ from the left we obtain

$$\begin{aligned}
\phi(\mathbf{y}^{(i)})\left(\alpha^{(i+1)}(t_{11} - \sigma^{(i)}s_{11})q^{(i+1)}\mathbf{e}_1 + \alpha^{(i+1)}\bar{\mathbf{I}}_{n-1}(\mathbf{T}_{22} - \sigma^{(i)}\mathbf{S}_{22})\mathbf{p}^{(i+1)}\right) \\
= \alpha^{(i)}s_{11}q^{(i)}\mathbf{e}_1 + \alpha^{(i)}\bar{\mathbf{I}}_{n-1}\mathbf{S}_{22}\mathbf{p}^{(i)} - \mathbf{U}^{-1}\mathbf{d}^{(i)}.
\end{aligned} \tag{3.28}$$

Multiplying (3.28) by $\mathbf{e}_1^H$ and $\bar{\mathbf{I}}_{n-1}^H$ from the left we split (3.28) into two equations, namely,

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

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

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

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

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

Using (3.25) we obtain

$$T^{(i+1)} = \frac{\|\mathbf{S}_{22}\mathbf{p}^{(i+1)}\|}{|s_{11}q^{(i+1)}|} \leq \frac{|\lambda_1 - \sigma^{(i)}|\|\mathbf{S}_{22}\|}{\|(\mathbf{T}_{22} - \sigma^{(i)}\mathbf{S}_{22})^{-1}\|^{-1}} \frac{\left(\|\alpha^{(i)}\mathbf{S}_{22}\mathbf{p}^{(i)}\| + \|\mathbf{U}_2^H\mathbf{d}^{(i)}\|\right)}{(1-\beta)|\alpha^{(i)}s_{11}q^{(i)}|}. \tag{3.29}$$

Now $\|\mathbf{U}_2\| = 1$, since, using equation (3.10) we may write

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

and with the special form of $\mathbf{G}$ (see Lemma 3.3) we obtain

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

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

$$T^{(i+1)} = \frac{\|\mathbf{S}_{22}\mathbf{p}^{(i+1)}\|}{|s_{11}q^{(i+1)}|} \leq \frac{|\lambda_1 - \sigma^{(i)}|\|\mathbf{S}_{22}\|}{\|(\mathbf{T}_{22} - \sigma^{(i)}\mathbf{S}_{22})^{-1}\|^{-1}} \frac{\left(\|\alpha^{(i)}\mathbf{S}_{22}\mathbf{p}^{(i)}\| + \|\mathbf{d}^{(i)}\|\right)}{(1-\beta)|\alpha^{(i)}s_{11}q^{(i)}|}, \tag{3.30}$$

as required. □

This bound is a significant improvement over the corresponding results in [50, Lemma 2.2] and [12, Lemma 3.1] which have a bound involving the norm of the unknown eigenvector basis matrix. This matrix may be arbitrarily ill-conditioned, and hence may result in an unnecessarily severe restriction on the solve tolerance in the later theory.
Condition (3.25) asks that $\tau^{(i)}$ is small enough and bounded in terms of $|\alpha^{(i)}s_{11}q^{(i)}|$, which can be considered as a generalised cosine. In practice this means that if the eigenvector approximation $\mathbf{x}^{(i)}$ is coarse, $|s_{11}q^{(i)}|$ is close to zero and hence $\tau^{(i)}$ has to be chosen small enough.

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

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

As in [137], we introduce the function $\text{sep}(\lambda_1, (\mathbf{T}_{22}, \mathbf{S}_{22}))$, which measures the separation of the sought simple eigenvalue λ_1 from the eigenvalues $\lambda(\mathbf{T}_{22}, \mathbf{S}_{22})$ as follows

$$\begin{aligned} \text{sep}(\lambda_1, (\mathbf{T}_{22}, \mathbf{S}_{22})) &:= \inf_{\|\mathbf{a}\|=1} \|(\mathbf{T}_{22} - \lambda_1\mathbf{S}_{22})\mathbf{a}\|_2 \qquad (3.31) \\ &= \begin{cases} \|(\mathbf{T}_{22} - \lambda_1\mathbf{S}_{22})^{-1}\|^{-1}, & \lambda_1 \notin \lambda(\mathbf{T}_{22}, \mathbf{S}_{22}) \\ 0, & \lambda_1 \in \lambda(\mathbf{T}_{22}, \mathbf{S}_{22}) \end{cases}. \end{aligned}$$

Using this definition we get

$$\begin{aligned}\operatorname{sep}(\sigma^{(i)},(\mathbf{T}_{22},\mathbf{S}_{22})) &= \inf_{\|\mathbf{a}\|_2=1}\|(\mathbf{T}_{22}-\sigma^{(i)}\mathbf{S}_{22})\mathbf{a}\|_2 \\ &\geq \operatorname{sep}(\lambda_1,(\mathbf{T}_{22},\mathbf{S}_{22}))-|\lambda_1-\sigma^{(i)}|\|\mathbf{S}_{22}\|_2,\end{aligned}$$

and also

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

for the case of exact solves. Since $\operatorname{sep}(\sigma^{(i)},(\mathbf{T}_{22},\mathbf{S}_{22}))$ is a measure for the separation of the shift $\sigma^{(i)}$ from the rest of the spectrum, this means that the convergence rate depends on the ratio $\dfrac{|\lambda_1-\sigma^{(i)}|\|\mathbf{S}_{22}\|}{\operatorname{sep}(\sigma^{(i)},(\mathbf{T}_{22},\mathbf{S}_{22}))}$. For diagonalisable systems, where $\mathbf{T}_{22}$ is diagonal and $\mathbf{S}_{22}=\mathbf{I}_{n-1}$, this ratio becomes $\dfrac{|\lambda_1-\sigma^{(i)}|}{|\lambda_2-\sigma^{(i)}|}$, the familiar ratio obtained for inverse iteration. In the next subsection we give the convergence rate for inexact inverse iteration for certain choices of the shift and the solve tolerance, using Lemma 3.6.

3.3.3 Convergence rate for inexact inverse iteration

Assume that the shift $\sigma^{(i)}$ in Algorithm 4 satisfies

$$|\lambda_1-\sigma^{(i)}| < \frac{\operatorname{sep}(\lambda_1,(\mathbf{T}_{22},\mathbf{S}_{22}))}{2\|\mathbf{S}_{22}\|}, \tag{3.32}$$

that is $\sigma^{(i)}$ is close to λ_1 and certainly closer to λ_1 than to any other eigenvalue. Then, using (3.32), for the first factor on the right hand side of (3.29)

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

holds. Note that for diagonalisable systems with $\mathbf{S}_{22}=\mathbf{I}_{n-1}$ condition (3.32) becomes $|\lambda_1-\sigma^{(i)}| < \frac{1}{2}|\lambda_2-\lambda_1|$, where $|\lambda_2-\lambda_1| = \min_{j\neq 1}|\lambda_j-\lambda_1|$ and hence $|\lambda_1-\sigma^{(i)}| < |\lambda_2-\sigma^{(i)}|$, a familiar condition for the choice of the shift.

Using Lemma 3.6 we can prove convergence results for variable and fixed shifts $\sigma^{(i)}$ and for different choices of the tolerances $\tau^{(i)}$.

Theorem 3.7 (Convergence of Algorithm 4). *Let (3.1) be a generalised eigenproblem and consider the application of Algorithm 4 to find a simple eigenvalue λ_1 with corresponding right eigenvector $\mathbf{x}_1$. Let the assumptions of Lemma 3.6 hold and let $0<\|\mathbf{S}_{22}\mathbf{p}^{(0)}\|<1$, that is $\mathbf{x}^{(0)}$ is neither the solution itself nor deficient in the sought eigendirection.*

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

$$|\lambda_1-\sigma^{(i)}| < \frac{sep(\lambda_1,(\mathbf{T}_{22},\mathbf{S}_{22}))}{2\|\mathbf{S}_{22}\|}\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|. \tag{3.33}$$

and $\|\mathbf{d}^{(i)}\| \leq \tau^{(i)}\|\mathbf{M}\mathbf{x}^{(i)}\|$ *where* $\tau^{(i)} < \dfrac{\alpha^{(i)}}{\|\mathbf{M}\mathbf{x}^{(i)}\|\|\mathbf{u}_1\|}\beta|s_{11}q^{(i)}|$ *with* $0 \leq 2\beta < 1 - T^{(0)}$, *then Algorithm 4 converges linearly, that is*

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

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

2. *If* $\tau^{(i)} < \alpha^{(i)}\gamma\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|/\|\mathbf{M}\mathbf{x}^{(i)}\|$ *for some positive constant* γ *and furthermore (3.33) is replaced by*

$$|\lambda_1 - \sigma^{(i)}| < \frac{1-\beta}{2-\beta+\gamma+\delta}\frac{sep(\lambda_1,(\mathbf{T}_{22},\mathbf{S}_{22}))}{\|\mathbf{S}_{22}\|}, \tag{3.34}$$

where $\delta > 0$, *then Algorithm 4 converges linearly, that is*

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

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

Proof. 1. If (3.33) holds then

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

since $\|\mathbf{S}_{22}\mathbf{p}^{(i)}\| < 1$. Thus, from (3.30)

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

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

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

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

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

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

2. If (3.34) holds then

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

Further, if $\tau^{(i)} < \alpha^{(i)}\gamma\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|/\|\mathbf{M}\mathbf{x}^{(i)}\|$ in (3.30) then (with the results from the first part of the proof)

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

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

Thus we have proved Theorem 3.7. □

Note that if β is chosen close to zero, that is, more accurate solves are used for the inner iteration (see (3.25)), then according to Theorem 3.7, which requires $\beta < (1 - T^{(0)})/2$, $T^{(0)}$ is allowed to be close to one, and hence the initial eigenvector approximation is allowed to be coarse. In contrast, for a larger value of β, which allows the solve tolerance $\tau^{(i)}$ to be larger, we require that $T^{(0)}$ is very small and hence the initial eigenvector approximation $\mathbf{x}^{(0)}$ has to be very close to the sought eigenvector. Also, note that $\|\mathbf{u}_1\| = (1 + \|\mathbf{g}_{12}\|)$, so that if $\|\mathbf{g}_{12}\|$ is large then $\|\mathbf{u}_1\|$ is large and the solve tolerance satisfying (3.25) may be small. Note also that condition (3.25) is the same condition as $\tau^{(i)} < \beta|\mathbf{u}_1^H\mathbf{M}\mathbf{x}^{(i)}|/\ \|\mathbf{u}_1\|$ as in Lemma 3.1 of [12].

Remark 3.8. *The assumption $0 < \|\mathbf{S}_{22}\mathbf{p}^{(0)}\| < 1$ in Theorem 3.7 requires that the initial guess $\mathbf{x}^{(0)}$ is neither the solution itself nor deficient in the sought eigendirection. Condition (3.34) states that the shift $\sigma^{(i)}$ is closer to λ_1 than to any other eigenvalue, clearly this condition can be satisfied by a close enough fixed shift $\sigma^{(i)} = \sigma$, $\forall i$. Condition (3.33) is stronger than requirement (3.34) since it states not only that the shift $\sigma^{(i)}$ has to be close enough to λ_1 but also that it should converge to λ_1 as i increases (since $\|\mathbf{S}_{22}\mathbf{p}^{(i)}\| \to 0$). This condition can be satisfied by using, for example, a Rayleigh quotient shift given by (3.14).*

The condition on the solve tolerance $\tau^{(i)} < \alpha^{(i)}\beta|s_{11}q^{(i)}|/\|\mathbf{M}\mathbf{x}^{(i)}\|\|\mathbf{u}_1\|$ requires that the solve tolerance $\tau^{(i)}$ is small enough. This condition is satisfied if a small enough constant for $\tau^{(i)} = \tau$ is chosen, since $|s_{11}q^{(i)}|$ increases during the iteration. The stronger condition $\tau^{(i)} < \alpha^{(i)}\gamma\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|/\|\mathbf{M}\mathbf{x}^{(i)}\|$ is met for a decreasing tolerance $\tau^{(i)}$ where $\tau^{(i)} \to 0$ as i increases, since $\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|$ decreases during the iteration.

We note that the conditions on $\tau^{(i)}$ and $\sigma^{(i)}$ are only sufficient but not necessary requirements and convergence might be obtained by much larger values of $\tau^{(i)}$.

Remark 3.9. *One way of choosing $\tau^{(i)} < \alpha^{(i)}\gamma\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|/\|\mathbf{M}\mathbf{x}^{(i)}\|$ is to use*

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

where $\mathbf{r}^{(i)}$ is the eigenvalue residual which is given by (3.21) and satisfies $\|\mathbf{r}^{(i)}\| := \mathcal{O}(\|\mathbf{p}^{(i)}\|)$ and C is a small enough constant.

As mentioned in Remark 3.8 the condition $\tau^{(i)} < \alpha^{(i)}\gamma\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|/\|\mathbf{M}\mathbf{x}^{(i)}\|$ is probably too restrictive and also contains quantities which are unknown (for example $\mathbf{S}_{22}$). It is possible to bound the quantity $\alpha^{(i)}\gamma\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|/\|\mathbf{M}\mathbf{x}^{(i)}\|$ from below and above:

$$\gamma\frac{\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|}{\|\mathbf{U}\|} \le \alpha^{(i)}\gamma\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|/\|\mathbf{M}\mathbf{x}^{(i)}\| \le \gamma\|\mathbf{U}^{-1}\|\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|.$$

However, the choice of the lower bound of $\alpha^{(i)}\gamma\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|/\|\mathbf{M}\mathbf{x}^{(i)}\|$ for $\tau^{(i)}$ is likely to be too restrictive (since $\|\mathbf{U}\|$ could be large for non-normal matrices) and the upper bound is too coarse. Both upper and lower bounds contain unknown quantities $\mathbf{S}_{22}$, $\|\mathbf{U}\|$ and $\|\mathbf{U}^{-1}\|$. We note that the conditions on $\tau^{(i)}$ are only qualitative statements since in our experiments considerably larger values of $\tau^{(i)}$ have been used successfully.

Remark 3.10. *We point out two shift strategies;*

- *Fixed shift: With a decreasing tolerance $\tau^{(i)} = C_1\|\mathbf{r}^{(i)}\|$ for small enough $\tau^{(0)}$ and C_1 the second case in Theorem 3.7 arises. If the shift satisfies (3.34), that is the shift is close enough to the sought eigenvalue then Algorithm 4 exhibits linear convergence.*

- *Rayleigh quotient shift: A generalised Rayleigh quotient shift $\sigma^{(i)} = \rho(\mathbf{x}^{(i)})$ chosen as in (3.14) satisfies (see (3.22)) $|\sigma^{(i)} - \lambda_1| = C_2\|\mathbf{p}^{(i)}\|$ for some constant C_2. Hence, for small enough C_2 it will also satisfy (3.33). Therefore, with a decreasing tolerance $\tau^{(i)} = C_1\|\mathbf{r}^{(i)}\|$ quadratic convergence is achieved for small enough $\tau^{(0)}$.*

Finally we would like to discuss the application of Theorem 3.7 to the case of $\mathbf{M}$ is positive definite and $\mathbf{M}^{-1}\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ is diagonalisable, see [12]. In this case $\mathbf{S}$ is the identity matrix, and $\mathbf{T}$ can be represented by a diagonal matrix. Condition (3.33) then becomes

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

which is the same condition as used in [12].

3.4 A relation between the normalisation function and the eigenvalue residual

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

First we give an extension of Lemma 3.5 which provides a lower bound on the eigenvalue residual in terms of $\mathbf{p}^{(i)}$.

Lemma 3.11. *Let the assumptions of Lemma 3.5 be satisfied and let $\|\mathbf{M}\mathbf{x}^{(i)}\| = 1$. Then the following bound holds*

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

Proof. With $\|\mathbf{U}_2^H\| = 1$ (see remarks after Lemma 3.6) and $\mathbf{U}_2^H = \bar{\mathbf{I}}_{n-1}^H\mathbf{G}^{-1}\mathbf{Q}^H$ we have

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

With $\mathbf{H}^{-1}\mathbf{Z}^H = \mathbf{X}^{-1}$ and using (3.16) as well as the special structure of $\bar{\mathbf{I}}_{n-1}^H$ we then obtain

$$\begin{aligned}
\|\mathbf{r}^{(i)}\| &\geq \|\alpha^{(i)}\begin{bmatrix}\mathbf{0}^H \\ \mathbf{I}_{n-1}\end{bmatrix}^H\begin{bmatrix} t_{11}-\rho(\mathbf{x}^{(i)})s_{11} & \mathbf{0}^H \\ \mathbf{0} & \mathbf{T}_{22}-\rho(\mathbf{x}^{(i)})\mathbf{S}_{22}\end{bmatrix}\mathbf{X}^{-1}(\mathbf{x}_1q^{(i)}+\mathbf{X}_2\mathbf{p}^{(i)})\| \\
&= \alpha^{(i)}\|\begin{bmatrix}\mathbf{0}^H \\ \mathbf{I}_{n-1}\end{bmatrix}^H\begin{bmatrix} t_{11}-\rho(\mathbf{x}^{(i)})s_{11} & \mathbf{0}^H \\ \mathbf{0} & \mathbf{T}_{22}-\rho(\mathbf{x}^{(i)})\mathbf{S}_{22}\end{bmatrix}(q^{(i)}\mathbf{e}_1+\bar{\mathbf{I}}_{n-1}\mathbf{p}^{(i)})\| \\
&= \alpha^{(i)}\left\|\mathbf{I}_{n-1}(\mathbf{T}_{22}-\rho(\mathbf{x}^{(i)})\mathbf{S}_{22})\mathbf{p}^{(i)}\right\|
\end{aligned}$$

The definition of the separation (3.31) yields

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

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

Lemma 3.11 and Lemma 3.5 show that the eigenvalue residual is equivalent to $\|\mathbf{p}^{(i)}\|$ as a measure of convergence, provided λ_1 is a well-separated eigenvalue, though, of course, in practice, if $\|\mathbf{G}\|$ is large then a small residual does not necessarily imply a small error. The following Proposition gives upper and lower bounds on $\dfrac{1}{\phi(\mathbf{y}^{(i)})}$ in terms of $\|\mathbf{r}^{(i)}\|$.

Proposition 3.12. *Let $(\lambda^{(i)},\mathbf{x}^{(i)})$ be the current approximation to $(\lambda_1,\mathbf{x}_1)$. Let $\|\mathbf{M}\mathbf{x}^{(i)}\| = 1$ so that $\phi(\mathbf{y}^{(i)}) := \|\mathbf{M}\mathbf{y}^{(i)}\|$. Assume that $\mathbf{y}^{(i)}$ is such that*

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

Then

$$\|\mathbf{r}^{(i+1)}\| \leq \frac{1+\tau^{(i)}}{\phi(\mathbf{y}^{(i)})} \tag{3.35}$$

and

$$\frac{1-\tau^{(i)}}{\phi(\mathbf{y}^{(i)})} \leq \|\mathbf{r}^{(i+1)}\| + |\rho(\mathbf{x}^{(i+1)})-\sigma^{(i)}|, \tag{3.36}$$

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

Proof. We have
$$(\mathbf{A}-\sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{M}\mathbf{x}^{(i)} - \mathbf{d}^{(i)}$$
and, since $\mathbf{x}^{(i+1)} = \dfrac{\mathbf{y}^{(i)}}{\phi(\mathbf{y}^{(i)})}$,
$$\mathbf{A}\mathbf{x}^{(i+1)} - \sigma^{(i)}\mathbf{M}\mathbf{x}^{(i+1)} = \frac{1}{\phi(\mathbf{y}^{(i)})}((\mathbf{A}-\sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)}).$$
Hence
$$\frac{\|\mathbf{A}\mathbf{x}^{(i+1)} - \sigma^{(i)}\mathbf{M}\mathbf{x}^{(i+1)}\|}{\|\mathbf{M}\mathbf{x}^{(i)} - \mathbf{d}^{(i)}\|} = \frac{1}{\phi(\mathbf{y}^{(i)})}. \tag{3.37}$$
Finally, $\|\mathbf{M}\mathbf{x}^{(i)} - \mathbf{d}^{(i)}\| \le 1+\tau^{(i)}$ together with the minimising property of $\rho(\mathbf{x}^{(i+1)})$ (see (3.15)) yields the first bound (3.35). In order to obtain the second bound, equality (3.37) gives
$$\begin{aligned}\frac{1}{\phi(\mathbf{y}^{(i)})} \le & \frac{\|\mathbf{A}\mathbf{x}^{(i+1)} - \rho(\mathbf{x}^{(i+1)})\mathbf{M}\mathbf{x}^{(i+1)}\| + |\rho(\mathbf{x}^{(i+1)}) - \sigma^{(i)}|\|\mathbf{M}\mathbf{x}^{(i+1)}\|}{\|\mathbf{M}\mathbf{x}^{(i)}\| - \|\mathbf{d}^{(i)}\|} \\ & \le \frac{\|\mathbf{r}^{(i+1)}\| + |\rho(\mathbf{x}^{(i+1)}) - \sigma^{(i)})|}{1-\tau^{(i)}},\end{aligned} \tag{3.38}$$
which yields (3.36). ☐

Proposition 3.12 provides the following result. If we chose the shift to be $\sigma^{(i)} := \rho(\mathbf{x}^{(i)})$ then
$$\frac{1-\tau^{(i)}}{\phi(\mathbf{y}^{(i)})} - |\rho(\mathbf{x}^{(i+1)}) - \rho(\mathbf{x}^{(i)})| \le \|\mathbf{r}^{(i+1)}\| \le \frac{1+\tau^{(i)}}{\phi(\mathbf{y}^{(i)})}.$$
From Section 3.3, convergence of inexact inverse iteration yields $\|\mathbf{p}^{(i)}\| \to 0$. By Lemmas 3.5 and 3.11 convergence of inexact inverse iteration implies $\|\mathbf{r}^{(i)}\| \to 0$ as well as $|\rho(\mathbf{x}^{(i)}) - \lambda_1| \to 0$. The last property also yields $|\rho(\mathbf{x}^{(i+1)}) - \rho(\mathbf{x}^{(i)})| \to 0$, if inexact inverse iteration converges. Therefore Proposition 3.12 shows that inexact inverse iteration converges if and only if $\phi(\mathbf{y}^{(i)}) \to \infty$ as $i \to \infty$.

We end this section with an application of inexact inverse iteration to block structured systems of the form $\mathbf{A}\mathbf{x} = \lambda\mathbf{M}\mathbf{x}$, where
$$\mathbf{A} = \begin{bmatrix} \mathbf{K} & \mathbf{C} \\ \mathbf{C}^H & \mathbf{0}\end{bmatrix} \quad \text{and} \quad \mathbf{M} = \begin{bmatrix} \mathbf{M}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix},$$
and $\mathbf{M}_1$ is symmetric positive definite. Matrices with this block structure arise after a mixed finite element discretisation of the linearised incompressible Navier-Stokes equations, see for example [17, 18, 33, 85]. If the desired eigenvector is written in terms of the velocity and pressure components $\mathbf{x} = [\mathbf{x}_u\, \mathbf{x}_p]^H$, the incompressibility condition $\mathbf{C}^H\mathbf{x}_u = 0$ holds. If the system $(\mathbf{A}-\sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{M}\mathbf{x}^{(i)}$ is solved inexactly, we cannot guarantee that $\mathbf{C}^H\mathbf{x}_u^{(i)} = 0$, even if the starting guess satisfies $\mathbf{C}^H\mathbf{x}_u^{(0)} = 0$: we only know that $\|\mathbf{C}^H\mathbf{x}_u^{(i)}\| \le \tau^{(i)}$. Simoncini [117] considered the application of the inexact Shift-and-Invert Lanczos method to a generalised symmetric eigenproblem where a constraint is given in terms of null space orthogonality of the sought eigenvector. She showed that in exact arithmetic the constraint is maintained for exact solves. However,

for inexact solves depending on the iterative method used and on the preconditioning strategy applied for the inner solver, the approximate solution may not satisfy the constraint. She gave special preconditioning strategies so that the solution of the inner system satisfies the constraint in exact arithmetic. However, in finite precision arithmetic the constraint $\mathbf{C}^H\mathbf{x}_u = 0$ will be violated. Hence either a projection enforcing the orthogonality constraint (even if not in every single step of Lanczos method) or a so-called purification step introduced by Nour-Omid et al. [96] (see also [79, 85]) is necessary.

We show that inexact inverse iteration exhibits a different behaviour. Due to the structure of the algorithm, a projection after each inexact solve is not necessary, since, in the limit, as $i \to \infty$, the constraint $\|\mathbf{C}^H\mathbf{x}_u^{(i)}\| = 0$ is satisfied. The following Corollary shows that inexact inverse iteration need not enforce the incompressibility condition at each outer iteration. We have $\|\mathbf{C}^H\mathbf{x}_u^{(i)}\| \leq \tau^{(i)}$ at each outer iteration, however, as $i \to \infty$ we have $\|\mathbf{C}^H\mathbf{x}_u^{(i)}\| \to 0$ and the incompressibility condition holds in the limit.

Corollary 3.13. *Let the assumptions of Proposition 3.12 be satisfied and consider inexact inverse iteration applied to the block structured system*

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

Then

$$\|\mathbf{C}^H\mathbf{x}_u^{(i)}\| \to 0 \quad \textit{as} \quad i \to \infty,$$

that is, in the limit, as $i \to \infty$, *the constraint* $\|\mathbf{C}^H\mathbf{x}_u^{(i)}\| = 0$ *is satisfied.*

Proof. From Algorithm 4 and Proposition 3.12 we have

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

and hence $\|\mathbf{C}^H\mathbf{x}_u^{(i)}\| \to 0$ as $i \to \infty$ that is, as the outer iteration proceeds. □

3.5 Two numerical examples

Finally, we give two test problems for our theory. We chose problems $\mathbf{A}\mathbf{x} = \lambda\mathbf{M}\mathbf{x}$ which are not necessarily diagonalisable and with singular $\mathbf{M}$, since problems with positive definite $\mathbf{M}$ (including the standard problem $\mathbf{M} = \mathbf{I}$) have been extensively investigated by other authors (see, for example [10], [12]). Smit and Paardekooper [129] contains examples for the standard symmetric eigenproblem and Golub and Ye [50] discuss the standard diagonalisable problem $\mathbf{M}^{-1}\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$. A nuclear reactor problem similar to the one in the following example with $\mathbf{M}$ singular was considered in [75]. However, in [75] the problem was first transformed into a standard eigenproblem.

Example 3.14 (Nuclear Reactor Problem)**.** *We use the same example as in (2.18), with the same setup, regions and data as in Figure 2-5 and Table 2.4. For initial conditions, we take* $\mathbf{x}^{(0)} = [1, \ldots, 1]^H/\sqrt{n}$. *In fact, the exact eigenvalue is given by* $\lambda_1 = 0.9707$ *and* $\cos(\mathbf{x}_1, \mathbf{x}^{(0)}) \approx 0.44$.

We used a fixed shift and a variable shift strategy. The vector $\mathbf{x}^{(i)}$ *is normalised such that* $\|\mathbf{M}\mathbf{x}^{(i)}\| = 1$*, that is* $\phi(\mathbf{y}^{(i)}) = \sqrt{\mathbf{y}^{(i)H}\mathbf{M}^H\mathbf{M}\mathbf{y}^{(i)}}$ *in Algorithm 4. For the inner solver we use right-preconditioned GMRES with an incomplete LU-factorisation as preconditioner. We perform three different numerical experiments.*

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

$$\tau^{(i)} = \min\{0.1, \|\mathbf{r}^{(i)}\|\}, \tag{3.39}$$

where $\mathbf{r}^{(i)}$ *is defined by (3.21). The iteration stops once the eigenvalue residual satisfies* $\|\mathbf{r}^{(i)}\| < 10^{-9}$.

(b) *Inexact inverse iteration using a variable shift given by* $\rho(\mathbf{x}^{(i)})$ *from (3.14) and a decreasing solve tolerance* $\tau^{(i)}$ *for the inner solver which satisfies (3.39). The iteration stops once the eigenvalue residual satisfies* $\|\mathbf{r}^{(i)}\| < 10^{-14}$.

(c) *Inexact inverse iteration using a variable shift given by* $\rho(\mathbf{x}^{(i)})$ *from (3.14) with a fixed solve tolerance, which we chose to be* $\tau^{(i)} = \tau^{(0)} = 0.4$*. This iteration also stops once the eigenvalue residual satisfies* $\|\mathbf{r}^{(i)}\| < 10^{-9}$.

Figure 3-1 illustrates the convergence history of the eigenvalue residuals for the three different experiments described in (a), (b) and (c) above. The choice of (3.39) to provide

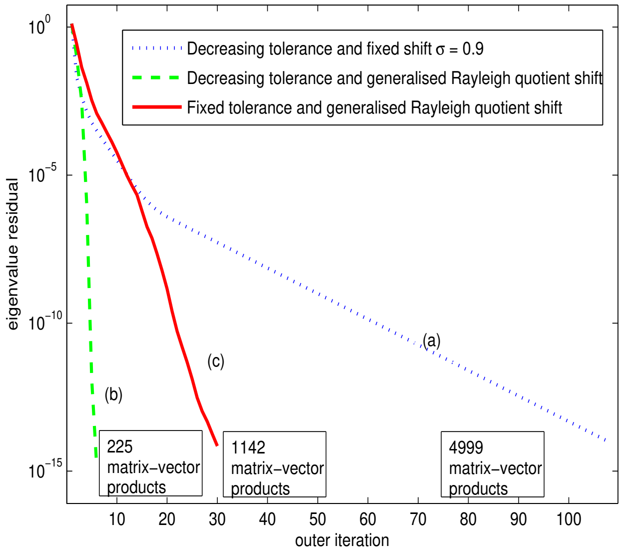


Figure 3-1: *Convergence history of the eigenvalue residuals for Example 3.14 using* fixed shift $\sigma = 0.9$ *and* variable shift *and fixed or decreasing tolerances (see tests (a), (b) and (c).*

a solve tolerance $\tau^{(i)}$ is consistent with the discussion in Remark 3.9 and the assumption in Theorem 3.7. We have used this decreasing tolerance throughout our computations. As proved in Theorem 3.7, case (2), inexact inverse iteration with a decreasing solve tolerance and with a fixed shift, chosen to be close enough to the desired eigenvalue, exhibits linear convergence, as show in Figure 3-1, case (a) (see also the discussion on the fixed shift in Remark 3.10). If we use a generalised Rayleigh quotient as a shift (where the Rayleigh quotient is close enough to the sought eigenvalue) and a fixed solve tolerance $\tau^{(0)}$ the Algorithm 4 converges linearly (case (c)), whereas for a decreasing tolerance quadratic convergence is readily observed (case (b)). This covers case (1) in Theorem 3.7, we also refer to the discussion on the Rayleigh quotient shift in Remark 3.10.

We would like to note that all three methods have the same initial eigenvalue residual. Both methods (a) and (c) exhibit linear convergence, but the method with a variable shift and fixed solve tolerance performs better than the fixed shift method with a decreasing solve tolerance. This improvement in the behaviour of method (c) over (a) may be explained by close examination of the asymptotic constants in the expressions for linear convergence in Theorem 3.7. For a good starting guess (that is a $T^{(0)}$ close to zero) and a small enough β with $\beta < (1 - T^{(0)})/2$ the constant of linear convergence for method (c) may be much smaller than one, and hence smaller than the convergence rate for method (a). In our particular computations the constants for linear convergence are about 0.82 for method (a) and about 0.32 for method (c).

The total amount of work is measured by the number of matrix-vector multiplications given in Figure 3-1. We can observe that method (b), inexact Rayleigh quotient iteration with a decreasing solve tolerance, achieves the fastest convergence rate with smallest amount of work.

Example 3.15 (The linearised steady Navier-Stokes equations). *For the stability analysis of the steady state solutions of the Navier-Stokes equations generalised eigenproblems of the form* $\mathbf{A}\mathbf{x} = \lambda\mathbf{M}\mathbf{x}$ *arise, where* $\mathbf{A}$ *and* $\mathbf{M}$ *have a special block structure, that is*

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

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

$$\lambda_1 = 0.21 + 0.16i, \quad \lambda_2 = 0.21 - 0.16i,$$

and we aim to find the complex eigenvalue λ_1 *nearest to* $0.21 + 0.16i$. *We normalise* $\mathbf{x}^{(i)}$ *such that* $\|\mathbf{M}\mathbf{x}^{(i)}\| = 1$, *that is,* $\phi(\mathbf{y}^{(i)}) = \sqrt{\mathbf{y}^{(i)H}\mathbf{M}^H\mathbf{M}\mathbf{y}^{(i)}}$ *as in the first example. The convergence performance of the three methods considered in the previous example is repeated in this example and we do not reproduce the results here. Rather, we look at the incompressibility condition* $\mathbf{C}^H\mathbf{x}_u^{(i)} = \mathbf{0}$ *and examine how it behaves under inexact inverse iteration. In particular we ask if there is any advantage to be gained by imposing the incompressibility condition after each inexact solve. To this end we carry out inexact inverse iteration using a variable shift given by* $\rho(\mathbf{x}^{(i)})$ *from (3.14) and a close enough starting guess. We use a fixed solve tolerance* $\tau^{(i)} = \tau^{(0)} = 0.1$. *The iteration stops once the eigenvalue residual satisfies* $\|\mathbf{r}^{(i)}\| < 10^{-7}$. *To impose the incompressibility* *condition after an inner iteration we replace* $\mathbf{x}_u^{(i)}$ *by* $\pi\mathbf{x}_u^{(i)}$ *where the projection* π *from* $\mathbb{C}^{1406}$ *onto* $\mathbf{C}^\perp$ *along range*$(\mathbf{C})$ *is defined by*

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

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

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

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

Table 3.1: *Incompressibility condition* $\|\mathbf{C}^H\mathbf{x}_u^{(i)}\|$ *in the course of inexact inverse iteration without the application of* π.

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

Tables 3.1 and 3.2 show the eigenvalue residual $\|\mathbf{r}^{(i)}\|$, $\|\mathbf{C}^H\mathbf{x}_u^{(i)}\|$ and $\|\mathbf{C}^H\mathbf{y}_u^{(i)}\|$ at each outer iteration i. The second column of Table 3.2 shows $\|\mathbf{C}^H\mathbf{x}_u^{(i)}\|$ before projection is applied for the beginning of the next outer iteration step. We observe that there is almost no difference between performing inexact inverse iteration with or without projection at the beginning of each outer step. We also see $\|\mathbf{C}^H\mathbf{x}_u^{(i)}\| \to 0$ as i increases, as predicted by Corollary 3.13, and hence, the application of the projection π at every step is not necessary. Also note that in both tables $\|\mathbf{C}^H\mathbf{y}_u^{(i)}\| \le \tau^{(0)} = 0.1$.

3.6 A convergence theory for inexact simple Jacobi-Davidson method

In this section we show how the convergence theory obtained in Section 3.3 may be applied to a simplified version of the inexact Jacobi-Davidson method. The Jacobi-Davidson method was introduced by Sleijpen and van der Vorst (see [124] and [126]) for the linear eigenproblem and it has been applied to the generalised eigenproblem

Table 3.2: *Incompressibility condition* $\|\mathbf{C}^H\mathbf{x}_u^{(i)}\|$ *in the course of inexact inverse iteration with the application of* π.

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

and matrix pencils (see [39] and [123]). A survey has been given in [63] (see also [4]). A convergence theory for Jacobi-Davidson applied to the Hermitian eigenproblem has been given in [147] and for a special inner solver, namely the conjugate gradient method, in [93]. The relationship between a simplified version of Jacobi-Davidson method and Newton's method for exact solves has been established in several papers, see for example [124], [126], [125] and [94]. Here we provide a convergence theory for a version of an inexact simplified Jacobi-Davidson method for the generalised eigenvalue problem (3.1), and also present some numerical results to illustrate our theory.

We give a version for Jacobi-Davidson method for our problem (3.1) and present an equivalence between inexact Rayleigh quotient iteration and the inexact simplified Jacobi-Davidson method.

3.6.1 A simplified Jacobi-Davidson method and equivalence to Rayleigh quotient iteration

First, we briefly describe one possible version of a simplified Jacobi-Davidson algorithm for the generalised eigenvalue problem (3.1) (see [93, Algorithm 2.1] and [147, Algorithm 3.1] for similar algorithms for standard Hermitian eigenproblems).

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

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

With $\mathbf{r}^{(i)}$ defined by (3.21) solve the correction equation

$$\mathbf{P}^{(i)}(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbf{Q}^{(i)}\mathbf{s}^{(i)} = -\mathbf{r}^{(i)}, \quad \text{where} \quad \mathbf{s}^{(i)} \perp \mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}, \tag{3.40}$$

for $\mathbf{s}^{(i)}$. This is the Jacobi-Davidson correction equation which maps $\text{span}\{\mathbf{M}^H\mathbf{M}\mathbf{x}\}^\perp$ onto $\text{span}\{\mathbf{M}\mathbf{x}\}^\perp$. An improved guess for the eigenvector is given by a suitably normalised $\mathbf{x}^{(i)} + \mathbf{s}^{(i)}$. For other choices of projections and discussions on the correction equation (3.40) we refer to [123]. The motivation behind the Jacobi-Davidson algorithm is that for large systems which are solved iteratively, the form of the correction equation (3.40) is more amenable to efficient solution than the corresponding system for inverse iteration. Also, in practice, a subspace version of Jacobi-Davidson is used with each new direction being added to increase the dimension of a search space, but we do not consider this version here. Algorithm 5 provides a precise description of the method we discuss in this chapter. The function ϕ is a normalisation, which for both

Algorithm 5 Simplified Jacobi-Davidson (Jacobi-Davidson without subspace acceleration)

Input: $\mathbf{x}^{(0)}$, i_{max}.
  for $i = 1, \ldots, i_{max}$ **do**
    Choose $\tau^{(i)}$,
    $\mathbf{r}^{(i)} = (\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbf{x}^{(i)}$,
    Find $\mathbf{s}^{(i)}$ such that

$$\|\mathbf{P}^{(i)}(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbf{Q}^{(i)}\mathbf{s}^{(i)} + \mathbf{r}^{(i)}\| \leq \tau^{(i)}\|\mathbf{r}^{(i)}\| \quad \text{for} \quad \mathbf{s}^{(i)} \perp \mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)},$$

    Set $\mathbf{x}^{(i+1)} = (\mathbf{x}^{(i)} + \mathbf{s}^{(i)})/\phi(\mathbf{x}^{(i)} + \mathbf{s}^{(i)})$,
    Test for convergence.
  end for
Output: $\mathbf{x}^{(i_{max})}$.

practical computations and theoretical comparisons between Rayleigh quotient iteration and Jacobi-Davidson, is taken to be the same as in Algorithm 4. The procedure of the simplified Jacobi-Davidson method may be seen as a worst-case scenario for the complete Jacobi-Davidson procedure with subspace acceleration, the Jacobi-Davidson method is expected to converge faster than simplified Jacobi-Davidson.

In this section we shall provide a convergence theory for the inexact simplified Jacobi-Davidson method given in Algorithm 5. To do this we shall first show the close connection of inexact simplified Jacobi-Davidson with inexact Rayleigh-quotient iteration and apply the convergence theory in Section 3.3. Though simplified Jacobi-Davidson is not used in practice its convergence may be considered as a worst-case scenario for the more usual subspace Jacobi-Davidson procedure, and the convergence results here can be similarly interpreted.

First, we point out the following well-known equivalence between the simplified Jacobi-Davidson method and Rayleigh quotient iteration for *exact* system solves, which has been proved in [126], [93], [95] and in [123] for the generalised eigenproblem.

Lemma 3.16. *Suppose the correction equation in Algorithm 5 has a unique solution* $\mathbf{s}^{(i)}$. *Then the Jacobi-Davidson solution* $\mathbf{x}_{JD}^{(i+1)} = \mathbf{x}^{(i)} + \mathbf{s}^{(i)}$ *satisfies*

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

where

$$\mathbf{z}^{(i+1)} = \frac{1}{\gamma^{(i)}}\mathbf{x}_{JD}^{(i+1)} \quad \textit{with} \quad \gamma^{(i)} = \frac{{\mathbf{x}^{(i)}}^H\mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}}{{\mathbf{x}^{(i)}}^H\mathbf{M}^H\mathbf{M}(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})^{-1}\mathbf{M}\mathbf{x}^{(i)}}, \tag{3.41}$$

From Lemma 3.16 it is clear that for exact solves one step of simplified Jacobi-Davidson produces an improved approximation to the desired eigenvector that has the same direction as that given by one step of Rayleigh quotient iteration. Hence, as observed in [126], if the correction equation is solved exactly, the method converges as fast as Rayleigh quotient iteration (that is quadratically for nonsymmetric systems). If subspace expansion is used even faster convergence is expected. The next section shows how we can find a similar equivalence between inexact Rayleigh quotient iteration and the inexact Jacobi-Davidson method.

3.6.2 Transforming inexact Jacobi-Davidson into inexact Rayleigh quotient iterations

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

$$(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbf{y}^{(i)} = \mathbf{M}\mathbf{x}^{(i)} - \mathbf{d}_I^{(i)}, \quad \text{where} \quad \|\mathbf{d}_I^{(i)}\| \le \tau_I^{(i)}\|\mathbf{M}\mathbf{x}^{(i)}\|, \quad \text{with} \quad \tau_I^{(i)} < 1, \tag{3.42}$$

with one step of inexact Jacobi-Davidson method, that is,

$$\mathbf{P}^{(i)}(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbf{Q}^{(i)}\mathbf{s}^{(i)} = -\mathbf{r}^{(i)} + \mathbf{d}_{JD}^{(i)}, \quad \text{for} \quad \mathbf{s}^{(i)} \perp \mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}, \tag{3.43}$$

$$\text{where} \quad \|\mathbf{d}_{JD}^{(i)}\| \le \tau_{JD}^{(i)}\|\mathbf{r}^{(i)}\|, \quad \text{and} \quad \tau_{JD}^{(i)} < 1.$$

First, we transform (3.43) into a system of the form (3.42), as follows. Since $\mathbf{Q}^{(i)}\mathbf{s}^{(i)} = \mathbf{s}^{(i)}$ and $\mathbf{r}^{(i)} = \mathbf{P}^{(i)}\mathbf{r}^{(i)} = \mathbf{P}^{(i)}(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbf{x}^{(i)}$, we can write (3.43) as

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

or

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

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

$$(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\frac{\mathbf{x}^{(i)} + \mathbf{s}^{(i)}}{\gamma^{(i)}} = \mathbf{M}\mathbf{x}^{(i)} + \frac{\mathbf{d}_{JD}^{(i)}}{\gamma^{(i)}}. \tag{3.44}$$

where (see (3.41))

$$\gamma^{(i)} = \frac{\mathbf{x}^{(i)H}\mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)} - \mathbf{x}^{(i)H}\mathbf{M}^H\mathbf{M}(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})^{-1}\mathbf{d}_{JD}^{(i)}}{\mathbf{x}^{(i)H}\mathbf{M}^H\mathbf{M}(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})^{-1}\mathbf{M}\mathbf{x}^{(i)}}. \tag{3.45}$$

This linear system (3.44) is of the form (3.42), and under the assumption that $\dfrac{\|\mathbf{d}_{JD}^{(i)}\|}{|\gamma^{(i)}|} \le \tau_I^{(i)}\|\mathbf{M}\mathbf{x}^{(i)}\|$ we can apply the theory in Section 3.3. Thus, we obtain the following Corollary from Theorem 3.7.

Corollary 3.17. *Let the assumptions and definitions of Theorem 3.7 hold and let*

$$\tau_I^{(i)} := \tau_{JD}^{(i)}\frac{\|\mathbf{r}^{(i)}\|}{|\gamma^{(i)}|\|\mathbf{M}\mathbf{x}^{(i)}\|}. \tag{3.46}$$

Then Algorithm 5 converges

- *linearly, if* $\tau_I^{(i)} < \dfrac{\alpha^{(i)}}{\|\mathbf{M}\mathbf{x}^{(i)}\|\|\mathbf{u}_1\|}\beta|s_{11}q^{(i)}|$ *with* $0 \leq 2\beta < 1 - T^{(0)}$ *and*

- *quadratically, if in addition* $\tau_I^{(i)} < \alpha^{(i)}\eta\dfrac{\|\mathbf{S}_{22}\mathbf{p}^{(i)}\|}{\|\mathbf{M}\mathbf{x}^{(i)}\|}$ *for some constant* $\eta > 0$.

Proof. Note that

$$\frac{\|\mathbf{d}_{JD}^{(i)}\|}{|\gamma^{(i)}|} \leq \tau_{JD}^{(i)}\frac{\|\mathbf{r}^{(i)}\|}{|\gamma^{(i)}|} := \tau_I^{(i)}\|\mathbf{M}\mathbf{x}^{(i)}\| \tag{3.47}$$

and using $\tau^{(i)} := \tau_I^{(i)}$ in Theorem 3.7 gives the result. □

Example 3.18 (Bounded Finline Dielectric Waveguide). *Consider the generalised eigenproblem* $\mathbf{A}\mathbf{x} = \lambda\mathbf{M}\mathbf{x}$, *where* $\mathbf{A}$ *and* $\mathbf{M}$ *are given by* `bfw782a.mtx` *and* `bfw782b.mtx` *in the Matrix Market library [13]. These are matrices of size* 782, *where* $\mathbf{A}$ *is real nonsymmetric and has* 7514 *non-zero entries,* $\mathbf{M}$ *is real symmetric indefinite and has* 5982 *non-zero entries. We seek the smallest eigenvalue in magnitude which is given by* $\lambda_1 = 564.6$. *Our only interest in this chapter is the* outer convergence rate, *(though, for information we use GMRES for the inner solves in Algorithm 5). We use a variable shift given by the generalised Rayleigh quotient* $\rho(\mathbf{x}^{(i)})$, *and either a decreasing tolerance which is given by* $\tau^{(i)} = \min\{0.05, 0.05\,\|\mathbf{r}^{(i)}\|\}$ *or a fixed tolerance given by* $\tau = 0.05$.

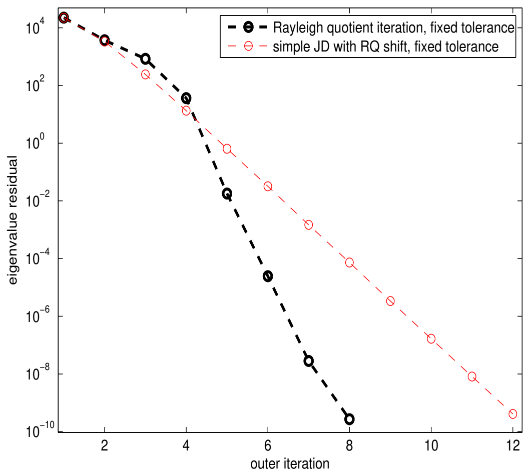


Figure 3-2: *Convergence history of the eigenvalue residuals for Example 3.18 using* Rayleigh quotient shift *and inexact solves with fixed tolerance.*

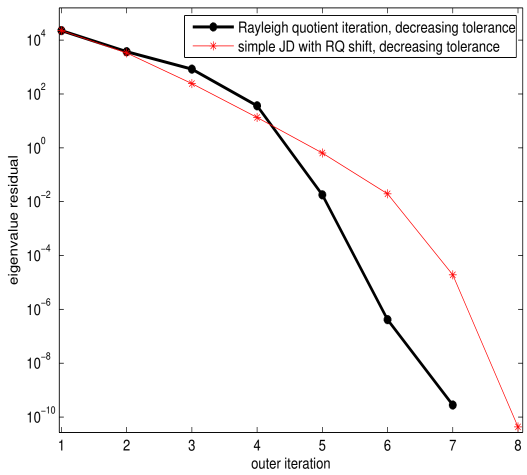


Figure 3-3: *Convergence history of the eigenvalue residuals for Example 3.18 using* Rayleigh quotient shift *and inexact solves with decreasing tolerance.*

Figures 3-2 and 3-3 illustrate the convergence history for inexact Rayleigh quotient iteration and simple Jacobi-Davidson. We observe that a decreasing solve tolerance in the simple Jacobi-Davidson method with generalised Rayleigh quotient shift leads to quadratic convergence (Figure 3-3) whereas with a fixed solve tolerance only linear convergence may be achieved with a small enough tolerance (Figure 3-2). For comparison we have also plotted the results for inexact inverse iteration with a generalised Rayleigh

quotient shift, where both the same decreasing tolerance $\tau^{(i)}$ and fixed tolerance τ were used as for the simple inexact Jacobi-Davidson method.
Since, in this chapter, we are only concerned about the outer convergence rate, from (3.47) we note that in theory the quantity $\|\mathbf{r}^{(i)}\|/|\gamma^{(i)}|$ is crucial for the comparison of the performance of the two methods. We note the following:

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

The following example illustrates this further.

Example 3.19. *We construct two simple test examples, one for which the quantity* $\|\mathbf{r}^{(i)}\|/|\gamma^{(i)}|$ *turns out to be greater than one, and one for which this quantity is less than one. We use a standard eigenproblem* $\mathbf{Ax} = \lambda\mathbf{x}$ *with* $\mathbf{A} = diag(1, 2, \ldots, 500)$ *and set either* $\mathbf{A}(1, 2 : 300) = 1$ *(case (a)) or* $\mathbf{A}(1, 2 : 300) = 10$ *(case (b)). Clearly, in the second problem the non-normality has been increased. We seek the smallest eigenvalue* $\lambda_1 = 1$ *and use GMRES for the inner solves. Further we use a variable shift given by the generalised Rayleigh quotient* $\rho(\mathbf{x}^{(i)})$ *and a fixed tolerance given by* $\tau = 0.1$*. We compare inexact Rayleigh quotient iteration and inexact simple Jacobi-Davidson. Both methods have linear convergence and stop once the eigenvalue residual satisfies* $\|\mathbf{r}^{(i)}\| < 10^{-10}$.

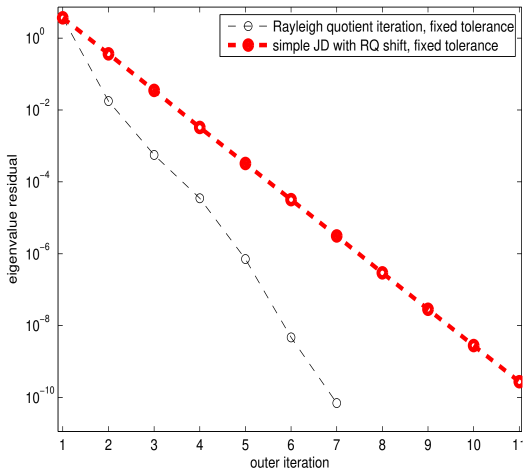


Figure 3-4: *Convergence history of the eigenvalue residuals for Example 3.19 where* $\|\mathbf{r}^{(i)}\|/|\gamma^{(i)}| > 1$ *(fixed tolerance)*

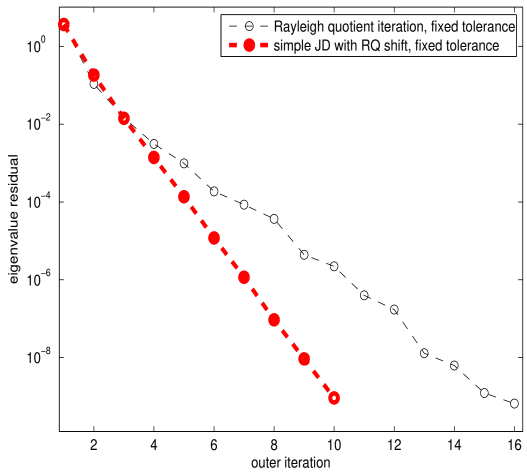


Figure 3-5: *Convergence history of the eigenvalue residuals for Example 3.19 where* $\|\mathbf{r}^{(i)}\|/|\gamma^{(i)}| < 1$ *(fixed tolerance)*

Figure 3-4 illustrates the convergence history of the eigenvalue residuals for the two methods discussed above for case (a), the mildly non-normal case. The corresponding values of $\|\mathbf{r}^{(i)}\|/|\gamma^{(i)}|$ are listed in the second row of Table 3.3 and turn out to be greater

Table 3.3: *Values for $\|\mathbf{r}^{(i)}\|/|\gamma^{(i)}|$ in Figures 3-4 and 3-5 for fixed tolerance solves (fixed tolerance)*

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

than one. As expected in this case, the convergence rate of inexact Rayleigh quotient iteration is better than the convergence rate of inexact simple Jacobi-Davidson with Rayleigh quotient shift. On the other hand, Figure 3-5 shows the convergence history of the eigenvalue residuals for case (b), there the nonnormality of the problem is larger. The corresponding values of $\|\mathbf{r}^{(i)}\|/|\gamma^{(i)}|$ are listed in the third row of Table 3.3 and are found to be less than one after the first iteration. As predicted, the convergence rate of inexact simple Jacobi-Davidson with Rayleigh quotient shift is better than inexact Rayleigh quotient iteration in this case. These numerical results suggest that the quantity, $\|\mathbf{r}^{(i)}\|/|\gamma^{(i)}|$ may depend on the nonnormality of the problem.

We remark that similar results as in Figures 3-4 and 3-5 and Table 3.3 are obtained if a decreasing tolerance is used, with the only difference that quadratic convergence is obtained for both Rayleigh quotient iteration and simplified Jacobi-Davidson with Rayleigh quotient shifts.

Finally, we note that for Example 3.18 the quantity $\|\mathbf{r}^{(i)}\|/|\gamma^{(i)}|$ was greater than one throughout the computations, leading to a faster convergence rate for inexact Rayleigh quotient iteration. Further investigation onto this quantity is future research.

3.6.3 Transforming inexact Rayleigh quotient iterations into inexact Jacobi-Davidson

Finally we would like to show how system (3.42) can be transformed into a system of the form (3.43). We may reformulate (3.42) as

$$(\mathbf{A}-\sigma^{(i)}\mathbf{M})\mathbf{x}^{(i+1)}\phi(\mathbf{y}^{(i)}) = \mathbf{M}\mathbf{x}^{(i)}-\mathbf{d}_I^{(i)}, \quad \text{where} \quad \|\mathbf{d}_I^{(i)}\| \leq \tau_I^{(i)}\|\mathbf{M}\mathbf{x}^{(i)}\|, \quad \text{with} \quad \tau_I^{(i)} < 1,$$

and $\mathbf{x}^{(i+1)}$ is just a normalised version of $\mathbf{x}^{(i)} + \delta\mathbf{x}^{(i)}$, where $\delta\mathbf{x}^{(i)}$ is chosen such that $\delta\mathbf{x}^{(i)} \perp \mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}$. Hence

$$(\mathbf{A}-\sigma^{(i)}\mathbf{M})(\mathbf{x}^{(i)}+\delta\mathbf{x}^{(i)}) = (\mathbf{M}\mathbf{x}^{(i)}-\mathbf{d}_I^{(i)})\beta^{(i)}, \tag{3.48}$$

where $\beta^{(i)}$ is determined using the condition $\delta\mathbf{x}^{(i)} \perp \mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}$:

$$\beta^{(i)} = \frac{\mathbf{x}^{(i)H}\mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}}{\mathbf{x}^{(i)H}\mathbf{M}^H\mathbf{M}(\mathbf{A}-\sigma^{(i)}\mathbf{M})^{-1}(\mathbf{M}\mathbf{x}^{(i)}-\mathbf{d}_I^{(i)})}.$$

Thus, using $(\mathbf{A}-\sigma^{(i)}\mathbf{M})\mathbf{x}^{(i)} = \mathbf{r}^{(i)}$ as well as $\delta\mathbf{x}^{(i)} \perp \mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}$ and multiplying equation (3.48) by $\mathbf{P}^{(i)}$ defined at the beginning of Section 3.6.1 we obtain

$$\mathbf{P}^{(i)}(\mathbf{A}-\sigma^{(i)}\mathbf{M})\mathbf{Q}^{(i)}\delta\mathbf{x}^{(i)} = -\mathbf{r}^{(i)} - \mathbf{P}^{(i)}\mathbf{d}_I^{(i)}\beta^{(i)}.$$

This system is of the form (3.43) and if we can show that $\|\mathbf{d}_I^{(i)}\beta^{(i)}\| < \tau_{JD}^{(i)}\|\mathbf{r}^{(i)}\|$, we have shown equivalence between (3.43) and (3.42).

We have the following Lemmata and Proposition which extend the results in Section 3.6.2.

Lemma 3.20. *Let (3.42) and (3.43) hold. If $\tau_{JD}^{(i)}$ is chosen such that*

$$\tau_{JD}^{(i)} = \frac{\tau_I^{(i)}}{1+\tau_I^{(i)}} \frac{\|\mathbf{M}\mathbf{x}^{(i)}\|}{\|\mathbf{M}(\mathbf{A}-\sigma^{(i)}\mathbf{M})^{-1}\|\|\mathbf{r}^{(i)}\|}, \tag{3.49}$$

then $\dfrac{\|\mathbf{d}_{JD}^{(i)}\|}{|\gamma^{(i)}|} \leq \tau_I^{(i)}\|\mathbf{M}\mathbf{x}^{(i)}\|$ *holds.*

Proof. With $\gamma^{(i)}$ chosen as in (3.45) and $\|\mathbf{d}_{JD}^{(i)}\| \leq \tau_{JD}^{(i)}\|\mathbf{r}^{(i)}\|$ we can bound

$$\frac{\|\mathbf{d}_{JD}^{(i)}\|}{|\gamma^{(i)}|} \leq \frac{\tau_{JD}^{(i)}\|\mathbf{r}^{(i)}\|\|\mathbf{M}(\mathbf{A}-\sigma^{(i)}\mathbf{M})^{-1}\|\|\mathbf{M}\mathbf{x}^{(i)}\|^2}{\|\mathbf{M}\mathbf{x}^{(i)}\|^2 - \|\mathbf{M}(\mathbf{A}-\sigma^{(i)}\mathbf{M})^{-1}\|\tau_{JD}^{(i)}\|\mathbf{r}^{(i)}\|\|\mathbf{M}\mathbf{x}^{(i)}\|}.$$

Choosing $\tau_{JD}^{(i)}$ as in (3.49) we obtain $\dfrac{\|\mathbf{d}_{JD}^{(i)}\|}{|\gamma^{(i)}|} \leq \tau_I^{(i)}\|\mathbf{M}\mathbf{x}^{(i)}\|$. □

Lemma 3.20 shows that if $\tau_{JD}^{(i)}$ is chosen as in (3.49), then one step of inexact Jacobi-Davidson with solve tolerance $\tau_{JD}^{(i)}$ can be expressed in terms of one step of inexact inverse iteration with solve tolerance $\tau_I^{(i)}$. Then the convergence theory for inexact inverse iteration of the previous sections can be applied to inexact simplified Jacobi-Davidson. Finally, we show the equivalence between $\tau_{JD}^{(i)}$ and $\tau_I^{(i)}$.

Proposition 3.21. *Let $\tau_{JD}^{(i)}$ be given as in (3.49) and assume that $\sigma^{(i)} := \rho(\mathbf{x}^{(i)})$. Then*

$$C\frac{\tau_I^{(i)}}{1+\tau_I^{(i)}} \leq \tau_{JD}^{(i)} \leq \frac{\tau_I^{(i)}}{1+\tau_I^{(i)}}, \tag{3.50}$$

where $C = \dfrac{|s_{11}|}{\|\mathbf{S}\|C^2_{\|\mathbf{g}_{12}\|}} < 1$ *is a constant independent of i.*

Proof. Multiplying $\mathbf{r}^{(i)} = (\mathbf{A}-\sigma^{(i)}\mathbf{M})\mathbf{x}^{(i)}$ by $\mathbf{M}(\mathbf{A}-\sigma^{(i)}\mathbf{M})^{-1}$ from the right we obtain $\|\mathbf{M}\mathbf{x}^{(i)}\| \leq \|\mathbf{M}(\mathbf{A}-\sigma^{(i)}\mathbf{M})^{-1}\|\|\mathbf{r}^{(i)}\|$ and hence the upper bound in (3.50) follows from (3.49). For the lower bound we have

$$\begin{aligned}
\|\mathbf{M}(\mathbf{A}-\sigma^{(i)}\mathbf{M})^{-1}\| &= \|\mathbf{M}(\mathbf{U}\mathbf{U}^{-1}(\mathbf{A}-\sigma^{(i)}\mathbf{M})\mathbf{X}\mathbf{X}^{-1})^{-1}\| \\
&\leq \|\mathbf{M}\mathbf{X}\|\|\mathbf{U}^{-1}\| \left\| \begin{bmatrix} \dfrac{1}{t_{11}-\sigma^{(i)}s_{11}} & \mathbf{0}^H \\ \mathbf{0} & (\mathbf{T}_{22}-\sigma^{(i)}\mathbf{S}_{22})^{-1} \end{bmatrix} \right\| \\
&\leq \|\mathbf{S}\|\|\mathbf{G}\|^2 \frac{1}{|\lambda_1-\sigma^{(i)}||s_{11}|} \\
&\leq \frac{\|\mathbf{S}\|C_{\|\mathbf{g}_{12}\|}}{|s_{11}|}\frac{1}{|\lambda_1-\sigma^{(i)}|},
\end{aligned}$$

where we have used the results in Lemma 3.3. $\mathbf{S}$ is the block-diagonal matrix in (3.10). Therefore, we have

$$\frac{\|\mathbf{M}(\mathbf{A}-\sigma^{(i)}\mathbf{M})^{-1}\|\|\mathbf{r}^{(i)}\|}{\|\mathbf{M}\mathbf{x}^{(i)}\|} \leq \frac{\|\mathbf{S}\|C_{\|\mathbf{g}_{12}\|}}{|s_{11}|}\frac{\|\mathbf{r}^{(i)}\|}{|\lambda_1-\sigma^{(i)}|\|\mathbf{M}\mathbf{x}^{(i)}\|}. \tag{3.51}$$

From Lemma 3.5 we have

$$\|\mathbf{r}^{(i)}\| \leq |\alpha^{(i)}|\|\mathbf{U}\|\|\bar{\mathbf{I}}_{n-1}(\mathbf{T}_{22} - \lambda_1 \mathbf{S}_{22})\mathbf{p}^{(i)}\|. \tag{3.52}$$

Furthermore, using the methods in Lemma 3.5, for $(\lambda_1 - \sigma^{(i)})\mathbf{M}\mathbf{x}^{(i)}$ we have

$$(\lambda_1 - \sigma^{(i)})\mathbf{M}\mathbf{x}^{(i)} = \alpha^{(i)} \frac{\mathbf{M}\mathbf{x}^{(i)}{\mathbf{x}^{(i)}}^H\mathbf{M}^H}{{\mathbf{x}^{(i)}}^H\mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}} \mathbf{U}\bar{\mathbf{I}}_{n-1}(\mathbf{T}_{22} - \lambda_1 \mathbf{S}_{22})\mathbf{p}^{(i)}.$$

Consider the pseudo-inverse of the rank one matrix $\dfrac{\mathbf{M}\mathbf{x}^{(i)}{\mathbf{x}^{(i)}}^H\mathbf{M}^H}{{\mathbf{x}^{(i)}}^H\mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}}$ and write

$$(\lambda_1 - \sigma^{(i)}) \left(\frac{\mathbf{M}\mathbf{x}^{(i)}{\mathbf{x}^{(i)}}^H\mathbf{M}^H}{{\mathbf{x}^{(i)}}^H\mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}} \right)^{\dagger} \mathbf{M}\mathbf{x}^{(i)} = \alpha^{(i)}\mathbf{U}\bar{\mathbf{I}}_{n-1}(\mathbf{T}_{22} - \lambda_1 \mathbf{S}_{22})\mathbf{p}^{(i)}.$$

This procedure is possible since $\mathbf{M}\mathbf{x}^{(i)}$ is in the range of $\dfrac{\mathbf{M}\mathbf{x}^{(i)}{\mathbf{x}^{(i)}}^H\mathbf{M}^H}{{\mathbf{x}^{(i)}}^H\mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}}$. Finally, multiplying by $\mathbf{U}^{-1}$ from the left and using $\left\| \left(\dfrac{\mathbf{M}\mathbf{x}^{(i)}{\mathbf{x}^{(i)}}^H\mathbf{M}^H}{{\mathbf{x}^{(i)}}^H\mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}} \right)^{\dagger} \right\| = 1$ we obtain the bound

$$\|\bar{\mathbf{I}}_{n-1}(\mathbf{T}_{22} - \lambda_1 \mathbf{S}_{22})\mathbf{p}^{(i)}\| \leq \frac{\|\mathbf{U}^{-1}\|}{|\alpha^{(i)}|} |\lambda_1 - \sigma^{(i)}|\|\mathbf{M}\mathbf{x}^{(i)}\|,$$

and hence

$$|\lambda_1 - \sigma^{(i)}|\|\mathbf{M}\mathbf{x}^{(i)}\| \geq \frac{|\alpha^{(i)}|}{\|\mathbf{U}^{-1}\|} \|\bar{\mathbf{I}}_{n-1}(\mathbf{T}_{22} - \lambda_1 \mathbf{S}_{22})\mathbf{p}^{(i)}\|. \tag{3.53}$$

Combining bounds (3.52) and (3.53) we obtain from (3.51)

$$\frac{\|\mathbf{M}(\mathbf{A} - \sigma^{(i)}\mathbf{M})^{-1}\|\|\mathbf{r}^{(i)}\|}{\|\mathbf{M}\mathbf{x}^{(i)}\|} \leq \frac{\|\mathbf{S}\|C_{\|\mathbf{g}_{12}\|}}{|s_{11}|} \frac{|\alpha^{(i)}|\|\mathbf{U}\|\|\mathbf{U}^{-1}\|}{|\alpha^{(i)}|} \leq \frac{\|\mathbf{S}\|C^2_{\|\mathbf{g}_{12}\|}}{|s_{11}|}.$$

Hence the lower bound in (3.50) follows from (3.49). □

The bounds in (3.50) show that $\tau_{JD}^{(i)}$ and $\tau_I^{(i)}$ are equivalent in the following sense: if $\tau_I^{(i)}$ is chosen to decrease then $\tau_{JD}^{(i)}$ decreases in the same manner, if $\tau_I^{(i)}$ is kept fixed then so is $\tau_{JD}^{(i)}$. Hence, if we choose $\tau_{JD}^{(i)}$ according to $\tau_I^{(i)}$ we may apply the convergence theory from Section 3.3 and obtain similar convergence rates for inexact Jacobi-Davidson as for inexact inverse iteration. Proposition 3.21 shows that $\tau_{JD}^{(i)}$ can be bounded below and above by terms only involving $\tau_I^{(i)}$, improving the result in 3.6.2, where the unknown $\gamma^{(i)}$ is used in the theory. With the following lemma and the remarks thereafter we see that the reverse also holds; $\tau_I^{(i)}$ can be bounded below and above by terms only involving $\tau_{JD}^{(i)}$.

Lemma 3.22. *Let (3.43) and (3.42) hold. If $\tau_I^{(i)}$ is chosen such that*

$$\tau_I^{(i)} = \frac{\tau_{JD}^{(i)}}{\tau_{JD}^{(i)} + \dfrac{\|\mathbf{M}\mathbf{x}^{(i)}\|}{\|\mathbf{M}(\mathbf{A} - \sigma^{(i)}\mathbf{M})^{-1}\|\|\mathbf{r}^{(i)}\|}}, \tag{3.54}$$

then $\left\| \left(\mathbf{I} - \frac{\mathbf{M}\mathbf{x}^{(i)}{\mathbf{x}^{(i)}}^H\mathbf{M}^H}{{\mathbf{x}^{(i)}}^H\mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}} \right) \mathbf{d}_I^{(i)}\beta^{(i)} \right\| < \tau_{JD}^{(i)}\|\mathbf{r}^{(i)}\|$ *holds.*

Proof. We have

$$\left\| \left(\mathbf{I} - \frac{\mathbf{M}\mathbf{x}^{(i)}{\mathbf{x}^{(i)}}^H\mathbf{M}^H}{{\mathbf{x}^{(i)}}^H\mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}} \right) \mathbf{d}_I^{(i)}\beta^{(i)} \right\| \leq \|\mathbf{d}_I^{(i)}\||\beta^{(i)}|$$

and using $\beta^{(i)}$ we get

$$\|\mathbf{d}_I^{(i)}\||\beta^{(i)}| \leq \tau_I^{(i)} \frac{\|\mathbf{M}\mathbf{x}^{(i)}\|\|\mathbf{M}\mathbf{x}^{(i)}\|^2}{\|\mathbf{M}(\mathbf{A}-\sigma^{(i)}\mathbf{M})^{-1}\|\|\mathbf{M}\mathbf{x}^{(i)}\|^2 - \tau_I^{(i)}\|\mathbf{M}\mathbf{x}^{(i)}\|^2\|\mathbf{M}(\mathbf{A}-\sigma^{(i)}\mathbf{M})^{-1}\|}.$$

With the choice of $\tau_I^{(i)}$ we obtain the result. □

Note that using the results from the proof of Proposition 3.21 we have

$$\frac{\tau_{JD}^{(i)}}{1+\tau_{JD}^{(i)}} \leq \tau_I^{(i)} \leq \frac{\tau_{JD}^{(i)}}{C+\tau_{JD}^{(i)}},$$

where $C = \frac{|s_{11}|}{\|\mathbf{S}\|C^2_{\|\mathbf{g}_{12}\|}} < 1$ is a constant independent of i.

3.7 Conclusions

In this chapter we have provided a full convergence theory for inexact inverse iteration for fixed and variable shifts applied to the generalised eigenproblem $\mathbf{A}\mathbf{x} = \lambda\mathbf{M}\mathbf{x}$ with minimal assumptions on $\mathbf{A}$ and $\mathbf{M}$ by introducing a new convergence measure.

Furthermore we have shown that convergence of inexact inverse iteration leads to an increase of the norm of the solution and hence no projection is necessary for inexact inverse iteration applied to a constraint eigenproblem.

Finally we have compared inexact Rayleigh quotient iteration to a simplified version of Jacobi-Davidson method with Rayleigh quotient shift and inexact solves, and shown that both methods are equivalent in a certain sense and hence provided convergence results for Jacobi-Davidson method.

CHAPTER 4

A tuned preconditioner for inexact inverse iteration for Hermitian eigenvalue problems

4.1 Introduction

In this chapter, we consider the problem of computing an eigenvalue and the corresponding eigenvector of a Hermitian positive definite matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$, that is

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{x}, \quad \lambda \in \mathbb{R}, \mathbf{x} \in \mathbb{C}^n \setminus \{\mathbf{0}\}, \tag{4.1}$$

using inexact inverse iteration with a fixed shift. We assume that the matrix $\mathbf{A}$ is very large and sparse and so to exploit the structure iterative techniques, in particular, preconditioned MINRES, may be used to solve the linear shifted systems

$$(\mathbf{A} - \sigma \mathbf{I})\mathbf{y} = \mathbf{x} \tag{4.2}$$

arising in inverse iteration, where the shift σ is chosen to be close to any eigenvalue. Whereas Chapters 2 and 3 concentrated on the convergence theory for inexact inverse iteration, this chapter deals with the efficiency of the preconditioned iterative solves of (4.2).

In order to reduce the number of inner iterations needed to solve (4.2), preconditioning becomes necessary. Since $\mathbf{A}$ is Hermitian positive definite, we use an incomplete Cholesky factorisation of $\mathbf{A}$ to construct a symmetrically preconditioned form of (4.2). Specifically, if $\mathbf{L}\mathbf{L}^H$ is an incomplete Cholesky factorisation of $\mathbf{A}$ then one applies an iterative solver (for example, MINRES) to

$$\mathbf{L}^{-1}(\mathbf{A} - \sigma \mathbf{I})\mathbf{L}^{-H}\tilde{\mathbf{y}} = \mathbf{L}^{-1}\mathbf{x}, \quad \mathbf{y} = \mathbf{L}^{-H}\tilde{\mathbf{y}}, \tag{4.3}$$

rather than to (4.2). For a fixed shift, it is known that (see for example [10], [75]) the number of inner iterations used by a Krylov solver applied to (4.3) increases steadily as the outer iteration proceeds, because the solve tolerance for the iterative solver has to be chosen to decrease in order to obtain convergence. We shall show that with a simple rank-one change to the preconditioner, which we call "tuning" the preconditioner, this steady increase in the number of inner iterations can be stopped, and indeed considerable improvements in the total inner iteration count can be achieved.

In Chapter 2 (see also [43]) the concept of "tuning" the preconditioner to improve the outer convergence of a variant of inverse iteration was introduced, but no analysis of the tuned preconditioner was given. Based on [43] and the technical report [41], [104] analysed a subspace version of tuning for the standard eigenproblem and introduced the concept of an "ideal" preconditioner. In this chapter we extend that analysis to obtain a detailed description of the performance of MINRES, and in particular show that the tuned preconditioner should not exhibit growth in the number of inner iterations as the outer iteration proceeds. Then we provide a careful spectral analysis that explains the differences between the iteration matrices for the tuned and standard cases. This involves the formulation of a nonstandard eigenvalue perturbation problem, which is analysed by a modification of the Bauer-Fike theorem (see [48]) and a novel interlacing property (in the spirit of [151, p. 94 ff] and [48]). These results show that the spectral properties of the tuned preconditioner are similar to those of the standard preconditioner.

For the case of Rayleigh quotient shifts, the idea from [119] is to modify the right hand side of the preconditioned system (4.3) so that the new right hand side is close to an approximate null-vector of the iteration matrix (see Section 4.5). This new strategy reduces the number of inner iterations for each solve of (4.2), but destroys the cubic outer convergence for Rayleigh quotient iteration, achieving only quadratic outer convergence. (Note that this strategy requires that the shifts tend to the desired eigenvalue and so is not an option when the shift is fixed.) We compare the use of the tuned preconditioner with the approach of [119] and find that the tuned preconditioner is also superior in terms of overall iteration count.

In Section 4.2 of this chapter we discuss the theory of inexact inverse iteration with a fixed shift, the convergence theory for MINRES, and then go over the use of the standard incomplete Cholesky preconditioner. We also discuss the application of these results to the solution of the shifted systems in inexact inverse iteration. In Section 4.3 we make a comparison with the "ideal" preconditioner of [104] and prove the main theorem (Theorem 4.11) about the performance of MINRES applied to the tuned preconditioned shifted system. Numerical results are presented to show the superiority of the tuned preconditioner over the standard preconditioner. In Section 4.4 we provide a detailed analysis of the spectra of both the tuned and untuned iteration matrices and discuss the consequences for MINRES as iterative solver for the inner iterations. In Section 4.5 the tuned preconditioner is applied to inexact Rayleigh quotient iteration and we compare the numerical performance of the standard and tuned preconditioners. Again, the tuned preconditioner is superior to the the standard preconditioner. Numerical results are also presented comparing the performance of the tuned preconditioner with the approach of [119]. Section 4.6 summarises the main results of the chapter.

We denote the eigenpairs of $\mathbf{A}$ by $(\lambda_j, \mathbf{x}_j)$, $j = 1, \ldots, n$, and use $\| \cdot \| = \| \cdot \|_2$.

4.2 Inexact inverse iteration with a fixed shift

In this section we review the theory for inexact inverse iteration with a fixed shift for the calculation of a simple eigenvalue of the standard Hermitian eigenvalue problem (4.1), and then go on to discuss the use of MINRES as the iterative solver and preconditioning. A fixed shift method is unlikely to be of interest on its own, but it might well be used to provide a good starting guess for the eigenvector to feed into the Rayleigh quotient iteration. Also, results for fixed shifts are of interest when using subspace based methods, like the Lanczos method.

The following algorithm is a version of inexact inverse iteration with a fixed shift to find any well-separated simple eigenvalue of a Hermitian matrix.

Algorithm 6 Inexact inverse iteration with a fixed shift

Input: Shift σ and $\mathbf{x}^{(0)}$ with $\|\mathbf{x}^{(0)}\|= 1$.
  for $i = 1, \ldots, i_{max}$ **do**
    Choose $\tau^{(i)}$,
    Solve $(\mathbf{A} - \sigma\mathbf{I})\mathbf{y}^{(i)} = \mathbf{x}^{(i)}$ inexactly, that is,

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

    Compute approximate eigenvector $\mathbf{x}^{(i+1)} = \dfrac{\mathbf{y}^{(i)}}{\|\mathbf{y}^{(i)}\|}$,
    Compute approximate eigenvalue $\lambda^{(i+1)} = {\mathbf{x}^{(i+1)}}^H \mathbf{A}\mathbf{x}^{(i+1)}$,
    Evaluate eigenvalue residual $\mathbf{r}^{(i+1)} = (\mathbf{A} - \lambda^{(i+1)}\mathbf{I})\mathbf{x}^{(i+1)}$,
    Test for convergence.
  end for
Output: $\mathbf{x}^{i_{\max}}$, $\lambda^{i_{\max}}$.

The following theorem states the convergence theory for inexact inverse iteration with a fixed shift. It follows directly from Theorem 2.2 in [10], where a detailed proof is given (see also Lemma 2.2 in [50]).

Theorem 4.1 (Convergence of inexact inverse iteration with a fixed shift). *Let (4.1) be the standard eigenvalue problem for a Hermitian matrix* $\mathbf{A}$ *and consider the application of Algorithm 6 to find a simple eigenpair* $(\lambda_1, \mathbf{x}_1)$ *of* $\mathbf{A}$*. Assume* σ *is closer to* λ_1 *than to any other eigenvalue of* $\mathbf{A}$*, and that* $\mathbf{x}^{(0)}$ *is close enough to the desired* $\mathbf{x}_1$*. Then, if a decreasing tolerance is chosen for the inexact solves in the inverse iteration Algorithm 6, say* $\tau^{(i)} = C_1 \|\mathbf{r}^{(i)}\|$ *in step (1), then linear convergence is achieved for small enough* $\tau^{(0)}$ *and* C_1*.*

Proof. Following [101], if we write $\mathbf{x}^{(i)}$ as orthogonal decomposition

$$\mathbf{x}^{(i)} = \cos\theta^{(i)}\mathbf{x}_1 + \sin\theta^{(i)}\mathbf{x}_\perp^{(i)}, \quad \mathbf{x}_\perp^{(i)} \perp \mathbf{x}_1, \tag{4.4}$$

with $\|\mathbf{x}_1\| = \|\mathbf{x}_\perp^{(i)}\| = 1$ and $\theta^{(i)} = \angle(\mathbf{x}^{(i)}, \mathbf{x}_1)$, then the eigenvalue residual defined by

$$\mathbf{r}^{(i)} = (\mathbf{A} - \lambda^{(i)}\mathbf{I})\mathbf{x}^{(i)}, \tag{4.5}$$

with $\lambda^{(i)} = {\mathbf{x}^{(i)}}^H\mathbf{A}\mathbf{x}^{(i)}$ satisfies (see [101, Theorem 11.7.1])

$$|\sin\theta^{(i)}||\lambda_2 - \lambda^{(i)}| \leq \|\mathbf{r}^{(i)}\| \leq |\sin\theta^{(i)}||\lambda_n - \lambda_1|. \tag{4.6}$$

Thus the choice of $\tau^{(i)} = C_1 \|\mathbf{r}^{(i)}\|$ asks that the solve tolerance in step (2) of Algorithm 6 decreases with the error angle $\theta^{(i)}$. From [10, Lemma 2.1] we have

$$|\tan\theta^{(i+1)}| \leq \frac{|\lambda_1 - \sigma|}{|\lambda_2 - \sigma|}\frac{|\sin\theta^{(i)}| + \tau^{(i)}}{|\cos\theta^{(i)}| - \tau^{(i)}},$$

which, with the choice of $\tau^{(i)}$ yields linear convergence for small enough C_1. □

Note that for the special case of a tolerance $\tau^{(i)} = 0$ we obtain the well known linear convergence achieved by exact inverse iteration.

4.2.1 Convergence theory of MINRES

In order to understand the performance of the inner iteration part of the inexact inverse iteration algorithm we review some convergence theory of MINRES.

First, we quote a theorem about the convergence of MINRES when applied to

$$\mathbf{B}\mathbf{z} = \mathbf{b} \tag{4.7}$$

for the case of interest here. This is a special case of Theorem 3.1 of [10], but similar results are well known in the literature (see, for example [55] and [56]).

Theorem 4.2. *Suppose that the Hermitian matrix $\mathbf{B}$ has eigenvalues $\mu_1, \ldots, \mu_n$ with corresponding eigenvectors $\mathbf{w}_1, \ldots, \mathbf{w}_n$. Let μ_1 be well separated from $\{\mu_j\}_{j=2}^n$. Furthermore, let $\kappa^1 = \dfrac{\max_{j=2,\ldots,n} |\mu_j|}{\min_{j=2,\ldots,n} |\mu_j|}$ be the reduced condition number of $\mathbf{B}$, assume $\max_{j=2,\ldots,n} |\mu_1 - \mu_j| = |\mu_1 - \mu_n|$ and define $\mathcal{P}^\perp$ to be the orthogonal projection along $\mathbf{w}_1$ onto span$\{\mathbf{w}_2, \ldots, \mathbf{w}_n\}$. If $\mathbf{z}_k$ is the result of applying MINRES to (4.7) with starting value $\mathbf{z}_0 = \mathbf{0}$ then*

$$\|\mathbf{b} - \mathbf{B}\mathbf{z}_k\| \leq 2 \max_{j=2,\ldots,n} \frac{|\mu_1 - \mu_j|}{|\mu_1|} \left(\frac{\sqrt{\kappa^1} - 1}{\sqrt{\kappa^1} + 1} \right)^{k-1} \|\mathcal{P}^\perp \mathbf{b}\|, \tag{4.8}$$

if all the elements of $\{\mu_j\}_{j=2}^n$ have the same sign and

$$\|\mathbf{b} - \mathbf{B}\mathbf{z}_k\| \leq 2 \max_{j=2,\ldots,n} \frac{|\mu_1 - \mu_j|}{|\mu_1|} \left(\sqrt{\frac{\kappa^1 - 1}{\kappa^1 + 1}} \right)^{k-2} \|\mathcal{P}^\perp \mathbf{b}\|$$

otherwise. In addition, if the number of iterations satisfies

$$k \geq 1 + \frac{\sqrt{\kappa^1}}{2} \left(\log 2 \frac{|\mu_1 - \mu_n|}{|\mu_1|} + \log \frac{\|\mathcal{P}^\perp \mathbf{b}\|}{\tau} \right), \tag{4.9}$$

or

$$k \geq 2 + \kappa^1 \left(\log 2 \frac{|\mu_1 - \mu_n|}{|\mu_1|} + \log \frac{\|\mathcal{P}^\perp \mathbf{b}\|}{\tau} \right), \tag{4.10}$$

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

Note that in Theorem 4.2 the eigenvalues μ_j, $j = 1, \ldots, n$ of $\mathbf{B}$ are not necessarily sorted, in particular, we allow μ_1 to be an interior eigenvalue of $\mathbf{B}$.

Remark 4.3. *Note that the bounds (4.9) and (4.10) in Theorem 4.2 are worst case bounds and may indeed be worse than the trivial bound $k^{(i)} \geq n$. In general these bounds are often used to give qualitative rather than quantitative information, since in practice convergence of MINRES can be much faster. Also, for simplicity we shall consider only the case of a simple extreme eigenvalue, since the convergence theory for MINRES is easiest. Therefore, in this chapter we concentrate on the first case, where $\{\mu_j\}_{j=2}^n$ have the same sign; all results generalise to the second case.*

We now apply this theorem to the solution of $(\mathbf{A}-\sigma\mathbf{I})\mathbf{y}^{(i)} = \mathbf{x}^{(i)}$, with $\mathbf{B} = \mathbf{A}-\sigma\mathbf{I}$, $\mathbf{b} = \mathbf{x}^{(i)}$, $\mu_j = \lambda_j - \sigma$, $\mathbf{w}_j = \mathbf{x}_j$ and

$$\mathcal{P}^{\perp}\mathbf{x}^{(i)} = \sin\theta^{(i)}\mathbf{x}_{\perp}^{(i)}, \tag{4.11}$$

using (4.4). Thus if $k^{(i)}$ denotes the number of inner iterations used by MINRES to solve $(\mathbf{A}-\sigma\mathbf{I})\mathbf{y}^{(i)} = \mathbf{x}^{(i)}$ inexactly as in step (2) of Algorithm 6, then

$$k^{(i)} \geq 1 + \frac{\sqrt{\kappa^1}}{2}\left(\log 2\frac{|\lambda_1-\lambda_n|}{|\lambda_1-\sigma|} + \log\frac{|\sin\theta^{(i)}|}{\tau^{(i)}}\right). \tag{4.12}$$

With $|\lambda_1-\sigma|$ fixed and $\tau^{(i)} = C_1\|\mathbf{r}^{(i)}\|$, we see, using (4.6), that

$$\frac{|\sin\theta^{(i)}|}{\tau^{(i)}} = \frac{|\sin\theta^{(i)}|}{C_1\|\mathbf{r}^{(i)}\|} \leq \frac{|\sin\theta^{(i)}|}{C_1|\sin\theta^{(i)}||\lambda_2-\lambda^{(i)}|} \leq \frac{1}{C_1(|\lambda_2-\lambda_1|-|\lambda_1-\lambda^{(i)}|)},$$

which can be bounded independently of i for large enough i (since $\lambda^{(i)} \to \lambda_1$). Therefore the right hand side of (4.12) is bounded independently of i for large enough i. Hence we infer that the number of inner iterations used by MINRES will not increase as the outer iteration proceeds. This nice property is not maintained when preconditioning is applied as we discuss next.

4.2.2 Preconditioned inexact inverse iteration with a fixed shift

In this subsection we consider the application of a preconditioner in the solution of the linear system in step (2) of Algorithm 6.

Let $\mathbf{A}$ in the standard eigenvalue problem (4.1) be Hermitian positive definite and consider the incomplete Cholesky factorisation $\mathbf{L}\mathbf{L}^H$, that is,

$$\mathbf{A} = \mathbf{L}\mathbf{L}^H + \mathbf{E}, \tag{4.13}$$

where $\mathbf{E}$ is the Hermitian error matrix associated with the incomplete decomposition of $\mathbf{A}$. Then, instead of solving $(\mathbf{A}-\sigma\mathbf{I})\mathbf{y}^{(i)} = \mathbf{x}^{(i)}$ in step (2) of Algorithm 6 inexactly, we solve the Hermitian system

$$\mathbf{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbf{L}^{-H}\tilde{\mathbf{y}}^{(i)} = \mathbf{L}^{-1}\mathbf{x}^{(i)}, \quad \mathbf{y}^{(i)} = \mathbf{L}^{-H}\tilde{\mathbf{y}}^{(i)}, \tag{4.14}$$

to a tolerance $\tau^{(i)}\|\mathbf{L}\|^{-1}$ so that $\|\mathbf{x}^{(i)} - (\mathbf{A}-\sigma\mathbf{I})\mathbf{y}^{(i)}\| \leq \tau^{(i)}$. This does not change the linear outer rate of convergence of the inexact inverse iteration algorithm. However, the right hand side $\mathbf{L}^{-1}\mathbf{x}^{(i)}$ is no longer close to the eigenvector corresponding to the eigenvalue of $\mathbf{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbf{L}^{-H}$ closest to zero and this changes the inner iteration behaviour as the outer iteration proceeds as we now explain. Apply Theorem 4.2 with $\mathbf{B} = \mathbf{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbf{L}^{-H}$, $\mathbf{b} = \mathbf{L}^{-1}\mathbf{x}^{(i)}$, $\tau = \dfrac{\tau^{(i)}}{\|\mathbf{L}\|}$ and with $\kappa^1_{\mathbf{L}}$ denoting the corresponding reduced condition number of $\mathbf{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbf{L}^{-H}$, to obtain the following bound on $k^{(i)}$:

$$k_{\mathbf{L}}^{(i)} \geq 1 + \frac{\sqrt{\kappa^1_{\mathbf{L}}}}{2}\left(\log 2\frac{|\mu_1-\mu_n|}{|\mu_1|} + \log\frac{\|\mathcal{P}^{\perp}\mathbf{L}^{-1}\mathbf{x}^{(i)}\|\|\mathbf{L}\|}{\tau^{(i)}}\right) \tag{4.15}$$

The key point to note is that there is no reason for $\|\mathcal{P}^{\perp}\mathbf{L}^{-1}\mathbf{x}^{(i)}\|$ to behave like $\sin\theta^{(i)}$ as is the case when there is no preconditioning. So, using $\|\mathcal{P}^{\perp}\mathbf{L}^{-1}\mathbf{x}^{(i)}\| \leq \|\mathbf{L}^{-1}\|$, (4.15) provides

$$k_{\mathbf{L}}^{(i)} \geq 1 + \frac{\sqrt{\kappa_{\mathbf{L}}^{1}}}{2}\left(\log 2\frac{|\mu_1-\mu_n|\|\mathbf{L}\|\|\mathbf{L}^{-1}\|}{|\mu_1|} + \log\frac{1}{\tau^{(i)}}\right), \tag{4.16}$$

and the right hand side increases with i for a decreasing $\tau^{(i)}$. This indicates that there will be growth in the number of inner iterations used by MINRES to solve (4.14). This is indeed observed in practice as is seen in Figure 4-1 (solid line with circles).

In order to recover the reassuring property of a constant number of inner iterations for preconditioned MINRES, a different approach has to be chosen. Simoncini and Eldén [119] alter the right hand side in (4.14), but for outer convergence this strategy requires that the shift tends to the desired eigenvalue as is the case for Rayleigh quotient iteration.

In this chapter we try the alternative approach of changing the preconditioner to recover the nice property of a constant number of inner iterations at each outer step. This idea is explained in the next section.

Remark 4.4. *In this chapter we shall assume that a good preconditioner for $\mathbf{A}$ is also a good preconditioner for $\mathbf{A}-\sigma\mathbf{I}$. This is the approach taken in [119] and it is likely to be the case if $\mathbf{A}$ arises from a discretised partial differential equation where a tailor-made preconditioner for $\mathbf{A}$ may be available.*

4.3 The tuned preconditioner

In this section we introduce a new preconditioner to be applied to $(\mathbf{A}-\sigma\mathbf{I})\mathbf{y}^{(i)} = \mathbf{x}^{(i)}$, so that the linear outer convergence is retained, but which provides the advantage of cheap inner solves. This approach is motivated by the tuned preconditioner that was introduced in Chapter 2 (see also [43]) for the nonsymmetric generalised eigenproblem but needs a more careful treatment to retain the Hermitian structure. Additionally, in this section and in Section 4.4 we are able to provide theoretical results for the tuned preconditioner that are not available in the nonsymmetric case discussed in Chapter 2.

4.3.1 An ideal preconditioner

In this subsection we discuss a rather hypothetical case. Assume we know the sought eigenvector $\mathbf{x}_1$ and that instead of solving (4.14) in step (2) of Algorithm 6 we solve the preconditioned Hermitian system

$$\mathbb{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}^{-H}\tilde{\mathbf{y}}_1 = \mathbb{L}^{-1}\mathbf{x}_1, \quad \mathbf{y}_1 = \mathbb{L}^{-H}\tilde{\mathbf{y}}_1, \tag{4.17}$$

where $\mathbb{L}$ is chosen such that the right hand side of (4.17) is an eigenvector of $\mathbb{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}^{-H}$ corresponding to the eigenvalue closest to zero. We shall see below that this is achieved if we ask that the preconditioner $\mathbb{L}\mathbb{L}^H$ should satisfy

$$\mathbb{L}\mathbb{L}^H\mathbf{x}_1 = \mathbf{A}\mathbf{x}_1, \tag{4.18}$$

and so $\mathbf{x}_1$ is an eigenvector of both $\mathbf{A}$ and $\mathbb{L}\mathbb{L}^H$. Hence, in addition to $\mathbb{L}\mathbb{L}^H$ being close to $\mathbf{A}$ as is usual in preconditioning we require that $\mathbb{L}\mathbb{L}^H$ acts exactly like $\mathbf{A}$

in the direction of $\mathbf{x}_1$. From (4.18) it is easy to see that $\mathbb{L}^{-1}\mathbf{A}\mathbb{L}^{-H}\mathbb{L}^{H}\mathbf{x}_1 = \mathbb{L}^{H}\mathbf{x}_1$, that is, $\mathbb{L}^{H}\mathbf{x}_1$ is an eigenvector of $\mathbb{L}^{-1}\mathbf{A}\mathbb{L}^{-H}$ corresponding to the eigenvalue 1. Also $\mathbb{L}^{H}\mathbf{x}_1 = \lambda_1\mathbb{L}^{-1}\mathbf{x}_1$, and so $\mathbb{L}^{-1}\mathbf{x}_1$ is an eigenvector of $\mathbb{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}^{-H}$ corresponding to the eigenvalue $(\lambda_1-\sigma)/\lambda_1$, which justifies the assertion made after (4.17).

We now have the following lemma, that tells us about the existence and construction of the ideal preconditioner $\mathbb{P} = \mathbb{L}\mathbb{L}^H$ and its (theoretical) impact on the solution of (4.17).

Lemma 4.5. *Let $\mathbf{P} = \mathbf{L}\mathbf{L}^H$ be the positive definite preconditioner given by (4.13) with $\mathbf{E}$ dropped and assume it has the eigendecomposition $\mathbf{V}\mathbf{P}\mathbf{V}^H = \mathbf{D} = diag(\eta_1,\ldots,\eta_n)$, where $0 < \eta_1 \leq \ldots \leq \eta_n$. Introduce $\mathbf{u}_1 := (\mathbf{A}-\mathbf{P})\mathbf{x}_1 = \mathbf{E}\mathbf{x}_1$. For $\mathbf{x}_1^H\mathbf{u}_1 \neq 0$ define*

$$\mathbb{P} = \mathbf{P} + \frac{\mathbf{u}_1\mathbf{u}_1^H}{\mathbf{x}_1^H\mathbf{u}_1}. \tag{4.19}$$

Then

(1) $\mathbf{x}_1^H\mathbf{u}_1 \in \mathbb{R}$.

(2) $\mathbb{P}\mathbf{x}_1 = \mathbf{A}\mathbf{x}_1$.

(3) If

(a) $\mathbf{x}_1^H\mathbf{u}_1 > 0$

(b) or

$$\mathbf{x}_1^H\mathbf{u}_1 < 0 \quad \textit{and} \quad \mathbf{x}_1^H\mathbf{u}_1 < -\frac{|(\mathbf{V}\mathbf{u}_1)_1|^2}{\eta_1} \tag{4.20}$$

where $(\mathbf{V}\mathbf{u}_1)_1$ is the first entry of $\mathbf{V}\mathbf{u}_1$,

then $\mathbb{P}$ is positive definite.

Now assume (4.20) holds and that

$$\mathbb{P} = \mathbb{L}\mathbb{L}^H \tag{4.21}$$

is the Cholesky factorisation of $\mathbb{P}$. Then

(4) $\mathbb{L}^H\mathbf{x}_1 = \lambda_1\mathbb{L}^{-1}\mathbf{x}_1$.

(5) $\left(\frac{\lambda_1-\sigma}{\lambda_1}, \mathbb{L}^{-1}\mathbf{x}_1\right)$ *is an eigenpair of* $\mathbb{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}^{-H}$.

(6) With starting guess set to zero, MINRES solves (4.17) in exactly one step.

Proof. (1) $\mathbf{x}_1^H\mathbf{u}_1 = \mathbf{x}_1^H(\mathbf{A}-\mathbf{P})\mathbf{x}_1$ is real since both $\mathbf{A}$ and $\mathbf{P}$ are Hermitian matrices.

(2) $\mathbb{P}\mathbf{x}_1 = \mathbf{P}\mathbf{x}_1 + \mathbf{u}_1 = \mathbf{A}\mathbf{x}_1$.

(3) (a) Obvious.

(b) Standard rank-one perturbation theory (see [48, Theorem 8.5.3] for the symmetric eigenproblem and [2] and [153] for extension to the Hermitian problem) shows that the eigenvalues of $\mathbb{P} = \mathbf{P} + \frac{\mathbf{u}_1\mathbf{u}_1^H}{\mathbf{x}_1^H\mathbf{u}_1} = \mathbf{V}^H(\mathbf{D} + \frac{\mathbf{V}\mathbf{u}_1\mathbf{u}_1^H\mathbf{V}^H}{\mathbf{x}_1^H\mathbf{u}_1})\mathbf{V}$ are given by the zeros of $h(\lambda) = 1 + \frac{1}{\mathbf{x}_1^H\mathbf{u}_1}\left(\sum_{i=1}^{n}\frac{|(\mathbf{V}\mathbf{u}_1)_i|^2}{\eta_i - \lambda}\right)$. With $\mathbf{x}_1^H\mathbf{u}_1 < 0$, the smallest zero of $h(\lambda)$ is less than η_1. Now the root of $1 + \frac{1}{\mathbf{x}_1^H\mathbf{u}_1}\left(\frac{|(\mathbf{V}\mathbf{u}_1)_1|^2}{\eta_1 - \lambda}\right)$ provides a lower bound for the smallest eigenvalue of $\mathbb{P}$, which, with $\mathbf{x}_1^H\mathbf{u}_1 < 0$ and the requirement $\lambda > 0$, gives the result.

(4) Follows from $\mathbb{L}\mathbb{L}^H\mathbf{x}_1 = \mathbf{A}\mathbf{x}_1 = \lambda_1\mathbf{x}_1$.

(5) $\mathbb{L}^{-1}(\mathbf{A} - \sigma\mathbf{I})\mathbb{L}^{-H}(\mathbb{L}^{-1}\mathbf{x}_1) = \frac{1}{\lambda_1}\mathbb{L}^{-1}(\mathbf{A} - \sigma\mathbf{I})\mathbb{L}^{-H}\mathbb{L}^H\mathbf{x}_1 = \frac{\lambda_1 - \sigma}{\lambda_1}(\mathbb{L}^{-1}\mathbf{x}_1)$.

(6) Using a Krylov subspace method the solution of $\mathbb{L}^{-1}(\mathbf{A} - \sigma\mathbf{I})\mathbb{L}^{-H}\tilde{\mathbf{y}}_1 = \mathbb{L}^{-1}\mathbf{x}_1$ is contained in the subspace given by

$$
\begin{aligned}
\mathcal{K}(\mathbb{L}^{-1}(\mathbf{A} - \sigma\mathbf{I})\mathbb{L}^{-H}, \mathbb{L}^{-1}\mathbf{x}_1) &= \mathrm{span}\{\mathbb{L}^{-1}\mathbf{x}_1, \mathbb{L}^{-1}(\mathbf{A} - \sigma\mathbf{I})\mathbb{L}^{-H}\mathbb{L}^{-1}\mathbf{x}_1\} \\
&= \mathrm{span}\{\mathbb{L}^{-1}\mathbf{x}_1, \frac{1}{\lambda_1}\mathbb{L}^{-1}(\mathbf{A} - \sigma\mathbf{I})\mathbf{x}_1\} \\
&= \mathrm{span}\{\mathbb{L}^{-1}\mathbf{x}_1, \frac{\lambda_1 - \sigma}{\lambda_1}\mathbb{L}^{-1}\mathbf{x}_1\},
\end{aligned}
$$

where we have used $\mathbf{A}\mathbf{x}_1 = \mathbb{L}\mathbb{L}^H\mathbf{x}_1 = \lambda_1\mathbf{x}_1$. Hence stagnation of MINRES occurs after one step ("lucky breakdown") and the solution of (4.17), which lies in $\mathrm{span}\{\mathbb{L}^{-1}\mathbf{x}_1\}$ is found after one step of MINRES.

□

Note that the roots of the polynomial $h(\lambda) = 1 + \frac{1}{\mathbf{x}_1^H\mathbf{u}_1}\left(\sum_{i=1}^{n}\frac{|(\mathbf{V}\mathbf{u}_1)_i|^2}{\eta_i - \lambda}\right)$ are given by $h(\lambda) = 0$, which is equal to

$$
\prod_{i=1}^{n}(\eta_i - \lambda) + \frac{1}{\mathbf{x}_1^H\mathbf{u}_1}\left(\sum_{i=1}^{n}|(\mathbf{V}\mathbf{u}_1)_i|^2\prod_{\substack{j=1\\ j\neq i}}^{n}(\eta_j - \lambda)\right) = 0,
$$

and hence for $\mathbf{u}_1 = \mathbf{0}$ we obtain $\lambda = \eta_i$, $i = 1, \ldots, n$.

Remark 4.6.

(a) *Given* $\mathbf{P}$, $\mathbb{P}$ *is the "ideal" preconditioner for MINRES, since MINRES (with this particular preconditioner) converges in one step. A similar "ideal" preconditioner is employed in [104] to analyse subspace iteration for the nonsymmetric eigenvalue problem.*

(b) *Lemma 4.5 shows that there is a range* $-\frac{|(\mathbf{V}\mathbf{u}_1)_1|^2}{\eta_1} \leq \mathbf{x}_1^H\mathbf{u}_1 \leq 0$ *where the ideal preconditioner is not positive definite. The lower bound of this range will be large, if*

η_1 is small, that is, if $\mathbf{P}$ is a poor preconditioner. Also if $\mathbf{P}\mathbf{x}_1$ is close to $\mathbf{A}\mathbf{x}_1$, say, for example, if $\mathbf{E}$ in (4.13) were small, then $\mathbf{x}_1^H\mathbf{u}_1$ would be close to zero. However, in this case, there would be no need for tuning. For the practical tuned preconditioner discussed in the next subsection the conditions corresponding to (4.20) are investigated in the examples in Section 4.3.3, and indeed, are shown to hold in all cases considered.

Of course, in practice the preconditioner (4.19) cannot be used since $\mathbf{x}_1$ is not available. However, its form suggests a practical tuned preconditioner.

4.3.2 The practical tuned preconditioner

At the ith step in Algorithm 6 define

$$\mathbf{u}^{(i)} = (\mathbf{A} - \mathbf{P})\mathbf{x}^{(i)}. \tag{4.22}$$

Assuming that ${\mathbf{x}^{(i)}}^H\mathbf{u}^{(i)} \neq 0$ the practical tuned preconditioner is obtained by replacing $\mathbb{P}$ in (4.19) by $\mathbb{P}_i$ given by

$$\mathbb{P}_i = \mathbf{P} + \frac{\mathbf{u}^{(i)}{\mathbf{u}^{(i)}}^H}{{\mathbf{x}^{(i)}}^H\mathbf{u}^{(i)}}, \tag{4.23}$$

where the unknown $\mathbf{x}_1$ is replaced by its approximation $\mathbf{x}^{(i)}$. Note that ${\mathbf{x}^{(i)}}^H\mathbf{u}^{(i)} \in \mathbb{R}$, since $\mathbf{A} - \mathbf{P}$ is Hermitian. Clearly $\mathbb{P}_i$ tends to $\mathbb{P}$ as $\mathbf{x}^{(i)} \to \mathbf{x}_1$. Assume also

$${\mathbf{x}^{(i)}}^H\mathbf{u}^{(i)} < -\frac{|(\mathbf{V}\mathbf{u}^{(i)})_1|^2}{\eta_1} \quad (\text{if} \quad {\mathbf{x}^{(i)}}^H\mathbf{u}^{(i)} < 0) \quad \text{or} \quad {\mathbf{x}^{(i)}}^H\mathbf{u}^{(i)} > 0, \tag{4.24}$$

where $\mathbf{V}$ and η_1 are defined as in Lemma 4.5. Then $\mathbb{P}_i$ is positive definite. If we can prove similar results to Lemma 4.5, we can expect to obtain a significant benefit in the iterative solution of

$$\mathbb{L}_i^{-1}(\mathbf{A} - \sigma\mathbf{I})\mathbb{L}_i^{-H}\tilde{\mathbf{y}}^{(i)} = \mathbb{L}_i^{-1}\mathbf{x}^{(i)}, \quad \mathbf{y}^{(i)} = \mathbb{L}_i^{-H}\tilde{\mathbf{y}}^{(i)}, \tag{4.25}$$

where $\mathbb{P}_i = \mathbb{L}_i\mathbb{L}_i^H$ is the Cholesky decomposition of $\mathbb{P}_i$. First note that the tuned preconditioner $\mathbb{P}_i$ satisfies the tuning condition

$$\mathbb{P}_i\mathbf{x}^{(i)} = \mathbf{A}\mathbf{x}^{(i)}, \tag{4.26}$$

and we will use this condition several times. We now state a Lemma about $\mathbb{P}_i$.

Lemma 4.7. *Let $\mathbb{P}$ be given by (4.19) and $\mathbb{P}_i$ be given by (4.23). Further let $\mathbf{u}_1$ be given as in Lemma 4.5 and $\mathbf{u}^{(i)}$ as in (4.22). Assume (4.24) holds and let*

$$\mathbf{R}^{(i)} = \frac{\mathbf{x}^{(i)}{\mathbf{x}^{(i)}}^H}{{\mathbf{x}^{(i)}}^H\mathbf{u}^{(i)}} - \frac{\mathbf{x}_1\mathbf{x}_1^H}{{\mathbf{x}_1}^H\mathbf{u}_1}.$$

Then

$$\|\mathbf{R}^{(i)}\| \leq C_1|\tan\theta^{(i)}|, \tag{4.27}$$

where $\theta^{(i)}$ is given in (4.4) and C_1 is independent of i for large enough i. Furthermore

$$\mathbb{P}_i - \mathbb{P} = \mathbf{\Delta}_i, \quad \textit{with} \quad \|\mathbf{\Delta}_i\| \leq C_2|\tan\theta^{(i)}|, \tag{4.28}$$

where C_2 is independent of i for large enough i. If $\mathbb{P}^{-1}$ exists then, for

$$C_2|\tan\theta^{(i)}| < \|\mathbb{P}^{-1}\|^{-1}, \tag{4.29}$$

$\mathbb{P}_i^{-1}$ exists and

$$\|\mathbb{P}_i^{-1}\| \leq \frac{\|\mathbb{P}^{-1}\|}{1 - C_2|\tan\theta^{(i)}|\|\mathbb{P}^{-1}\|}. \tag{4.30}$$

This term can be bounded independently of i for large enough i.

Proof. Write $\mathbf{x}^{(i)}$ as (4.4), then a straightforward but lengthy calculation gives (4.27). Clearly, we have $\mathbb{P}_i - \mathbb{P} = (\mathbf{A}-\mathbf{P})\mathbf{R}^{(i)}(\mathbf{A}-\mathbf{P})$ and (4.28) is readily obtained. Further, we can bound

$$\|\mathbb{P}_i^{-1}\| = \|(\mathbb{P}+\mathbf{\Delta}_i)^{-1}\| \leq \|(\mathbf{I}+\mathbb{P}^{-1}\mathbf{\Delta}_i)\|\|\mathbb{P}^{-1}\|,$$

and (4.29) gives (4.30). □

Next, we have a Lemma that provides bounds on $\|\mathbb{L}_i\|$ and $\|\mathbb{L}_i^{-1}\|$.

Lemma 4.8. *Let $\mathbb{P} = \mathbb{L}\mathbb{L}^H$ and $\mathbb{P}_i = \mathbb{L}_i\mathbb{L}_i^H$ be the Cholesky factorisations of $\mathbb{P}$ and $\mathbb{P}_i$ and assume (4.28) and (4.29) hold. Then*

$$\|\mathbb{L}_i\| \leq C_3 \quad \textit{and} \quad C_4 \leq \|\mathbb{L}_i^{-1}\| \leq C_5, \tag{4.31}$$

where C_3, C_4 and C_5 are independent of i for large enough i.

Proof. First note that

$$\mathbb{P}_i = \mathbb{P} + \mathbf{\Delta}_i = \mathbb{L}(\mathbf{I}+\mathbf{D}^{(i)})\mathbb{L}^H,$$

where

$$\mathbf{D}^{(i)} = \mathbb{L}^{-1}\mathbf{\Delta}_i\mathbb{L}^{-H}. \tag{4.32}$$

For large enough i, $\mathbf{I}+\mathbf{D}^{(i)}$ is Hermitian positive definite, and $\|\mathbf{D}^{(i)}\| \leq C_6|\tan\theta^{(i)}| \leq C_7$. [28] show that the Cholesky factorisations $\mathbf{I}+\mathbf{D}^{(i)} = (\mathbf{I}+\mathbf{F}^{(i)})(\mathbf{I}+\mathbf{F}^{(i)})^H$ exist with $\|\mathbf{F}^{(i)}\| \leq C_8\|\mathbf{D}^{(i)}\|$ where C_8 depends on the matrix dimension but is independent of i. Hence, we may write the Cholesky factor of $\mathbb{P}_i$ as $\mathbb{L}_i = \mathbb{L}(\mathbf{I}+\mathbf{F}^{(i)})$ with

$$\|\mathbb{L}_i\| \leq \|\mathbb{L}\|(\|\mathbf{I}+\|\mathbf{F}^{(i)}\|) \leq \|\mathbb{L}\|(1+C_8|\tan\theta^{(i)}|) \leq C_3,$$

for some constant C_3 independent of i. For the upper bound on $\|\mathbb{L}_i^{-1}\|$ observe that $\mathbb{L}_i^{-1} = (\mathbf{I}+\mathbf{F}^{(i)})^{-1}\mathbb{L}^{-1}$ and so

$$\|\mathbb{L}_i^{-1}\| \leq \frac{1}{1-\|\mathbf{F}^{(i)}\|}\|\mathbb{L}^{-1}\| \leq \frac{1}{1-C_8\|\mathbf{D}^{(i)}\|}\|\mathbb{L}^{-1}\| \leq C_5$$

since $\|\mathbf{D}\|^{(i)} \leq C_7$ for large enough i. For the lower bound use $(\mathbf{I}+\mathbf{F}^{(i)})\mathbb{L}_i^{-1} = \mathbb{L}^{-1}$ and hence

$$\|\mathbb{L}^{-1}\| \leq \|(\mathbf{I}+\mathbf{F}^{(i)})\|\|\mathbb{L}_i^{-1}\| \leq (1+\|\mathbf{F}^{(i)}\|)\|\mathbb{L}_i^{-1}\|.$$

Reordering gives

$$\|\mathbb{L}_i^{-1}\| \geq \frac{\|\mathbb{L}^{-1}\|}{1+\|\mathbf{F}^{(i)})\|} \geq \frac{\|\mathbb{L}^{-1}\|}{1+C_7C_8} \geq C_4$$

for large enough i from which the stated result holds. □

The following proposition shows that the eigenvalues of $\mathbb{L}_i^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}_i^{-H}$ and $\mathbb{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}^{-H}$ are close to each other, where $\mathbb{L}_i\mathbb{L}_i^H$ is the Cholesky factorisation of $\mathbb{P}_i$.

Proposition 4.9. *Let $\mathbb{L}_i^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}_i^{-H}$ have eigenvalues $\hat{\xi}_j^{(i)}$ and corresponding eigenvectors $\hat{\mathbf{w}}_j^{(i)}$, and $\mathbb{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}^{-H}$ have eigenvalues $\hat{\xi}_j$ with eigenvectors $\hat{\mathbf{w}}_j$. Assume σ is not an eigenvalue of $\mathbf{A}$. Then, for each j, $\hat{\xi}_j \neq 0$ and*

$$\frac{|\hat{\xi}_j^{(i)}-\hat{\xi}_j|}{|\hat{\xi}_j|} \leq \|\mathbf{D}^{(i)}\| \leq C_6|\tan\theta^{(i)}|,$$

with $\|\mathbf{D}^{(i)}\|$ given by (4.32) and C_6 independent of i.

Proof. With $\mathbb{L}_i = \mathbb{L}(\mathbf{I}+\mathbf{F}^{(i)})$ we have that

$$\mathbb{L}_i^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}_i^{-H}\hat{\mathbf{w}}_j^{(i)} = \hat{\xi}_j^{(i)}\hat{\mathbf{w}}_j^{(i)}$$

may be written as

$$\mathbb{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}^{-H}\tilde{\mathbf{z}}_j^{(i)} = \hat{\xi}_j^{(i)}(\mathbf{I}+\mathbf{D}^{(i)})\tilde{\mathbf{z}}_j^{(i)}, \quad \tilde{\mathbf{z}}_j^{(i)} = (\mathbf{I}+\mathbf{F}^{(i)})^{-H}\hat{\mathbf{w}}_j^{(i)},$$

where $\|\mathbf{D}^{(i)}\| \leq C_6|\tan\theta^{(i)}|$. This eigenvalue problem is a perturbation of

$$\mathbb{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}^{-H}\mathbf{z}_j = \hat{\xi}_j\mathbf{z}_j,$$

and an analysis similar to the proof of Theorem 4.14 provides the stated results. □

The following Theorem shows that if (4.26) holds then the right hand side $\mathbb{L}_i^{-1}\mathbf{x}^{(i)}$ in (4.25) is an approximation to the eigenvector of the iteration matrix $\mathbb{L}_i^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}_i^{-H}$ corresponding to the eigenvalue nearest zero. The idea is to show first that $\left(\dfrac{\lambda^{(i)}-\sigma}{\lambda^{(i)}}, \mathbb{L}_i^H\mathbf{x}^{(i)}\right)$ is an approximate eigenpair of $\mathbb{L}_i^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}_i^{-H}$ (cf. (4) in Lemma (4.5)) and then use the fact that $\mathbb{L}_i^H\mathbf{x}^{(i)}$ is approximately in the direction of $\mathbb{L}_i^{-1}\mathbf{x}^{(i)}$. This follows since (4.5) and (4.26) give $\mathbb{L}_i\mathbb{L}_i^H\mathbf{x}^{(i)} = \lambda^{(i)}\mathbf{x}^{(i)} + \mathbf{r}^{(i)}$ and hence

$$\mathbb{L}_i^H\mathbf{x}^{(i)} - \lambda^{(i)}\mathbb{L}_i^{-1}\mathbf{x}^{(i)} = \mathbb{L}_i^{-1}\mathbf{r}^{(i)},$$

with $\|\mathbf{r}^{(i)}\| \leq C_9|\tan\theta^{(i)}|$ for some constant C_9 using (4.6).

Theorem 4.10. *Let $\mathbb{L}_i^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}_i^{-H}$ have eigenvalues $\hat{\xi}_j^{(i)}$ and corresponding eigenvectors $\hat{\mathbf{w}}_j^{(i)}$, with $\hat{\xi}_1^{(i)}$ the eigenvalue nearest zero. Let $\mathcal{P}_i^{\perp}$ denote the orthogonal projection onto span$\{\hat{\mathbf{w}}_2^{(i)}, \ldots, \hat{\mathbf{w}}_n^{(i)}\}$. Assume (4.26) holds, let $\mathbf{r}^{(i)}$ be defined by (4.5) and assume $\lambda^{(i)} \neq 0$. Then, for small enough $\|\mathbf{r}^{(i)}\|$ we have*

$$\|\mathbb{L}_i^{-1}\mathbf{x}^{(i)} - c_3^{(i)}\hat{\mathbf{w}}_1^{(i)}\| \leq C_{10}\|\mathbf{r}^{(i)}\| \tag{4.33}$$

and

$$\|\mathcal{P}_i^{\perp}\mathbb{L}_i^{-1}\mathbf{x}^{(i)}\| \leq C_{10}\|\mathbf{r}^{(i)}\| \tag{4.34}$$

for some C_{10} independent of i for large enough i.

Proof. Using (4.5) and (4.26) with $\mathbb{P}_i = \mathbb{L}_i\mathbb{L}_i^H$ we have

$$\begin{aligned}
\mathbb{L}_i^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}_i^{-H}(\mathbb{L}_i^{-1}\mathbf{x}^{(i)}) &= \mathbb{L}_i^{-1}(\mathbf{A}-\sigma\mathbf{I})(\mathbb{L}_i\mathbb{L}_i^H)^{-1}\mathbf{x}^{(i)} \\
&= \frac{1}{\lambda^{(i)}}\mathbb{L}_i^{-1}(\mathbf{A}-\sigma\mathbf{I})(\mathbf{x}^{(i)}-(\mathbb{L}_i\mathbb{L}_i^H)^{-1}\mathbf{r}^{(i)}) \\
&= \frac{1}{\lambda^{(i)}}\mathbb{L}_i^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbf{x}^{(i)} - \frac{1}{\lambda^{(i)}}\mathbb{L}_i^{-1}(\mathbf{A}-\sigma\mathbf{I})(\mathbb{L}_i\mathbb{L}_i^H)^{-1}\mathbf{r}^{(i)} \\
&= \frac{\lambda^{(i)}-\sigma}{\lambda^{(i)}}(\mathbb{L}_i^{-1}\mathbf{x}^{(i)}) + \frac{1}{\lambda^{(i)}}(\mathbf{I}-\mathbb{L}_i^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}_i^{-H})\mathbb{L}_i^{-1}\mathbf{r}^{(i)},
\end{aligned}$$

that is $\mathbb{L}_i^{-1}\mathbf{x}^{(i)}$ is an approximate eigenvector of $\mathbb{L}_i^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}_i^{-H}$ with approximate eigenvalue $\dfrac{\lambda^{(i)}-\sigma}{\lambda^{(i)}}$ for small enough $\|\mathbf{r}^{(i)}\|$. To quantify the size of the perturbation we can rewrite

$$\mathbb{L}_i^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}_i^{-H}\left(\mathbb{L}_i^{-1}\mathbf{x}^{(i)}+\frac{\mathbb{L}_i^{-1}\mathbf{r}^{(i)}}{\lambda^{(i)}}\right) = \frac{\lambda^{(i)}-\sigma}{\lambda^{(i)}}\left(\mathbb{L}_i^{-1}\mathbf{x}^{(i)}+\frac{\mathbb{L}_i^{-1}\mathbf{r}^{(i)}}{\lambda^{(i)}}\right)+\frac{\sigma}{{\lambda^{(i)}}^2}\mathbb{L}_i^{-1}\mathbf{r}^{(i)}.$$

Equivalently we have

$$(\mathbb{L}_i^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}_i^{-H})\mathbb{L}_i^H\mathbf{x}^{(i)} = \frac{\lambda^{(i)}-\sigma}{\lambda^{(i)}}\mathbb{L}_i^H\mathbf{x}^{(i)} + \frac{\sigma}{\lambda^{(i)}}\mathbb{L}_i^{-1}\mathbf{r}^{(i)},$$

and that $\dfrac{\lambda^{(i)}-\sigma}{\lambda^{(i)}}$ is the Rayleigh quotient of $\mathbb{L}_i^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}_i^{-H}$ with respect to the vector $\mathbb{L}_i^H\mathbf{x}^{(i)}$. Then standard perturbation theory for simple eigenvalues of symmetric matrices (see [101, Chapter 11] for symmetric matrices or [137, page 250] for Hermitian matrices) shows that $\mathbb{L}_i^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}_i^{-H}$ has a simple eigenvalue $\hat{\xi}_1^{(i)}$ near $\dfrac{\lambda^{(i)}-\sigma}{\lambda^{(i)}}$ with corresponding eigenvector $\hat{\mathbf{w}}_1^{(i)}$ near $\mathbb{L}_i^H\mathbf{x}^{(i)}$. We obtain

$$\begin{aligned}
\sin\angle(\hat{\mathbf{w}}_1^{(i)},\mathbb{L}_i^H\mathbf{x}^{(i)}) &\leq \frac{1}{\delta^{(i)}}\frac{|\sigma|}{\lambda^{(i)}}\|\mathbb{L}_i^{-1}\mathbf{r}^{(i)}\| \quad \text{where} \quad \delta^{(i)} = \min_{j=2,\dots,n}\left|\hat{\xi}_j^{(i)} - \frac{\lambda^{(i)}-\sigma}{\lambda^{(i)}}\right| \\
&\leq \frac{1}{\delta}\frac{|\sigma|}{\lambda_1}\|\mathbb{L}_i^{-1}\mathbf{r}^{(i)}\|,
\end{aligned}$$

where in the last bound we have used $\lambda^{(i)} > \lambda_1$ due to the properties of the Rayleigh quotient, and the results of Proposition 4.9 and $|\lambda^{(i)}-\lambda_1| = c|\sin(\theta^{(i)})^2|$ for some constant c which yield

$$\delta^{(i)} = \min_{j=2,\dots,n}\left|\hat{\xi}_j^{(i)} - \frac{\lambda^{(i)}-\sigma}{\lambda^{(i)}}\right| \geq \min_{j=2,\dots,n}\left|\hat{\xi}_j - \frac{\lambda_1-\sigma}{\lambda_1}\right| - C_6\hat{\xi}_j|\tan\theta^{(i)}| > C_{11}$$

where C_{11} is a constant independent of i for large enough i. After normalising the vectors we obtain

$$\|c^{(i)}\hat{\mathbf{w}}_1^{(i)} - \frac{\mathbb{L}_i^H\mathbf{x}^{(i)}}{\|\mathbb{L}_i^H\mathbf{x}^{(i)}\|}\| \leq \frac{1}{\delta}\frac{|\sigma|}{\lambda_1}\|\mathbb{L}_i^{-1}\mathbf{r}^{(i)}\|,$$

where $c^{(i)} := \cos\angle(\hat{\mathbf{w}}_1^{(i)}, \mathbb{L}_i^H\mathbf{x}^{(i)})$. Multiplying by $\frac{\|\mathbb{L}_i^H\mathbf{x}^{(i)}\|}{\lambda^{(i)}}$ and using $\frac{\mathbb{L}_i^H\mathbf{x}^{(i)}}{\lambda^{(i)}} = \mathbb{L}_i^{-1}\mathbf{x}^{(i)} + \frac{\mathbb{L}_i^{-1}\mathbf{r}^{(i)}}{\lambda^{(i)}}$ we get

$$\|\tilde{c}^{(i)}\hat{\mathbf{w}}_1^{(i)} - \mathbb{L}_i^{-1}\mathbf{x}^{(i)} - \frac{\mathbb{L}_i^{-1}\mathbf{r}^{(i)}}{\lambda^{(i)}}\| \leq \frac{\|\mathbb{L}_i^H\mathbf{x}^{(i)}\|}{\delta}\frac{|\sigma|}{\lambda^{(i)}\lambda_1}\|\mathbb{L}_i^{-1}\mathbf{r}^{(i)}\|,$$

where $\tilde{c}^{(i)}$ is chosen appropriately. Hence

$$\|c_3^{(i)}\hat{\mathbf{w}}_1^{(i)} - \mathbb{L}_i^{-1}\mathbf{x}^{(i)}\| \leq \left(\frac{\|\mathbb{L}_i^H\mathbf{x}^{(i)}\|}{\delta}\frac{|\sigma|}{\lambda^{(i)}\lambda_1} + \frac{1}{\lambda^{(i)}}\right)\|\mathbb{L}_i^{-1}\mathbf{r}^{(i)}\|. \tag{4.35}$$

With $\lambda^{(i)} > \lambda_1$ and (4.31) we obtain (4.33), since all the terms in the brackets of (4.35) can be bounded independently of i for large enough i. Finally, we have

$$\|\mathcal{P}_i^{\perp}\mathbb{L}_i^{-1}\mathbf{x}^{(i)}\| = \|\mathcal{P}_i^{\perp}(\mathbb{L}_i^{-1}\mathbf{x}^{(i)} - c_3^{(i)}\hat{\mathbf{w}}_1^{(i)})\| \leq C_{11}\|\mathbf{r}^{(i)}\|$$

since $\mathcal{P}_i^{\perp}\hat{\mathbf{w}}_1^{(i)} = 0$ and $\|\mathcal{P}_i^{\perp}\| = 1$. □

For our purposes, the important result in Lemma 4.10 is (4.34), which with (4.6) implies that $\|\mathcal{P}_i^{\perp}\mathbb{L}_i^{-1}\mathbf{x}^{(i)}\| = \mathcal{O}(|\sin\theta^{(i)}|)$. This is similar to the corresponding result in the unpreconditioned case given by (4.11) and is important when analysing the lower bound for the number of iterations needed by MINRES. We have the following consequence of Theorem 4.2 (compare with (4.12) for the unpreconditioned case).

Theorem 4.11. *Assume the conditions of Lemma 4.10 and that (4.24) and (4.26) hold. Consider the application of MINRES to the inexact solution of*

$$\mathbb{L}_i^{-1}(\mathbf{A} - \sigma\mathbf{I})\mathbb{L}_i^{-H}\tilde{\mathbf{y}}^{(i)} = \mathbb{L}_i^{-1}\mathbf{x}^{(i)}. \tag{4.36}$$

Assume $\mathbb{L}_i^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}_i^{-H}$ *satisfies the conditions on* $\mathbf{B}$ *in Theorem 4.2. Further assume that we seek the smallest eigenvalue of* $\mathbf{A}$ *with the shift* σ *being closer to* λ_1 *than to any other eigenvalue of* $\mathbf{A}$*, such that* $\{\hat{\xi}_j^{(i)}\}_{j=2}^n > 0$*. Denote the reduced condition number of* $\mathbb{L}_i^{-1}(\mathbf{A} - \sigma\mathbf{I})\mathbb{L}_i^{-H}$ *by* $\kappa_{\mathbb{L}_i}^1$*. Then the number of inner iterations needed by MINRES to solve (4.36) to a tolerance* $\tau^{(i)}\|\mathbb{L}_i\|^{-1}$*, where* $\tau^{(i)} = C_1\|\mathbf{r}^{(i)}\|$*, satisfies*

$$k^{(i)} \geq 1 + \frac{\sqrt{\kappa_{\mathbb{L}_i}^1}}{2}\left(\log 2\frac{|\hat{\xi}_1^{(i)} - \hat{\xi}_n^{(i)}|}{|\hat{\xi}_1^{(i)}|} + \log\frac{\|\mathcal{P}_i^{\perp}\mathbb{L}_i^{-1}\mathbf{x}^{(i)}\|\|\mathbb{L}_i\|}{\tau^{(i)}}\right) \tag{4.37}$$

and the right hand side of (4.37) can be bounded independently of i *for large enough* i*.*

Proof. The bound on the iteration number (4.37) follows from (4.9) applied to (4.36), with τ replaced by $\tau^{(i)}\|\mathbb{L}_i\|^{-1}$ and $\mathcal{P}^{\perp}\mathbf{b}$ replaced by $\mathcal{P}_i^{\perp}\mathbb{L}_i^{-1}\mathbf{x}^{(i)}$. The bound (4.34) and the first bound in (4.31) show that

$$\log\frac{\|\mathcal{P}_i^{\perp}\mathbb{L}_i^{-1}\mathbf{x}^{(i)}\|}{\tau^{(i)}\|\mathbb{L}_i\|^{-1}} \leq \log\frac{C_{11}\|\mathbf{r}^{(i)}\|\|\mathbb{L}_i\|}{\tau^{(i)}} \leq \log\frac{C_{11}C_3}{C_1}$$

is independent of i for large enough i. The first term in the brackets in (4.34) can be bounded using Proposition 4.9:

$$\frac{|\hat{\xi}_1^{(i)} - \hat{\xi}_n^{(i)}|}{|\hat{\xi}_1^{(i)}|} \leq \frac{|\hat{\xi}_1 - \hat{\xi}_n| + C_5(\hat{\xi}_1 + \hat{\xi}_n)|\tan\theta^{(i)}|}{|\hat{\xi}_1| - C_5\hat{\xi}_1|\tan\theta^{(i)}|}.$$

Since $\tan\theta^{(i)}$ is decreasing, the first term in (4.37) can be bounded independently of i for large enough i. Furthermore we have

$$\kappa_{\mathbb{L}_i^1} = \frac{\max_{j=2,\dots,n}|\hat{\xi}_j^{(i)}|}{\min_{j=2,\dots,n}|\hat{\xi}_j^{(i)}|} \leq \frac{\max_{j=2,\dots,n}|\hat{\xi}_j| + C_5|\tan\theta^{(i)}|}{\min_{j=2,\dots,n}|\hat{\xi}_j| - C_5|\tan\theta^{(i)}|}, \tag{4.38}$$

which can also be bounded independently of i for large enough i. □

Theorem 4.11 indicates that if we can find a positive definite preconditioner that satisfies (4.26) then we expect no growth in the inner iteration count for MINRES using the tuned preconditioner as the outer iteration proceeds. Numerical results confirming this effect are given in Figures 4-1and 4-13.

We shall return in Section 4.4 to the assumption about the eigenvalues of $\mathbb{L}_i^{-1}(\mathbf{A} - \sigma\mathbf{I})\mathbb{L}_i^{-H}$ satisfying the conditions on $\mathbf{B}$ in Theorem 4.2. In the rest of this section we illustrate the performance of the tuned preconditioner by two numerical examples.

Note that by applying the second case in Theorem 4.2 a modification of Theorem 4.11 also holds for interior eigenvalues though we do not give examples of this case here.

It is important to note that replacing $\mathbf{P}$ by $\mathbb{P}_i$ involves minimal extra computational work. Indeed for the implementation of $\mathbb{P}_i$ rather than $\mathbf{P}$ at each (i) (that is, at each outer iteration) only a single extra back substitution with $\mathbf{P}$ is needed for the tuned preconditioner $\mathbb{P}_i$. This is proved using the Sherman-Morrison formula (see [23, p. 95]) for the inverse of a matrix with a rank-one change. In particular for the inverse of $\mathbb{P}_i$ in (4.23) with (4.22) which is used in preconditioned MINRES we have

$$\mathbb{P}_i^{-1} = \mathbf{P}^{-1} - \frac{(\mathbf{P}^{-1}\mathbf{A}\mathbf{x}^{(i)} - \mathbf{x}^{(i)})(\mathbf{P}^{-1}\mathbf{A}\mathbf{x}^{(i)} - \mathbf{x}^{(i)})^H}{(\mathbf{P}^{-1}\mathbf{A}\mathbf{x}^{(i)} - \mathbf{x}^{(i)})^H\mathbf{A}\mathbf{x}^{(i)}}.$$

The single extra back substitution $\mathbf{P}^{-1}\mathbf{A}\mathbf{x}^{(i)}$ can be carried out before the inner iterative solve of the linear system and we are only left with some additional inner products.

4.3.3 Numerical examples

We now present two numerical examples to illustrate the theory in this section.

Example 4.12 (Problem from [13]). *Here we consider the matrix* `nos5.mtx` *from the Matrix market library [13]. It is a real symmetric positive definite matrix of size* 468×468 *with* 5172 *nonzero entries. Its first five eigenvalues are given by*

	1st	*2nd*	*3rd*	*4th*	*5th*	*6th*
eigenvalue	52.8995	67.5430	115.5912	131.5636	185.7169	229.0844.

We consider a fixed shift strategy and seek the smallest and the 5*th eigenvalue. We compare the costs of the following two different methods:*

(a) Standard incomplete Cholesky preconditioner: Algorithm 6 with step (2) implemented by solving (4.14), where $\mathbf{L}\mathbf{L}^H$ *is the incomplete Cholesky factorisation of* $\mathbf{A}$.

(b) Tuned incomplete Cholesky preconditioner: Algorithm 6 with step (2) implemented by solving (4.25), where $\mathbb{P}_i$ *is given by (4.23).*

For the inexact solves we use preconditioned MINRES with

$$\tau^{(i)} = \min\{\tau, \tau \,\|\mathbf{r}^{(i)}\|\}, \quad \tau = 0.1. \tag{4.39}$$

and for the incomplete Cholesky decomposition we use a drop tolerance of 0.1*. Then the number of nonzero entries in* $\mathbf{L}$ *is* 1032 *and* $\|\mathbf{E}\| \approx 7.7e+04$ *(with* $\|\mathbf{A}\| \approx 5.8e+05$*). We use a starting guess* $\mathbf{x}^{(0)}$ *of all ones and a fixed shift of* $\sigma = 58$ *if the smallest eigenvalue is sought and* $\sigma = 200$ *if the fifth eigenvalue is sought. The computations stop once the eigenvalue residual satisfies* $\|\mathbf{r}^{(i)}\| < 10^{-8}$.

Figure 4-1 shows the number of inner iterations used by methods (a) and (b), where we see the steady increase in inner iterations needed by the standard preconditioner in method (a), but essentially constant number of inner iterations needed to solve (4.25) using the tuned preconditioner as in method (b) . This supports the result of Theorem 4.11. Figure 4-2 plots the eigenvalue residual norms against the total number of iterations, which again shows the superiority of the tuned preconditioner in terms of the total number of iterations. In Figure 4-3 we plot the right hand sides of the lower bounds (4.16) and (4.37) respectively, which again agrees with the theory, though, as noted in Remark 4.3, these bounds should not be used quantitatively. Indeed, the bound for method (a) exceeds the trivial bound $k^{(i)} \geq n$ for i large enough. However, the bound for method (b) only overestimates the actual number of inner iterations by a factor of roughly 1.5. Figure 4-4 shows that both methods with standard and tuned preconditioner exhibit the same eigenvalue residuals at each outer iteration step. Figures 4-5 and 4-6 plot the relative MINRES residual where, for the standard preconditioner the linear system residuals for MINRES shows an initial plateau before slow convergence whereas for the tuned preconditioner rapid convergence can be observed immediately. The behaviour as the outer iteration proceeds can be read out from top to bottom of Figures 4-5 and 4-6. Note that there is almost complete stagnation after the first three iterations in Figure 4-6. If the tuned preconditioner is used within the MINRES solve, the residual decreases rapidly during the first three iterations, due to the right hand side of the linear system being a good approximate eigenvector of the system matrix and hence the bound (4.8) in Theorem 4.2 suggest fast convergence, since $\|\mathcal{P}^{\perp}\mathbf{b}\|$ is small for that special right hand side $\mathbf{b}$. This decrease in the linear system residual is even more rapid as the outer iteration proceeds due to an increasingly better approximation of the right hand side to an eigenvector of the system matrix. After the first few iterations the residual almost stagnates since the expansion of the Krylov subspace for MINRES does not yield any new information (in the limit, for large enough i we should have only one iteration, see Lemma 4.5, part (6)).

The stagnation after the first few iterations in Figure 4-6 suggests that the inner iteration could be stopped earlier (that is, a larger $\tau^{(i)}$ is sufficient if a tuned preconditioner is used within MINRES). However, we have used this (tighter) tolerance here since we required convergence for inexact inverse iteration with the standard preconditioner used within the inner system. To make both approaches (the use of the standard preconditioner versus the use of the tuned preconditioner) comparable both methods used the same decreasing sequence of solve tolerances for the inner linear system. We note that it is possible to introduce even more significant savings in the total number of inner iterations if inexact inverse iteration with the tuned preconditioner would be used with a coarser inner solve tolerance.

Next, in Table 4.1, we present the values in condition (4.24), which ensures that $\mathbb{P}_i$ is positive definite. We see the (4.24) holds at each outer iteration. Also, we see that $\kappa^1_{\mathbb{L}_i}$ quickly becomes independent of i, as stated after (4.38).

Both methods have a relatively high number of outer iterations, since we have only linear convergence with convergence rate $\dfrac{|\lambda_1 - \sigma|}{|\lambda_2 - \sigma|} \approx 0.534$. As we would expect there is almost no difference between methods (a) and (b) as regards the number of outer iterations and the overall outer convergence rate. These results are not presented here since we are primarily interested in the inner iterations used by preconditioned MINRES. The tuned preconditioner is clearly much better than the standard preconditioner in terms of the total number of iterations.

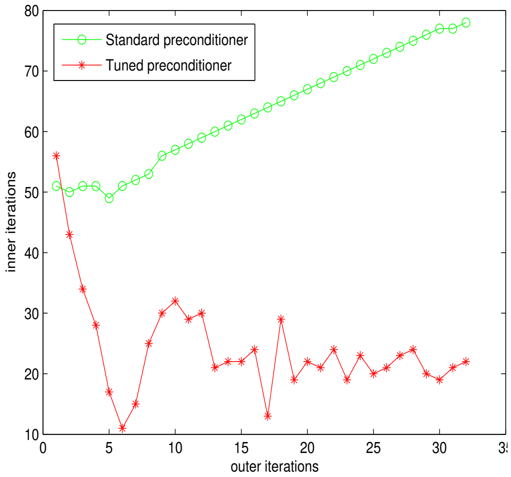


Figure 4-1: *Number of inner iterations against outer iterations for methods (a) and (b)*

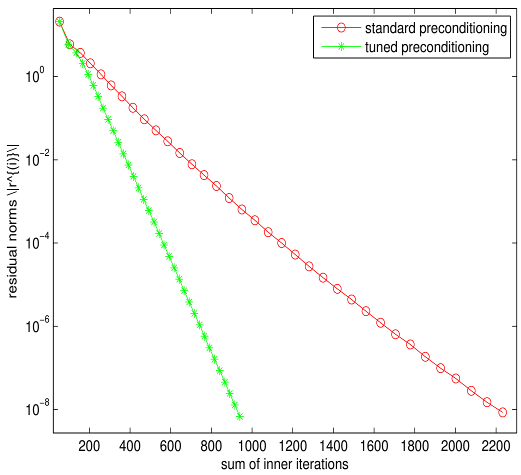


Figure 4-2: *Eigenvalue residual norms against total sum of iterations for methods (a) and (b)*

The results for an interior eigenvalue, namely the fifth eigenvalue, are shown in Figures 4-7 to 4-12 and show very similar behaviour as the results for the extreme eigenvalues.

To summarise, the number of inner iterations per outer iteration grows steadily for the standard incomplete Cholesky preconditioner as expected, whilst it stays roughly constant for the tuned preconditioner. Also, the tuned preconditioner requires about half the total number of inner iterations than the standard preconditioner. This suggests that tuned preconditioner has a clear advantage over the standard incomplete Cholesky preconditioner. Similar behaviour is observed in our second example.

Example 4.13 (Elliptic operator problem from [83], [119]). *The matrix $\mathbf{A}(t)$ is a symmetry-preserving central finite difference approximation of the self-adjoint elliptic operator*

$$\mathcal{A}(t)u = ((1 + tx)u_x)_x + ((1 + ty)u_y)_y$$

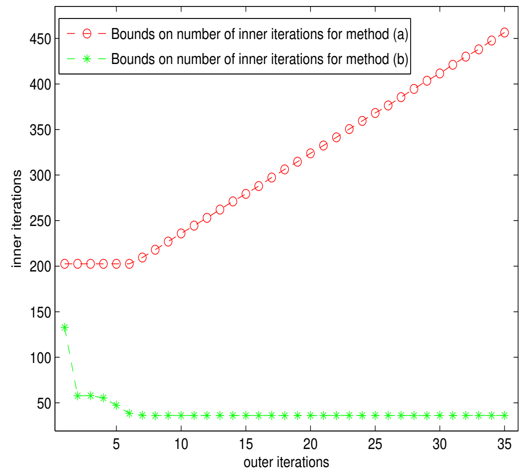


Figure 4-3: *The bound (4.16) for method (a) and the bound (4.37) for method (b)*

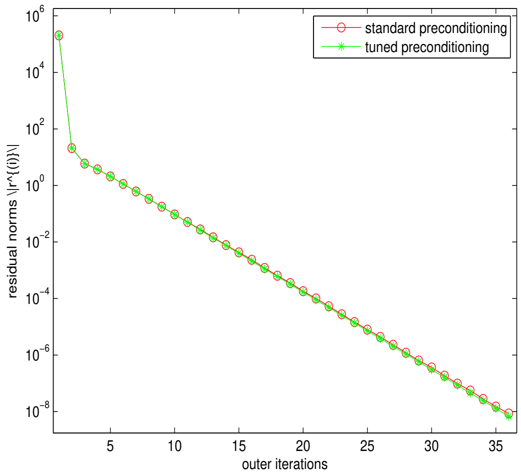


Figure 4-4: *Residual norms against outer iterations for methods (a) and (b)*

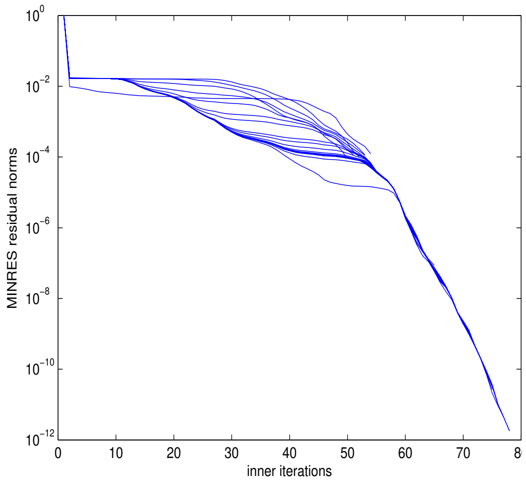


Figure 4-5: *Evolution of relative MINRES residual norms for method (a) (standard preconditioner)*

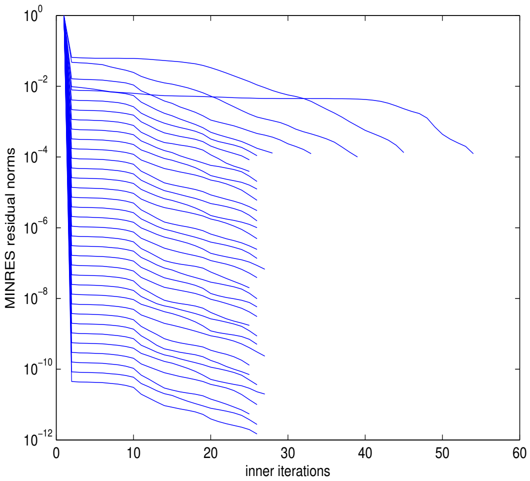


Figure 4-6: *Evolution of relative MINRES residual norms for method (b) (tuned preconditioner)*

on an equidistant grid on the unit square with Dirichlet boundary conditions and 50 *nodes in each dimension. This leads to a matrix* $\mathbf{A}(t)$ *of size* 2500 *with* 12300 *nonzero entries. The two smallest eigenvalues of* $\mathbf{A}(1)$ *are as follows.*

	1st	*2nd*
eigenvalue	0.01102	0.02758

We are interested in approximating the smallest eigenpair of $\mathbf{A}(1)$*. Our starting approximation* $\mathbf{x}^{(0)}$ *is given by the vector of all ones. We will compare the costs of methods (a) and (b) from Example 4.12 For the inexact solves we use preconditioned MINRES with (4.39) and for the incomplete Cholesky decomposition we use a drop tolerance of*

Table 4.1: *Values of* ${\mathbf{x}^{(i)}}^H\mathbf{u}^{(i)}$ *and* $-|(\mathbf{V}\mathbf{u}^{(i)})_1|^2/\eta_1$ *as well as reduced condition number* $\kappa^1_{\mathbb{L}_i}$.

Iteration	${\mathbf{x}^{(i)}}^H\mathbf{u}^{(i)}$	$-\lvert(\mathbf{V}\mathbf{u}^{(i)})_1\rvert^2/\eta_1$	$\kappa^1_{\mathbb{L}_i}$
1	5711	-1153.8	786.96
2	-3412.1	-0.94218	161.46
3	-3624.9	-7.9332	380.14
4	-3679	-3.4465	657.13
5	-3724	-6.651	751.79
6	-3723.6	-4.9865	783.28
7	-3731	-5.924	790.89
8	-3729.1	-5.4301	793.9
9	-3730.7	-5.692	794.31
10	-3730.1	-5.5508	794.66
11	-3730.5	-5.6241	794.63
12	-3730.3	-5.5842	794.69
13	-3730.4	-5.606	794.68
14	-3730.3	-5.5944	794.69
⋮	⋮	⋮	⋮
35	-3730.3	-5.5986	794.68

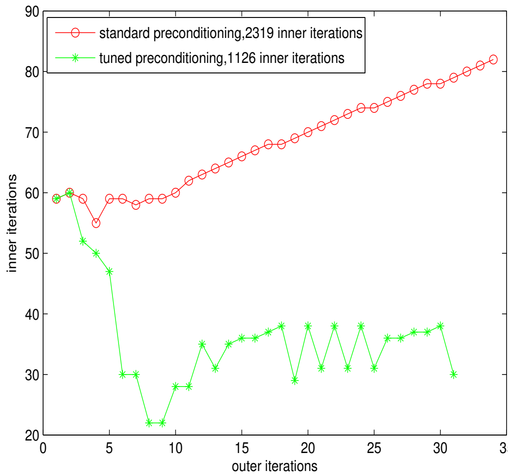


Figure 4-7: *Number of inner iterations against outer iterations for methods (a) and (b)*

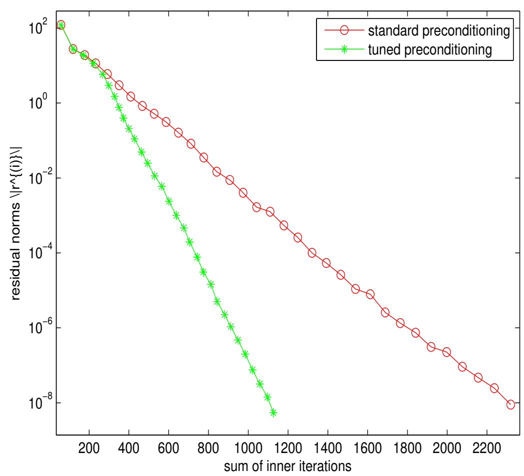


Figure 4-8: *Eigenvalue residual norms against total sum of iterations for methods (a) and (b)*

0.1. *We use a fixed shift of* $\sigma = 0.015$. *Again, the computations stop once the eigenvalue residual satisfies* $\|\mathbf{r}^{(i)}\| < 10^{-8}$.

Figure 4-13 shows the number of inner iterations used by methods (a) and (b). Figure 4-14 plots the residual norms against the total number of iterations. Furthermore, Table 4.2 shows that conditions (4.24) are satisfied for each i. Methods (a) and (b) require the same number of outer iterations and the same outer convergence rate.

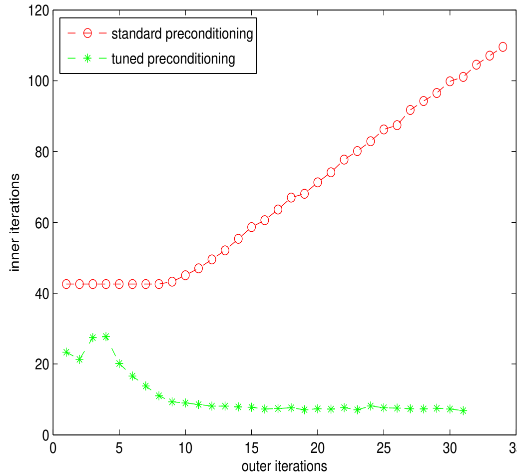


Figure 4-9: *The bound (4.16) for method (a) and the bound (4.37) for method (b)*

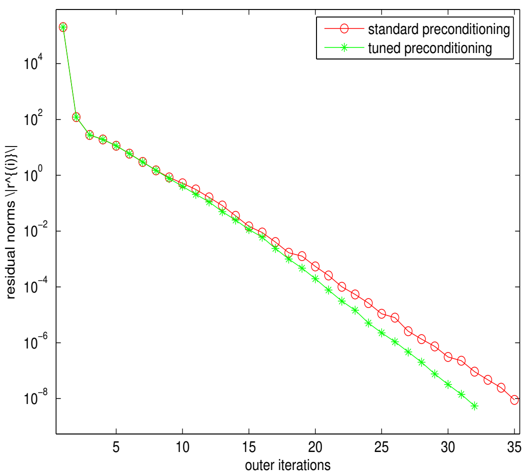


Figure 4-10: *Residual norms against outer iterations for methods (a) and (b)*

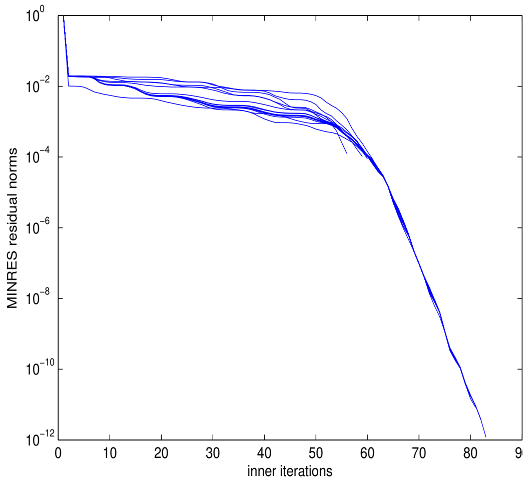


Figure 4-11: *Evolution of relative MINRES residual norms for method (a) (standard preconditioner)*

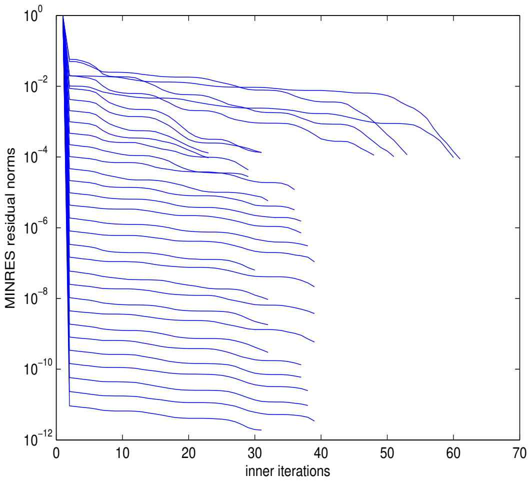


Figure 4-12: *Evolution of relative MINRES residual norms for method (b) (tuned preconditioner)*

In terms of the total number of iterations the tuned preconditioner is clearly much better than the standard preconditioner. We also observe that the theoretical bounds in Figure 4-15 overestimate the actual number of inner iterations by a factor of around 2 or less.

Figure 4-16 shows that the eigenvalue residuals at the outer iteration for both methods are the same.

The number of inner iterations per outer iteration grows steadily for the standard incomplete Cholesky preconditioner as expected, whilst it stays roughly constant for the tuned preconditioner (see Figure 4-13). Also, the tuned preconditioner requires about half the total number of inner iterations than the standard preconditioner. In-

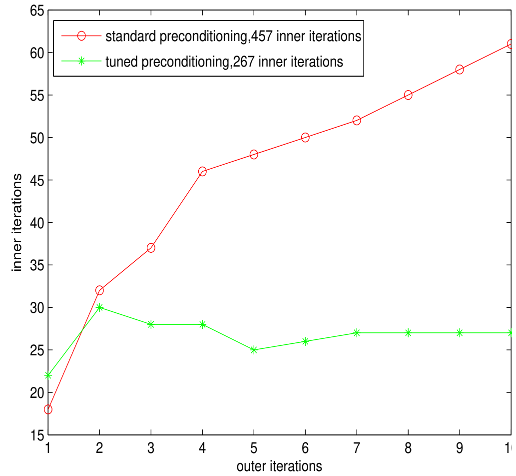


Figure 4-13: *Number of inner iterations against outer iterations for methods (a) and (b)*

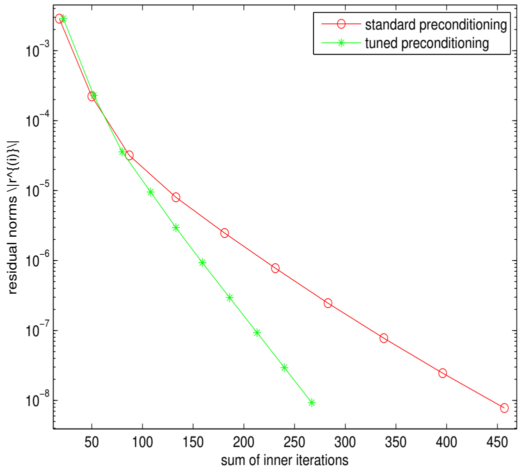


Figure 4-14: *Residual norms against total sum of iterations for methods (a) and (b)*

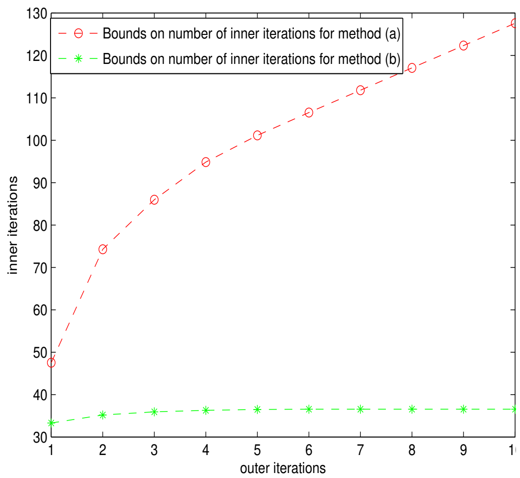


Figure 4-15: *The bound (4.16) for method (a) and the bound (4.37) for method (b)*

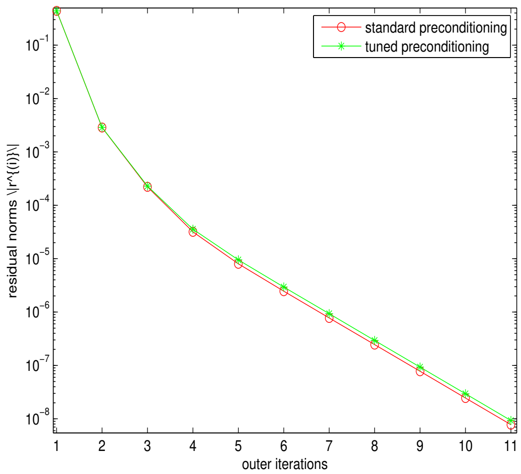


Figure 4-16: *Eigenvalue residual norms against outer iterations for methods (a) and (b)*

deed this superiority is seen in other numerical experiments not reproduced here, and overall, it appears that the tuned preconditioner has a clear advantage over the standard incomplete Cholesky preconditioner.

4.4 Spectral analysis for the tuned preconditioner

In Section 4.3 we proved various properties of the tuned preconditioner $\mathbb{P}_i = \mathbb{L}_i \mathbb{L}_i^H$ given by (4.23) by comparison with the ideal (but unknown) preconditioner given by (4.19). In this section we shall present a direct comparison of the spectral properties

Table 4.2: *Values of* ${\mathbf{x}^{(i)}}^H\mathbf{u}^{(i)}$ *and* $-|(\mathbf{V}\mathbf{u}^{(i)})_1|^2/\eta_1$ *as well as reduced condition number* $\kappa^1_{\mathbb{L}_i}$.

Iteration	${\mathbf{x}^{(i)}}^H\mathbf{u}^{(i)}$	$-\lvert(\mathbf{V}\mathbf{u}^{(i)})_1\rvert^2/\eta_1$	$\kappa^1_{\mathbb{L}_i}$
1	-0.80298	-0.0000039	82.112
2	-0.81856	-0.0001141	83.52
3	-0.82117	-0.0001049	83.267
4	-0.82029	-0.0001058	83.388
5	-0.82057	-0.0001055	83.356
6	-0.82048	-0.0001056	83.366
7	-0.82051	-0.0001056	83.363
⋮	⋮	⋮	⋮
10	-0.8205	-0.0001056	83.364

of $\mathbf{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbf{L}^{-H}$ (where $\mathbf{L}$ is the Cholesky factor of the standard preconditioner $\mathbf{P}$, see Lemma 4.5) and $\mathbb{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}^{-H}$ (where $\mathbb{L}$ is the Cholesky factor of the perfect preconditioner $\mathbb{P}$, see (4.21) in Lemma 4.5). Since the analysis does not involve i, it is identical to a comparison of the spectral properties of $\mathbf{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbf{L}^{-H}$ and $\mathbb{L}_i^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}_i^{-H}$ (which we do not repeat), where $\mathbb{L}_i$ is the Cholesky factor of the perfect preconditioner $\mathbb{P}_i$. The numerical results presented are for the practical preconditioner $\mathbb{P}_i=\mathbb{L}_i\mathbb{L}_i^H$.

We shall show that there is a close relationship between the respective spectra, and so if $\mathbf{L}\mathbf{L}^H$ is a good preconditioner for $\mathbf{A}-\sigma\mathbf{I}$ then $\mathbb{L}\mathbb{L}^H$ will also be a good preconditioner. Specifically, we make the comparison using both a perturbation analysis and an interlacing analysis leading to the main results in Theorem 4.14 and Theorem 4.19.

First recall that for $j=1,\dots,n$, $\mathbf{A}$ has eigenpairs $(\lambda_j,\mathbf{x}_j)$, $\mathbf{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbf{L}^{-H}$ has eigenpairs $(\mu_j,\mathbf{w}_j)$ and $\mathbb{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}^{-H}$ has eigenpairs $(\xi_j,\hat{\mathbf{w}}_j)$. Note that both μ_j and ξ_j are real $\forall j$ since $\mathbf{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbf{L}^{-H}$ and $\mathbb{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}^{-H}$ are Hermitian. If we consider the problem of finding the smallest eigenvalue of $\mathbf{A}$, say λ_1, using a shift σ between λ_1 and λ_2 (the next smallest eigenvalue of $\mathbf{A}$) then $\mathbf{A}-\sigma\mathbf{I}$ has one negative eigenvalue, $\lambda_1-\sigma$, and $n-1$ positive eigenvalues, $\{\lambda_j-\sigma\}_{j=2}^n$. Sylvester's Inertia Theorem readily shows that both $\mathbf{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbf{L}^{-H}$ and $\mathbb{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}^{-H}$ have one negative eigenvalue and $n-1$ positive eigenvalues. Thus, in this case, the assumption in Theorem 4.11 that $\mathbb{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}^{-H}$ satisfies the conditions on $\mathbf{B}$ in Theorem 4.2 is satisfied. We emphasise that our theory is applicable to an interior eigenvalue, in which case Theorem 4.2 can be altered to apply to a $\mathbf{B}$ with a more general spectrum.

First, we note that if

$$\mathbb{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}^{-H}\hat{\mathbf{w}}_j=\xi_j\hat{\mathbf{w}}_j, \tag{4.40}$$

then

$$\mathbf{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbf{L}^{-H}\hat{\mathbf{w}}_j'=\xi_j(\mathbf{I}+\gamma\mathbf{v}\mathbf{v}^H)\hat{\mathbf{w}}_j', \tag{4.41}$$

where $\hat{\mathbf{w}}_j'=\mathbf{L}^H\mathbb{L}^{-H}\hat{\mathbf{w}}_j$, $\mathbf{v}=\mathbf{L}^{-1}\mathbf{u}$ and $\gamma=\dfrac{1}{\mathbf{x}^H\mathbf{u}}$. Note that $\gamma\in\mathbb{R}$, since $\mathbf{u}=(\mathbf{A}-\mathbf{P})\mathbf{x}$. Hence, we find that (4.40) is equivalent to the generalised eigenvalue problem (4.41) and compare the eigenvalues of

$$\mathbf{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbf{L}^{-H}\mathbf{w}_j = \mu_j\mathbf{w}_j \tag{4.42}$$

with those of (4.41). Also, Sylvester's inertia theorem shows that if (4.20) holds, then $1+\gamma\mathbf{v}\mathbf{v}^H$ is positive definite and

$$1+\gamma\mathbf{v}^H\mathbf{v} > 0. \tag{4.43}$$

In Section 4.4.1 we will present a perturbation result comparing the eigenvalues ξ of (4.40) to the eigenvalues μ of (4.42), which is a modification of the theorem by Bauer and Fike (see, for example [48, Theorem 7.2.2]). In Section 4.4.2 we obtain a nonstandard interlacing result to compare the spectra of the standard and tuned preconditioned systems.

4.4.1 Perturbation theory

The following theorem yields a perturbation result for the eigenvalues μ and ξ of (4.42) and (4.40).

Theorem 4.14 (Perturbation Property)**.** *Assume σ is not an eigenvalue $\mathbf{A}$. Define $\mathbf{S}=\mathbf{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbf{L}^{-H}$ and consider the two eigenvalue problems*

$$\mathbf{S}\mathbf{w} = \mu\mathbf{w} \tag{4.44}$$

and

$$\mathbf{S}\mathbf{w}' = \xi(\mathbf{I}+\gamma\mathbf{v}\mathbf{v}^H)\mathbf{w}'. \tag{4.45}$$

Then μ and ξ are nonzero. Also, let ξ be a solution of (4.45). Then

$$\min_{\mu\in\Lambda(\mathbf{S})}\left|\frac{\mu-\xi}{\xi}\right| \le |\gamma\mathbf{v}^H\mathbf{v}|. \tag{4.46}$$

Proof. If σ is not an eigenvalue of $\mathbf{A}$ then $\mathbf{A}-\sigma\mathbf{I}$ is nonsingular, and Sylvester's Inertia Theorem shows that μ and ξ in (4.44) and (4.45) respectively cannot be zero.

Write equation (4.45) as

$$(\mathbf{S}-\xi\mathbf{I})\mathbf{w}' = \xi\gamma\mathbf{v}\mathbf{v}^H\mathbf{w}'.$$

Now, let $\mu\neq\xi$ (for $\mu=\xi$ the result (4.46) follows immediately). Then $\mathbf{S}-\xi\mathbf{I}$ is nonsingular and

$$\mathbf{w}' = \xi(\mathbf{S}-\xi\mathbf{I})^{-1}\gamma\mathbf{v}\mathbf{v}^H\mathbf{w}'.$$

Taking norms we obtain

$$\|\mathbf{w}'\| \le |\xi|\|(\mathbf{S}-\xi\mathbf{I})^{-1}\||\gamma|\|\mathbf{v}\mathbf{v}^H\|\|\mathbf{w}'\|$$

and hence

$$1\le|\xi|\frac{1}{\min\limits_{\mu\in\Lambda(\mathbf{S})}|\mu-\xi|}|\gamma|\|\mathbf{v}^H\mathbf{v}\|,$$

yielding (4.46) after rearrangement. ☐

Corollary 4.15. *Interchanging the roles of (4.44) and (4.45) we have*

$$\min_{\xi\in\Lambda(\mathbf{S},(\mathbf{I}+\gamma\mathbf{v}\mathbf{v}^H))}\left|\frac{\xi-\mu}{\mu}\right| \leq \left|\gamma\mathbf{v}^H\mathbf{v}\right|, \tag{4.47}$$

where $\Lambda(\mathbf{S},(\mathbf{I}+\gamma\mathbf{v}\mathbf{v}^H))$ *is the spectrum of the generalised eigenproblem (4.45).*

In Section 4.4.3 we use this perturbation result to estimate the change in the condition number of the system matrix of (4.25) compared to the condition number of the system matrix of (4.14), which is important for the performance of the iterative solver.

4.4.2 Interlacing property

The following two Lemmata lead to an interlacing result (Theorem 4.19) between the eigenvalues μ of (4.44) and ξ of (4.45), which leads to an interlacing result between the eigenvalues of the matrices in (4.42) and (4.40). Here we use ideas from Wilkinson (see [151]) and Golub and van Loan [48, Lemma 8.5.2, Theorem 8.5.3]. Note that both [151] and GolubvanLoan96 use symmetric matrices, however, the ideas can be easily extended to Hermitian problems (see, for example [2] and [153]).

Lemma 4.16. *Consider the eigenvalue problems*

$$\mathbf{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbf{L}^{-H}\mathbf{w}=\mu\mathbf{w} \tag{4.48}$$

and

$$\mathbb{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}^{-H}\hat{\mathbf{w}}=\xi\hat{\mathbf{w}}, \tag{4.49}$$

where $\mathbb{L}$ *is the Cholesky factor of* $\mathbb{P}$ *given by (4.19). Then we can rewrite the second equation as*

$$\mathbf{D}\mathbf{t}=\xi(\mathbf{I}+\gamma\mathbf{z}\mathbf{z}^H)\mathbf{t} \tag{4.50}$$

or

$$(\mathbf{D}-\xi\gamma\mathbf{z}\mathbf{z}^H)\mathbf{t}=\xi\mathbf{t}, \tag{4.51}$$

where $\mathbb{L}^H\mathbf{L}^{-H}\mathbf{Q}\mathbf{t}=\hat{\mathbf{w}}$, $\mathbf{z}=\mathbf{Q}^H\mathbf{v}$ *with* $\mathbf{v}=\mathbf{L}^{-1}\mathbf{u}$ *as in (4.41) and* $\mathbf{S}=\mathbf{Q}\mathbf{D}\mathbf{Q}^H$ *is the Schur decomposition of* $\mathbf{S}=\mathbf{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbf{L}^{-H}$, *i.e.* $\mathbf{D}$ *is a diagonal matrix* $\mathbf{D}=diag(\mu_1,\ldots,\mu_n)$ *containing the eigenvalues of* $\mathbf{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbf{L}^{-H}$.

Proof. We already know from (4.42) and (4.41) that with $\mathbf{S}=\mathbf{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbf{L}^{-H}$ we can rewrite equations (4.48) and (4.49) as

$$\mathbf{S}\mathbf{w}=\mu\mathbf{w} \tag{4.52}$$

and

$$\mathbf{S}\mathbf{w}'=\xi(\mathbf{I}+\gamma\mathbf{v}\mathbf{v}^H)\mathbf{w}', \tag{4.53}$$

Then, by using the Schur decomposition of $\mathbf{S}=\mathbf{Q}\mathbf{D}\mathbf{Q}^H$, where $\mathbf{D}=\mathrm{diag}(\mu_1,\ldots,\mu_n)$ and setting $\mathbf{Q}^H\mathbf{w}'=\mathbf{t}$ and $\mathbf{Q}^H\mathbf{v}=\mathbf{z}$ we obtain (4.50), that is

$$\mathbf{D}\mathbf{t}=\xi(\mathbf{I}+\gamma\mathbf{z}\mathbf{z}^H)\mathbf{t}.$$

□

Remark 4.17. *Note, that in (4.53) the matrix of the right hand side* $\mathbf{I} + \gamma\mathbf{v}\mathbf{v}^H$ *has all eigenvalues* 1 *except for one eigenvalue at* $1 + \gamma\mathbf{v}^H\mathbf{v}$*. Thus the matrix* $(\mathbf{I} + \gamma\mathbf{v}\mathbf{v}^H)$ *is positive definite if* $1 + \gamma\mathbf{v}^H\mathbf{v} > 0$ *(see condition (4.43)).*

In [48] the efficient computation of the eigenvalues and eigenvectors of a diagonal plus rank-1 matrix was described establishing also an interlacing property between the eigenvalues of the diagonal matrix and the perturbed matrix (see also [151]). Here, problem (4.50) is a generalised eigenvalue problem rather than a standard eigenproblem with rank-1 change but we shall prove that for this problem an interlacing property also holds.

The proofs of the following Lemma and Theorem follow the lines of the proofs of Lemma 8.5.2 and Theorem 8.5.3 in [48].

Lemma 4.18. *Suppose* $\mathbf{D} = diag(\mu_1, \ldots, \mu_n) \in \mathbb{R}^{n\times n}$ *has the property that* $\mu_1 < \ldots < \mu_n$*. Assume that* $\gamma \neq 0$ *and that* $\mathbf{z}$ *has no zero components. If*

$$(\mathbf{D} - \xi\gamma\mathbf{z}\mathbf{z}^H)\mathbf{t} = \xi\mathbf{t}, \quad \mathbf{t} \neq 0$$

then $\mathbf{z}^H\mathbf{t} \neq 0$ *and* $\mathbf{D} - \xi\mathbf{I}$ *is nonsingular.*

Proof. If ξ were an eigenvalue of $\mathbf{D}$ then $\xi = \mu_j$ for some j and hence with $\mathbf{e}_j$ being the jth canonical vector we have

$$0 = \mathbf{e}_j^H[(\mathbf{D} - \xi\mathbf{I})\mathbf{t} - \xi\gamma(\mathbf{z}^H\mathbf{t})\mathbf{z}] = \xi\gamma(\mathbf{z}^H\mathbf{t})\mathbf{z}_j.$$

Since γ, $\mathbf{z}_i$ and ξ are nonzero (if ξ were zero then $\mathbf{D}$ would be singular and σ would be an eigenvalue of $\mathbf{A}$) we must have $\mathbf{z}^H\mathbf{t} = 0$ and so $\mathbf{D}\mathbf{t} = \xi\mathbf{t}$. However $\mathbf{D}$ has distinct eigenvalues μ_j and therefore $\mathbf{t} \in \text{span}\{\mathbf{e}_j\}$. But then $0 = \mathbf{z}^H\mathbf{t} = \mathbf{z}_j$, yielding a contradiction. Thus ξ is not an eigenvalue of $\mathbf{D}$ and hence $\mathbf{D} - \xi\mathbf{I}$ is nonsingular and $\mathbf{z}^H\mathbf{t} \neq 0$. ☐

We use this result to prove the following theorem.

Theorem 4.19 (Interlacing Property)**.** *Consider the two eigenvalue problems*

$$\mathbf{L}^{-1}(\mathbf{A} - \sigma\mathbf{I})\mathbf{L}^{-H}\mathbf{w} = \mu\mathbf{w} \tag{4.54}$$

and

$$\mathbb{L}^{-1}(\mathbf{A} - \sigma\mathbf{I})\mathbb{L}^{-H}\hat{\mathbf{w}} = \xi\hat{\mathbf{w}}, \tag{4.55}$$

and assume condition (4.20) holds. Suppose $\mathbf{D} = diag(\mu_1, \ldots, \mu_n) \in \mathbb{R}^{n\times n}$ *and that the diagonal entries satisfy* $\mu_1 < \ldots < \mu_n$*. Let* $\gamma = \dfrac{1}{\mathbf{x}^H\mathbf{u}} \in \mathbb{R}$*. Furthermore let* $\mathbf{z}$ *and* $\mathbf{t}$ *be defined as in Lemma 4.16. Assume that* $\gamma \neq 0$ *and that* $\mathbf{z}$ *has no zero components. Let*

$$\mathbf{D}\mathbf{t}_j = \xi_j(\mathbf{I} + \gamma\mathbf{z}\mathbf{z}^H)\mathbf{t}_j, \tag{4.56}$$

where ξ_j *are the eigenvalues, with* $\xi_1 \leq \ldots \leq \xi_n$ *and* $\mathbf{t}_j$ *are the corresponding eigenvectors. Also, let* $\mu_1 < \ldots < \mu_p < 0 < \mu_{p+1} < \ldots < \mu_n$*, where* p *is the number of negative eigenvalues of* $\mathbf{L}^{-1}(\mathbf{A} - \sigma\mathbf{I})\mathbf{L}^{-H}$*. Then*

(a) The ξ_j *are the* n *zeros of* $f(\xi) = 1 - \xi\gamma\mathbf{z}^H(\mathbf{D} - \xi\mathbf{I})^{-1}\mathbf{z}$*.*

(b) The eigenvector $\mathbf{t}_j$ *is a multiple of* $(\mathbf{D} - \xi_j \mathbf{I})^{-1}\mathbf{z}$.

(c) If $\gamma > 0$*, then*

$$\mu_1 < \xi_1 < \mu_2 < \xi_2 < \ldots < \mu_p < \xi_p < 0$$

and

$$0 < \xi_{p+1} < \mu_{p+1} < \xi_{p+2} < \mu_{p+2} < \ldots < \xi_n < \mu_n,$$

while, if $\gamma < 0$ *then*

$$\xi_1 < \mu_1 < \xi_2 < \mu_2 < \ldots < \xi_p < \mu_p < 0$$

and

$$0 < \mu_{p+1} < \xi_{p+1} < \mu_{p+2} < \xi_{p+2} < \ldots < \mu_n < \xi_n.$$

Proof. From Lemma 4.16 we know that we can reduce problems (4.54) and (4.55) to (4.56). If $(\mathbf{D} - \xi\gamma\mathbf{z}\mathbf{z}^H)\mathbf{t} = \xi\mathbf{t}$, then

$$(\mathbf{D} - \xi\mathbf{I})\mathbf{t} - \xi\gamma(\mathbf{z}^H\mathbf{t})\mathbf{z} = 0. \tag{4.57}$$

From Lemma 4.18 we know that $(\mathbf{D} - \xi\mathbf{I})$ is nonsingular. Thus

$$\mathbf{t} \in \operatorname{span}((\mathbf{D} - \xi\mathbf{I})^{-1}\mathbf{z})$$

thereby establishing (b). Applying $\mathbf{z}^H(\mathbf{D} - \xi\mathbf{I})^{-1}$ to both sides of equation (4.57) we get

$$\mathbf{z}^H\mathbf{t}(1 - \xi\gamma\mathbf{z}^H(\mathbf{D} - \xi\mathbf{I})^{-1}\mathbf{z}) = 0.$$

By Lemma (4.18), $\mathbf{z}^H\mathbf{t} \neq 0$ and so this shows that if ξ is an eigenvalue of the generalized problem (4.56) then $f(\xi) = 0$, establishing (a). To show the interlacing property (c) we need to look more carefully at the equation

$$f(\xi) = 1 - \xi\gamma\sum_{j=1}^{n}\frac{|\mathbf{z}_j|^2}{\mu_j - \xi}.$$

In order to find the roots of $f(\xi) = 0$ the following equality has to be satisfied

$$\frac{1}{\xi} = \gamma\sum_{j=1}^{n}\frac{|\mathbf{z}_j|^2}{\mu_j - \xi}. \tag{4.58}$$

Hence, the roots of $f(\xi) = 0$ can be found by determining the intersection points of

$$f_1(\xi) = \frac{1}{\xi} \quad \text{and} \quad f_2(\xi) = \gamma\sum_{j=1}^{n}\frac{|\mathbf{z}_j|^2}{\mu_j - \xi}.$$

Note that the derivative of $f_2(\xi)$ is given by

$$f_2'(\xi) = \gamma\sum_{j=1}^{n}\frac{|\mathbf{z}_j|^2}{(\mu_j - \xi)^2}$$

and thus the derivative is either strictly positive or strictly negative, depending on the sign of γ. Also, note that for $\xi \to \pm\infty$ we get $f_2(\xi) \to 0$. Furthermore, since $\gamma \neq 0$

and the μ_j are distinct, $f(\xi)$ has n zeroes. Depending on the sign of γ we have the following situations:

If $\gamma > 0$, then $f_2'(\xi) > 0$, that is $f_2(\xi)$ is monotonely increasing between its poles at $\xi = \mu_j$, where μ_j are the eigenvalues of (4.44). The hyperbola $f_1(\xi)$ is monotonely decreasing for all ξ in $(-\infty, 0)$ and $(0, \infty)$. The plot in Figure 4-17 illustrates the situation for $n = 4$ and $p = 2$. We see that due to the monotonicity properties of $f_1(\xi)$ and $f_2(\xi)$ there is exactly one intersection point of $f_1(\mu)$ and $f_2(\xi)$ between each of the poles at $\xi = \mu_j$ except in the interval containing zero. In this case there is an intersection point between μ_p and zero and a second intersection point between zero and μ_{p+1}. Next, we show that there are no intersection points $\xi > \mu_n$ and $\xi < \mu_1$, that is that the intersection points are shifted towards the origin with respect to the poles. For $\xi \to \pm\infty$ we get $f_2(\xi) \to 0$ and since $f_2(\xi)$ is monotonely increasing $f_2(\xi)$ approaches zero from below (for $\xi \to \infty$) or from above (for $\xi \to -\infty$). The decreasing hyperbola $f_1(\xi)$ does exactly the opposite and therefore the two curves cannot intersect for $\xi > \mu_n$ and $\xi < \mu_1$.

On the other hand, if $\gamma < 0$, then $f_2'(\xi) < 0$ and therefore $f_2(\xi)$ is monotonely decreasing between its poles at $\xi = \mu_j$ and the hyperbola $f_1(\xi)$ is monotonely decreasing for all ξ in $(-\infty, 0)$ and $(0, \infty)$. The plot in Figure 4-18 illustrates the situation for $n = 4$ and $p = 2$. Again we observe that due to the monotonicity properties of $f_1(\xi)$ and $f_2(\xi)$ there is exactly one intersection point of $f_1(\xi)$ and $f_2(\xi)$ between each of the poles at $\xi = \mu_j$ with the exception that there is no intersection between the poles $\mu_p < 0$ and $\mu_{p+1} > 0$. Next, we show that there are two further intersection points, one for $\xi > \mu_n$ and one for $\xi < \mu_1$, and hence the intersection points are shifted away from the origin with respect to the poles. Consider $\xi \to \infty$. Both functions $f_1(\xi)$ and $f_2(\xi)$ are monotonely decreasing and approaching zero. In order to show that they intersect we need to show that $f_1(\xi) > f_2(\xi)$ for $\xi \to \infty$, since, obviously close to the pole $\xi = \mu_n + \delta$, $\delta \to 0$, $f_1(\xi) < f_2(\xi)$. Hence, for $f_1(\xi) > f_2(\xi)$ for $\xi \to \infty$, we have to

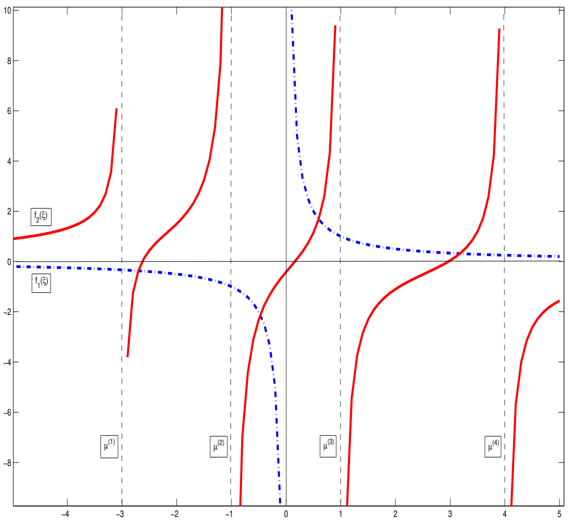


Figure 4-17: *Intersection points of $f_1(\xi)$ and $f_2(\xi)$ for $\gamma > 0$*

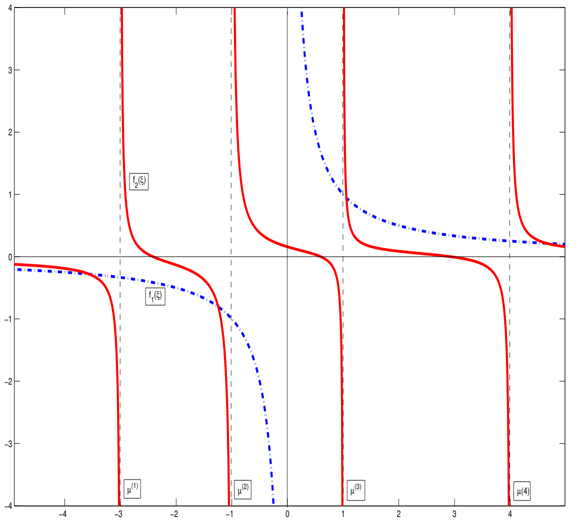


Figure 4-18: *Intersection points of $f_1(\xi)$ and $f_2(\xi)$ for $\gamma < 0$*

show

$$\frac{1}{\xi} > \gamma \sum_{j=1}^{n} \frac{|\mathbf{z}_j|^2}{\mu_j - \xi} \quad \text{for} \quad \xi \to \infty$$

which is equivalent to

$$1 > \gamma\xi \sum_{j=1}^{n} \frac{|\mathbf{z}_j|^2}{\mu_j - \xi} \quad \text{for} \quad \xi \to \infty.$$

Taking the limit we obtain

$$1 > -\gamma \sum_{j=1}^{n} |\mathbf{z}_j|^2 = -\gamma \mathbf{z}^H \mathbf{z}$$

and using $\mathbf{Q}^H\mathbf{v} = \mathbf{z}$ this is equivalent to

$$1 + \gamma \mathbf{v}^H \mathbf{v} > 0$$

which holds from (4.43). In order to show that $f_1(\xi) < f_2(\xi)$ for $\xi \to -\infty$ a similar analysis applies. Thus we have shown that the eigenvalues are shifted away from the origin for $\gamma < 0$. □

Hence, we see that for $\gamma > 0$ the eigenvalues ξ are moved towards the origin, interlacing the eigenvalues μ, whereas for $\gamma < 0$ the eigenvalues ξ are moved away from the origin interlacing the eigenvalues μ.

Theorem 4.19 is proved in the special case of no multiple eigenvalues μ and no zero components of $\mathbf{z}$. Just as in [48, Theorem 8.5.4] these restrictions are easily removed.

Theorem 4.20. *Consider the two eigenvalue problems*

$$\mathbf{L}^{-1}(\mathbf{A} - \sigma\mathbf{I})\mathbf{L}^{-H}\mathbf{w} = \mu\mathbf{w} \tag{4.59}$$

and

$$\mathbb{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}^{-H}\hat{\mathbf{w}} = \xi\hat{\mathbf{w}}, \tag{4.60}$$

and assume condition (4.20) holds. Suppose $\mathbf{D} = diag(\mu_1,\ldots,\mu_n) \in \mathbb{R}^{n\times n}$ *and let* $\gamma = \dfrac{1}{\mathbf{x}^H\mathbf{u}}$. *Furthermore let* $\mathbf{z}$ *and* $\mathbf{t}$ *be defined as in Lemma 4.16. Assume that* $\gamma \neq 0$ *and let*

$$\mathbf{D}\mathbf{t}_j = \xi_j(\mathbf{I}+\gamma\mathbf{z}\mathbf{z}^H)\mathbf{t}_j, \tag{4.61}$$

where ξ_j *are the eigenvalues, with* $\xi_1 \leq \ldots \leq \xi_n$ *and* $\mathbf{t}_j$ *are the corresponding eigenvectors. Also, let* $\mu_1 \leq \ldots \leq \mu_p < 0 < \mu_{p+1} \leq \ldots \leq \mu_n$, *where* p *is the number of negative eigenvalues of* $\mathbf{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbf{L}^{-H}$ *. Then the same interlacing result as in Theorem 4.19 (c) holds, except that the strict inequalities change to equalities for* $\mathbf{z}_j = 0$ *and in case of multiple eigenvalues* μ_j *of* $\mathbf{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbf{L}^{-H}$.

Proof. We only need to show the result for $\mathbf{z}_j = 0$ and in case of multiple μ_j. For other cases the result follows from Theorem 4.19.

If $\mathbf{z}_j = 0$ then from (4.50) we obtain

$$\mathbf{D}\mathbf{e}_j = \xi(\mathbf{I}+\gamma\mathbf{z}\mathbf{z}^H)\mathbf{e}_j = \xi\mathbf{e}_j,$$

where $\mathbf{e}_j$ is the jth canonical vector. Hence $\xi_j = \mu_j$ with corresponding eigenvector $\mathbf{e}_j$ which is even better than interlacing. Furthermore, if $\mu_j = \mu_{j+1}$ we can transform the problem to a problem with a zero component of $\mathbf{z}$. Let $\mathbf{U} = \mathbf{G}(j, j+1, \theta)$ be a (unitary) Givens rotation in the $(j, j+1)$ plane with the property that $\tilde{\mathbf{z}}_{j+1} = 0$, that is

$$\mathbf{U}\mathbf{z} = [\mathbf{z}_1,\ldots,\tilde{\mathbf{z}}_j, 0, \mathbf{z}_{j+2},\ldots,\mathbf{z}_n]^H = \tilde{\mathbf{z}}.$$

It is not hard to show that $\mathbf{U}^H\mathbf{D}\mathbf{U} = \mathbf{D}$. Hence

$$\mathbf{U}^H(\mathbf{D}-\xi\gamma\mathbf{z}\mathbf{z}^H)\mathbf{U} = \mathbf{D}-\xi\gamma\tilde{\mathbf{z}}\tilde{\mathbf{z}}^H$$

and using the previous observation for $\tilde{\mathbf{z}}_{j+1} = 0$ we get $\mu_{j+1} = \mu_j$ is an eigenvalue ξ of the generalized problem (4.56) with corresponding eigenvector $\mathbf{U}\mathbf{e}_{j+1}$. □

Remark 4.21. *Combining the results of Theorem 4.14 and 4.19 we obtain one sided bounds for the largest and smallest eigenvalues of* $\mathbb{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}^{-H}$ *in terms of the largest and smallest eigenvalues of* $\mathbf{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbf{L}^{-H}$. *Thus we also obtain bounds on the condition number of* $\mathbb{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}^{-H}$ *in terms of the condition number of* $\mathbf{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbf{L}^{-H}$. *Furthermore we can conclude that any eigenvalue clustering properties of* $\mathbf{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbf{L}^{-H}$ *are preserved in* $\mathbb{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}^{-H}$.

Thus we are able to obtain qualitative and quantitative information about the quality of $\mathbb{L}$ *as a preconditioner compared with* $\mathbf{L}$. *We note that all the results in this subsection hold identically for the practical tuned preconditioner* $\mathbb{P}_i = \mathbb{L}_i\mathbb{L}_i^H$, *provided (4.24) holds. Numerical results are given for this case below.*

4.4.3 Consequences for the tuned preconditioner

Here we merely compare the various terms which appear in (4.15) and (4.37), which give bounds for the inner iterations in MINRES using the standard and tuned preconditioners respectively.

As before, we assume that in our investigation all the eigenvalues $\mu_2, \ldots, \mu_n$ of $\mathbf{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbf{L}^{-H}$ are positive and μ_1, the extremal eigenvalue is negative. Thus (using Silvester's Inertia Theorem) the eigenvalues $\xi_2, \ldots, \xi_n$ of $\mathbb{L}^{-1}(\mathbf{A}-\sigma\mathbf{I})\mathbb{L}^{-H}$ are positive and ξ_1 is negative. We have to compare the reduced condition numbers

$$\kappa^1_{\mathbf{L}} = \frac{|\mu_n|}{|\mu_2|} \quad \text{and} \quad \kappa^1_{\mathbb{L}} = \frac{|\xi_n|}{|\xi_2|}$$

as well as the terms

$$\frac{|\mu_1 - \mu_n|}{|\mu_1|} \quad \text{and} \quad \frac{|\xi_1 - \xi_n|}{|\xi_1|}.$$

Situation for $\gamma > 0$

If $\gamma > 0$ then, from Theorem 4.19, the eigenvalues ξ are shifted towards the origin with respect to the eigenvalues μ. Hence

$$\xi_n \leq \mu_n$$

holds and from (4.46) we get

$$\frac{\mu_2}{1 + |\gamma \mathbf{v}^H \mathbf{v}|} \leq \xi_2.$$

Combining both bounds yields

$$\kappa^1_{\mathbb{L}} = \frac{|\xi_n|}{|\xi_2|} \leq \frac{|\mu_n|}{|\mu_2|}(1 + |\gamma \mathbf{v}^H \mathbf{v}|) = \kappa^1_{\mathbf{L}}(1 + |\gamma \mathbf{v}^H \mathbf{v}|), \tag{4.62}$$

which is an upper bound on the change to the reduced condition number due to tuning.

Using a similar consideration we obtain

$$\frac{|\mu_1 - \mu_n|}{|\mu_1|} \leq (1 + |\gamma \mathbf{v}^H \mathbf{v}|)\frac{|\xi_1 - \xi_n|}{|\xi_1|} \tag{4.63}$$

for $\gamma > 0$.

Situation for $\gamma < 0$

For $\gamma < 0$ a similar discussion also yields (4.62) and

$$\frac{|\xi_1 - \xi_n|}{|\xi_1|} \leq (1 + |\gamma \mathbf{v}^H \mathbf{v}|)\frac{|\mu_1 - \mu_n|}{|\mu_1|}. \tag{4.64}$$

4.4.4 Numerical example

We consider a numerical example to support our theory in this section and compare the reduced condition numbers.

Example 4.22. *We consider the matrix* `nos5.mtx` *from the Matrix Market library [13]. This is a matrix of size* 468. *We use a shift* $\sigma = 55$ *which is close to the smallest eigenvalue of* $(\mathbf{A}-\sigma\mathbf{I})$ *and which leads to exactly one negative eigenvalue of* $(\mathbf{A}-\sigma\mathbf{I})$. *Again, we choose* $\mathbf{x}$ *to be a random perturbation from the eigenvector belonging to the smallest eigenvalue. Note that in this case* $\gamma > 0$ *and condition (4.20) is ensured.*

Table 4.3: *Results for Example 4.22. The table gives values for* $\mathbf{u}^H\mathbf{x} = \frac{1}{\gamma}$, $1 + |\gamma\mathbf{v}^H\mathbf{v}|$, $\kappa^1_{\mathbf{L}}$ *and* $\kappa^1_{\mathbb{L}}$ *and for different drop tolerances*

Drop Tolerance	$\mathbf{u}^H\mathbf{x}$	$1 + \|\gamma\mathbf{v}^H\mathbf{v}\|$	$\kappa^1_{\mathbf{L}}$	$\kappa^1_{\mathbb{L}}$
0.5	24837.2	3.1230	10393.2	12068.2
0.2	31020.6	2.1342	588.9	595.7

Table 4.3 shows the results for Example 4.22. With regard to the solution of the preconditioned shifted linear systems using MINRES, we observe that the change in the condition numbers and thus the change in the convergence rate is moderate. In fact, it only changes in the third or fourth significant digit. We also observe that the perturbation of the reduced condition number (4.62) is not sharp. The actual perturbation of the reduced condition number is rather small, with for both drop tolerances $\kappa^1_{\mathbf{L}} \leq \kappa^1_{\mathbb{L}} \ll \kappa^1_{\mathbf{L}}(1 + |\gamma\mathbf{v}^H\mathbf{v}|)$.

4.5 Numerical examples for inexact Rayleigh quotient iteration

In this section we present some numerical results to show the use of tuning when Rayleigh quotient shifts are used. Also, we compare the performance of the tuned preconditioner with the technique introduced by [119].

Algorithm 7 Inexact Rayleigh quotient iteration

Input: Initial guess $\mathbf{x}^{(0)}$ with $\|\mathbf{x}^{(0)}\|= 1$.
  Compute $\lambda^{(0)} = {\mathbf{x}^{(0)}}^H\mathbf{A}\mathbf{x}^{(0)}$.
  for $i = 1, \dots, i_{max}$ **do**
    Choose $\tau^{(i)}$,
    Solve $(\mathbf{A} - \lambda^{(i)}\mathbf{I})\mathbf{y}^{(i)} = \mathbf{x}^{(i)}$ inexactly, that is,

$$\| (\mathbf{A} - \lambda^{(i)}\mathbf{I})\mathbf{y}^{(i)} - \mathbf{x}^{(i)} \| \leq \tau^{(i)},$$

    Compute approximate eigenvector $\mathbf{x}^{(i+1)} = \dfrac{\mathbf{y}^{(i)}}{\|\mathbf{y}^{(i)}\|}$,
    Compute approximate eigenvalue $\lambda^{(i+1)} = {\mathbf{x}^{(i+1)}}^H\mathbf{A}\mathbf{x}^{(i+1)}$,
    Evaluate eigenvalue residual $\mathbf{r}^{(i+1)} = (\mathbf{A} - \lambda^{(i+1)}\mathbf{I})\mathbf{x}^{(i+1)}$,
    Test for convergence.
  end for
Output: $\mathbf{x}^{i_{\max}}$, $\lambda^{i_{\max}}$.

First, we summarise the theory of the inexact Rayleigh quotient iteration for the standard Hermitian positive definite eigenvalue problem (4.1). Rayleigh quotient iteration is a special version of inverse iteration where the variable shift in (4.2) is chosen to be the Rayleigh quotient

$$\lambda^{(i)} = \rho(\mathbf{x}^{(i)}) = \frac{\mathbf{x}^{(i)^H}\mathbf{A}\mathbf{x}^{(i)}}{\mathbf{x}^{(i)^H}\mathbf{x}^{(i)}}. \tag{4.65}$$

Algorithm 7 gives a version of inexact inverse iteration with Rayleigh quotient shifts. The following theorem states the convergence theory for inexact Rayleigh quotient iteration.

Theorem 4.23 (Convergence of inexact Rayleigh quotient iteration). *Let (4.1) be the standard eigenvalue problem for a Hermitian matrix $\mathbf{A}$ and consider the application of Algorithm 7 to find a simple eigenpair $(\lambda_1, \mathbf{x}_1)$. Depending on the tolerance $\tau^{(i)}$ the following rates of convergence are obtained by inexact Rayleigh quotient iteration (Algorithm 7) with a sufficiently close starting guess.*

1. Decreasing tolerance. *If $\tau^{(i)} \leq C_1 \, \| \mathbf{r}^{(i)} \|$ in step (1) Algorithm 7, then cubic convergence is achieved by Algorithm 7.*
2. Fixed tolerance. *If $\tau^{(i)} = \tau$ in step (1) Algorithm 7, then quadratic convergence is achieved by Algorithm 7.*

For a tolerance $\tau^{(i)} = 0$ (exact RQI) we obtain cubic convergence.

Proof. Proofs for exact Rayleigh quotient iteration can be found, for example in [101], and for inexact Rayleigh Quotient iteration in [10] and [129]. □

Consider step (2) of Algorithm 6 where the shift σ is chosen as

$$\rho^{(i)} = \mathbf{x}^{(i)^H}\mathbf{A}\mathbf{x}^{(i)}. \tag{4.66}$$

For the solution by MINRES of the preconditioned system in the inexact Rayleigh quotient method we have for the standard preconditioner

$$\mathbf{L}^{-1}(\mathbf{A} - \rho^{(i)}\mathbf{I})\mathbf{L}^{-H}\tilde{\mathbf{y}}^{(i)} = \mathbf{L}^{-1}\mathbf{x}^{(i)}, \quad \mathbf{y}^{(i)} = \mathbf{L}^{-H}\tilde{\mathbf{y}}^{(i)}, \tag{4.67}$$

where $\mathbf{A} = \mathbf{L}\mathbf{L}^H + \mathbf{E}$ as in (4.13), and

$$\mathbb{L}_i^{-1}(\mathbf{A} - \rho^{(i)}\mathbf{I})\mathbb{L}_i^{-H}\tilde{\mathbf{y}}^{(i)} = \mathbb{L}_i^{-1}\mathbf{x}^{(i)}, \quad \mathbf{y}^{(i)} = \mathbb{L}^{-H}\tilde{\mathbf{y}}^{(i)}, \tag{4.68}$$

for the tuned preconditioner $\mathbb{P}_i = \mathbb{L}_i\mathbb{L}_i^H$ given by (4.23). Clearly the outer rate of convergence of Algorithm 7 is unaltered by the choice of $\mathbf{L}$ or $\mathbb{L}_i$, but we shall see in Example 4.24 that $\mathbb{L}_i$ shows significant numerical advantages over $\mathbf{L}$ because of the fact that $\mathbb{L}_i^{-1}\mathbf{x}^{(i)}$ is roughly in the direction of the eigenvector of $\mathbb{L}_i^{-1}(\mathbf{A} - \rho^{(i)}\mathbf{I})\mathbb{L}_i^{-H}$ corresponding to the eigenvalue nearest zero.

As already noted in Section 4.2.2, where a fixed shift is considered, Simoncini and Eldén [119] suggest that the right hand side of (4.67) be altered so that one solves the modified Hermitian system

$$\mathbf{L}^{-1}(\mathbf{A} - \rho^{(i)}\mathbf{I})\mathbf{L}^{-H}\tilde{\mathbf{y}}^{(i)} = \mathbf{L}^{H}\mathbf{x}^{(i)}, \quad \mathbf{y}^{(i)} = \mathbf{L}^{-H}\tilde{\mathbf{y}}^{(i)}, \tag{4.69}$$

in step (2) of Algorithm 6 rather than (4.67). We remark that this strategy has been used before to enhance the Lanczos process [87], [116], however the motivation in

Simoncini & Eldén [119] is new. They noted that $\mathbf{L}^H\mathbf{x}^{(i)}$ is an approximation to the eigenvector of the coefficient matrix corresponding to the eigenvalue closest to zero (see Section 6 of [119]). This method is analysed in [10] where the advantages of (4.69) over (4.67) from the point of view of the inner iteration count are discussed carefully. However (4.69) only gives quadratic outer convergence, since (4.69) is equivalent to $(\mathbf{A}-\rho^{(i)}\mathbf{I})\mathbf{y}^{(i)} = \mathbf{L}\mathbf{L}^H\mathbf{x}^{(i)}$ and so the right hand side is altered from the traditional $\mathbf{x}^{(i)}$. Also, if this approach is used there is no advantage in using a decreasing tolerance for the inexact solves applied to (4.69), since any method based on (4.69) would normally only converge quadratically due to the quadratic convergence of the Rayleigh quotient to the desired eigenvalue for a close enough starting guess.

In fact there is a close relationship between tuning the preconditioner and the approach of [119] as we now show. Equation (4.68) can be written as

$$\mathbb{L}_i^{-1}(\mathbf{A}-\rho^{(i)}\mathbf{I})\mathbb{L}_i^{-H}\tilde{\mathbf{y}}^{(i)} = \mathbb{L}_i^H\mathbb{L}_i^{-H}\mathbb{L}_i^{-1}\mathbf{x}^{(i)},$$

and using $\mathbb{L}_i\mathbb{L}_i^H\mathbf{x}^{(i)} = \lambda^{(i)}\mathbf{x}^{(i)} + \mathbf{r}^{(i)}$,

$$\mathbb{L}^{-1}(\mathbf{A}-\rho^{(i)}\mathbf{I})\mathbb{L}^{-H}\tilde{\mathbf{y}}^{(i)} = \frac{1}{\lambda^{(i)}}\mathbb{L}^H\mathbf{x}^{(i)} - \frac{1}{\lambda^{(i)}}\mathbb{L}^{-1}\mathbf{r}^{(i)}. \tag{4.70}$$

We see that (4.70) has the form of (4.69), but with a perturbed and scaled right hand side.

We consider a numerical example to compare the methods discussed above.

Example 4.24 (Problem from the Matrix Market library [13]). *Consider the same matrix* `nos5.mtx` *and setup as in Example 4.12 but use Rayleigh quotient shift (4.66). We seek the third smallest eigenvalue, given by $\lambda_3 \approx 115.5912$. The starting approximation $\mathbf{x}^{(0)}$ is chosen to be sufficiently close to $\mathbf{x}_3$. We compare the costs of the following methods:*

(a) Standard incomplete Cholesky preconditioner: Algorithm 6 with shift (4.66) and step (2) implemented by solving (4.67), where $\mathbf{L}\mathbf{L}^H$ is the incomplete Cholesky factorisation of $\mathbf{A}$.

(b) Tuned incomplete Cholesky preconditioner: Algorithm 6 with shift (4.66) and (2) implemented by solving (4.68), where $\mathbb{L}_i$ is the Cholesky factor of $\mathbb{P}_i$ given by (4.23).

For the inexact solves we use the decreasing tolerance $\tau^{(i)} = \min\{0.1, 0.1\|\mathbf{r}^{(i)}\|\}$. We use the incomplete Cholesky factorisation of $\mathbf{A}$ given by $\mathbf{L}\mathbf{L}^H$ with drop tolerances 0.25 (leading to 662 nonzero entries in $\mathbf{L}$) and 0.1 (leading to 1032 nonzero entries in $\mathbf{L}$). Note that other similar drop tolerances give comparable results. The computations stop once the eigenvalue residual satisfies

$$\|\mathbf{r}^{(i)}\| < 10^{-10}.$$

Table 4.4 gives the iteration count for methods (a) and (b) and Table 4.5 shows the outer convergence rates. Both methods show the same outer rate of convergence, which should be cubic, as indicated in the error reduction from step 2 to step 3. The tuned preconditioner, method (b), requires fewer inner iterations than the standard preconditioner. The gain is not as significant as in the fixed shift case, see Figure 4-1 and 4-13, but the tuned preconditioner still produces a saving in the total number of iterations of over 25%.

Table 4.4: *Iteration numbers for Example 4.24 using* Rayleigh quotient shifts. *The total number of iterations and number of inner iterations for inexact Rayleigh quotient iteration using either the standard incomplete Cholesky preconditioner (a), or the tuned preconditioner (b).*

	Standard preconditioner		*Tuned preconditioner*	
	DROP TOLERANCES			
OUTER ITERATION	0.25	0.1	0.25	0.1
1	68	62	61	56
2	85	76	69	63
3	148	132	90	78
total	301	270	220	197

Table 4.5: *Error propagation* $\|\mathbf{A}\mathbf{x}^{(i)} - \rho^{(i)}\mathbf{x}^{(i)}\|_2$ *for Example 4.24 using* Rayleigh quotient shift *for inexact RQI with preconditioned solves using methods (a) and (b)*

	Standard preconditioner		*Tuned preconditioner*	
	DROP TOLERANCES			
OUTER ITERATION	0.25	0.1	0.25	0.1
1	17.45	17.45	17.45	17.45
2	6.3e-2	6.2e-2	6.2e-2	6.2e-2
3	1.7e-7	1.3e-7	2.1e-7	1.5e-7
4	8.5e-11	6.3e-11	2.4e-11	2.1e-11

Example 4.25 (Comparison between [119] and the tuned preconditioner)**.** *We use the same matrix and setup as in Example 4.24 and look for the same eigenvalue. We compare the tuned preconditioner using (4.68) to the modified right hand side approach of Simoncini & Eldén using (4.69). We solve both (4.69) and (4.68) to the fixed tolerance of* $\tau = 0.01$*, so both methods exhibit a quadratic outer convergence rate.*

Table 4.6: *Iteration numbers for Example 4.25 using* Rayleigh quotient shifts. *The total number of iterations and number of inner iterations for inexact Rayleigh quotient iteration using either the the modified right hand side approach by Simoncini & Eldén or the tuned preconditioner*

	Simoncini & Eldén		*Tuned preconditioner*	
	DROP TOLERANCES			
OUTER ITERATION	0.25	0.1	0.25	0.1
1	67	62	29	26
2	74	66	56	55
3	85	75	71	67
4	63		18	
total	289	203	174	148

From Table 4.6 we observe that the tuned preconditioner requires fewer inner iterations than the modified right hand side approach, which again shows the advantage of using the tuned preconditioner.

Remark 4.26. *Note that further computations showed that there is not always a gain in using the tuned preconditioner in favour of the approach by Simoncini & Eldén: Clearly the cubic convergence of the tuned preconditioner is superior to quadratic convergence of the modified right hand side approach, but this will only be of importance if we want to solve the eigenvalue problem to very high precision. In those cases the stopping condition for both methods becomes crucial. Further research is needed.*

4.6 Conclusions

We have analysed the behaviour of a new "tuned" preconditioner for MINRES in inexact inverse iteration using a fixed shift for computing eigenpairs of a given Hermitian positive definite matrix.

We prove that, for a fixed shift and decreasing solve tolerance, a bound on the number of inner iterations needed by MINRES at each outer step should not increase as the outer iteration converges. This is confirmed by numerical experiments which indicate that a tuned Cholesky preconditioner is superior to the standard Cholesky preconditioner. Numerical results are also presented for a method with Rayleigh quotient shifts where again the tuned preconditioner is superior to the standard preconditioner.

A detailed analysis is given to compare the spectral properties of the tuned preconditioner with the standard preconditioner. This analysis shows that tuning the preconditioner should not greatly effect the condition number of the preconditioned system.

Finally, the use of the tuned preconditioner is compared with the method of [119], again with favourable results.

In summary, the tuned preconditioner used in an inexact iterative method with decreasing tolerance, combines the advantages of maintaining the outer convergence rate achieved by exact solves, with the efficient inner iteration performance exhibited by the Simoncini & Eldén method.

CHAPTER 5

Rayleigh Quotient iteration and simplified Jacobi-Davidson method with preconditioned iterative solves

5.1 Introduction

In this chapter we use the idea of tuning the preconditioner to obtain an equivalence result between Rayleigh Quotient iteration and simplified Jacobi-Davidson method with preconditioned iterative solves for the standard and generalised non-Hermitian eigenproblem.

Consider the problem of computing a simple, well-separated eigenvalue and corresponding eigenvector of a large, sparse, non-Hermitian matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$, that is,

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{x}, \quad \mathbf{x}^H \mathbf{x} = 1.$$

We assume $\mathbf{A}$ is large and sparse, and that good approximations of the wanted eigenvalue and eigenvector are available. Many popular methods involve the inexact solution of a shifted linear system: examples are inexact inverse iteration, [10, 12, 50] inexact Rayleigh quotient iteration [119] and the Jacobi-Davidson method [63, 124]. As a practical tool, the Jacobi-Davidson method builds a subspace from which the approximate eigenvector is chosen. In this chapter, we shall consider only the simplified version, (in [119] the simplified Jacobi-Davidson method is called the Newton-Grassmann method) where no use is made of previous information.

In [119] it is proved that for Hermitian matrices, simplified Jacobi-Davidson is equivalent to Rayleigh quotient iteration if no preconditioner is used in the inner solve. This equivalence is based on a Lemma from [132] which also holds for the non-Hermitian case, though no use of this fact is made in [119]. In [64] this equivalence is generalised to two-sided Jacobi-Davidson and accelerated two-sided Rayleigh quotient iteration. However, as noted in [64] these results do not hold if a preconditioner is used to speed up the iterative solves.

In this chapter we extend the result of [119] to preconditioned iterative solves for non-Hermitian problems where we utilise the "tuning" of any standard preconditioner as introduced in [42, 43]. Specifically, we shall show in Section 5.2 that, assuming exact arithmetic, the inexact Rayleigh quotient iteration with the altered preconditioner and the inexact simplified Jacobi-Davidson method with the standard preconditioner produce equivalent approximate eigenvectors. Numerical results that support the theory are presented in Section 5.3. In Section 5.4 we give an extension of the theory to generalised non-Hermitian eigenproblems.

The equivalence result proved here is of interest since, in most applications, preconditioned iterative solves will be applied. Additionally, there is the possibility of further equivalence results for subspace based methods.

5.2 Inexact Rayleigh quotient iteration and inexact Jacobi-Davidson method

In this section we describe the inexact Rayleigh quotient algorithm and the inexact Jacobi-Davidson algorithm to find a simple eigenvalue of a large and sparse non-Hermitian matrix $\mathbf{A}$.

Let $\mathbf{x}$ be an approximate unit eigenvector and let the corresponding approximate eigenvalue be given by $\rho(\mathbf{x}) = \mathbf{x}^H\mathbf{A}\mathbf{x}$. The Rayleigh quotient iteration gives a new approximate eigenvector by normalising the solution $\mathbf{y}$ of the system

$$(\mathbf{A} - \rho(\mathbf{x})\mathbf{I})\mathbf{y} = \mathbf{x}. \tag{5.1}$$

Alternatively, the simplified Jacobi-Davidson method (which can also be considered as a Newton-Grassmann method) produces a correction $\mathbf{s}$ to $\mathbf{x}$, which satisfies $\mathbf{s} \perp \mathbf{x}$, from the correction equation

$$(\mathbf{I} - \mathbf{x}\mathbf{x}^H)(\mathbf{A} - \rho(\mathbf{x})\mathbf{I})(\mathbf{I} - \mathbf{x}\mathbf{x}^H)\mathbf{s} = -\mathbf{r}, \tag{5.2}$$

where

$$\mathbf{r} = (\mathbf{A} - \rho(\mathbf{x})\mathbf{I})\mathbf{x} \tag{5.3}$$

is the current eigenvalue residual. The new eigenvector approximation is then given by the normalisation of $\mathbf{x} + \mathbf{s}$. In practice the Jacobi-Davidson approach builds up a subspace, from which an improved eigendirection is obtained, but in this chapter we concentrate on the simplified version which ignores previous information. It has been shown that, if both systems (5.1) and (5.2) are solved exactly, then $\mathbf{y}$ and $\mathbf{x}+\mathbf{s}$ have the same direction (see [119, 126]). Hence, in exact arithmetic both methods produce the same sequence of eigenvector approximations. For inexact solves this property need not hold. However, Simoncini and Eldén [119] have shown that if the same Galerkin-Krylov subspace method is applied to solve (5.1) and (5.2), then there exists a constant $c \in \mathbb{C}$, such that

$$\mathbf{y}_{k+1} = c(\mathbf{x} + \mathbf{s}_k),$$

where $\mathbf{y}_{k+1}$ and $\mathbf{s}_k$ denote the approximate solution of (5.1) and (5.2) after $k+1$ and k steps respectively. (Note that the proof of [119, Proposition 3.2] applies to non-Hermitian matrices, even thought the paper only considers Hermitian positive definite matrices). This means that if $k+1$ steps of a Galerkin-Krylov method were applied to (5.1) and k steps of the same Galerkin-Krylov method were applied to (5.2) then the resulting approximate eigenvectors would be the same. The results in Figure 5-1 in the next section support this equivalence. Hochstenbach and Sleijpen [64] have extended these results to the case of a two-sided Rayleigh quotient iteration and a two-sided Jacobi-Davidson, when BiCG is used as the iterative solver. However, both papers also observe that these results do not hold if preconditioned Krylov methods are used with the inexact iterative solve. In this chapter, we extend these results to the case of preconditioned solves, where a special "tuned" preconditioner is applied to the Rayleigh quotient iteration.

5.2.1 Preconditioned Rayleigh-quotient iteration and Jacobi-Davidson

First, we give an account of how both inexact Rayleigh quotient iteration and inexact simplified Jacobi-Davidson are preconditioned. We restrict ourselves to right-preconditioned methods here, although the results extend to left-preconditioned methods. (Note that in order to preserve symmetry for Hermitian eigenproblems a split preconditioner may be used for the inner iteration. However, a split preconditioner may be transformed to either a left- or a right-preconditioner using a different inner product, (see [111]).)

Let $\mathbf{P}$ be any preconditioner for $\mathbf{A} - \rho(\mathbf{x})\mathbf{I}$. Then right-preconditioned (5.1) has the form

$$(\mathbf{A} - \rho(\mathbf{x})\mathbf{I})\mathbf{P}^{-1}\tilde{\mathbf{y}} = \mathbf{x}, \quad \text{with} \quad \mathbf{y} = \mathbf{P}^{-1}\tilde{\mathbf{y}}. \tag{5.4}$$

Hence, for a Krylov method applied to (5.4) the solution $\tilde{\mathbf{y}}$ lies in the Krylov subspace

$$\operatorname{span}\{\mathbf{x}, (\mathbf{A} - \rho(\mathbf{x})\mathbf{I})\mathbf{P}^{-1}\mathbf{x}, ((\mathbf{A} - \rho(\mathbf{x})\mathbf{I})\mathbf{P}^{-1})^2\mathbf{x}, \ldots\}. \tag{5.5}$$

The preconditioning of an iterative solver for the approximate solution of (5.2) has to be discussed more carefully. The preconditioner $\mathbf{P}$ is restricted to the subspace orthogonal to $\mathbf{x}$, so that,

$$\tilde{\mathbf{P}} := (\mathbf{I} - \mathbf{x}\mathbf{x}^H)\mathbf{P}(\mathbf{I} - \mathbf{x}\mathbf{x}^H), \tag{5.6}$$

is used instead of $\mathbf{P}$. Clearly $\tilde{\mathbf{P}}$ is singular on $\mathbb{C}^n$, but is assumed to be nonsingular on the subspace $\mathbb{C}^n_\perp := \{\mathbf{v} \in \mathbb{C}^n : \mathbf{v} \perp \mathbf{x}\}$. Let $\tilde{\mathbf{P}}^\dagger$ denote the pseudo-inverse of $\tilde{\mathbf{P}}$. Right preconditioned (5.2) then has the form

$$(\mathbf{I} - \mathbf{x}\mathbf{x}^H)(\mathbf{A} - \rho(\mathbf{x})\mathbf{I})(\mathbf{I} - \mathbf{x}\mathbf{x}^H)\tilde{\mathbf{P}}^\dagger\tilde{\mathbf{s}} = -\mathbf{r}, \quad \text{with} \quad \mathbf{s} = \tilde{\mathbf{P}}^\dagger\tilde{\mathbf{s}}. \tag{5.7}$$

The solution of (5.7) using a Krylov solver requires the action of the matrix $(\mathbf{I} - \mathbf{x}\mathbf{x}^H)(\mathbf{A} - \rho(\mathbf{x})\mathbf{I})(\mathbf{I} - \mathbf{x}\mathbf{x}^H)\tilde{\mathbf{P}}^\dagger$. First we need the efficient implementation of $\tilde{\mathbf{P}}^\dagger\tilde{\mathbf{s}}$ for some $\tilde{\mathbf{s}} \in \mathbb{C}^n_\perp$. This is discussed in [4, page 90] and [127] as we now describe. Consider $\mathbf{v} = \tilde{\mathbf{P}}^\dagger\tilde{\mathbf{s}}$, where $\mathbf{v}$ and $\tilde{\mathbf{s}}$ in $\mathbb{C}^n_\perp$. Then $\tilde{\mathbf{P}}\mathbf{v} = \tilde{\mathbf{s}}$, and using (5.6) we have

$$\begin{aligned}
(\mathbf{I} - \mathbf{x}\mathbf{x}^H)\mathbf{P}\mathbf{v} &= \tilde{\mathbf{s}} \\
\mathbf{P}\mathbf{v} - \mathbf{x}\mathbf{x}^H\mathbf{P}\mathbf{v} &= \tilde{\mathbf{s}} \\
\mathbf{v} - \mathbf{P}^{-1}\mathbf{x}\mathbf{x}^H\mathbf{P}\mathbf{v} &= \mathbf{P}^{-1}\tilde{\mathbf{s}}.
\end{aligned}$$

Hence with $\mathbf{v} \perp \mathbf{x}$ we obtain

$$\mathbf{v} = \left(\mathbf{I} - \frac{\mathbf{P}^{-1}\mathbf{x}\mathbf{x}^H}{\mathbf{x}^H\mathbf{P}^{-1}\mathbf{x}}\right)\mathbf{P}^{-1}\tilde{\mathbf{s}}. \tag{5.8}$$

If $\mathbf{t} = (\mathbf{I} - \mathbf{x}\mathbf{x}^H)(\mathbf{A} - \rho(\mathbf{x})\mathbf{I})(\mathbf{I} - \mathbf{x}\mathbf{x}^H)\tilde{\mathbf{P}}^\dagger\tilde{\mathbf{s}}$, that is $\mathbf{t}$ denotes the action of $(\mathbf{I} - \mathbf{x}\mathbf{x}^H)(\mathbf{A} - \rho(\mathbf{x})\mathbf{I})(\mathbf{I} - \mathbf{x}\mathbf{x}^H)\tilde{\mathbf{P}}^\dagger$ on the vector $\tilde{\mathbf{s}}$, we have

$$\mathbf{t} = (\mathbf{I} - \mathbf{x}\mathbf{x}^H)(\mathbf{A} - \rho(\mathbf{x})\mathbf{I})\mathbf{v}.$$

So with $\tilde{\mathbf{s}}$ denoting the approximate solution of the preconditioned linear system in (5.7), $\mathbf{s} = \tilde{\mathbf{P}}^{\dagger}\tilde{\mathbf{s}}$ is recovered using (5.8):

$$\mathbf{s} = \left(\mathbf{I} - \frac{\mathbf{P}^{-1}\mathbf{x}\mathbf{x}^H}{\mathbf{x}^H\mathbf{P}^{-1}\mathbf{x}}\right)\mathbf{P}^{-1}\tilde{\mathbf{s}}. \tag{5.9}$$

If we introduce the projectors

$$\Pi_1 = \mathbf{I} - \mathbf{x}\mathbf{x}^H \quad \text{and} \quad \Pi_2^{\mathbf{P}} = \left(\mathbf{I} - \frac{\mathbf{P}^{-1}\mathbf{x}\mathbf{x}^H}{\mathbf{x}^H\mathbf{P}^{-1}\mathbf{x}}\right), \tag{5.10}$$

a Krylov solver applied to (5.7) generates the subspace

$$\operatorname{span}\{\mathbf{r}, \Pi_1(\mathbf{A} - \rho(\mathbf{x})\mathbf{I})\Pi_2^{\mathbf{P}}\mathbf{P}^{-1}\mathbf{r}, (\Pi_1(\mathbf{A} - \rho(\mathbf{x})\mathbf{I})\Pi_2^{\mathbf{P}}\mathbf{P}^{-1})^2\mathbf{r}, \ldots\}. \tag{5.11}$$

Clearly, the subspaces (5.5) and (5.11) are not the same and the numerical results in Section 5.3 confirm that there is no equivalence between the eigenvector approximations obtained from (5.4) and (5.7). However, we shall show that if a small modification is made to the standard preconditioner $\mathbf{P}$ in (5.4) then we obtain an equivalence between the inexact versions of Rayleigh quotient iteration and the simplified Jacobi-Davidson method.

5.2.2 Equivalence between preconditioned Jacobi-Davidson and Rayleigh quotient iteration

In Chapter 2, Section 2.5.2 (see also [43]) and in Chapter 4 (see also [42]) a "tuned" preconditioner, $\mathbb{P}$, was introduced. $\mathbb{P}$ is merely a rank-one change to $\mathbf{P}$, a standard preconditioner and has the additional property

$$\mathbb{P}\mathbf{x} = \mathbf{A}\mathbf{x}. \tag{5.12}$$

It is shown in (4) (see also [42]) that for Hermitian problems the use of $\mathbb{P}$ instead of $\mathbf{P}$ leads to an overall reduction of the number of matrix-vector multiplications within the inner solve, since the right hand side of the system in (5.4) with $\mathbf{P}$ replaced by $\mathbb{P}$ is approximately in the kernel of the system matrix.

In this chapter we employ a slightly different choice for $\mathbb{P}$. Specifically, we ask that

$$\mathbb{P}\mathbf{x} = \mathbf{x}, \tag{5.13}$$

and in this chapter we will achieve this by making the choice

$$\mathbb{P} = \mathbf{P} + (\mathbf{I} - \mathbf{P})\mathbf{x}\mathbf{x}^H. \tag{5.14}$$

An immediate consequence of (5.13) is that for the projector $\Pi_2^{\mathbb{P}}$ in (5.10) we have

$$\Pi_2^{\mathbb{P}} = \Pi_1. \tag{5.15}$$

Using the Sherman-Morrison formula and assuming $\mathbf{x}^H\mathbf{P}^{-1}\mathbf{x} \neq 0$ we obtain

$$\mathbb{P}^{-1} = \mathbf{P}^{-1} - \frac{(\mathbf{P}^{-1}\mathbf{x} - \mathbf{x})\mathbf{x}^H\mathbf{P}^{-1}}{\mathbf{x}^H\mathbf{P}^{-1}\mathbf{x}}. \tag{5.16}$$

The application of $\mathbb{P}^{-1}$ involves only one extra solve per outer iteration, since $\mathbf{P}^{-1}\mathbf{x}$ has to be computed only once in the iteration process. This observation is similar to iteration process of the Jacobi-Davidson method, which also only involves one extra solve within the projection process of the preconditioner. Note that for a symmetric positive definite preconditioner a slightly different tuning has to be applied to ensure symmetry and positive definiteness of the tuned preconditioner, see Chapter 4 and [42] for details.

The following Lemma is a generalisation of [132, Lemma 4.1] for preconditioned iterative solves.

Lemma 5.1. *Let $\mathbf{x}$ be a unit-norm vector and let $\rho(\mathbf{x}) = \mathbf{x}^H\mathbf{A}\mathbf{x}$. Let $\mathbf{P}$ be a preconditioner for $\mathbf{A}$ and let Π_1 be defined as in (5.10). Let the tuned preconditioner $\mathbb{P}$ satisfy (5.13) and let $\mathbf{r} = \mathbf{A}\mathbf{x} - \rho(\mathbf{x})\mathbf{x} = \Pi_1\mathbf{r}$. Introduce*

$$\mathcal{K}_k = span\{\mathbf{x}, \mathbf{A}\mathbb{P}^{-1}\mathbf{x}, (\mathbf{A}\mathbb{P}^{-1})^2\mathbf{x}, \ldots, (\mathbf{A}\mathbb{P}^{-1})^k\mathbf{x}\}$$

and

$$\mathcal{L}_k = span\{\mathbf{x}, \mathbf{r}, \Pi_1\mathbf{A}\Pi_2^{\mathbb{P}}\mathbb{P}^{-1}\mathbf{r}, \ldots, (\Pi_1\mathbf{A}\Pi_2^{\mathbb{P}}\mathbb{P}^{-1})^{k-1}\mathbf{r}\}.$$

Then, for every $k \geq 1$, we have $\mathcal{L}_k = \mathcal{K}_k$.

Proof. As noted in (5.15), $\Pi_2^{\mathbb{P}} = \Pi_1$, and

$$\Pi_2^{\mathbb{P}}\mathbb{P}^{-1} = \Pi_1\mathbb{P}^{-1} = \mathbb{P}^{-1}(\mathbf{I} - \mathbf{x}\mathbf{x}^H\mathbb{P}^{-1}). \tag{5.17}$$

In order to prove the equivalence between $\mathcal{L}_k$ and $\mathcal{K}_k$ in the non-Hermitian case we use induction over k. Note that by construction $\mathcal{K}_k$ and $\mathcal{L}_k$ are $k+1$-dimensional subspaces. Clearly $\mathcal{L}_0 = \mathcal{K}_0$ and since $\mathbf{A}\mathbb{P}^{-1}\mathbf{x} = \mathbf{A}\mathbf{x}$ we also have $\mathcal{L}_1 = \mathcal{K}_1$. Assume that $\mathcal{L}_i = \mathcal{K}_i$ for $i < k$.

For $\mathbf{z} \in \mathcal{L}_k$, there exists a $\mathbf{z}_1 \in \mathcal{L}_{k-1} = \mathcal{K}_{k-1}$ and $\gamma \in \mathbb{C}$ such that

$$\begin{aligned}\mathbf{z} &= \mathbf{z}_1 + \gamma\left(\Pi_1\mathbf{A}\mathbb{P}^{-1}(\mathbf{I} - \mathbf{x}\mathbf{x}^H\mathbb{P}^{-1})\right)^{k-1}\mathbf{r}\\ &= \mathbf{z}_1 + \Pi_1\mathbf{A}\mathbb{P}^{-1}(\mathbf{I} - \mathbf{x}\mathbf{x}^H\mathbb{P}^{-1})\mathbf{z}_2,\end{aligned}$$

where $\mathbf{z}_2 = \gamma\left(\Pi_1\mathbf{A}\mathbb{P}^{-1}(\mathbf{I} - \mathbf{x}\mathbf{x}^H\mathbb{P}^{-1})\right)^{k-2}\mathbf{r} \in \mathcal{L}_{k-1} = \mathcal{K}_{k-1}$. Then we obtain

$$\begin{aligned}\mathbf{z} &= \mathbf{z}_1 + (\mathbf{I} - \mathbf{x}\mathbf{x}^H)\mathbf{A}\mathbb{P}^{-1}\left(\mathbf{z}_2 - \mathbf{x}\mathbf{x}^H\mathbb{P}^{-1}\mathbf{z}_2\right)\\ &= \mathbf{z}_1 + \mathbf{A}\mathbb{P}^{-1}\mathbf{z}_2 - \mathbf{x}^H\mathbf{A}\mathbb{P}^{-1}\mathbf{z}_2\mathbf{x} - \mathbf{x}^H\mathbb{P}^{-1}\mathbf{z}_2\mathbf{A}\mathbb{P}^{-1}\mathbf{x} + \mathbf{x}^H\mathbf{A}\mathbb{P}^{-1}\mathbf{x}\mathbf{x}^H\mathbb{P}^{-1}\mathbf{z}_2\mathbf{x}.\end{aligned}$$

We have $\mathbf{z}_1 \in \mathcal{K}_{k-1}$, $\mathbf{x} \in \mathcal{K}_1$, $\mathbf{A}\mathbb{P}^{-1}\mathbf{x} \in \mathcal{K}_2$ and, by the induction hypothesis $\mathbf{A}\mathbb{P}^{-1}\mathbf{z}_2 \in \mathcal{K}_k$. Thus $\mathbf{z} \in \mathcal{K}_k$ and $\mathcal{L}_k \subseteq \mathcal{K}_k$. Finally, if $\mathcal{L}_k$ is of full rank, then its dimension is $k+1$, the same as $\mathcal{K}_k$ and hence the two spaces must be equal and the lemma is proved. If $\mathcal{L}_k$ is not of full dimension, then let i be the largest index such that $\mathcal{L}_i$ is full rank, then $\mathcal{L}_{i+1} = \mathcal{L}_i = \mathcal{K}_i$. Now let $\mathbf{u}_3 \in \mathcal{K}_i$, then, we deduce that $\Pi_1\mathbf{A}\mathbb{P}^{-1}\left(\mathbf{I} - \mathbf{x}\mathbf{x}^H\mathbb{P}^{-1}\right)\mathbf{u}_3 \in \mathcal{K}_i$. Using similar equations to the ones displayed above we obtain that $\mathbf{A}\mathbb{P}^{-1}\mathbf{u}_3 \in \mathcal{K}_i$, so that $\mathcal{K}_{i+1} = \mathcal{K}_i$. By induction we have $\mathcal{L}_k = \mathcal{L}_i = \mathcal{K}_i = \mathcal{K}_k$ for all $k \geq i$, which completes the proof. □

Remark 5.2. *If the tuned $\mathbb{P}$ satisfies (5.13) and is also constructed to be Hermitian then $\mathbb{P}^{-1}$ commutes with Π_1, and the equivalence of $\mathcal{L}_k$ and $\mathcal{K}_k$ is a corollary of [132, Lemma 4.1] applied to $\mathbf{A}\mathbb{P}^{-1}$.*

However, as we now show, a wider result is possible, in that, there is an equivalence between $\mathcal{L}_k$ and the subspace built by the Jacobi-Davidson method using the standard preconditioner, rather than the tuned preconditioner.

Lemma 5.3. *Let the assumptions of Lemma 5.1 hold. With $\mathbb{P}$ given by (5.14),*

$$\mathcal{L}_k = \mathit{span}\{\mathbf{x}, \mathbf{r}, \Pi_1 \mathbf{A} \Pi_2^{\mathbb{P}} \mathbb{P}^{-1}\mathbf{r}, \ldots, (\Pi_1 \mathbf{A} \Pi_2^{\mathbb{P}} \mathbb{P}^{-1})^{k-1}\mathbf{r}\},$$

and

$$\mathcal{M}_k = \mathit{span}\{\mathbf{x}, \mathbf{r}, \Pi_1 \mathbf{A} \Pi_2^{\mathbf{P}} \mathbf{P}^{-1}\mathbf{r}, \ldots, (\Pi_1 \mathbf{A} \Pi_2^{\mathbf{P}} \mathbf{P}^{-1})^{k-1}\mathbf{r}\},$$

we have $\mathcal{L}_k = \mathcal{M}_k$ for every $k > 1$.

Proof. In order to prove this equivalence it is sufficient to show that

$$\Pi_2^{\mathbb{P}} \mathbb{P}^{-1} = \Pi_2^{\mathbf{P}} \mathbf{P}^{-1}.$$

With (5.16) we have

$$\begin{aligned}
\Pi_2^{\mathbb{P}} \mathbb{P}^{-1} &= \Pi_1 \mathbb{P}^{-1} = \mathbb{P}^{-1} - \mathbf{x}\mathbf{x}^H \mathbb{P}^{-1} \\
&= \mathbf{P}^{-1} - \frac{(\mathbf{P}^{-1}\mathbf{x} - \mathbf{x})\mathbf{x}^H \mathbf{P}^{-1}}{\mathbf{x}^H \mathbf{P}^{-1}\mathbf{x}} - \mathbf{x}\mathbf{x}^H \left(\mathbf{P}^{-1} - \frac{(\mathbf{P}^{-1}\mathbf{x} - \mathbf{x})\mathbf{x}^H \mathbf{P}^{-1}}{\mathbf{x}^H \mathbf{P}^{-1}\mathbf{x}} \right) \\
&= \mathbf{P}^{-1} - \frac{\mathbf{P}^{-1}\mathbf{x}\mathbf{x}^H \mathbf{P}^{-1}}{\mathbf{x}^H \mathbf{P}^{-1}\mathbf{x}}
\end{aligned}$$

and hence

$$\Pi_2^{\mathbb{P}} \mathbb{P}^{-1} = \left(\mathbf{I} - \frac{\mathbf{P}^{-1}\mathbf{x}\mathbf{x}^H}{\mathbf{x}^H \mathbf{P}^{-1}\mathbf{x}} \right) \mathbf{P}^{-1} = \Pi_2^{\mathbf{P}} \mathbf{P}^{-1}.$$

which gives the required result. ☐

Combining Lemma 5.1 and Lemma 5.3 we have that $\mathcal{K}_k = \mathcal{L}_k = \mathcal{M}_k$ for every $k > 1$. Note, that the space $\mathcal{K}_k := \mathcal{K}_k(\mathbf{A}\mathbb{P}^{-1}, \mathbf{x})$ is a Krylov subspace. A Galerkin-Krylov method to solve the right preconditioned system $\mathbf{A}\mathbb{P}^{-1}\tilde{\mathbf{y}} = \mathbf{x}$, constructs an approximate solution $\tilde{\mathbf{y}}_k \in \mathcal{K}_k(\mathbf{A}\mathbb{P}^{-1}, \mathbf{x})$ such that the residual $\mathbf{x} - \mathbf{A}\mathbb{P}^{-1}\tilde{\mathbf{y}}_k$ is orthogonal to the Krylov subspace $\mathcal{K}_k(\mathbf{A}\mathbb{P}^{-1}, \mathbf{x})$, assuming the starting guess is zero. An example of such a method is the preconditioned conjugate gradient method (for symmetric systems) or preconditioned FOM (for nonsymmetric linear systems), see [111]. Note that Lemma 5.1 and Lemma 5.3 also hold for shifted systems $\mathbf{A} - \sigma\mathbf{I}$ for any $\sigma \in \mathbb{C}$, by simply replacing $\mathbf{A}$ by $\mathbf{A} - \sigma\mathbf{I}$ in Lemmata 5.1 and 5.3. The next theorem, which is the main result of this chapter, is an extension of [119, Proposition 3.2] and will make use of Lemma 5.1 and Lemma 5.3 applied to shifted systems.

Theorem 5.4. *Let the unit vector $\mathbf{x}$ be an approximate eigenvector of the non-Hermitian matrix $\mathbf{A}$ and set $\rho(\mathbf{x}) = \mathbf{x}^H \mathbf{A}\mathbf{x}$. Let the assumptions of Lemma 5.1 hold and let $\mathbf{y}_{k+1}^{RQ}$ and $\mathbf{s}_k^{JD}$ be the approximate solutions to*

$$(\mathbf{A} - \rho(\mathbf{x})\mathbf{I})\mathbb{P}^{-1}\tilde{\mathbf{y}} = \mathbf{x}, \quad \text{with} \quad \mathbf{y} = \mathbb{P}^{-1}\tilde{\mathbf{y}}, \tag{5.18}$$

and

$$(\mathbf{I} - \mathbf{x}\mathbf{x}^H)(\mathbf{A} - \rho(\mathbf{x})\mathbf{I})(\mathbf{I} - \mathbf{x}\mathbf{x}^H)\tilde{\mathbf{P}}^{\dagger}\tilde{\mathbf{s}} = -\mathbf{r}, \quad \text{with} \quad \mathbf{s} = \tilde{\mathbf{P}}^{\dagger}\tilde{\mathbf{s}}, \tag{5.19}$$

respectively, obtained by $k+1$ (k, respectively) steps of the same Galerkin-Krylov method with starting vector zero. Then there exists a constant $c \in \mathbb{C}$ such that

$$\mathbf{y}_{k+1}^{RQ} = c(\mathbf{x} + \mathbf{s}_k^{JD}). \tag{5.20}$$

Proof. The proof consists of two parts. First we compute the solution $\mathbf{s}_k^{JD}$ to (5.19) and then the solution $\mathbf{y}_{k+1}^{RQ}$ to (5.18) and then we compare both solutions.

(a) The solution $\mathbf{s}_k^{JD}$ to (5.19).
Let $\mathbf{r} = (\mathbf{A} - \rho(\mathbf{x})\mathbf{I})\mathbf{x}$. The Krylov subspace for the solution $\tilde{\mathbf{s}}_k^{JD}$ of (5.19) is given by

$$\mathrm{span}\{\mathbf{r}, \Pi_1(\mathbf{A} - \rho(\mathbf{x})\mathbf{I})\Pi_2^{\mathbf{P}}\mathbf{P}^{-1}\mathbf{r}, \ldots, (\Pi_1(\mathbf{A} - \rho(\mathbf{x})\mathbf{I})\Pi_2^{\mathbf{P}}\mathbf{P}^{-1})^{k-1}\mathbf{r}\}.$$

which, by Lemma 5.3 (with $\mathbf{A}$ replaced by $\mathbf{A} - \rho(\mathbf{x})\mathbf{I}$) and $\Pi_2^{\mathbb{P}} = \Pi_1$ is equal to

$$\mathrm{span}\{\mathbf{r}, \Pi_1(\mathbf{A} - \rho(\mathbf{x})\mathbf{I})\Pi_1\mathbb{P}^{-1}\mathbf{r}, \ldots, (\Pi_1(\mathbf{A} - \rho(\mathbf{x})\mathbf{I})\Pi_1\mathbb{P}^{-1})^{k-1}\mathbf{r}\}.$$

Let $\mathbf{V}_k$ be an orthogonal basis of this subspace. Note that $\mathbf{x} \perp \mathbf{V}_k$, so that $\mathbf{V}_k^H\mathbf{x} = 0$ and $\mathbf{V}_k^H\Pi_1 = \mathbf{V}_k^H$. Then the Galerkin-Krylov solution is given by $\tilde{\mathbf{s}}_k^{JD} = \mathbf{V}_k\mathbf{w}^{JD}$, with $\mathbf{w}^{JD} \in \mathbb{C}^k$, and where the Galerkin condition imposes

$$\mathbf{V}_k^H\Pi_1(\mathbf{A} - \rho(\mathbf{x})\mathbf{I})\Pi_1\mathbb{P}^{-1}\mathbf{V}_k\mathbf{w}^{JD} = -\mathbf{V}_k^H\mathbf{r},$$

or $\mathbf{V}_k^H(\mathbf{A} - \rho(\mathbf{x})\mathbf{I})\Pi_1\mathbb{P}^{-1}\mathbf{V}_k\mathbf{w}^{JD} = -\mathbf{V}_k^H\mathbf{A}\mathbf{x}$. Thus

$$\mathbf{w}^{JD} = -(\mathbf{V}_k^H(\mathbf{A} - \rho(\mathbf{x})\mathbf{I})\Pi_1\mathbb{P}^{-1}\mathbf{V}_k)^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x},$$

and hence

$$\tilde{\mathbf{s}}_k^{JD} = -\mathbf{V}_k(\mathbf{V}_k^H(\mathbf{A} - \rho(\mathbf{x})\mathbf{I})\Pi_1\mathbb{P}^{-1}\mathbf{V}_k)^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}.$$

Using (5.9) with $\mathbb{P}$ instead of $\mathbf{P}$, and $\Pi_2^{\mathbb{P}} = \Pi_1$ we obtain

$$\mathbf{s}_k^{JD} = -\Pi_1\mathbb{P}^{-1}\mathbf{V}_k(\mathbf{V}_k^H(\mathbf{A} - \rho(\mathbf{x})\mathbf{I})\Pi_1\mathbb{P}^{-1}\mathbf{V}_k)^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x} \tag{5.21}$$

as an approximate Galerkin solution to (5.19) after k steps of the method. We can rewrite $\mathbf{s}_k^{JD}$ in the following way. Using the definition of Π_1 we can write

$$\mathbf{w}^{JD} = -(\mathbf{V}_k^H(\mathbf{A} - \rho(\mathbf{x})\mathbf{I})\mathbb{P}^{-1}\mathbf{V}_k - \mathbf{V}_k^H\mathbf{A}\mathbf{x}\mathbf{x}^H\mathbb{P}^{-1}\mathbf{V}_k)^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x},$$

and with the Sherman-Morrison formula and assuming $\mathbf{x}^H\mathbb{P}^{-1}\mathbf{V}_k\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x} \neq 1$ we can determine the inverse in order to get

$$\mathbf{w}^{JD} = -\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}\left(1 + \frac{\mathbf{x}^H\mathbb{P}^{-1}\mathbf{V}_k\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}}{1 - \mathbf{x}^H\mathbb{P}^{-1}\mathbf{V}_k\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}}\right),$$

where

$$\mathbf{S}_k = \mathbf{V}_k^H(\mathbf{A} - \rho(\mathbf{x})\mathbf{I})\mathbb{P}^{-1}\mathbf{V}_k.$$

Then, with $\mathbf{s}_k^{JD}$ from (5.21) we obtain

$$\mathbf{s}_k^{JD} = -\Pi_1\mathbb{P}^{-1}\mathbf{V}_k\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}\left(1 + \frac{\mathbf{x}^H\mathbb{P}^{-1}\mathbf{V}_k\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}}{1 - \mathbf{x}^H\mathbb{P}^{-1}\mathbf{V}_k\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}}\right).$$

Using again the definition of Π_1 we get

$$\begin{aligned}
\mathbf{s}_k^{JD} &= -\mathbb{P}^{-1}\mathbf{V}_k\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}\left(1+\frac{\mathbf{x}^H\mathbb{P}^{-1}\mathbf{V}_k\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}}{1-\mathbf{x}^H\mathbb{P}^{-1}\mathbf{V}_k\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}}\right)\\
&\qquad+\mathbf{x}\mathbf{x}^H\mathbb{P}^{-1}\mathbf{V}_k\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}\left(1+\frac{\mathbf{x}^H\mathbb{P}^{-1}\mathbf{V}_k\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}}{1-\mathbf{x}^H\mathbb{P}^{-1}\mathbf{V}_k\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}}\right)\\
&= -\mathbb{P}^{-1}\mathbf{V}_k\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}\left(1+\frac{\mathbf{x}^H\mathbb{P}^{-1}\mathbf{V}_k\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}}{1-\mathbf{x}^H\mathbb{P}^{-1}\mathbf{V}_k\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}}\right)\\
&\qquad+\mathbf{x}\left(\mathbf{x}^H\mathbb{P}^{-1}\mathbf{V}_k\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}+\frac{(\mathbf{x}^H\mathbb{P}^{-1}\mathbf{V}_k\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x})^2}{1-\mathbf{x}^H\mathbb{P}^{-1}\mathbf{V}_k\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}}\right)\\
&= -\mathbb{P}^{-1}\mathbf{V}_k\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}-\left(\mathbb{P}^{-1}\mathbf{V}_k\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}-\mathbf{x}\right)\xi,
\end{aligned}$$

where ξ is a constant given by

$$\xi=\frac{\mathbf{x}^H\mathbb{P}^{-1}\mathbf{V}_k\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}}{1-\mathbf{x}^H\mathbb{P}^{-1}\mathbf{V}_k\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}}. \tag{5.22}$$

Finally, using (5.13) and the definition of $\mathbf{S}_k$ we obtain

$$\begin{aligned}
\mathbf{s}_k^{JD} = -\mathbb{P}^{-1}\mathbf{V}_k(\mathbf{V}_k^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{I})\mathbb{P}^{-1}\mathbf{V}_k)^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}-&\\
\left(\mathbb{P}^{-1}\mathbf{V}_k(\mathbf{V}_k^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{I})\mathbb{P}^{-1}\mathbf{V}_k)^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}-\mathbf{x}\right)\xi.&
\end{aligned} \tag{5.23}$$

(b) The solution $\mathbf{y}_{k+1}^{RQ}$ to (5.18).
Consider now the solution of (5.18). According to Lemma 5.1 (with $\mathbf{A}$ replaced by $\mathbf{A}-\rho(\mathbf{x})\mathbf{I}$), the columns of $[\mathbf{x},\mathbf{V}_k]$ form an orthogonal basis of

$$\operatorname{span}\{\mathbf{x},(\mathbf{A}-\rho(\mathbf{x})\mathbf{I})\mathbb{P}^{-1}\mathbf{x},((\mathbf{A}-\rho(\mathbf{x})\mathbf{I})\mathbb{P}^{-1})^2\mathbf{x},\ldots,((\mathbf{A}-\rho(\mathbf{x})\mathbf{I})\mathbb{P}^{-1})^k\mathbf{x}\},$$

which is the same space as generated by the Krylov subspace method applied to (5.4). Then the approximate solution to (5.4) is given by $\tilde{\mathbf{y}}_{k+1}^{RQ}=h\mathbf{x}+\mathbf{V}_k\mathbf{h}$, where $h\in\mathbb{C}$ and $\mathbf{h}\in\mathbb{C}^k$. The values of h and $\mathbf{h}$ are determined by imposing the Galerkin condition on (5.4):

$$\begin{bmatrix}\mathbf{x}^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{I})\mathbb{P}^{-1}\mathbf{x} & \mathbf{x}^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{I})\mathbb{P}^{-1}\mathbf{V}_k\\ \mathbf{V}_k^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{I})\mathbb{P}^{-1}\mathbf{x} & \mathbf{V}_k^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{I})\mathbb{P}^{-1}\mathbf{V}_k\end{bmatrix}\begin{bmatrix}h\\ \mathbf{h}\end{bmatrix}=\begin{bmatrix}1\\ \mathbf{0}\end{bmatrix}.$$

Note that $\mathbf{x}^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{I})\mathbb{P}^{-1}\mathbf{x}=\mathbf{x}^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{I})\mathbf{x}=0$. From the second row we obtain

$$\mathbf{V}_k^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{I})\mathbb{P}^{-1}\mathbf{x}h+\mathbf{V}_k^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{I})\mathbb{P}^{-1}\mathbf{V}_k\mathbf{h}=0,$$

and hence

$$\mathbf{V}_k^H\mathbf{A}\mathbf{x}h+\mathbf{V}_k^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{I})\mathbb{P}^{-1}\mathbf{V}_k\mathbf{h}=0,$$

where we have used that $\mathbb{P}\mathbf{x}=\mathbf{x}$ and $\mathbf{V}_k^H\mathbf{x}=0$. Therefore we have

$$\mathbf{h}=-(\mathbf{V}_k^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{I})\mathbb{P}^{-1}\mathbf{V}_k)^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}h,$$

and thus from $\tilde{\mathbf{y}}_{k+1}^{RQ} = h\mathbf{x} + \mathbf{V}_k\mathbf{h}$

$$\tilde{\mathbf{y}}_{k+1}^{RQ} = h(\mathbf{x} - \mathbf{V}_k(\mathbf{V}_k^H(\mathbf{A} - \rho(\mathbf{x})\mathbf{I})\mathbb{P}^{-1}\mathbf{V}_k)^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}).$$

Finally from (5.4) with the tuned preconditioner $\mathbb{P}$ we obtain

$$\mathbf{y}_{k+1}^{RQ} = h(\mathbf{x} - \mathbb{P}^{-1}\mathbf{V}_k(\mathbf{V}_k^H(\mathbf{A} - \rho(\mathbf{x})\mathbf{I})\mathbb{P}^{-1}\mathbf{V}_k)^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}), \tag{5.24}$$

where we have used $\mathbb{P}^{-1}\mathbf{x} = \mathbf{x}$.

Combining both the results of (a) and (b), (5.24) and (5.23) and setting

$$\mathbf{t}_k := \mathbb{P}^{-1}\mathbf{V}_k(\mathbf{V}_k^H(\mathbf{A} - \rho(\mathbf{x})\mathbf{I})\mathbb{P}^{-1}\mathbf{V}_k)^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}$$

we obtain

$$\mathbf{y}_{k+1}^{RQ} = h(\mathbf{x} - \mathbf{t}_k)$$

and

$$\mathbf{s}_k^{JD} = -\mathbf{t}_k - (\mathbf{t}_k - \mathbf{x})\xi.$$

Rewriting these equations and using $\xi \neq -1$ yields

$$\mathbf{y}_{k+1}^{RQ} = \frac{h}{1+\xi}(\mathbf{x} + \mathbf{s}_k^{JD}).$$

The condition $\xi \neq -1$ follows straight from (5.22). The required result follows with $c = \dfrac{h}{1+\xi}$. □

Note that for a Hermitian tuned preconditioner, the result follows straight from (5.24) using (5.21) and $\mathbb{P}^H = \mathbb{P}$ as well as $\mathbf{V}_k^H\Pi_1 = \mathbf{V}_k^H$ such that $\Pi_1\mathbb{P}^{-1}\mathbf{V}_k = \mathbb{P}\Pi_1\mathbf{V}_k = \mathbb{P}\mathbf{V}_k$.
Theorem 5.4 shows that, in exact arithmetic, solving (5.4) and (5.7) with the same preconditioned Galerkin-Krylov method where in (5.4) the tuned preconditioner and in (5.7) the standard preconditioner is used, are equivalent. Note that Rayleigh quotient iteration uses one step more than Jacobi-Davidson ($k+1$ instead of k) because simplified Jacobi-Davidson already uses a matrix-vector multiplication to compute the residual.

Remark 5.5. *Theorem 5.4 also holds if a fixed shift σ is used for both methods (5.4) and (5.7) instead of a Rayleigh quotient shift $\rho(\mathbf{x})$.*

Remark 5.6. *For left-preconditioning we can prove similar results to Lemma 5.1 and Theorem 5.4.*

These observations become clear by examining the proofs of Lemma 5.1 and 5.4. We present numerical results to illustrate Theorem 5.4 in Example 5.8, where the preconditioned Full Orthogonalisation Method (FOM) is used as an iterative solver.

5.2.3 A remark on Petrov-Galerkin methods and tuning with $\mathbb{P}\mathbf{x} = \mathbf{A}\mathbf{x}$

For indefinite systems it is likely that one would use preconditioned GMRES (or preconditioned MINRES for Hermitian systems) instead of FOM/CG and Lanczos as iterative solver. These are Petrov-Galerkin methods which minimise the residual norm at each iteration and which compute the approximate solution of the linear system by requiring that the associated residual is orthogonal to $\mathbf{AP}^{-1}\mathcal{K}_k(\mathbf{AP}^{-1}, \mathbf{x})$. However, it is well-known that Galerkin and norm-minimising Petrov-Galerkin methods are related to each other. Specifically, the residuals of the approximate solutions of FOM (CG or Lanczos in the Hermitian case) and GMRES (MINRES in the Hermitian case) are related as in the shown in the following theorem (see [20]).

Theorem 5.7. *In exact arithmetic, at iteration k, the FOM and the GMRES residuals $\mathbf{r}_k^{FOM}$ and $\mathbf{r}_k^{GMRES}$ are related by*

$$\|\mathbf{r}_k^{FOM}\| = \frac{\|\mathbf{r}_k^{GMRES}\|}{\sqrt{1 - (\|\mathbf{r}_k^{GMRES}\|/\|\mathbf{r}_{k-1}^{GMRES}\|)^2}}.$$

This theorem shows that stagnation of the GMRES/MINRES residual corresponds to peaks in the FOM residual. Furthermore, if the GMRES/MINRES residual norm is reduced significantly at step k, then the FOM/CG residual norm will be approximately equal to the MINRES residual norm. Hence, not so large differences between the application of the Galerkin-Krylov method and the more common norm-minimising methods GMRES and MINRES should be expected. Therefore the equivalence results established in this chapter should be approximately true. The next section gives some numerical evidence.

We conclude this section with a final remark. It is noted at the beginning of subsection (5.2.2) that we make use of the tuning strategy $\mathbb{P}\mathbf{x} = \mathbf{x}$ instead of $\mathbb{P}\mathbf{x} = \mathbf{A}\mathbf{x}$, which was introduced in Chapter 4. Now, since $\mathbb{P}\mathbf{x} = \mathbf{A}\mathbf{x} = \rho(\mathbf{x})\mathbf{x} + \mathbf{r}$, we expect only little differences between the two tuning strategies if $\mathbf{x}$ becomes close to an eigenvector.

The next section gives some numerical evidence to support Theorem 5.4 and the remarks in this subsection.

5.3 Numerical examples

We provide several numerical examples, including one for the fixed shift and one using Rayleigh quotient shifts, where FOM and GMRES are used as iterative solvers for the nonsymmetric problem and CG and MINRES are used as iterative solver for the symmetric problem.

Example 5.8 (Problem from the Matrix Market library [13]). *Consider the nonsymmetric matrix* `sherman5.mtx` *from the Matrix Market library [13]. It is a real matrix of size* 3312×3312 *with* 20793 *nonzero entries. We seek the eigenvector belonging to the smallest eigenvalue* $4.692e-02$. *We use a fixed shift* $\sigma = 0$ *and an initial starting guess of all ones and compare inexact inverse iteration with simplified inexact Jacobi-Davidson method and investigate the following approaches to preconditioning:*

(a) no preconditioner is used for the inner iteration.

(b) a standard preconditioner is used for the inner iteration.

(c) a tuned preconditioner with $\mathbb{P}\mathbf{x} = \mathbf{x}$ *is used for the inner iteration.*

We use FOM as a solver with incomplete LU factorisation with drop tolerance 0.005 *as preconditioner where appropriate. Furthermore, we carry out exactly* 4 *steps of preconditioned FOM for the inner solve in the simplified Jacobi-Davidson method, while precisely* 5 *steps of preconditioned FOM are taken for each inner solve in the inexact inverse iteration. If no preconditioner is used* 124 *steps of FOM are carried out in each inner step of simplified Jacobi-Davidson whilst* 125 *steps of FOM are used in each inner step of inverse iteration. We do this in order to verify (5.20). We also restrict the number of total outer solves to* 20*.*

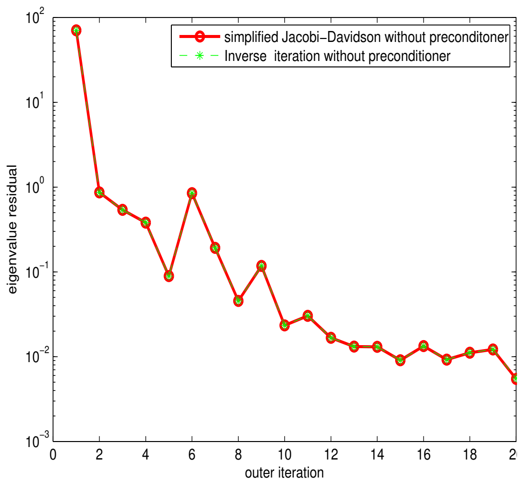


Figure 5-1: *Convergence history of the eigenvalue residuals for Example 5.8, case (a)*

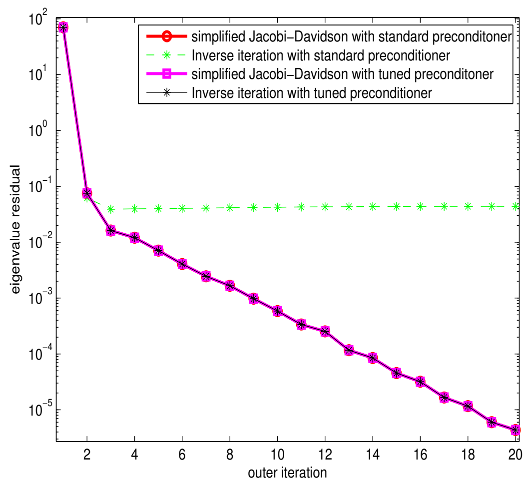


Figure 5-2: *Convergence history of the eigenvalue residuals for Example 5.8, cases (b) and (c)*

Figures 5-1 to 5-2 and Tables 5.1 to 5.2 show the results for Example 5.8.

For unpreconditioned solves (Figure 5-1 and Table 5.1) we observe that inexact simplified Jacobi-Davidson exhibits the same convergence behaviour as inexact inverse iteration, which confirms the results in [119]. Differences in the eigenvalue residuals in the seventh digit only arise after about 14 outer iterations (see Table 5.1) due to accumulated rounding error. For preconditioned solves with a standard preconditioner this property is lost, as it can be readily observed in Figure 5-2 and Table 5.2 . For inexact inverse iteration with the standard preconditioner the eigenvalue residual stagnates! This can be observed in the third column of Table 5.2. The slight increase in the final outer iterations is due to accumulated rounding error.

For the tuned preconditioner which satisfies $\mathbb{P}\mathbf{x} = \mathbf{x}$, we see in Figure 5-2 that with inexact inverse iteration we obtain the same convergence behaviour as for the simplified inexact Jacobi-Davidson method with standard or tuned preconditioner (see columns two, three and five in Table 5.2), which confirms the results in Theorem 5.4.

Example 5.9. *We use the same matrix as in Example 5.8, but a Rayleigh quotient shift is employed to find the eigenvector belonging to the smallest eigenvalue. The initial eigenvector approximation is close enough to that eigenvector. Again methods*

Table 5.1: *Eigenvalue residuals for Example 5.8, case (a), comparing inexact simplified Jacobi-Davidson with inexact inverse iteration when no preconditioner is used for the inner iteration.*

Outer it. i	$\|\mathbf{r}^{(i)}\|$ for simplified JD without preconditioner	$\|\mathbf{r}^{(i)}\|$ for inverse iteration without preconditioner
1	7.044916×10^{1}	7.044916×10^{1}
2	8.657000×10^{-1}	8.657000×10^{-1}
3	5.381041×10^{-1}	5.381041×10^{-1}
4	3.802910×10^{-1}	3.802910×10^{-1}
5	8.916415×10^{-2}	8.916415×10^{-2}
6	8.488819×10^{-1}	8.488819×10^{-1}
7	1.922275×10^{-1}	1.922275×10^{-1}
8	4.550823×10^{-2}	4.550823×10^{-2}
9	1.177346×10^{-1}	1.177346×10^{-1}
10	2.339614×10^{-2}	2.339614×10^{-2}
11	3.027985×10^{-2}	3.027985×10^{-2}
12	1.669518×10^{-2}	1.669518×10^{-2}
13	1.313751×10^{-2}	1.313751×10^{-2}
14	1.306215×10^{-2}	1.306217×10^{-2}
15	9.081211×10^{-3}	9.081264×10^{-3}
16	1.331963×10^{-2}	1.331905×10^{-2}
17	9.221573×10^{-3}	9.220900×10^{-3}
18	1.113956×10^{-2}	1.115966×10^{-2}
19	1.215220×10^{-2}	1.214959×10^{-2}
20	5.457105×10^{-3}	5.564940×10^{-3}

Table 5.2: *Eigenvalue residuals for Example 5.8, cases (b) and (c), comparing inexact simplified Jacobi-Davidson with inexact inverse iteration when the standard and the tuned preconditioner are used within the inner iteration.*

Outer it. i	$\|\mathbf{r}^{(i)}\|$ for simplified JD with standard preconditioner	$\|\mathbf{r}^{(i)}\|$ for inverse iteration with standard preconditioner	$\|\mathbf{r}^{(i)}\|$ for simplified JD with tuned preconditioner	$\|\mathbf{r}^{(i)}\|$ for inverse iteration with tuned preconditioner
1	7.044916×10^{1}	7.044916×10^{1}	7.044916×10^{1}	7.044916×10^{1}
2	7.514239×10^{-2}	6.213435×10^{-2}	7.514239×10^{-2}	7.514239×10^{-2}
3	1.615790×10^{-2}	3.902732×10^{-2}	1.615790×10^{-2}	1.615790×10^{-2}
4	1.206399×10^{-2}	3.961465×10^{-2}	1.206399×10^{-2}	1.206399×10^{-2}
5	7.081815×10^{-3}	3.993356×10^{-2}	7.081815×10^{-3}	7.081815×10^{-3}
6	4.077065×10^{-3}	4.036687×10^{-2}	4.077065×10^{-3}	4.077065×10^{-3}
7	2.445578×10^{-3}	4.090303×10^{-2}	2.445578×10^{-3}	2.445578×10^{-3}
8	1.663492×10^{-3}	4.146019×10^{-2}	1.663492×10^{-3}	1.663492×10^{-3}
9	9.757668×10^{-4}	4.197703×10^{-2}	9.757668×10^{-4}	9.757668×10^{-4}
10	5.836878×10^{-4}	4.242340×10^{-2}	5.836878×10^{-4}	5.836878×10^{-4}
11	3.356904×10^{-4}	4.279152×10^{-2}	3.356904×10^{-4}	3.356904×10^{-4}
12	2.534805×10^{-4}	4.308597×10^{-2}	2.534805×10^{-4}	2.534805×10^{-4}
13	1.166279×10^{-4}	4.331667×10^{-2}	1.166279×10^{-4}	1.166279×10^{-4}
14	8.458152×10^{-5}	4.349488×10^{-2}	8.458152×10^{-5}	8.458152×10^{-5}
15	4.539860×10^{-5}	4.363118×10^{-2}	4.539860×10^{-5}	4.539860×10^{-5}
16	3.180194×10^{-5}	4.373472×10^{-2}	3.180194×10^{-5}	3.180194×10^{-5}
17	1.656168×10^{-5}	4.381299×10^{-2}	1.656168×10^{-5}	1.656168×10^{-5}
18	1.160085×10^{-5}	4.387194×10^{-2}	1.160085×10^{-5}	1.160085×10^{-5}
19	5.999517×10^{-6}	4.391622×10^{-2}	5.999517×10^{-6}	5.999517×10^{-6}
20	4.344489×10^{-6}	4.394944×10^{-2}	4.344489×10^{-6}	4.344489×10^{-6}

(a), (b) and (c) from Example 5.8 are tested and we used (un)preconditioned FOM as iterative inner solver. We carry out exactly 4 *steps of preconditioned FOM for the inner solve in the simplified Jacobi-Davidson method, while precisely* 5 *steps of preconditioned FOM are taken for each inner solve in the inexact Rayleigh quotient*

iteration. If no preconditioner is used 124 *steps of FOM are carried out in each inner step of simplified Jacobi-Davidson whilst* 125 *steps of FOM are used in each inner step of Rayleigh quotient iteration. The maximum number of outer iterations is taken to be* 20.

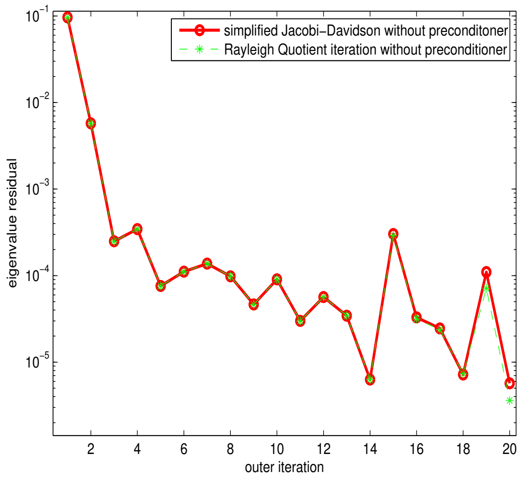


Figure 5-3: *Convergence history of the eigenvalue residuals for Example 5.9, case (a)*

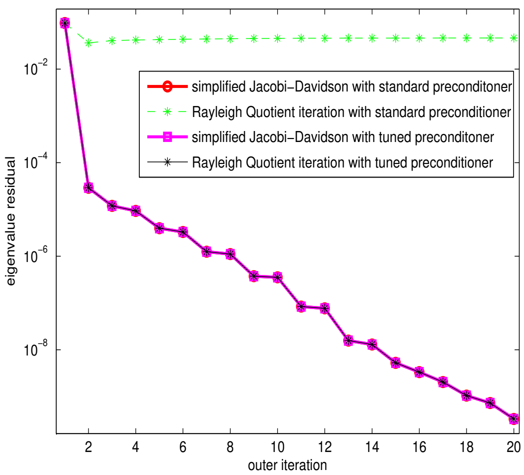


Figure 5-4: *Convergence history of the eigenvalue residuals for Example 5.9, cases (b) and (c)*

Figures 5-3 to 5-4 show the results for Example 5.9. Note that outer convergence is much faster than in Example 5.8, reaching about 10^{-6} instead of 10^{-3} for unpreconditioned solves and 10^{-10} instead of 10^{-5} for preconditioned solves as is seen by comparing the size of the eigenvalue residuals on the vertical axis of Figures 5-1-5-4. For unpreconditioned solves (Figure 5-3) we observe that inexact Rayleigh quotient iteration shows the same convergence behaviour as the simplified Jacobi-Davidson method. If a preconditioner is used, this equivalence holds only if a tuned preconditioner is used for the inexact Rayleigh quotient iteration and either a tuned or the standard preconditoner is used for the simplified Jacobi-Davidson method (Figure 5-4). For the Jacobi-Davidson method the tuned and the standard preconditioner exhibit the same behaviour, that is tuning does not give any benefits if the Jacobi-Davidson method is used (see Lemma 5.3). For the standard preconditioner stagnation is observed in this example (Figure 5-4). This again supports the theoretical results in Theorem 5.4.

The next two examples consider the Hermitian case. We present numerical results in Example 5.10, where preconditioned CG is the iterative solver. However, $\mathbf{A}-\rho(\mathbf{x})\mathbf{I}$ will not be positive definite, and so the preconditioned Lanczos method for linear systems is used for the indefinite system. One can show CG and Lanczos are mathematically equivalent, but the implementation of the CG iteration itself might be unstable, due to the fact that in the case of an indefinite system the Cholesky factorisation in the tridiagonal Lanczos method might not exist (see [86]). Hence we use the direct Lanczos method for the case where the system matrix is indefinite.

We also note that for Hermitian problems, where a Hermitian preconditioner is used for either CG or Lanczos, we have that Lemma 5.3 does not hold. Hence only equivalence between simplified Jacobi-Davidson and inexact inverse iteration can be proved when both methods use the tuned preconditioner (see Figure 5-8).

Example 5.10 (Problem from the Matrix Market library [13]). *Consider matrix* `nos5.mtx` *from the Matrix Market library [13]. It is a real symmetric positive definite matrix of size* 468×468 *with* 5172 *nonzero entries. We seek the eigenvector belonging to the smallest eigenvalue* 52.8995*. We use a fixed shift* $\sigma = 50$ *and an initial starting guess of all ones and compare inexact inverse iteration with the simplified inexact Jacobi-Davidson method and investigate the same three cases (a), (b) and (c) as in Example 5.8. Since* $\mathbf{A} - \sigma\mathbf{I}$ *is positive definite we use CG as a solver with incomplete Cholesky preconditioner where appropriate. Furthermore, we carry out exactly* 20 *steps of (P)CG for the inner solve in the simplified Jacobi-Davidson method, while precisely* 21 *steps of (P)CG are taken for each inner solve in the inexact inverse iteration. We do this in order to confirm (5.20). We also restrict the number of total outer solves to* 20.

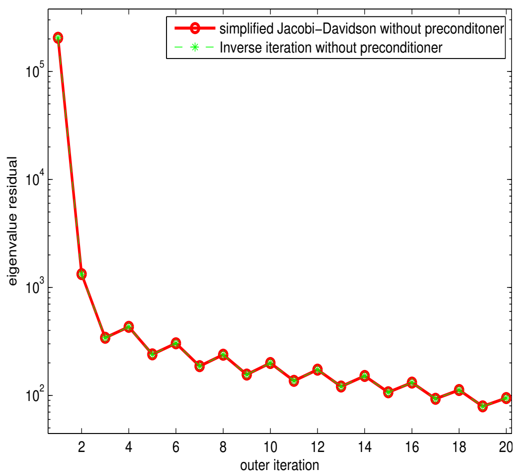


Figure 5-5: *Convergence history of the eigenvalue residuals for Example 5.10, case (a)*

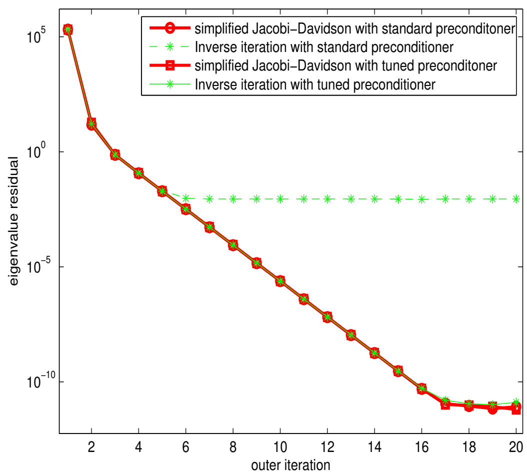


Figure 5-6: *Convergence history of the eigenvalue residuals for Example 5.10, cases (b) and (c)*

Figures 5-5 to 5-6 show the results for Example 5.10. For unpreconditioned solves (Figure 5-5) we observe that inexact simplified Jacobi-Davidson exhibits the same convergence behaviour as inexact inverse iteration, which confirms the results in [119]. For preconditioned solves with a standard preconditioner this property is lost, as it can be readily observed in Figure 5-6. We note that for the Hermitian case, with a Hermitian preconditioner Lemma 5.3 does not hold as can be readily checked. Hence only equivalence holds for inexact simplified Jacobi-Davidson and inexact inverse iteration if both methods use the tuned preconditioner. However, the difference between the use of the standard preconditioner and the tuned preconditioner in the Jacobi-Davidson method is minor and can only be observed by close examination of Figure 5-6 from the 17th to the 20th outer iteration.

For the tuned preconditioner which satisfies $\mathbb{P}\mathbf{x} = \mathbf{x}$, we see in Figure 5-6 that inexact inverse iteration iteration recovers the same convergence behaviour as the simplified inexact Jacobi-Davidson method, which confirms the results in Theorem 5.4.

Example 5.11. *We use the same matrix as in Example 5.10, but a Rayleigh quotient shift is employed to find the eigenvector belonging to the smallest eigenvalue. The initial eigenvector approximation is close enough to that eigenvector. Again methods (a), (b) and (c) from Example 5.8 are tested. This time we use (un)preconditioned Lanczos as iterative inner solver. We carry out exactly* 10 *steps of (P)Lanczos for the inner solve in the simplified Jacobi-Davidson method, while precisely* 11 *steps of (P)Lanczos are taken for each inner solve in the inexact Rayleigh quotient iteration.*

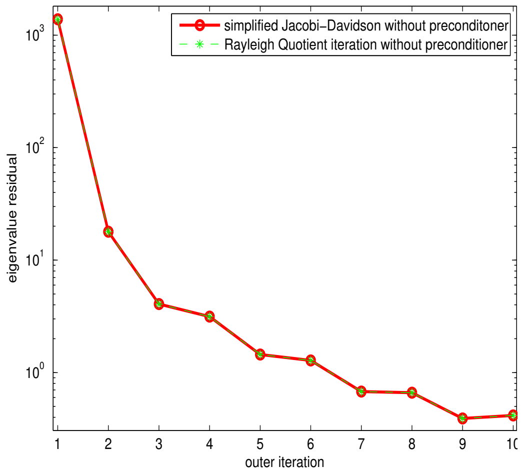


Figure 5-7: *Convergence history of the eigenvalue residuals for Example 5.11, case (a)*

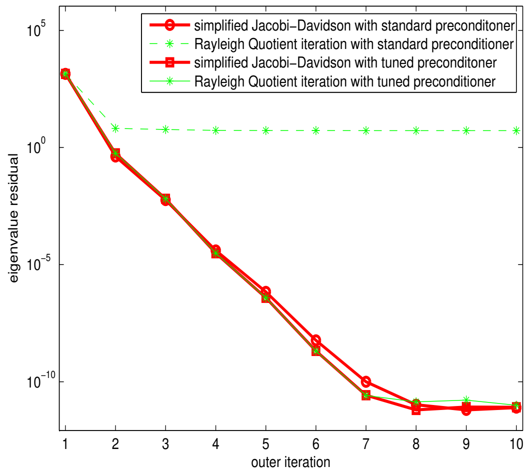


Figure 5-8: *Convergence history of the eigenvalue residuals for Example 5.11, cases (b) and (c)*

Figures 5-7 to 5-8 show the results for Example 5.11. For unpreconditioned solves (Figure 5-7) and for the tuned preconditioner (Figure 5-8) we observe that inexact Rayleigh quotient iteration shows the same convergence behaviour as the simplified inexact Jacobi-Davidson method, which supports the theoretical results in Theorem 5.4. As noted in the previous example before with a Hermitian preconditioner Lemma 5.3 does not hold. Therefore equivalence holds only for inexact simplified Jacobi-Davidson and inexact inverse iteration if both methods use the tuned preconditioner, see Figure 5-8. However, the difference between the use of the standard preconditioner and the tuned preconditioner in the Jacobi-Davidson method is minor and can be observed in Figure 5-8 from the 4th to the 20th outer iteration. Furthermore slight deviations of the order of machine precision arise due to inexact arithmetic (see final iterations in Figure 5-8).

Example 5.12. *The same matrix and setup is used as in Example 5.8. However, instead of tuning with* $\mathbb{P}\mathbf{x} = \mathbf{x}$ *we apply* $\mathbb{P}\mathbf{x} = \mathbf{A}\mathbf{x}$.

Results for Example 5.12 are shown in Figure 5-9. We only show the results in the preconditioned case since the plot for the unpreconditioned case would be as in Figure 5-1. We also only show the case of fixed shifts. Observe that for the inexact simplified Jacobi-Davidson method with the standard and the tuned preconditioner using $\mathbb{P}\mathbf{x} = \mathbf{A}\mathbf{x}$ we obtain the same eigenvalue residuals at each step, see Figure 5-9, thick solid line with circles or triangles. Furthermore, the tuning with $\mathbb{P}\mathbf{x} = \mathbf{A}\mathbf{x}$ performs better than

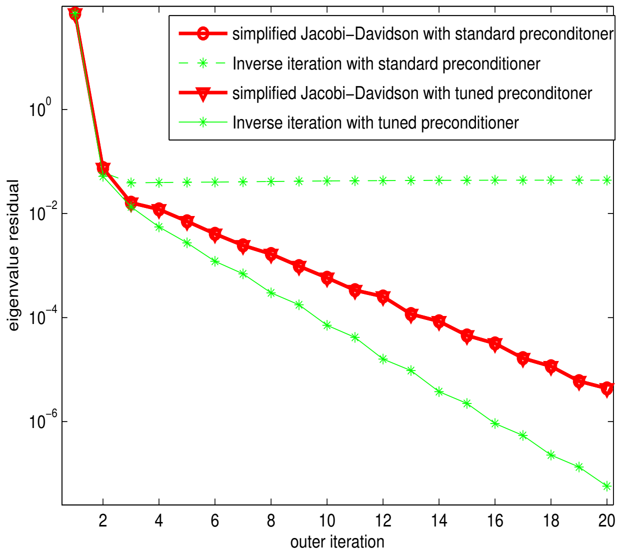


Figure 5-9: *Convergence history of the eigenvalue residuals for Example 5.12, cases (b) and (c), where tuning with $\mathbb{P}\mathbf{x} = \mathbf{A}\mathbf{x}$ is applied*

the tuning with $\mathbb{P}\mathbf{x} = \mathbf{x}$ for the inexact inverse iteration with the tuned preconditioner, as can be seen in the starred solid line.

The following example shows the minor differences if instead of Galerkin methods norm-minimising methods are used, as described in Theorem 5.7.

Example 5.13. *The same matrix and setup is used as in Example 5.8. However, instead of FOM we employ GMRES within the inner iterative solve.*

In order to illustrate the results for Example 5.13 we plot the errors in the eigenvalue residuals as seen in Figure 5-10. Let $\|\mathbf{r}_{JD}^{(i)}\|$ be the eigenvalue residuals at outer iteration i when using the inexact simplified Jacobi-Davidson method with the standard preconditioner and let $\|\mathbf{r}_{II}^{(i)}\|$ be the eigenvalue residual at outer iteration i when using inexact inverse iteration with the tuned preconditioner. The plot in Figure 5-10 shows the difference $\left|\|\mathbf{r}_{JD}^{(i)}\| - \|\mathbf{r}_{II}^{(i)}\|\right|$ if FOM or GMRES is used within the inner preconditioned iterative solve. From that Figure we see that for FOM, the difference between inexact simplified Jacobi-Davidson with standard preconditioner and inexact inverse iteration with tuned preconditioner is of the order of machine precision $\mathcal{O}(10^{-16})$, as expected from the results in this chapter and already verified by Examples 5.8 and 5.9. However, a norm-minimising method like GMRES yields a larger difference between inexact simplified Jacobi-Davidson with standard preconditoner and inexact inverse iteration with tuned preconditioner, which is of the order $\mathcal{O}(10^{-2})$ to $\mathcal{O}(10^{-7})$. This is still small and cannot be seen if logarithmic plots like in the previous examples were used. This behaviour of GMRES as opposed to FOM can be explained using Theorem 5.7.

The final Figures 5-11 and 5-12 show the results for Example 5.8 combined with GMRES and tuning with $\mathbb{P}\mathbf{x} = \mathbf{A}\mathbf{x}$, that is Example 5.8 merged with Example 5.13 and Example 5.12.

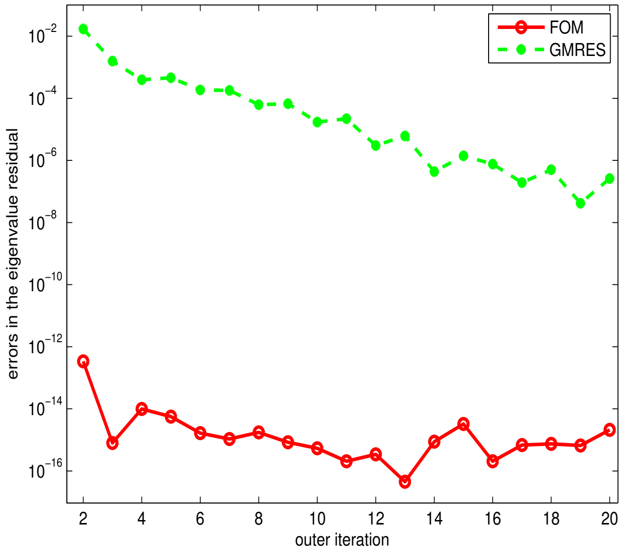


Figure 5-10: *Difference between simplified Jacobi-Davidson with standard preconditoner and inverse iteration with tuned preconditioner as the outer iteration proceeds when using FOM (difference in the order of machine precision) and when using GMRES (larger, but still minor differences)*

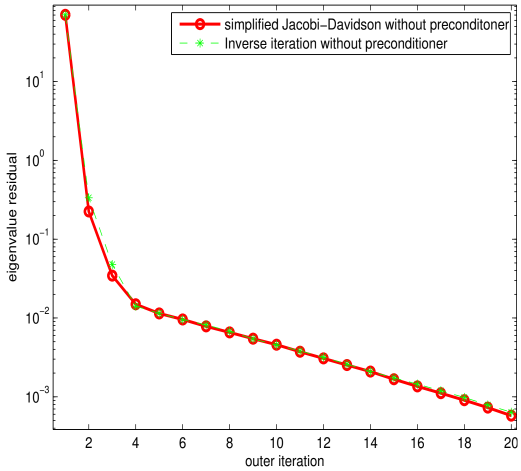


Figure 5-11: *Convergence history of the eigenvalue residuals for Example 5.13, case (a), no preconditioner*

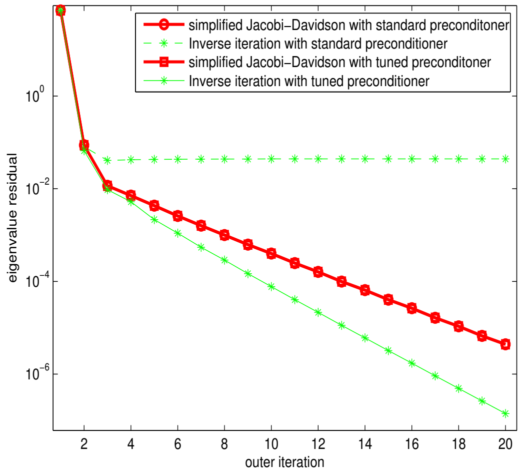


Figure 5-12: *Convergence history of the eigenvalue residuals for Example 5.13, cases (b) and (c), tuning with* $\mathbb{P}\mathbf{x} = \mathbf{A}\mathbf{x}$

5.4 An extension to the generalised non-Hermitian eigenproblem

The results in this chapter can be extended to the non-Hermitian generalised eigenproblem

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{M}\mathbf{x}.$$

For the generalised eigenproblem, Sleijpen et al. [123] introduced a Jacobi-Davidson type method which we describe briefly. Assume $(\rho(\mathbf{x}), \mathbf{x})$ is an approximation to $(\lambda_1, \mathbf{x}_1)$ and introduce the orthogonal projections

$$\Pi_1 = \mathbf{I} - \frac{\mathbf{M}\mathbf{x}\mathbf{w}^H}{\mathbf{w}^H\mathbf{M}\mathbf{x}} \quad \text{and} \quad \Pi_2 = \mathbf{I} - \frac{\mathbf{x}\mathbf{u}^H}{\mathbf{u}^H\mathbf{x}},$$

where $\mathbf{u}^H\mathbf{x} \neq 0$ and $\mathbf{w}^H\mathbf{M}\mathbf{x} \neq 0$. Note that in Chapter 3 we used $\mathbf{w} = \mathbf{M}\mathbf{x}$, $\mathbf{u} = \mathbf{M}^H\mathbf{M}\mathbf{x}$ and $\rho(\mathbf{x}) = \dfrac{\mathbf{x}^H\mathbf{M}^H\mathbf{A}\mathbf{x}}{\mathbf{x}^H\mathbf{M}^H\mathbf{M}\mathbf{x}}$ (see (3.14)) for the eigenvalue approximation. With $\mathbf{r}$ defined by

$$\mathbf{r} = \mathbf{A}\mathbf{x} - \rho(\mathbf{x})\mathbf{M}\mathbf{x}$$

solve the correction equation

$$\Pi_1(\mathbf{A} - \rho(\mathbf{x})\mathbf{M})\Pi_2\mathbf{s} = -\mathbf{r}, \quad \text{where} \quad \mathbf{s} \perp \mathbf{u}, \tag{5.25}$$

for $\mathbf{s}$. This is the Jacobi-Davidson correction equation which maps $\text{span}\{\mathbf{u}\}^\perp$ onto $\text{span}\{\mathbf{w}\}^\perp$. An improved guess for the eigenvector is given by a suitably normalised $\mathbf{x}+\mathbf{s}$. Sleijpen et al. [123, Theorem 3.2] have shown that if (5.25) is solved exactly then $\mathbf{x}^{(i)}$ converges quadratically to the right eigenvector $\mathbf{x}_1$.

Several choices for the projectors Π_1 and Π_2 are possible, depending on the choice of $\mathbf{w}$ and $\mathbf{u}$. We show that if a certain tuned preconditioner is used in inexact inverse iteration applied to the generalised eigenproblem then this method is equivalent to the simple Jacobi-Davidson method with correction equation (5.25) and a standard preconditioner. From now on we assume without loss of generality that $\mathbf{x}$ is normalised such that $\mathbf{x}^H\mathbf{u} = 1$. Let $\mathbf{P}$ be any preconditioner for $\mathbf{A} - \rho(\mathbf{x})\mathbf{M}$, then a system of the form

$$(\mathbf{A} - \rho(\mathbf{x})\mathbf{M})\mathbf{P}^{-1}\tilde{\mathbf{y}} = \mathbf{M}\mathbf{x}, \quad \text{with} \quad \mathbf{y} = \mathbf{P}^{-1}\tilde{\mathbf{y}} \tag{5.26}$$

has to be solved at each inner iteration for inexact inverse iteration whilst a system of the form

$$\left(\mathbf{I} - \frac{\mathbf{M}\mathbf{x}\mathbf{w}^H}{\mathbf{w}^H\mathbf{M}\mathbf{x}}\right)(\mathbf{A} - \rho(\mathbf{x})\mathbf{M})\left(\mathbf{I} - \mathbf{x}\mathbf{u}^H\right)\tilde{\mathbf{P}}^\dagger\tilde{\mathbf{s}} = -\mathbf{r}, \quad \text{with} \quad \mathbf{s} = \mathbf{P}^{-1}\tilde{\mathbf{s}} \tag{5.27}$$

needs to be solved at each inner iteration of the simplified Jacobi-Davidson method, where the preconditioner is restricted such that

$$\tilde{\mathbf{P}} = \left(\mathbf{I} - \frac{\mathbf{M}\mathbf{x}\mathbf{w}^H}{\mathbf{w}^H\mathbf{M}\mathbf{x}}\right)\mathbf{P}\left(\mathbf{I} - \mathbf{x}\mathbf{u}^H\right).$$

Following a similar analysis as in Section 5.2.1 (see also [123, Proposition 7.2]) and introducing the projectors Π_1 and $\Pi_2^{\mathbf{P}}$ given by

$$\Pi_1 = \mathbf{I} - \frac{\mathbf{M}\mathbf{x}\mathbf{w}^H}{\mathbf{w}^H\mathbf{M}\mathbf{x}} \quad \text{and} \quad \Pi_2^{\mathbf{P}} = \mathbf{I} - \frac{\mathbf{P}^{-1}\mathbf{M}\mathbf{x}\mathbf{u}^H}{\mathbf{u}^H\mathbf{P}^{-1}\mathbf{M}\mathbf{x}}, \tag{5.28}$$

we have that a Krylov solve applied to (5.27) generates the subspace

$$\text{span}\{\mathbf{r}, \Pi_1(\mathbf{A} - \rho(\mathbf{x})\mathbf{M})\Pi_2^{\mathbf{P}}\mathbf{P}^{-1}\mathbf{r}, (\Pi_1(\mathbf{A} - \rho(\mathbf{x})\mathbf{M})\Pi_2^{\mathbf{P}}\mathbf{P}^{-1})^2\mathbf{r}, \ldots\},$$

whereas the solution $\tilde{\mathbf{y}}$ to (5.26) lies in the Krylov subspace

$$\text{span}\{\mathbf{M}\mathbf{x}, (\mathbf{A} - \rho(\mathbf{x})\mathbf{M})\mathbf{P}^{-1}\mathbf{M}\mathbf{x}, ((\mathbf{A} - \rho(\mathbf{x})\mathbf{M})\mathbf{P}^{-1})\mathbf{M}\mathbf{x}, \ldots\}.$$

Similarly to the case $\mathbf{M} = \mathbf{I}$ these subspaces are not equal, but if a tuned version of the preconditioner is applied within the inner solve arising at inverse iteration we can show an equivalence between the inexact simplified Jacobi-Davidson method and inexact inverse iteration.

Instead of tuning using $\mathbb{P}\mathbf{x} = \mathbf{A}\mathbf{x}$ we use a slightly different choice for $\mathbb{P}$, namely

$$\mathbb{P}\mathbf{x} = \mathbf{M}\mathbf{x}, \tag{5.29}$$

which, assuming the normalisation $\mathbf{u}^H\mathbf{x} = 1$, is achieved by

$$\mathbb{P} = \mathbf{P} + (\mathbf{M} - \mathbf{P})\mathbf{x}\mathbf{u}^H.$$

Using the Sherman-Morrison formula and assuming $\mathbf{u}^H\mathbf{P}^{-1}\mathbf{M}\mathbf{x} \neq 0$ its inverse $\mathbb{P}^{-1}$ is given by

$$\mathbb{P}^{-1} = \mathbf{P}^{-1} - \frac{(\mathbf{P}^{-1}\mathbf{M}\mathbf{x} - \mathbf{x})\mathbf{u}^H\mathbf{P}^{-1}}{\mathbf{u}^H\mathbf{P}^{-1}\mathbf{M}\mathbf{x}}. \tag{5.30}$$

We can then generalise Lemmata 5.1 and 5.3.

Lemma 5.14. *Consider vectors $\mathbf{w}$ and $\mathbf{u}$ for which $\mathbf{u}^H\mathbf{x} \neq 0$ and $\mathbf{w}^H\mathbf{M}\mathbf{x} \neq 0$. Let $\mathbf{x}$ be a vector normalised such that $\mathbf{x}^H\mathbf{u} = 1$ and let $\rho(\mathbf{x}) = \dfrac{\mathbf{w}^H\mathbf{A}\mathbf{x}}{\mathbf{w}^H\mathbf{M}\mathbf{x}}$ be the Rayleigh quotient. Let $\mathbf{P}$ be a preconditioner for $\mathbf{A}$ and let Π_1 be defined as in (5.28). Further, let the tuned preconditioner $\mathbb{P}$ satisfy (5.29) and let $\mathbf{r} = \mathbf{A}\mathbf{x} - \rho(\mathbf{x})\mathbf{M}\mathbf{x} = \Pi_1\mathbf{r}$. Introduce the subspaces*

$$\mathcal{K}_k = span\{\mathbf{M}\mathbf{x}, \mathbf{A}\mathbb{P}^{-1}\mathbf{M}\mathbf{x}, (\mathbf{A}\mathbb{P}^{-1})^2\mathbf{M}\mathbf{x}, \ldots, (\mathbf{A}\mathbb{P}^{-1})^k\mathbf{M}\mathbf{x}\},$$

$$\mathcal{L}_k = span\{\mathbf{M}\mathbf{x}, \mathbf{r}, \Pi_1\mathbf{A}\Pi_2^{\mathbb{P}}\mathbb{P}^{-1}\mathbf{r}, \ldots, (\Pi_1\mathbf{A}\Pi_2^{\mathbb{P}}\mathbb{P}^{-1})^{k-1}\mathbf{r}\}$$

and

$$\mathcal{M}_k = span\{\mathbf{M}\mathbf{x}, \mathbf{r}, \Pi_1\mathbf{A}\Pi_2^{\mathbf{P}}\mathbf{P}^{-1}\mathbf{r}, \ldots, (\Pi_1\mathbf{A}\Pi_2^{\mathbf{P}}\mathbf{P}^{-1})^{k-1}\mathbf{r}\}$$

Then, for every $k \geq 1$, we have $\mathcal{K}_k = \mathcal{L}_k = \mathcal{M}_k$.

Proof. We first show equivalence between $\mathcal{L}_k$ and $\mathcal{K}_k$. The proof is similar to the proof of Lemma 5.1, but here $\Pi_2^{\mathbb{P}} \neq \Pi_1$. Instead, with (5.28) and (5.29) we have

$$\Pi_2^{\mathbb{P}}\mathbb{P}^{-1} = (\mathbf{I} - \mathbf{x}\mathbf{u}^H)\mathbb{P}^{-1} = \mathbb{P}^{-1}(\mathbf{I} - \mathbf{M}\mathbf{x}\mathbf{u}^H\mathbb{P}^{-1}). \tag{5.31}$$

We use induction over k. Clearly $\mathcal{L}_0 = \mathcal{K}_0$ and since $\mathbf{A}\mathbb{P}^{-1}\mathbf{M}\mathbf{x} = \mathbf{A}\mathbf{x}$ we also have $\mathcal{L}_1 = \mathcal{K}_1$. Assume that $\mathcal{L}_i = \mathcal{K}_i$ for $i < k$.

For $\mathbf{z} \in \mathcal{L}_k$, there exists a $\mathbf{z}_1 \in \mathcal{L}_{k-1} = \mathcal{K}_{k-1}$ and $\gamma \in \mathbb{C}$ such that

$$\begin{aligned}\mathbf{z} &= \mathbf{z}_1 + \gamma\left(\Pi_1\mathbf{A}\mathbb{P}^{-1}(\mathbf{I} - \mathbf{M}\mathbf{x}\mathbf{u}^H\mathbb{P}^{-1})\right)^{k-1}\mathbf{r}\\ &= \mathbf{z}_1 + \Pi_1\mathbf{A}\mathbb{P}^{-1}(\mathbf{I} - \mathbf{M}\mathbf{x}\mathbf{u}^H\mathbb{P}^{-1})\mathbf{z}_2,\end{aligned}$$

where $\mathbf{z}_2 = \gamma\left(\Pi_1\mathbf{A}\mathbb{P}^{-1}(\mathbf{I} - \mathbf{M}\mathbf{x}\mathbf{u}^H\mathbb{P}^{-1})\right)^{k-2}\mathbf{r} \in \mathcal{L}_{k-1} = \mathcal{K}_{k-1}$. Then

$$\begin{aligned}\mathbf{z} &= \mathbf{z}_1 + \left(\mathbf{I} - \frac{\mathbf{M}\mathbf{x}\mathbf{w}^H}{\mathbf{w}^H\mathbf{M}\mathbf{x}}\right)\mathbf{A}\mathbb{P}^{-1}\left(\mathbf{z}_2 - \mathbf{M}\mathbf{x}\mathbf{u}^H\mathbb{P}^{-1}\mathbf{z}_2\right)\\ &= \mathbf{z}_1 + \mathbf{A}\mathbb{P}^{-1}\mathbf{z}_2 - \frac{\mathbf{w}^H\mathbf{A}\mathbb{P}^{-1}\mathbf{z}_2}{\mathbf{w}^H\mathbf{M}\mathbf{x}}\mathbf{M}\mathbf{x} - (\mathbf{u}^H\mathbb{P}^{-1}\mathbf{z}_2)\mathbf{A}\mathbb{P}^{-1}\mathbf{M}\mathbf{x} + \rho(\mathbf{x})\mathbf{u}^H\mathbb{P}^{-1}\mathbf{z}_2\mathbf{M}\mathbf{x}.\end{aligned}$$

We have $\mathbf{z}_1 \in \mathcal{K}_{k-1}$, $\mathbf{Mx} \in \mathcal{K}_1$, $\mathbf{A}\mathbb{P}^{-1}\mathbf{Mx} \in \mathcal{K}_2$ and, by the induction hypothesis $\mathbf{A}\mathbb{P}^{-1}\mathbf{z}_2 \in \mathcal{K}_k$. Thus $\mathbf{z} \in \mathcal{K}_k$ and $\mathcal{L}_k \subseteq \mathcal{K}_k$. Showing $\mathcal{K}_k \subseteq \mathcal{L}_k$ and hence equality follows similar to the proof of Lemma 5.1.

In order to show that $\mathcal{L}_k = \mathcal{M}_k$ it is sufficient to show that $\Pi_2^{\mathbb{P}}\mathbb{P}^{-1} = \Pi_2^{\mathbf{P}}\mathbf{P}^{-1}$. We have $\Pi_2^{\mathbb{P}}\mathbb{P}^{-1} = \mathbb{P}^{-1} - \mathbf{xu}^H\mathbb{P}^{-1}$ and using (5.30) we get

$$\begin{aligned}\Pi_2^{\mathbb{P}}\mathbb{P}^{-1} &= \mathbf{P}^{-1} - \frac{(\mathbf{P}^{-1}\mathbf{Mx}-\mathbf{x})\mathbf{u}^H\mathbf{P}^{-1}}{\mathbf{u}^H\mathbf{P}^{-1}\mathbf{Mx}} - \mathbf{xu}^H\mathbf{P}^{-1} + \mathbf{xu}^H\frac{(\mathbf{P}^{-1}\mathbf{Mx}-\mathbf{x})\mathbf{u}^H\mathbf{P}^{-1}}{\mathbf{u}^H\mathbf{P}^{-1}\mathbf{Mx}}\\ &= \mathbf{P}^{-1} - \frac{\mathbf{P}^{-1}\mathbf{Mxu}^H\mathbf{P}^{-1}}{\mathbf{u}^H\mathbf{P}^{-1}\mathbf{Mx}} = \Pi_2^{\mathbf{P}}\mathbf{P}^{-1}.\end{aligned}$$

Hence we have the claimed equivalence of $\mathcal{K}_k$, $\mathcal{L}_k$ and $\mathcal{M}_k$. □

Finally, adding the assumption $\mathbf{w} = \mathbf{Mx}$, Theorem 5.4 can be generalised to the following result:

Theorem 5.15. *Let the assumptions of Lemma 5.14 hold and furthermore choose* $\mathbf{w} = \mathbf{Mx}$. *Let* $\mathbf{y}_{k+1}^{RQ}$ *and* $\mathbf{s}_k^{JD}$ *be the approximate solutions to*

$$(\mathbf{A} - \rho(\mathbf{x})\mathbf{M})\mathbb{P}^{-1}\tilde{\mathbf{y}} = \mathbf{Mx}, \quad \textit{with} \quad \mathbf{y} = \mathbb{P}^{-1}\tilde{\mathbf{y}}, \tag{5.32}$$

and

$$(\mathbf{I} - \frac{\mathbf{Mxx}^H\mathbf{M}^H}{\mathbf{x}^H\mathbf{M}^H\mathbf{Mx}})(\mathbf{A} - \rho(\mathbf{x})\mathbf{M})(\mathbf{I} - \mathbf{xu}^H)\tilde{\mathbf{P}}^{\dagger}\tilde{\mathbf{s}} = -\mathbf{r}, \quad \textit{with} \quad \mathbf{s} = \tilde{\mathbf{P}}^{\dagger}\tilde{\mathbf{s}}, \tag{5.33}$$

respectively, obtained by $k+1$ *(*k*, respectively) steps of the same Galerkin-Krylov method with starting vector zero. Then there exists a constant* $c \in \mathbb{C}$ *such that*

$$\mathbf{y}_{k+1}^{RQ} = c(\mathbf{x} + \mathbf{s}_k^{JD}). \tag{5.34}$$

Proof. The argument is similar to the one for Theorem 5.4. Hence we give only a sketch of the proof here. We compare the solution $\mathbf{s}_k^{JD}$ to (5.33) and then the solution $\mathbf{y}_{k+1}^{RQ}$ to (5.32).

(a) The solution $\mathbf{s}_k^{JD}$ to (5.33).
With $\mathbf{r} = (\mathbf{A} - \rho(\mathbf{x})\mathbf{M})\mathbf{x}$ and $\mathcal{L}_k = \mathcal{M}_k$ from Lemma 5.14 the Krylov subspace for the solution $\tilde{\mathbf{s}}_k^{JD}$ of (5.33) is given by

$$\mathrm{span}\{\mathbf{r}, \Pi_1(\mathbf{A} - \rho(\mathbf{x})\mathbf{M})\Pi_2^{\mathbb{P}}\mathbb{P}^{-1}\mathbf{r}, \ldots, (\Pi_1(\mathbf{A} - \rho(\mathbf{x})\mathbf{M})\Pi_2^{\mathbb{P}}\mathbb{P}^{-1})^{k-1}\mathbf{r}\}.$$

If $\mathbf{V}_k$ is an orthogonal basis of this subspace then $\mathbf{Mx} \perp \mathbf{V}_k$, so that $\mathbf{V}_k^H\mathbf{Mx} = 0$ and $\mathbf{V}_k^H\Pi_1 = \mathbf{V}_k^H$. Similar to Theorem 5.4 the Galerkin-Krylov solution is then given by

$$\tilde{\mathbf{s}}_k^{JD} = -\mathbf{V}_k(\mathbf{V}_k^H(\mathbf{A} - \rho(\mathbf{x})\mathbf{M})\Pi_2^{\mathbb{P}}\mathbb{P}^{-1}\mathbf{V}_k)^{-1}\mathbf{V}_k^H\mathbf{Ax}.$$

and

$$\mathbf{s}_k^{JD} = -\Pi_2^{\mathbb{P}}\mathbb{P}^{-1}\mathbf{V}_k(\mathbf{V}_k^H(\mathbf{A} - \rho(\mathbf{x})\mathbf{M})\Pi_2^{\mathbb{P}}\mathbb{P}^{-1}\mathbf{V}_k)^{-1}\mathbf{V}_k^H\mathbf{Ax} \tag{5.35}$$

is an approximate Galerkin solution to (5.33) after k steps of the method. We can rewrite $\mathbf{s}_k^{JD}$ in the following way. First $\mathbf{V}_k^H\mathbf{Mx} = 0$ applied to (5.35) gives

$$\mathbf{s}_k^{JD} = -\Pi_2^{\mathbb{P}}\mathbb{P}^{-1}\mathbf{V}_k(\mathbf{V}_k^H(\mathbf{A} - \rho(\mathbf{x})\mathbf{M})\mathbb{P}^{-1}\mathbf{V}_k - \mathbf{V}_k^H\mathbf{Axu}^H\mathbb{P}^{-1}\mathbf{V}_k)^{-1}\mathbf{V}_k^H\mathbf{Ax},$$

and with the Sherman-Morrison formula we obtain

$$\mathbf{s}_k^{JD} = -\Pi_2^{\mathbb{P}}\mathbb{P}^{-1}\mathbf{V}_k\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}\left(1 + \frac{\mathbf{u}^H\mathbb{P}^{-1}\mathbf{V}_k\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}}{1-\mathbf{u}^H\mathbb{P}^{-1}\mathbf{V}_k\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}}\right),$$

where $\mathbf{S}_k = (\mathbf{V}_k^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{M})\mathbb{P}^{-1}\mathbf{V}_k$ and we assume that $\mathbf{u}^H\mathbb{P}^{-1}\mathbf{V}_k\mathbf{S}_k^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x} \neq 1$. Finally, using the definition of $\Pi_2^{\mathbb{P}}$ we can rewrite $\mathbf{s}_k^{JD}$ as

$$\begin{aligned}\mathbf{s}_k^{JD} = &-\mathbb{P}^{-1}\mathbf{V}_k(\mathbf{V}_k^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{M})\mathbb{P}^{-1}\mathbf{V}_k)^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}- \\ &\left(\mathbb{P}^{-1}\mathbf{V}_k(\mathbf{V}_k^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{M})\mathbb{P}^{-1}\mathbf{V}_k)^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}-\mathbf{x}\right)\xi,\end{aligned}$$

where ξ is a constant given by

$$\xi = \frac{\mathbf{u}^H\mathbb{P}^{-1}\mathbf{V}_k(\mathbf{V}_k^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{M})\mathbb{P}^{-1}\mathbf{V}_k)^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}}{1-\mathbf{u}^H\mathbb{P}^{-1}\mathbf{V}_k(\mathbf{V}_k^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{M})\mathbb{P}^{-1}\mathbf{V}_k)^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}}. \tag{5.36}$$

(b) The solution $\mathbf{y}_{k+1}^{RQ}$ to (5.32).
With Lemma 5.14 (with $\mathbf{A}$ replaced by $\mathbf{A}-\rho(\mathbf{x})\mathbf{M}$), the columns of $[\mathbf{M}\mathbf{x}, \mathbf{V}_k]$ form an orthogonal basis of

$$\operatorname{span}\{\mathbf{M}\mathbf{x}, (\mathbf{A}-\rho(\mathbf{x})\mathbf{M})\mathbb{P}^{-1}\mathbf{x}, ((\mathbf{A}-\rho(\mathbf{x})\mathbf{M})\mathbb{P}^{-1})^2\mathbf{x}, \dots, ((\mathbf{A}-\rho(\mathbf{x})\mathbf{M})\mathbb{P}^{-1})^k\mathbf{x}\},$$

which is the same space as generated by the Krylov subspace method applied to (5.32). Then the approximate solution to (5.32) is given by $\tilde{\mathbf{y}}_{k+1}^{RQ} = h\mathbf{M}\mathbf{x} + \mathbf{V}_k\mathbf{h}$, where $h \in \mathbb{C}$ and $\mathbf{h} \in \mathbb{C}^k$. The values of h and $\mathbf{h}$ are determined by imposing the Galerkin condition

$$\begin{bmatrix} \mathbf{x}^H\mathbf{M}^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{M})\mathbb{P}^{-1}\mathbf{M}\mathbf{x} & \mathbf{x}^H\mathbf{M}^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{M})\mathbb{P}^{-1}\mathbf{V}_k \\ \mathbf{V}_k^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{M})\mathbb{P}^{-1}\mathbf{M}\mathbf{x} & \mathbf{V}_k^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{M})\mathbb{P}^{-1}\mathbf{V}_k \end{bmatrix}\begin{bmatrix} h \\ \mathbf{h}\end{bmatrix} = \begin{bmatrix} 1 \\ \mathbf{0}\end{bmatrix}.$$

Note that $\mathbf{x}^H\mathbf{M}^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{M})\mathbb{P}^{-1}\mathbf{M}\mathbf{x} = \mathbf{x}^H\mathbf{M}^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{M})\mathbf{x} = 0$. From the second row we have

$$\mathbf{V}_k^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{M})\mathbb{P}^{-1}\mathbf{M}\mathbf{x}h + \mathbf{V}_k^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{M})\mathbb{P}^{-1}\mathbf{V}_k\mathbf{h} = 0$$

and hence with $\mathbb{P}^{-1}\mathbf{M}\mathbf{x} = \mathbf{x}$ and $\mathbf{V}_k^H\mathbf{M}\mathbf{x} = 0$

$$\mathbf{h} = -(\mathbf{V}_k^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{M})\mathbb{P}^{-1}\mathbf{V}_k)^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}h$$

and therefore from $\tilde{\mathbf{y}}_{k+1}^{RQ} = h\mathbf{M}\mathbf{x} + \mathbf{V}_k\mathbf{h}$

$$\tilde{\mathbf{y}}_{k+1}^{RQ} = h(\mathbf{M}\mathbf{x} - \mathbf{V}_k(\mathbf{V}_k^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{M})\mathbb{P}^{-1}\mathbf{V}_k)^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x})$$

and

$$\mathbf{y}_{k+1}^{RQ} = h(\mathbf{x} - \mathbb{P}^{-1}\mathbf{V}_k(\mathbf{V}_k^H(\mathbf{A}-\rho(\mathbf{x})\mathbf{M})\mathbb{P}^{-1}\mathbf{V}_k)^{-1}\mathbf{V}_k^H\mathbf{A}\mathbf{x}), \tag{5.37}$$

where we have used (5.29).

As in the proof of Theorem 5.4 (a) and (b) can be combined to obtain the result. ☐

Note that there is no restriction on the choice of $\mathbf{u}$ used to normalise $\mathbf{x}$. Indeed, we give results for two different choices in the following example.

Example 5.16. *Consider a generalised eigenproblem* $\mathbf{Ax} = \lambda\mathbf{Mx}$*, where the matrix* $\mathbf{A}$ *is given by the matrix* `sherman5.mtx` *from the Matrix Market library [13], the same matrix as in Example 5.8. The matrix* $\mathbf{M}$ *is given by a tridiagonal matrix with entries* $2/3$ *on the diagonal and entries* $1/6$ *on the sub- and superdiagonal. We seek the eigenvector belonging to the smallest eigenvalue, use a fixed shift* $\sigma = 0$ *and an initial starting guess of all ones. We compare inexact inverse iteration with simplified inexact Jacobi-Davidson method and investigate the following approaches to preconditioning:*

(a) no preconditioner is used for the inner iteration.

(b) a standard preconditioner is used for the inner iteration.

(c) a tuned preconditioner with $\mathbb{P}\mathbf{x} = \mathbf{Mx}$ *is used for the inner iteration.*

We use FOM as a solver with incomplete LU factorisation with drop tolerance 0.005 *as preconditioner where appropriate. Furthermore, we carry out exactly* 10 *steps of preconditioned FOM for the inner solve in the simplified Jacobi-Davidson method, while precisely* 11 *steps of preconditioned FOM are taken for each inner solve in the inexact inverse iteration. We do this in order to verify (5.34). We also restrict the number of total outer solves to* 20*. Furthermore, we use two different choices for* $\mathbf{u}$*, namely*

(i) a constant $\mathbf{u}$ *given by a vector of all ones,*

(ii) a variable $\mathbf{u}^{(i)}$ *given by* $\mathbf{u}^{(i)} = \mathbf{M}^H\mathbf{Mx}^{(i)}$*, which changes at each outer iteration.*

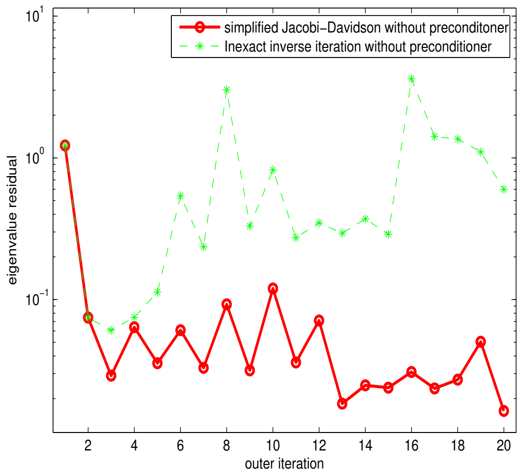


Figure 5-13: *Convergence history of the eigenvalue residuals for Example 5.16, case (a) and a constant* $\mathbf{u}$

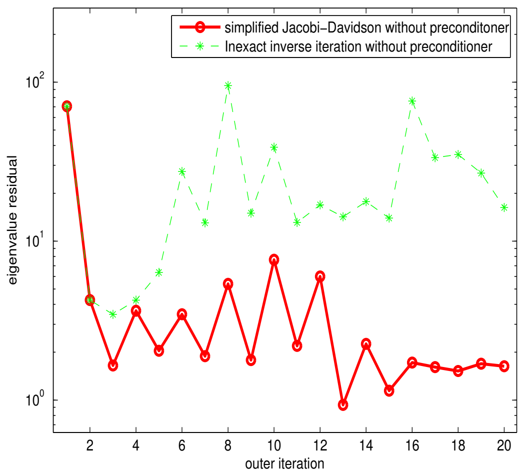


Figure 5-14: *Convergence history of the eigenvalue residuals for Example 5.16, case (a) and a variable* $\mathbf{u}^{(i)} = \mathbf{M}^H\mathbf{Mx}^{(i)}$

Figures 5-13 to 5-18 show the results for Example 5.16. We can make two observations: first of all, we see that only for case (c), when the tuned preconditioner is applied to inexact inverse iteration and a standard preconditioner is used with a simplified Jacobi-Davidson method, the convergence history of the eigenvalue residuals is the same (see Figures 5-17 and 5-18), as we would expect from Theorem 5.15. If no preconditioner is used (see Figures 5-13 and 5-14) or a standard preconditioner is applied (see Figures

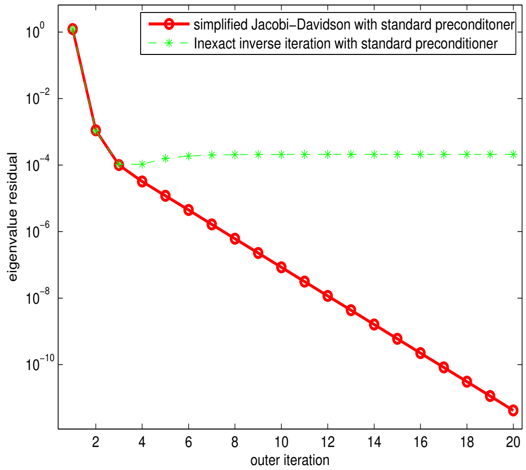


Figure 5-15: *Convergence history of the eigenvalue residuals for Example 5.16, case (b) and a constant* $\mathbf{u}$

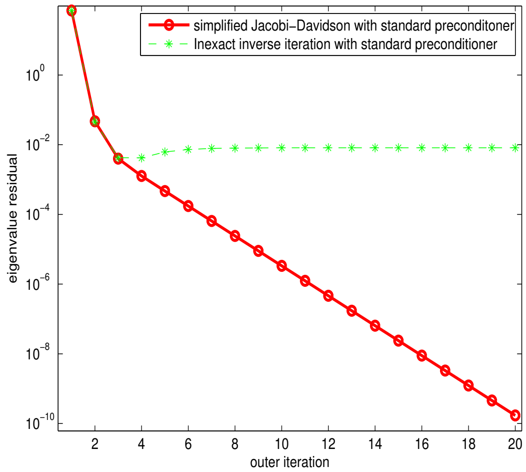


Figure 5-16: *Convergence history of the eigenvalue residuals for Example 5.16, case (b) and a variable* $\mathbf{u}^{(i)} = \mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}$

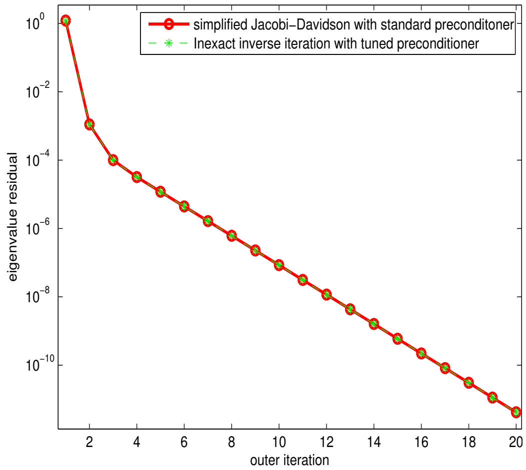


Figure 5-17: *Convergence history of the eigenvalue residuals for Example 5.16, case (c) and a constant* $\mathbf{u}$

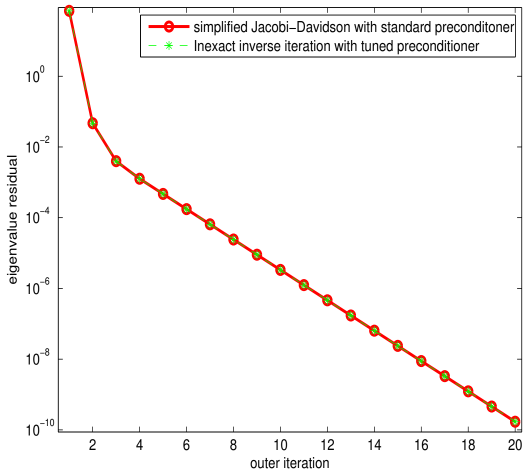


Figure 5-18: *Convergence history of the eigenvalue residuals for Example 5.16, case (c) and a variable* $\mathbf{u}^{(i)} = \mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}$

5-15 and 5-16), then inexact inverse iteration and the simplified Jacobi-Davidson are not equivalent. Secondly, we can use any vector $\mathbf{u}$ within the Jacobi-Davidson method (see Figures on the left compared to Figures on the right) and will get the same results.

As a summary, an advantage of inverse iteration with the tuned preconditioner in comparison to the Jacobi-Davidson method with the standard preconditioner is that for inexact inverse iteration one does not have to worry about the choice of the projections Π_1 and Π_2 in (5.25) to obtain equivalent results.

5.5 Conclusions

This section has provided an equivalence result for simplified Jacobi-Davidson and inverse iteration (or Rayleigh quotient iteration) if inexact solves are applied for the inner iteration. The results show that inexact inverse iteration if a simple modification to the preconditioner is implemented, is at least as efficient (in terms of total number of inner iterations) as simplified Jacobi-Davidson.

CHAPTER 6

Tuning the preconditioner for inexact inverse iteration applied to the generalised nonsymmetric eigenproblem

6.1 Introduction

In this chapter we consider the computation of a simple eigenvalue and corresponding eigenvector of the generalised eigenproblem $\mathbf{Ax} = \lambda\mathbf{Mx}$ where $\mathbf{A}$ and $\mathbf{M}$ are large sparse nonsymmetric matrices using inexact inverse iteration with fixed or variable shifts. We concentrate on iterative techniques for solving the inner linear system

$$(\mathbf{A} - \sigma\mathbf{M})\mathbf{y} = \mathbf{Mx} \tag{6.1}$$

arising at each step of inverse iteration, where the shift σ (fixed or variable) is chosen to be close to an eigenvalue.

The convergence theory for inverse iteration with inexact solves has been considered in Chapter 3 (see also [44]) for general shift strategies. This theory covers the most general setting, where $\mathbf{A}$ and $\mathbf{M}$ are nonsymmetric with $\mathbf{M}$ allowed to be singular. It was shown that, for a fixed shift strategy, a decreasing tolerance provides linear convergence, for an appropriately chosen variable shift fixing the solve tolerance gives linear convergence whereas a decreasing tolerance achieves quadratic convergence.

This chapter concentrates on the performance of the inner solver for the linear system in (6.1). In particular we consider unpreconditioned and preconditioned GMRES, although other Krylov methods are possible (see, for example [111]). For inexact inverse iteration the costs of the inner solves using Krylov methods has been investigated in [10] and Chapter 4 (see also [42]) for the symmetric solvers CG/MINRES and in [12] for GMRES. In these papers it was shown that, for the standard eigenvalue problem, the number of inner iterations remained approximately constant as the outer iteration proceeded if no preconditioner was used but increased if a standard preconditioner was applied. A so-called tuned preconditioner has been introduced in Chapter 4 (see also [42]) for the Hermitian standard positive definite eigenproblem and in Chapter 2 (see also [43]) for the generalised eigenproblem. Here we extend the results from Chapter 4 to the generalised nonsymmetric eigenproblem and give a detailed analysis of the costs of the inner solves as the outer iteration proceeds. We also extend the analysis to variable shift strategies. For the generalised eigenproblem it turns out that for both unpreconditioned and preconditioned GMRES a tuning strategy gives significant improvement.

The chapter is organised as follows. In Section 6.2 we briefly recall the convergence theory of Chapter 3 and give some further preliminary results. We also give a modified convergence theory of GMRES, where the right-hand side is taken into consideration. Section 6.3 describes a phenomenon that arises for the unpreconditioned generalised eigenproblem, namely an increase in the iteration numbers as the outer iteration proceeds. We introduce a tuning strategy and show how tuning decreases the overall costs significantly. In Sections 6.4 and 6.5 we consider preconditioned GMRES as a solver for (6.1) and prove how the tuned preconditioner increases efficiency in the inner solves. We analyse both fixed and variable shift strategies. In Section 6.6 we compare the concept of the tuned preconditioner to a simplified version of the preconditioned Jacobi-Davidson method.

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

6.2 Some preliminary results

This section recalls results on the outer convergence rate of inexact inverse iteration, and convergence theory for GMRES for the inner iteration.

6.2.1 Convergence of inexact inverse iteration

Consider the general nonsymmetric eigenproblem

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{M}\mathbf{x}, \quad \mathbf{x} \neq \mathbf{0}, \tag{6.2}$$

with $\mathbf{A} \in \mathbb{C}^{n\times n}$ and $\mathbf{M} \in \mathbb{C}^{n\times n}$, where we assume that $(\lambda_1, \mathbf{x}_1)$ is a simple, well-separated finite eigenpair with corresponding left eigenvector $\mathbf{u}_1^H$, that is $\mathbf{A}\mathbf{x}_1 = \lambda_1\mathbf{M}\mathbf{x}_1$ and $\mathbf{u}_1^H\mathbf{A} = \lambda_1\mathbf{u}_1^H\mathbf{M}$. We use inexact inverse iteration as described in Algorithm 4, which we do not repeat here.

As in Chapter 3 the eigenvalue is updated using the generalised Rayleigh quotient given by

$$\rho(\mathbf{x}^{(i)}) = \frac{\mathbf{x}^{(i)H}\mathbf{M}^H\mathbf{A}\mathbf{x}^{(i)}}{\mathbf{x}^{(i)H}\mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}}, \tag{6.3}$$

which has the desirable property that the eigenvalue residual

$$\mathbf{r}^{(i)} = \mathbf{A}\mathbf{x}^{(i)} - \rho(\mathbf{x}^{(i)})\mathbf{M}\mathbf{x}^{(i)} \tag{6.4}$$

is minimised.

This chapter is concerned with the costs in the inner iteration but for completeness we state the result on the convergence theory of the outer iteration, which is necessary to understand the inner iteration. For a simple eigenvalue λ_1 we can block-diagonalise $\mathbf{A} - \lambda\mathbf{M}$ as

$$\mathbf{U}^{-1}(\mathbf{A} - \lambda\mathbf{M})\mathbf{X} = \begin{bmatrix} t_{11} & \mathbf{0}^H \\ \mathbf{0} & \mathbf{T}_{22} \end{bmatrix} - \lambda \begin{bmatrix} s_{11} & \mathbf{0}^H \\ \mathbf{0} & \mathbf{S}_{22} \end{bmatrix}, \tag{6.5}$$

where $\lambda_1 = t_{11}/s_{11}$. Hence $\mathbf{x}_1 = \mathbf{X}\mathbf{e}_1$ and $\mathbf{u}_1 = \mathbf{U}^{-H}\mathbf{e}_1$ are the right and left eigenvectors corresponding to λ_1. We measure the deviation of $\mathbf{x}^{(i)}$ from $\mathbf{x}_1$ using the decomposition

$$\mathbf{x}^{(i)} = \alpha^{(i)}(\mathbf{x}_1 q^{(i)} + \mathbf{X}_2\mathbf{p}^{(i)}), \tag{6.6}$$

where $\alpha^{(i)} := \|\mathbf{U}^{-1}\mathbf{M}\mathbf{x}^{(i)}\|$, $q^{(i)} \in \mathbb{C}$, $\mathbf{p}^{(i)} \in \mathbb{C}^{(n-1)\times 1}$, $\mathbf{X}_2 = \mathbf{X}\bar{\mathbf{I}}_{n-1}$ and $\bar{\mathbf{I}}_{n-1} = \begin{bmatrix} \mathbf{0}^H \\ \mathbf{I}_{n-1} \end{bmatrix} \in \mathbf{C}^{n\times(n-1)}$ with $\mathbf{I}_{n-1}$ being the identity matrix of size $n-1$. Clearly $q^{(i)}$ and $\mathbf{p}^{(i)}$ measure how close the approximate eigenvector $\mathbf{x}^{(i)}$ is to the sought eigenvector $\mathbf{x}_1$. If we define the separation between λ_1 and the matrix pair $(\mathbf{T}_{22}, \mathbf{S}_{22})$ by the function (see [137])

$$\begin{aligned} \mathrm{sep}(\lambda_1, (\mathbf{T}_{22}, \mathbf{S}_{22})) &:= \inf_{\|\mathbf{a}\|_2=1} \|(\mathbf{T}_{22} - \lambda_1\mathbf{S}_{22})\mathbf{a}\| \\ &= \begin{cases} \|(\mathbf{T}_{22} - \lambda_1\mathbf{S}_{22})^{-1}\|^{-1}, & \lambda_1 \notin \lambda(\mathbf{T}_{22}, \mathbf{S}_{22}) \\ 0, & \lambda_1 \in \lambda(\mathbf{T}_{22}, \mathbf{S}_{22}) \end{cases}. \end{aligned}$$

we have the following lemma (which has been proved in Chapter 3, Lemmata 3.5 and 3.11) which provides bounds on the absolute error in the eigenvalue approximation $|\rho(\mathbf{x}^{(i)}) - \lambda_1|$ and on the eigenvalue residual, defined by (6.4) in terms of $\|\mathbf{p}^{(i)}\|$. For convenience, we repeat these results here.

Lemma 6.1 (Lemma 3.5 and Lemma 3.11). *For the generalised Rayleigh quotient $\rho(\mathbf{x}^{(i)})$ we have*

$$|\rho(\mathbf{x}^{(i)}) - \lambda_1| \le \kappa(\mathbf{U})\|\mathbf{T}_{22} - \lambda_1\mathbf{S}_{22}\|\|\mathbf{p}^{(i)}\| = \mathcal{O}(\|\mathbf{p}^{(i)}\|), \tag{6.7}$$

where $\kappa(\mathbf{U}) = \|\mathbf{U}\|\|\mathbf{U}^{-1}\|$ and the eigenvalue residual (6.4) satisfies

$$\|\mathbf{r}^{(i)}\| \le \kappa(\mathbf{U})\|\mathbf{T}_{22} - \lambda_1\mathbf{S}_{22}\|\|\mathbf{p}^{(i)}\| = \mathcal{O}(\|\mathbf{p}^{(i)}\|). \tag{6.8}$$

Furthermore

$$\|\mathbf{p}^{(i)}\| \le \frac{\|\mathbf{U}\|}{sep(\rho(\mathbf{x}^{(i)}), (\mathbf{T}_{22}, \mathbf{S}_{22}))}\|\mathbf{r}^{(i)}\|, \tag{6.9}$$

where $\mathbf{p}^{(i)}$ is given in (6.6).

If inexact inverse iteration converges, then $\|\mathbf{r}^{(i)}\| \to 0$ and Lemma 6.1 implies that

$$\|\mathbf{r}^{(i)}\| \to 0 \quad \text{if and only if} \quad \|\mathbf{p}^{(i)}\| \to 0. \tag{6.10}$$

We summarise the convergence of inexact inverse iteration in the following theorem.

Theorem 6.2 (see Theorem 3.7). *Consider the application of inexact inverse iteration to find a simple eigenvalue λ_1 with corresponding right eigenvector $\mathbf{x}_1$ of (6.2). Assume $\sigma^{(i)}$ is closer to λ_1 than to any other eigenvalue and $\mathbf{x}^{(0)}$ is close enough to $\mathbf{x}_1$. Let*

$$\mathbf{M}\mathbf{x}^{(i)} - (\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{d}^{(i)}, \quad \textit{with} \quad \|\mathbf{d}^{(i)}\| \le \tau^{(i)}\|\mathbf{M}\mathbf{x}^{(i)}\|$$

with $\tau^{(0)}$ small enough. Then,

1. *if the shift $\sigma^{(i)} := \sigma$ is fixed and the tolerance is decreasing $\tau^{(i)} = \delta\|\mathbf{r}^{(i)}\|$ then linear convergence is obtained in Algorithm 4 for small enough δ.*

2. *if the shift is chosen to be the Rayleigh quotient (6.3) $\sigma^{(i)} := \rho(\mathbf{x}^{(i)})$ and the tolerance is decreasing $\tau^{(i)} = \delta\|\mathbf{r}^{(i)}\|$ then quadratic convergence is obtained in Algorithm 4 for small enough δ. For a fixed $\tau^{(i)} := \tau^{(0)}$ we obtain linear convergence.*

We will see that the choice of $\tau^{(i)}$ is crucial for the efficiency of the inner iterations.

6.2.2 The inner iteration

Inexact inverse iteration requires solving a system

$$(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbf{M}\mathbf{x}^{(i)} \tag{6.11}$$

at each outer step i. For the moment we assume that we use a fixed shift strategy, that is, $\sigma^{(i)} = \sigma$. A popular method to solve the linear system (6.11) iteratively is GMRES. In Section 6.2.3 we state a convergence result for GMRES applied to the system

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

where $\mathbf{B}$ has a well-separated simple eigenvalue near zero. This theory is general in the sense that it does not need the system matrix $\mathbf{B}$ to be diagonalisable. Furthermore we give results on the number of inner iterations per outer iteration for GMRES, depending on the right hand side $\mathbf{b}$.

This subsection contains some technical results on the system matrix $\mathbf{B}$, that we need for the convergence theory of GMRES in the next subsection. We summarise some theoretical results assuming that $\mathbf{B}$ has an algebraically simple eigenpair $(\mu_1, \mathbf{w}_1)$.

Schur's theorem [48, page 313] ensures the existence of a unitary matrix $[\mathbf{w}_1, \mathbf{W}_1^{\perp}]$ such that

$$\mathbf{B} = \begin{bmatrix} \mathbf{w}_1 & \mathbf{W}_1^{\perp} \end{bmatrix} \begin{bmatrix} \mu_1 & \mathbf{n}_{12}^H \\ \mathbf{0} & \mathbf{N}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 & \mathbf{W}_1^{\perp} \end{bmatrix}^H, \tag{6.12}$$

where $\mathbf{w}_1 \in \mathbb{C}^{n\times 1}$, $\mathbf{W}_1^{\perp} \in \mathbb{C}^{n\times(n-1)}$, $\mathbf{n}_{12} \in \mathbb{C}^{(n-1)\times 1}$ and $\mathbf{N}_{22} \in \mathbb{C}^{(n-1)\times(n-1)}$. If μ_1 is not contained in the spectrum of $\mathbf{N}_{22}$ the equation

$$\mathbf{f}^H\mathbf{N}_{22} - \mu_1\mathbf{f}^H = \mathbf{n}_{12}^H \tag{6.13}$$

has a unique solution $\mathbf{f} \in \mathbb{C}^{(n-1)\times 1}$ (see [48, Lemma 7.1.5]) and with

$$\mathbf{W}_2 = (-\mathbf{w}_1\mathbf{f}^H + \mathbf{W}_1^{\perp})(\mathbf{I} + \mathbf{f}\mathbf{f}^H)^{-\frac{1}{2}} \tag{6.14}$$

and $\mathbf{C} = (\mathbf{I} + \mathbf{f}\mathbf{f}^H)^{-\frac{1}{2}}\mathbf{N}_{22}(\mathbf{I} + \mathbf{f}\mathbf{f}^H)^{\frac{1}{2}}$ we obtain the following block-diagonalisation of $\mathbf{B}$:

$$\mathbf{B} = \begin{bmatrix} \mathbf{w}_1 & \mathbf{W}_2 \end{bmatrix} \begin{bmatrix} \mu_1 & \mathbf{0}^H \\ \mathbf{0} & \mathbf{C} \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^H \\ \mathbf{V}_2^H \end{bmatrix}, \tag{6.15}$$

where $[\mathbf{v}_1, \mathbf{V}_2]^H$ is the inverse of the nonsingular matrix $[\mathbf{w}_1, \mathbf{W}_2]$, see [136, Theorem 1.18]. We have

$$\mathbf{v}_1 = \mathbf{w}_1 + \mathbf{W}_1^{\perp}\mathbf{f} \quad \text{and} \quad \mathbf{V}_2 = \mathbf{W}_1^{\perp}(\mathbf{I} + \mathbf{f}\mathbf{f}^H)^{\frac{1}{2}}.$$

Note that $\mathbf{C}$ and $\mathbf{N}_{22}$ have the same spectrum.

Hence, $\mathbf{B}\mathbf{w}_1 = \mu_1\mathbf{w}_1$ and $\mathbf{v}_1^H\mathbf{B} = \mu_1\mathbf{v}_1^H$, that is, $\mathbf{v}_1$ is the left eigenvector of $\mathbf{B}$ corresponding to μ_1. Note that $\mathbf{v}_1^H\mathbf{w}_1 = 1$, $\mathbf{V}_2^H\mathbf{W}_2 = \mathbf{I}$, $\mathbf{v}_1^H\mathbf{W}_2 = \mathbf{0}^H$ and $\mathbf{V}_2^H\mathbf{w}_1 = \mathbf{0}$. Also $\|\mathbf{w}_1\| = 1$, and $\mathbf{W}_2$ has orthonormal columns.

Further, define the separation function $\operatorname{sep}(\mu_1, \mathbf{C})$ (see, for example [134] or [137]) by

$$\operatorname{sep}(\mu_1, \mathbf{C}) := \begin{cases} \|(\mu_1\mathbf{I} - \mathbf{C})^{-1}\|_2^{-1}, & \mu_1 \notin \Lambda(\mathbf{C}) \\ 0, & \mu_1 \in \Lambda(\mathbf{C}) \end{cases}$$

and note that by definition

$$\operatorname{sep}(\mu_1, \mathbf{C}) = \sigma_{\min}(\mu_1 \mathbf{I} - \mathbf{C}),$$

where $\sigma_{\min}$ is the minimum singular value. We may say that the function sep roughly measures the separation of μ_1 from the eigenvalues of $\mathbf{C}$, since (see for example [48])

$$\operatorname{sep}(\mu_1, \mathbf{C}) \leq \min_{\mu \in \Lambda(\mathbf{C})} |\mu_1 - \mu|.$$

Also, the quantities $\operatorname{sep}(\mu_1, \mathbf{C})$ and $\operatorname{sep}(\mu_1, \mathbf{N}_{22})$ are related by (see [137])

$$\frac{\operatorname{sep}(\mu_1, \mathbf{N}_{22})}{\kappa} \leq \operatorname{sep}(\mu_1, \mathbf{C}) \leq \kappa \operatorname{sep}(\mu_1, \mathbf{N}_{22}), \quad \text{where} \quad \kappa = \sqrt{1 + \mathbf{f}^H \mathbf{f}}.$$

In order to provide a general convergence theory for GMRES applied to $\mathbf{Bz} = \mathbf{b}$ depending on the right hand side $\mathbf{b}$ of the system we further need the definition of an oblique projector (see, for example, [110, 140] and [137]).

Definition 6.3 (Oblique Projections). *An oblique projector $\mathcal{P}$ is a linear transformation from $\mathbb{C}^n$ to itself which satisfies*

$$\mathcal{P}^2 = \mathcal{P}.$$

Let $\mathbf{X} \in \mathbb{C}^{n \times k}$ and $\mathbf{Y} \in \mathbb{C}^{n \times k}$ be rectangular matrices and let $\mathcal{X} = \mathcal{R}(\mathbf{X})$ and $\mathcal{Y} = \mathcal{R}(\mathbf{Y})$ define k-dimensional subspaces of $\mathbb{C}^n$. Any oblique projector $\mathcal{P}$ can be written in the form

$$\mathcal{P} = \mathbf{X}\mathbf{Y}^H, \quad \mathbf{Y}^H \mathbf{X} = \mathbf{I},$$

where $\mathcal{X} = \mathcal{R}(\mathbf{X})$ and $\mathcal{Y} = \mathcal{R}(\mathbf{Y})$ uniquely define $\mathcal{P}$, which is said to project onto $\mathcal{X}$ along the orthogonal complement $\mathcal{Y}_\perp$ of $\mathcal{Y}$.

If $\mathcal{P}$ is an oblique projector onto $\mathcal{X}$ along $\mathcal{Y}_\perp$, then $\mathbf{I} - \mathcal{P}$ is its complementary projector and it projects onto $\mathcal{Y}_\perp$ along $\mathcal{X}$. Any vector $\mathbf{z}$ can be represented as the sum of a vector $\mathcal{P}\mathbf{z} \in \mathcal{X}$ and a vector $(\mathbf{I} - \mathcal{P})\mathbf{z} \in \mathcal{Y}_\perp$. As a result the space $\mathbb{C}^n$ can be decomposed as the direct sum

$$\mathbb{C}^n = \mathcal{N}(\mathcal{P}) \oplus \mathcal{R}(\mathcal{P}).$$

Note that $\mathbf{X}$ and $\mathbf{Y}$ are chosen as general matrices in Definition 6.3: we make this choice specific after the following remark.

Remark 6.4. *In the special case of $\mathcal{X} = \mathcal{Y}$ we have $\mathcal{P} = \mathbf{X}\mathbf{X}^H$, where $\mathbf{X}$ is an orthonormal basis for $\mathcal{R}(\mathcal{P})$. In this case we speak of orthogonal projections instead of oblique projections and we may set $\mathcal{P} := \mathcal{P}^\perp$ to emphasize the orthogonality in the projections for this special case. They are useful for Hermitian problems, see Chapter 4.*

Our oblique projection of interest is

$$\mathcal{P} = \mathbf{I} - \mathbf{w}_1 \mathbf{v}_1^H, \tag{6.16}$$

which projects onto $\mathcal{R}(\mathbf{W}_2)$ along $\mathcal{R}(\mathbf{w}_1)$ and $\mathbf{I} - \mathcal{P}$ projects onto $\mathcal{R}(\mathbf{w}_1)$ along the orthogonal complement $\mathcal{R}(\mathbf{W}_2)$ of $\mathcal{R}(\mathbf{v}_1)$. With the results after (6.15) we see that

$$\|\mathcal{P}\| = \sqrt{1 + \|\mathbf{f}\|^2}.$$

Before analysing the GMRES iteration we state a proposition which follows from the perturbation theory of eigenvectors belonging to simple eigenvalues (see [136], [137] and [134] and Appendix C) and holds for small enough perturbations $\mathbf{E}$ of the matrix $\mathbf{B}$.

Proposition 6.5. *Let μ_1 be a simple eigenvalue of $\mathbf{B}$ with corresponding right eigenvector $\mathbf{w}_1$ and let $\mathbf{W} = [\mathbf{w}_1, \mathbf{W}_1^\perp]$ be unitary such that (6.12) holds. Let*

$$\mathbf{B}\hat{\mathbf{w}} = \xi\hat{\mathbf{w}} + \hat{\mathbf{e}} \tag{6.17}$$

be a perturbed problem with $\hat{\mathbf{e}}$ small enough such that $\|\hat{\mathbf{e}}\| < \frac{1}{2}sep(\mu_1, \mathbf{N}_{22})$ and where $\|\hat{\mathbf{w}}\| = 1$. Then

$$\hat{\mathbf{w}} = \frac{\mathbf{w}_1 + \mathbf{W}_1^\perp \mathbf{p}}{\sqrt{1 + \mathbf{p}^H\mathbf{p}}},$$

where

$$\|\mathbf{p}\| \leq 2\frac{\|\hat{\mathbf{e}}\|}{sep(\mu_1, \mathbf{N}_{22}) - 2\|\hat{\mathbf{e}}\|}.$$

Proof. Write (6.17) as

$$(\mathbf{B} - \hat{\mathbf{e}}\hat{\mathbf{w}}^H)\hat{\mathbf{w}} = \xi\hat{\mathbf{w}}. \tag{6.18}$$

Using $\mathbf{E} = -\hat{\mathbf{e}}\hat{\mathbf{w}}^H$ we can apply Theorem C.1 and with $\|\mathbf{E}\| = \|\hat{\mathbf{e}}\hat{\mathbf{w}}^H\| = \|\hat{\mathbf{e}}\|$ as well as Remark C.2 the result follows. □

Proposition 6.5 shows that the eigenvector $\hat{\mathbf{w}}$ of the perturbed problem (6.18) compared to the exact problem $\mathbf{B}\mathbf{w}_1 = \mu_1\mathbf{w}_1$ depends on the size of the norm of the perturbation $\hat{\mathbf{e}}$ and on the separation of the eigenvalue μ_1 from the rest of the spectrum.

6.2.3 Convergence theory for GMRES

Using the block factorisation (6.15) and the oblique projector (6.16) we have the following convergence result for GMRES applied to the linear system $\mathbf{B}\mathbf{z} = \mathbf{b}$.

Theorem 6.6 (GMRES convergence). *Suppose the nonsymmetric matrix $\mathbf{B} \in \mathbb{C}^{n\times n}$ has a simple eigenpair $(\mu_1, \mathbf{w}_1)$ with block diagonalisation (6.15). Let μ_1 be well-separated from the eigenvalues of $\mathbf{C}$ and let $\mathcal{P} = \mathbf{I} - \mathbf{w}_1\mathbf{v}_1^H$. Let $\mathbf{z}_k$ be the result of applying GMRES to $\mathbf{B}\mathbf{z} = \mathbf{b}$ with starting value $\mathbf{z}_0 = \mathbf{0}$. Then*

$$\|\mathbf{b} - \mathbf{B}\mathbf{z}_k\| \leq \min_{p_{k-1}\in\Pi_{k-1}} \|p_{k-1}(\mathbf{C})\| \left(\frac{\|\mu_1\mathbf{I} - \mathbf{C}\|}{|\mu_1|}\right) \|\mathbf{V}_2\|\|\mathcal{P}\mathbf{b}\|, \tag{6.19}$$

where Π_{k-1} is the set of complex polynomials of degree $k-1$ normalised such that $p(0) = 1$ and $\|\mathbf{V}_2\| = \sqrt{1 + \|\mathbf{f}\|^2}$ where $\mathbf{f}$ is given by (6.13).

Proof. The residual for GMRES satisfies (see [55])

$$\|\mathbf{b} - \mathbf{B}\mathbf{z}_k\|_2 = \min_{p_k\in\Pi_k} \|p_k(\mathbf{B})\mathbf{b}\|_2,$$

where Π_k is the set of polynomials of degree k with $p(0) = 1$. Introduce special polynomials $\hat{p}_k \in \Pi_k$, given by

$$\hat{p}_k(z) = p_{k-1}(z)\left(1 - \frac{z}{\mu_1}\right),$$

where $p_{k-1} \in \Pi_{k-1}$. Note that similar polynomials were introduced by Campbell et al. [15]. Then we can write

$$\begin{aligned}
\|\mathbf{b} - \mathbf{B}\mathbf{z}_k\|_2 &= \min_{p_k \in \Pi_k} \|p_k(\mathbf{B})\mathcal{P}\mathbf{b} + p_k(\mathbf{B})(\mathbf{I} - \mathcal{P})\mathbf{b}\|_2 \\
&\leq \min_{\hat{p}_k \in \Pi_k} \|\hat{p}_k(\mathbf{B})\mathcal{P}\mathbf{b} + \hat{p}_k(\mathbf{B})(\mathbf{I} - \mathcal{P})\mathbf{b}\|_2 \\
&\leq \min_{p_{k-1} \in \Pi_{k-1}} \|p_{k-1}(\mathbf{B}) \left(\mathbf{I} - \frac{\mathbf{B}}{\mu_1}\right) \mathcal{P}\mathbf{b} + p_{k-1}(\mathbf{B}) \left(\mathbf{I} - \frac{\mathbf{B}}{\mu_1}\right) (\mathbf{I} - \mathcal{P})\mathbf{b}\|_2.
\end{aligned}$$

For the second term we have

$$\left(\mathbf{I} - \frac{\mathbf{B}}{\mu_1}\right)(\mathbf{I} - \mathcal{P})\mathbf{b} = \left(\mathbf{I} - \frac{\mathbf{B}}{\mu_1}\right)\mathbf{w}_1\mathbf{v}_1^H\mathbf{b} = \mathbf{0},$$

using (6.16) and $\mathbf{B}\mathbf{w}_1 = \mu_1\mathbf{w}_1$. Therefore

$$\|\mathbf{b} - \mathbf{B}\mathbf{z}_k\| \leq \min_{p_{k-1} \in \Pi_{k-1}} \|p_{k-1}(\mathbf{B}) \left(\mathbf{I} - \frac{\mathbf{B}}{\mu_1}\right) \mathcal{P}\mathbf{b}\|. \tag{6.20}$$

With $\mathcal{P}^2 = \mathcal{P}$ and $\mathcal{P}\mathbf{B} = \mathbf{B}\mathcal{P}$ we have

$$\begin{aligned}
p_{k-1}(\mathbf{B}) \left(\mathbf{I} - \frac{\mathbf{B}}{\mu_1}\right) \mathcal{P}\mathbf{b} &= p_{k-1}(\mathbf{W}_2\mathbf{C}\mathbf{V}_2^H) \left(\mathbf{I} - \frac{\mathbf{B}}{\mu_1}\right) \mathcal{P}\mathbf{b} \\
&= \mathbf{W}_2 p_{k-1}(\mathbf{C}) \left(\frac{\mu_1\mathbf{I} - \mathbf{C}}{\mu_1}\right) \mathbf{V}_2^H \mathcal{P}\mathbf{b},
\end{aligned}$$

and hence

$$\|\mathbf{b} - \mathbf{B}\mathbf{z}_k\|_2 \leq \min_{p_{k-1} \in \Pi_{k-1}} \|p_{k-1}(\mathbf{C})\| \left(\frac{\|\mu_1\mathbf{I} - \mathbf{C}\|}{|\mu_1|}\right) \|\mathbf{V}_2\| \|\mathcal{P}\mathbf{b}\|, \tag{6.21}$$

since $\mathbf{W}_2$ has orthonormal columns. □

Note that the minimum in (6.19) is taken with respect to the smaller matrix $\mathbf{C}$ instead of $\mathbf{B}$. In order to bound $\min_{p_{k-1} \in \Pi_{k-1}} \|p_{k-1}(\mathbf{C})\|$ we apply some further theory.

For a general convergence theory of GMRES we will use the definition of the ε-pseudospectrum (see, for example [35]):

Definition 6.7. *The ε-pseudospectrum $\Lambda_\varepsilon(\mathbf{C})$ of a matrix $\mathbf{C}$ is defined by*

$$\Lambda_\varepsilon(\mathbf{C}) := \{z \in \mathbb{C} : \|(z\mathbf{I} - \mathbf{C})^{-1}\|_2 \geq \varepsilon^{-1}\}. \tag{6.22}$$

We would like to note that other approaches for the following proposition, like the use of the field of values (see [65] and Appendix B), would also be possible. We can further bound the term in the brackets of (6.19) using results from complex analysis and Faber polynomials (see Appendix B). We have the following proposition.

Proposition 6.8. *Let E be a convex closed bounded set in the complex plane satisfying $0 \notin E$, containing the ε-pseudospectrum $\Lambda_\varepsilon(\mathbf{C})$. Let Ψ be the conformal mapping that carries the exterior of E onto the exterior of the unit circle $\{|w| > 1\}$ and that takes infinity to infinity. Then*

$$\min_{p_{k-1} \in \Pi_{k-1}} \|p_{k-1}(\mathbf{C})\| \leq S \left(\frac{1}{|\Psi(0)|}\right)^{k-1}, \quad \textit{where} \quad S = \frac{3\mathcal{L}(\Gamma_\varepsilon)}{2\pi\varepsilon} \tag{6.23}$$

and $|\Psi(0)| > 1$ *and hence*

$$\|\mathbf{b} - \mathbf{B}\mathbf{z}_k\| \leq S \left(\frac{1}{|\Psi(0)|} \right)^{k-1} \|\mathbf{V}_2\| \left(\frac{\|\mu_1 \mathbf{I} - \mathbf{C}\|}{|\mu_1|} \right) \|\mathcal{P}\mathbf{b}\|, \tag{6.24}$$

for any choice of the parameter ε*, where* $\mathcal{L}(\Gamma_\varepsilon)$ *is the contour length of* Γ_ε *and* Γ_ε *is the contour or union of contours enclosing* $\Lambda_\varepsilon(\mathbf{C})$*.*

Proof. With (B.4) in Appendix B (with $\mathbf{B}$ replaced by $\mathbf{C}$ and k replaced by $k-1$) we obtain

$$\|p_{k-1}(\mathbf{C})\| \leq \frac{\mathcal{L}(\Gamma_\varepsilon)}{2\pi\varepsilon} \max_{z \in \Lambda_\varepsilon(\mathbf{C})} |p_{k-1}(z)|,$$

where $\mathcal{L}(\Gamma_\varepsilon)$ is the contour length of Γ_ε. An approximation problem

$$\min_{p_{k-1} \in \Pi_{k-1}} \max_{z \in \Lambda_\varepsilon(\mathbf{C})} |p_{k-1}(z)|$$

remains to be solved and by Theorem B.1 in Appendix B with k replaced by $k-1$ we obtain

$$\min_{p_{k-1} \in \Pi_{k-1}} \max_{z \in \Lambda_\varepsilon(\mathbf{C})} |p_{k-1}(z)| \leq \frac{3}{|\Psi(0)|^{k-1}} \tag{6.25}$$

As $0 \notin E$ and as Ψ maps the exterior of E onto the exterior of a unit disc we have $|\Psi(0)| > 1$ and hence, with $\Lambda_\varepsilon(\mathbf{C}) \subset E$ and (6.23) we obtain (6.24) from (6.19). □

The following corollary is immediately obtained from Proposition 6.8.

Corollary 6.9. *Let* $\mathbf{C}$ *be perturbed to* $\mathbf{C} + \delta\mathbf{C}$*, where* $\|\delta\mathbf{C}\| < \varepsilon$*. Then*

$$\min_{p_{k-1} \in \Pi_{k-1}} \|p_{k-1}(\mathbf{C} + \delta\mathbf{C})\| \leq S_\delta \left(\frac{1}{|\Psi(0)|} \right)^{k-1}, \tag{6.26}$$

where $S_\delta = \dfrac{3\mathcal{L}(\Gamma_\varepsilon)}{2\pi(\varepsilon - \|\delta\mathbf{C}\|)}$.

Note that the bound in (6.24) describes the convergence behaviour in the worst-case sense and is by no means sharp. For further details we refer to [82]. Furthermore simpler bounds using Chebychev polynomials can be derived if $\mathbf{B}$ is diagonalisable and the eigenvalues are located in an ellipse or circle (see Saad [111] and Appendix B). Also Proposition 6.8 remains valid if the ε-pseudospectrum of $\mathbf{C}$ is replaced by the field of values of $\mathbf{C}$. Then the constant S in (6.23) is replaced by the smaller $S = \dfrac{3\mathcal{L}(\partial E)}{2\pi d(\partial E)}$, where $d(\partial E)$ is the minimal distance between the field of values of $\mathbf{C}$ and ∂E, the boundary of E. However, the advantage of the pseudospectral approach is that the set $\Lambda_\varepsilon(\mathbf{C})$ is generally smaller than the field of values of $\mathbf{C}$, see [145], [35], and hence the set E may be chosen further away from zero, leading to $|\Psi(0)| \gg 1$.
Proposition 6.8 leads to a bound on the number of iterations used by GMRES.

Proposition 6.10 (Number of inner iterations)**.** *Let the assumptions of Theorem 6.6 hold and let* $\mathbf{z}_k$ *be the approximate solution of* $\mathbf{B}\mathbf{z} = \mathbf{b}$ *obtained after* k *iterations of GMRES with starting value* $\mathbf{z}_0 = \mathbf{0}$*. If the number of inner iterations satisfies*

$$k \geq 1 + \frac{1}{\log |\Psi(0)|} \left(\log \frac{S\|\mu_1 \mathbf{I} - \mathbf{C}\|}{|\mu_1|} \|\mathbf{V}_2\| + \log \frac{\|\mathcal{P}\mathbf{b}\|}{\tau} \right), \tag{6.27}$$

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

Proof. Taking log's in (6.24) gives the required result. $\square$

The bound in (6.27) is only a sufficient condition, the desired accuracy might be reached for a much smaller value of k.

In the next subsection we apply Proposition 6.10 to the iterative solution of

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

using GMRES and in particular analyse the term $\mathcal{P}\mathbf{b}$ for both unpreconditioned and preconditioned GMRES. In this context Proposition 6.10 states that after $k^{(i)}$ iterations the system residual $\|\mathbf{d}_k^{(i)}\|$ is less than $\tau^{(i)}$ if (6.27) is satisfied.

6.3 Analysis of the right hand side term and tuning

In Chapter 4 (see also [42]) we have seen that for the standard symmetric eigenproblem inexact inverse iteration with a fixed shift leads to a constant number of inner iterations as the outer iteration proceeds, even though the solve tolerance is decreased in every step (see (4.12) and remarks therein). This somehow surprising outcome is a result of the right hand side of the linear system being approximately in the eigendirection of the system matrix. As we shall see this finding does not hold for the generalised eigenproblem.

6.3.1 The solution of the linear system using unpreconditioned GMRES

We analyse the projected right hand side term $\mathcal{P}\mathbf{b} = (\mathbf{I} - \mathbf{w}_1\mathbf{v}_1^H)\mathbf{b}$ for different values of the right hand side $\mathbf{b}$. For the moment we only consider the fixed shift approach $\sigma^{(i)} := \sigma$, so that a system of the form

$$(\mathbf{A} - \sigma\mathbf{M})\mathbf{y}^{(i)} = \mathbf{M}\mathbf{x}^{(i)}, \tag{6.28}$$

has to be solved using unpreconditioned GMRES at each inverse iteration step. Hence we take $\mathbf{B} = \mathbf{A} - \sigma\mathbf{M}$ in Theorem 6.6, where $(\mu_1, \mathbf{w}_1)$ is an eigenpair of $\mathbf{B}$ with left eigenvector $\mathbf{v}_1$. In this section we will show that

$$C_0 \leq \|\mathcal{P}\mathbf{M}\mathbf{x}^{(i)}\| \leq C_1, \tag{6.29}$$

for some positive constants C_0 and C_1 independent of i.

Theorem 6.11. *Let $\mathcal{P} = \mathbf{I} - \mathbf{w}_1\mathbf{v}_1^H$ where $\mathbf{v}_1$ and $\mathbf{w}_1$ are left and right eigenvectors of $\mathbf{A} - \sigma\mathbf{M}$. Furthermore, let any vector $\mathbf{z}$ be decomposed as $\mathbf{z} = z_1\mathbf{w}_1 + \mathbf{W}_2\mathbf{z}_2$, where $\mathbf{w}_1$ and $\mathbf{W}_2$ are as in (6.15), $z_1 \in \mathbb{C}$ and $\mathbf{z}_2 \in \mathbb{C}^{(n-1)\times 1}$. Then*

$$\|\mathcal{P}\mathbf{z}\| = \|\mathbf{z}_2\|. \tag{6.30}$$

Proof. For any vector $\mathbf{z}$ we have

$$\mathcal{P}\mathbf{z} = (\mathbf{I} - \mathbf{w}_1\mathbf{v}_1^H)\mathbf{z} = (\mathbf{z} - \mathbf{v}_1^H\mathbf{z}\mathbf{w}_1).$$

Using the decomposition $\mathbf{z} = z_1\mathbf{w}_1 + \mathbf{W}_2\mathbf{z}_2$ we have

$$\mathcal{P}\mathbf{z} = z_1\mathbf{w}_1 + \mathbf{W}_2\mathbf{z}_2 - \mathbf{v}_1^H(z_1\mathbf{w}_1 + \mathbf{W}_2\mathbf{z}_2)\mathbf{w}_1 = \mathbf{W}_2\mathbf{z}_2,$$

where we have used $\mathbf{v}_1^H\mathbf{w}_1 = 1$ and $\mathbf{v}_1^H\mathbf{W}_2 = \mathbf{0}^H$ (see remarks after (6.15)). Hence $\|\mathcal{P}\mathbf{z}\| = \|\mathbf{W}_2\|\|\mathbf{z}_2\| = \|\mathbf{z}_2\|$, where we have used that $\mathbf{W}_2$ has orthonormal columns. $\square$

Hence, for any vector $\mathbf{z}$, which is not parallel to $\mathbf{w}_1$ (that is, $\|\mathbf{z}_2\| \neq 0$) we have that $\|\mathcal{P}\mathbf{z}\|$ is nonzero. In particular, setting $\mathbf{z} = \dfrac{\mathbf{M}\mathbf{x}^{(i)}}{\|\mathbf{M}\mathbf{x}^{(i)}\|}$ in inequality (6.30) we obtain (6.29) for positive constants C_0 and C_1 since $\mathbf{M}\mathbf{x}^{(i)}$ (and also the limit vector for $\mathbf{x}^{(i)} \to \mathbf{x}_1$, $\mathbf{M}\mathbf{x}_1$) is not an eigenvector of $\mathbf{A} - \sigma\mathbf{M}$ and hence not parallel to $\mathbf{w}_1$. We use the result in Theorem 6.11 to obtain bounds on the number of iterations needed to solve (6.28).
The number of inner iterations $k^{(i)}$ per outer iteration i of unpreconditioned GMRES using the result in Proposition 6.8 is given by

$$k^{(i)} \geq 1 + \frac{1}{\log|\Psi(0)|}\left(\log\frac{S\|\mu_1\mathbf{I} - \mathbf{C}\|}{|\mu_1|}\|\mathbf{V}_2\| + \log\frac{\|\mathcal{P}\mathbf{M}\mathbf{x}^{(i)}\|}{\tau^{(i)}}\right), \tag{6.31}$$

using Lemma 6.10. Choosing a decreasing tolerance $\tau^{(i)} = \delta\|\mathbf{r}^{(i)}\|$, which is required for linear convergence of Algorithm 4, and using (6.29) only the second term depends on i and can be bounded by

$$\frac{\|\mathcal{P}\mathbf{M}\mathbf{x}^{(i)}\|}{\tau^{(i)}} \leq \frac{C_1}{\delta\|\mathbf{r}^{(i)}\|},$$

which increases as $\|\mathbf{r}^{(i)}\| \to 0$. Hence, the lower bound on $k^{(i)}$, the number of inner iterations per outer iteration, increases as the iteration proceeds and as convergence occurs. This behaviour can be observed in Figure 6-1 (dotted circled line).

6.3.2 The concept of tuning and its implementation

The increase in the number of inner iterations arises from the fact that the right hand side of the system $(\mathbf{A} - \sigma\mathbf{M})\mathbf{y}^{(i)} = \mathbf{M}\mathbf{x}^{(i)}$ is not an eigenvector (or eigenvector approximation) of the system matrix $\mathbf{A} - \sigma\mathbf{M}$ (as it is, for example in the case of the standard eigenproblem, see Chapter 4). A possible way of changing that is to use a rank one change of the identity as a tuning operator. This approach is a simplification of the symmetric preconditioning in Chapter 4 (see also [42]).

First, let us introduce an "ideal" tuned operator which assumes we know the eigenvector $\mathbf{x}_1$. Define the tuning operator given by

$$\mathbb{T} = \mathbf{I} + (\mathbf{M} - \mathbf{I})\frac{\mathbf{x}_1}{\|\mathbf{x}_1\|}\frac{\mathbf{x}_1^H}{\|\mathbf{x}_1\|}, \tag{6.32}$$

and note that $\mathbb{T}\mathbf{x}_1 = \mathbf{M}\mathbf{x}_1$. Then

$$(\mathbf{A} - \sigma\mathbf{M})\mathbb{T}^{-1}\mathbf{M}\mathbf{x}_1 = (\mathbf{A} - \sigma\mathbf{M})\mathbf{x}_1 = (\lambda_1 - \sigma)\mathbf{M}\mathbf{x}_1,$$

that is, $\mathbf{M}\mathbf{x}_1$ is an exact eigenvector of $(\mathbf{A} - \sigma\mathbf{M})\mathbb{T}^{-1}$. Now define

$$\mathcal{P} = \left(\mathbf{I} - \bar{\mathbf{w}}_1\bar{\mathbf{v}}_1^H\right), \quad \text{where} \quad \bar{\mathbf{v}}_1^H\bar{\mathbf{w}}_1 = 1,$$

where $\bar{\mathbf{w}}_1 = \dfrac{\mathbf{M}\mathbf{x}_1}{\|\mathbf{M}\mathbf{x}_1\|}$ is the normalised right eigenvector and $\bar{\mathbf{v}}_1$ the left eigenvector of $(\mathbf{A} - \sigma\mathbf{M})\mathbb{T}^{-1}$. Hence $\mathcal{P}\mathbf{M}\mathbf{x}_1 = \mathbf{0}$, and GMRES applied to the system

$$(\mathbf{A} - \sigma\mathbf{M})\mathbb{T}^{-1}\tilde{\mathbf{y}} = \mathbf{M}\mathbf{x}_1, \quad \mathbb{T}^{-1}\tilde{\mathbf{y}} = \mathbf{y}$$

would converge in just one iteration, since the right hand side $\mathbf{M}\mathbf{x}_1$ is an eigenvector of the system matrix $(\mathbf{A} - \sigma\mathbf{M})\mathbb{T}^{-1}$ and any Krylov subspace method stops after one step.

This ideal tuning given by (6.32) cannot be used in practice, but we can replace it by the approximation

$$\mathbb{T}_i = \mathbf{I} + (\mathbf{M} - \mathbf{I})\frac{\mathbf{x}^{(i)}}{\|\mathbf{x}^{(i)}\|}\frac{{\mathbf{x}^{(i)}}^H}{\|\mathbf{x}^{(i)}\|} \tag{6.33}$$

so that

$$\mathbb{T}_i\mathbf{x}^{(i)} = \mathbf{M}\mathbf{x}^{(i)}. \tag{6.34}$$

and clearly $\mathbb{T}_i \to \mathbb{T}$ as $\mathbf{x}^{(i)} \to \mathbf{x}_1$. In order to prove the efficiency of the tuning strategy we need the following Lemma. For convenience we set $\hat{\mathbf{x}}^{(i)} = \mathbf{x}^{(i)}/\|\mathbf{x}^{(i)}\|$ and

$$\varepsilon^{(i)} := \|\mathbf{p}^{(i)}\|. \tag{6.35}$$

Lemma 6.12. *Let (6.6) hold. Then*

$$\hat{\mathbf{x}}^{(i)}{\hat{\mathbf{x}}^{(i)}}^H - \hat{\mathbf{x}}_1\hat{\mathbf{x}}_1^H = \mathcal{E}^{(i)}, \tag{6.36}$$

where $\|\mathcal{E}^{(i)}\| \leq C_3\varepsilon^{(i)}$ *with* $C_3 := \|\mathbf{M}\|\|\mathbf{X}\|\|\mathbf{U}^{-1}\|$. *Furthermore we have*

$$\|(\mathbf{A} - \sigma\mathbf{M})\mathbb{T}^{-1} - (\mathbf{A} - \sigma\mathbf{M}){\mathbb{T}_i}^{-1}\| \leq \beta_1\varepsilon^{(i)}. \tag{6.37}$$

where β_1 *is independent of* i *for large enough* i.

Proof. We use the sine of the largest canonical angle and have (see [47, p. 76])

$$\sin\angle(\hat{\mathbf{x}}^{(i)}, \hat{\mathbf{x}}_1) = \|{\mathbf{X}_1^{\perp}}^H\hat{\mathbf{x}}^{(i)}\| = \|\hat{\mathbf{x}}^{(i)}{\hat{\mathbf{x}}^{(i)}}^H - \hat{\mathbf{x}}_1\hat{\mathbf{x}}_1^H\| = \|\mathcal{E}^{(i)}\|,$$

where ${\mathbf{X}_1}^{\perp} \in \mathbb{C}^{n\times(n-1)}$ is an orthonormal matrix consisting of unit vectors orthogonal to $\mathbf{x}_1$. Hence

$$\|\mathcal{E}^{(i)}\| = \|{\mathbf{X}_1^{\perp}}^H\hat{\mathbf{x}}^{(i)}\| = \left\|{\mathbf{X}_1^{\perp}}^H\left(\frac{\mathbf{x}^{(i)} - \alpha^{(i)}q^{(i)}\mathbf{x}_1}{\|\mathbf{x}^{(i)}\|}\right)\right\|,$$

and using (6.6) as well as $\|{\mathbf{X}_1}^{\perp}\| = 1$, $\|\mathbf{X}_2\| \leq \|\mathbf{X}\|$, $\alpha^{(i)} \leq \|\mathbf{U}^{-1}\|\|\mathbf{M}\|\|\mathbf{x}^{(i)}\|$ this yields

$$\|\mathcal{E}^{(i)}\| = \left\|{\mathbf{X}_1}^{\perp}\left(\frac{\alpha^{(i)}\mathbf{X}_2\mathbf{p}^{(i)}}{\|\mathbf{x}^{(i)}\|}\right)\right\| \leq \|\mathbf{M}\|\|\mathbf{X}\|\|\mathbf{U}^{-1}\|\|\mathbf{p}^{(i)}\| =: C_3\varepsilon^{(i)}.$$

Furthermore, we have $\mathbb{T}_i = \mathbb{T} + (\mathbf{M} - \mathbf{I})\mathcal{E}^{(i)}$ and therefore we can write

$$\begin{aligned}(\mathbf{A} - \sigma\mathbf{M})\mathbb{T}^{-1} - (\mathbf{A} - \sigma\mathbf{M}){\mathbb{T}_i}^{-1} &= (\mathbf{A} - \sigma\mathbf{M})\mathbb{T}^{-1}(\mathbb{T}_i - \mathbb{T}){\mathbb{T}_i}^{-1} \\ &= (\mathbf{A} - \sigma\mathbf{M})\mathbb{T}^{-1}(\mathbf{M} - \mathbf{I})\mathcal{E}^{(i)}(\mathbb{T} + (\mathbf{M} - \mathbf{I})\mathcal{E}^{(i)})^{-1}\end{aligned}$$

and with $\|(\mathbf{A} - \sigma\mathbf{M})\mathbb{T}^{-1}(\mathbf{M} - \mathbf{I})\| = C_4$,

$$\begin{aligned}\|(\mathbf{A} - \sigma\mathbf{M})\mathbb{T}^{-1} - (\mathbf{A} - \sigma\mathbf{M}){\mathbb{T}_i}^{-1}\| &\leq C_4\|\mathcal{E}^{(i)}\|\|\mathbb{T}^{-1}\|\|(\mathbf{I} + \mathbb{T}^{-1}(\mathbf{M} - \mathbf{I})\mathcal{E}^{(i)})^{-1}\| \\ &\leq \frac{C_4\|\mathbb{T}^{-1}\|}{1 - \|\mathbb{T}^{-1}(\mathbf{M} - \mathbf{I})\|\|\mathcal{E}^{(i)}\|}\|\mathcal{E}^{(i)}\| \\ &\leq \frac{C_4\|\mathbb{T}^{-1}\|}{1 - \|\mathbb{T}^{-1}(\mathbf{M} - \mathbf{I})\|C_3\varepsilon^{(i)}}C_3\varepsilon^{(i)}.\end{aligned}$$

Finally, since $\varepsilon^{(i)}$ is decreasing there exists $\beta_1 > 0$ independent of i for i large enough such that

$$\frac{C_4\|\mathbb{T}^{-1}\|}{1-\|\mathbb{T}^{-1}(\mathbf{M}-\mathbf{I})\|C_3\varepsilon^{(i)}}C_3 \leq \beta_1,$$

and hence (6.37) follows. □

Let $\bar{\mu}_1 = \lambda_1 - \sigma$ be a simple eigenvalue of $(\mathbf{A}-\sigma\mathbf{M})\mathbb{T}^{-1}$ with corresponding unit right eigenvector $\bar{\mathbf{w}}_1 = \dfrac{\mathbf{M}\mathbf{x}_1}{\|\mathbf{M}\mathbf{x}_1\|}$ and let $\bar{\mathbf{W}} = [\bar{\mathbf{w}}_1, \bar{\mathbf{W}}_1^{\perp}]$ be unitary such that

$$\begin{bmatrix} \bar{\mathbf{w}}_1 & \bar{\mathbf{W}}_1^{\perp} \end{bmatrix}^H (\mathbf{A}-\sigma\mathbf{M})\mathbb{T}^{-1} \begin{bmatrix} \bar{\mathbf{w}}_1 & \bar{\mathbf{W}}_1^{\perp} \end{bmatrix} = \begin{bmatrix} \bar{\mu}_1 & \bar{\mathbf{n}}_{12}^H \\ \mathbf{0} & \bar{\mathbf{N}}_{22} \end{bmatrix}.$$

is the Schur decomposition of $(\mathbf{A}-\sigma\mathbf{M})\mathbb{T}^{-1}$. Finally, $\bar{\mu}_1$ being a simple eigenvalue ensures the existence of the block diagonalisation (see (6.15))

$$(\mathbf{A}-\sigma\mathbf{M})\mathbb{T}^{-1} = \begin{bmatrix} \bar{\mathbf{w}}_1 & \bar{\mathbf{W}}_2 \end{bmatrix} \begin{bmatrix} \bar{\mu}_1 & \mathbf{0}^H \\ \mathbf{0} & \bar{\mathbf{C}} \end{bmatrix} \begin{bmatrix} \bar{\mathbf{v}}_1^H \\ \bar{\mathbf{V}}_2^H \end{bmatrix}.$$

We have the following lemma:

Lemma 6.13. *Assume that $\bar{\mathbf{w}}_1$ is a simple eigenvector of $(\mathbf{A}-\sigma\mathbf{M})\mathbb{T}^{-1}$. Then for i large enough, $(\mathbf{A}-\sigma\mathbf{M})\mathbb{T}_i^{-1}$ can be block diagonalised as*

$$(\mathbf{A}-\sigma\mathbf{M})\mathbb{T}_i^{-1} = \begin{bmatrix} \bar{\mathbf{w}}_1^{(i)} & \bar{\mathbf{W}}_2^{(i)} \end{bmatrix} \begin{bmatrix} \bar{\mu}_1^{(i)} & \mathbf{0}^H \\ \mathbf{0} & \bar{\mathbf{C}}^{(i)} \end{bmatrix} \begin{bmatrix} \bar{\mathbf{v}}_1^{(i)} & \bar{\mathbf{V}}_2^{(i)} \end{bmatrix}^H, \tag{6.38}$$

with

$$\begin{aligned} |\bar{\mu}_1 - \bar{\mu}_1^{(i)}| &\leq c_1\varepsilon^{(i)}, \\ \|\bar{\mathbf{C}} - \bar{\mathbf{C}}^{(i)}\| &\leq c_2\varepsilon^{(i)}, \\ \|\bar{\mathbf{V}}_2 - \bar{\mathbf{V}}_2^{(i)}\| &\leq c_3\varepsilon^{(i)}, \end{aligned}$$

where c_1, c_2 and c_3 are positive constants independent of i for i large enough and $\varepsilon^{(i)}$ is defined by (6.35).

Proof. Since $\bar{\mathbf{w}}_1$ is a simple eigenvector the block diagonalisation exists and Theorem C.3 can be used to compare the invariant subspaces of $(\mathbf{A}-\sigma\mathbf{M})\mathbb{T}^{-1}$ and $(\mathbf{A}-\sigma\mathbf{M})\mathbb{T}_i^{-1}$ and the representation of $(\mathbf{A}-\sigma\mathbf{M})\mathbb{T}^{-1}$ and $(\mathbf{A}-\sigma\mathbf{M})\mathbb{T}_i^{-1}$ with respect to these invariant subspaces. With Lemma 6.12 the result follows. □

Using Lemma 6.12 and Lemma 6.13 we are able to prove the following theorem about the tuning strategy for the generalised eigenproblem.

Theorem 6.14 (Right tuning)**.** *Let the assumptions of Theorem 6.6 hold and consider the solution of*

$$(\mathbf{A}-\sigma\mathbf{M})\mathbb{T}_i{}^{-1}\tilde{\mathbf{y}}^{(i)} = \mathbf{M}\mathbf{x}^{(i)}, \quad \textit{where} \quad \mathbf{y}^{(i)} = \mathbb{T}_i{}^{-1}\tilde{\mathbf{y}}^{(i)}. \tag{6.39}$$

with $\mathbb{T}_i$ chosen as in (6.33). Let $\bar{\mu}_1^{(i)}$ be a simple eigenvalue of $(\mathbf{A}-\sigma\mathbf{M})\mathbb{T}_i^{-1}$ with corresponding right eigenvector $\bar{\mathbf{w}}_1^{(i)}$ and let $[\bar{\mathbf{w}}_1^{(i)}, \bar{\mathbf{W}}_1^{(i)\perp}]$ be unitary such that

$$\begin{bmatrix} \bar{\mathbf{w}}_1^{(i)} & \bar{\mathbf{W}}_1^{(i)\perp} \end{bmatrix}^H (\mathbf{A}-\sigma\mathbf{M})\mathbb{T}_i^{-1} \begin{bmatrix} \bar{\mathbf{w}}_1^{(i)} & \bar{\mathbf{W}}_1^{(i)\perp} \end{bmatrix} = \begin{bmatrix} \bar{\mu}_1^{(i)} & \bar{\mathbf{n}}_{12}^{(i)^H} \\ \mathbf{0} & \bar{\mathbf{N}}_{22}^{(i)} \end{bmatrix}.$$

is the Schur decomposition of $(\mathbf{A}-\sigma\mathbf{M})\mathbb{T}_i^{-1}$. Furthermore, let $\mathcal{P}^{(i)} = \mathbf{I} - \bar{\mathbf{w}}_1^{(i)}\bar{\mathbf{v}}_1^{(i)^H}$ where $\bar{\mathbf{v}}_1^{(i)}$ is the right eigenvector of $(\mathbf{A}-\sigma\mathbf{M})\mathbb{T}_i^{-1}$. Then, for large enough i we have

$$\|\mathcal{P}^{(i)}\mathbf{M}\mathbf{x}^{(i)}\| \leq C_5\|\mathcal{P}^{(i)}\|\|\mathbf{r}^{(i)}\|, \tag{6.40}$$

for some positive constant C_5 independent of i.

Proof. Using the eigenvalue residual (6.4) and the condition (6.33) we obtain

$$\begin{aligned}(\mathbf{A}-\sigma\mathbf{M})\mathbb{T}_i^{-1}\mathbf{M}\mathbf{x}^{(i)} &= (\mathbf{A}-\sigma\mathbf{M})\mathbf{x}^{(i)} \\ &= (\rho(\mathbf{x}^{(i)})-\sigma)\mathbf{M}\mathbf{x}^{(i)} + \mathbf{r}^{(i)}.\end{aligned}$$

Thus $\mathbf{M}\mathbf{x}^{(i)}$ is an approximate eigenvector of $(\mathbf{A}-\sigma\mathbf{M})\mathbb{T}_i^{-1}$ with approximate eigenvalue $\rho(\mathbf{x}^{(i)})-\sigma$. For large enough i the eigenvalue residual $\|\mathbf{r}^{(i)}\|$ is small enough and the perturbation theory in Proposition 6.5 (now depending on i) applied to $(\mathbf{A}-\sigma\mathbf{M})\mathbb{T}_i^{-1}$ and $\hat{\mathbf{w}} = \dfrac{\mathbf{M}\mathbf{x}^{(i)}}{\|\mathbf{M}\mathbf{x}^{(i)}\|}$ yields

$$\left\| \frac{\mathbf{M}\mathbf{x}^{(i)}}{\|\mathbf{M}\mathbf{x}^{(i)}\|} - \frac{\bar{\mathbf{w}}_1^{(i)}}{\sqrt{1+\bar{\mathbf{p}}_i^H\bar{\mathbf{p}}_i}} \right\| \leq \frac{2\|\mathbf{r}^{(i)}\|}{\operatorname{sep}(\bar{\mu}_1^{(i)}, \bar{\mathbf{N}}_{22}^{(i)}) - 2\|\mathbf{r}^{(i)}\|}.$$

From Remark C.2 we have, for small enough $\|\mathbf{r}^{(i)}\|$,

$$\left\| \frac{\mathbf{M}\mathbf{x}^{(i)}}{\|\mathbf{M}\mathbf{x}^{(i)}\|} - \frac{\bar{\mathbf{w}}_1^{(i)}}{\sqrt{1+\bar{\mathbf{p}}_i^H\bar{\mathbf{p}}_i}} \right\| \leq \frac{2}{\operatorname{sep}(\bar{\mu}_1^{(i)}, \bar{\mathbf{N}}_{22}^{(i)})}\|\mathbf{r}^{(i)}\| + \mathcal{O}(\|\mathbf{r}^{(i)}\|^2),$$

and hence, for i large enough we have

$$\left\| \frac{\mathbf{M}\mathbf{x}^{(i)}}{\|\mathbf{M}\mathbf{x}^{(i)}\|} - \frac{\bar{\mathbf{w}}_1^{(i)}}{\sqrt{1+\bar{\mathbf{p}}_i^H\bar{\mathbf{p}}_i}} \right\| \leq \frac{C_6}{\operatorname{sep}(\bar{\mu}_1^{(i)}, \bar{\mathbf{N}}_{22}^{(i)})}\|\mathbf{r}^{(i)}\|,$$

for some constant $C_6 > 2$. Finally, we use $\mathcal{P}^{(i)}\dfrac{\mathbf{M}\mathbf{x}^{(i)}}{\|\mathbf{M}\mathbf{x}^{(i)}\|} = \mathcal{P}^{(i)}\left(\dfrac{\mathbf{M}\mathbf{x}^{(i)}}{\|\mathbf{M}\mathbf{x}^{(i)}\|} - \alpha\bar{\mathbf{w}}_1^{(i)}\right)$, $\forall\alpha\in\mathbb{C}$, since $\mathcal{P}^{(i)}\bar{\mathbf{w}}_1^{(i)} = 0$, and therefore, with $\alpha = \dfrac{1}{\sqrt{1+\bar{\mathbf{p}}_i^H\bar{\mathbf{p}}_i}}$, we obtain

$$\|\mathcal{P}^{(i)}\mathbf{M}\mathbf{x}^{(i)}\| \leq \|\mathcal{P}^{(i)}\|\frac{C_6\|\mathbf{M}\mathbf{x}^{(i)}\|}{\operatorname{sep}(\bar{\mu}_1^{(i)}, \bar{\mathbf{N}}_{22}^{(i)})}\|\mathbf{r}^{(i)}\|.$$

Lemma 6.13 implies $|\bar{\mu}_1^{(i)} - \bar{\mu}_1| \leq c_1\varepsilon^{(i)}$ and with a similar result we also have $\|\bar{\mathbf{N}}_{22}^{(i)} - \bar{\mathbf{N}}_{22}\| \leq c_4\varepsilon^{(i)}$ for an appropriately chosen constant c_4 (see Theorem C.1). For $\varepsilon^{(i)}$ small enough (see [137, p. 234]) we have

$$\frac{C_6\|\mathbf{M}\mathbf{x}^{(i)}\|}{\operatorname{sep}(\bar{\mu}_1^{(i)}, \bar{\mathbf{N}}_{22}^{(i)})} \leq \frac{C_6\|\mathbf{M}\mathbf{x}^{(i)}\|}{\operatorname{sep}(\bar{\mu}_1, \bar{\mathbf{N}}_{22}) - c_1\varepsilon^{(i)} - c_4\varepsilon^{(i)}} \leq C_5,$$

where C_5 is independent of i since $\varepsilon^{(i)}$ is decreasing (as $\|\mathbf{r}^{(i)}\|$ is decreasing) and where we have also used that $\|\mathbf{M}\mathbf{x}^{(i)}\|$ is bounded since $\mathbf{x}^{(i)}$ is normalised. Hence (6.40) follows for i large enough. $\square$

Before proving the main result of this section we need another lemma.

Lemma 6.15. *Let $\mathbf{B}_1$ and $\mathbf{B}_2$ be two matrices of the same dimensions and let $\mathcal{P}_\gamma(\mathbf{B}_1)$ and $\mathcal{P}_\gamma(\mathbf{B}_2)$ be the spectral projections onto the eigenvectors of $\mathbf{B}_1$ and $\mathbf{B}_2$ corresponding to the eigenvalues inside a closed contour γ. Assume that $\|\mathbf{B}_1 - \mathbf{B}_2\| \leq \xi$ and let $m_\gamma(\mathbf{B}_1) = \max_{\lambda\in\gamma} \|(\lambda\mathbf{I} - \mathbf{B}_1)^{-1}\|$. If $\xi m_\gamma(\mathbf{B}_1) < 1$ then*

$$\|\mathcal{P}_\gamma(\mathbf{B}_1) - \mathcal{P}_\gamma(\mathbf{B}_2)\| \leq \frac{1}{2\pi}\mathcal{L}(\gamma)\frac{\xi m_\gamma^2(\mathbf{B}_1)}{1 - \xi m_\gamma(\mathbf{B}_1)},$$

where $\mathcal{L}(\gamma)$ is the length of γ.

Proof. See [69] and [46, Section 8.2]. $\square$

We are now able to prove the following theorem which provides the main result of this section.

Theorem 6.16. *Let the assumptions of Theorem 6.14 be satisfied. Then the number $k^{(i)}$ of inner iterations used by GMRES to compute $\tilde{\mathbf{y}}_{k^{(i)}}$ satisfying the stopping criterion*

$$\|(\mathbf{A} - \sigma\mathbf{M})\mathbb{T}_i^{-1}\tilde{\mathbf{y}}_{k^{(i)}} - \mathbf{M}\mathbf{x}^{(i)}\| \leq \tau^{(i)} = \delta\|\mathbf{r}^{(i)}\|,$$

is bounded independently of i for large enough i.

Proof. Let Ψ and $\mathbf{E}$ be given by Proposition 6.8 applied to $\bar{\mathbf{C}}$ instead of $\mathbf{C}$. For large enough i Lemma 6.13 shows that decomposition (6.38) exists. We can use Corollary 6.9 to obtain a constant S_δ independent of i for small enough $\varepsilon^{(i)}$ such that

$$\min_{p_{k-1}\in\Pi_{k-1}} \|p_{k-1}(\bar{\mathbf{C}}^{(i)})\| \leq S_\delta \left(\frac{1}{|\Psi(0)|}\right)^{k-1}.$$

Then, by Proposition 6.10 the residual obtained after $k^{(i)}$ iterations of GMRES starting with 0 is less than $\tau^{(i)} = \delta\|\mathbf{r}^{(i)}\|$ if

$$k^{(i)} \geq 1 + \frac{1}{\log|\Psi(0)|}\left(\log\frac{S_\delta\|\bar{\mu}_1^{(i)}\mathbf{I} - \bar{\mathbf{C}}^{(i)}\|}{|\bar{\mu}_1^{(i)}|}\|\bar{\mathbf{V}}_2^{(i)}\| + \log\frac{\|\mathcal{P}^{(i)}\mathbf{M}\mathbf{x}^{(i)}\|}{\delta\|\mathbf{r}^{(i)}\|}\right).$$

Using Lemma 6.13 the argument of the first log term in the brackets can be bounded by

$$\frac{\|\bar{\mu}_1^{(i)}\mathbf{I} - \bar{\mathbf{C}}^{(i)}\|}{|\bar{\mu}_1^{(i)}|}\|\bar{\mathbf{V}}_2^{(i)}\| \leq \frac{\|\bar{\mu}_1\mathbf{I} - \bar{\mathbf{C}}\| + c_1\varepsilon^{(i)} + c_2\varepsilon^{(i)}}{|\bar{\mu}_1| - c_1\varepsilon^{(i)}}(\|\bar{\mathbf{V}}_2\| + c_3\varepsilon^{(i)}) \tag{6.41}$$

Since $\varepsilon^{(i)}$ is decreasing (6.41) can be bounded independently of i for small enough $\varepsilon^{(i)}$. For the second log term in the brackets we use Theorem 6.14 and obtain

$$\frac{\|\mathcal{P}^{(i)}\mathbf{M}\mathbf{x}^{(i)}\|}{\delta\|\mathbf{r}^{(i)}\|} \leq \frac{C_5\|\mathcal{P}^{(i)}\|}{\delta} \tag{6.42}$$

The term $\|\mathcal{P}^{(i)}\|$ can be bounded as follows

$$\|\mathcal{P}^{(i)}\| \leq \|\mathcal{P}^{(i)} - \mathcal{P}\| + \|\mathcal{P}\|, \tag{6.43}$$

where $\mathcal{P} = \mathbf{I} - \bar{\mathbf{w}}_1\bar{\mathbf{v}}_1^H$ is the spectral projection of $(\mathbf{A} - \sigma\mathbf{M})\mathbb{T}^{-1}$ onto $\bar{\mathbf{W}}_2$. For small enough $\varepsilon^{(i)}$ we use (6.37) and apply Lemma 6.15 with $\mathbf{B}_1 = (\mathbf{A} - \sigma\mathbf{M})\mathbb{T}^{-1}$ and $\mathbf{B}_2 = (\mathbf{A} - \sigma\mathbf{M})\mathbb{T}_i^{-1}$. Taking γ as a circle of centre $\lambda_1 - \sigma$ and radius $\varepsilon^{(i)}$, we obtain

$$\|\mathcal{P}^{(i)} - \mathcal{P}\| \leq \frac{\beta_1 m_\gamma^2((\mathbf{A} - \sigma\mathbf{M})\mathbb{T}^{-1})\varepsilon^{(i)}}{1 - \beta_1 m_\gamma((\mathbf{A} - \sigma\mathbf{M})\mathbb{T}^{-1})\varepsilon^{(i)}}\varepsilon^{(i)}.$$

Since $\varepsilon^{(i)}$ is decreasing we have for a small enough $\varepsilon^{(i)}$

$$\frac{\beta_1 m_\gamma^2((\mathbf{A} - \sigma\mathbf{M})\mathbb{T}^{-1})\varepsilon^{(i)}}{1 - \beta_1 m_\gamma((\mathbf{A} - \sigma\mathbf{M})\mathbb{T}^{-1})\varepsilon^{(i)}} \leq C_7,$$

where C_7 is independent of i. Hence (6.42) and (6.43) imply

$$\frac{\|\mathcal{P}^{(i)}\mathbf{M}\mathbf{x}^{(i)}\|}{\delta\|\mathbf{r}^{(i)}\|} \leq \frac{C_5\|\mathcal{P}\| + C_5C_7\varepsilon^{(i)}}{\delta}$$

and since $\varepsilon^{(i)}$ is decreasing, this inequality shows that $\dfrac{\|\mathcal{P}\mathbf{M}\mathbf{x}^{(i)}\|}{\delta\|\mathbf{r}^{(i)}\|}$ can be bounded independently of i for i large enough. Hence the number of inner iterations per outer iteration $k^{(i)}$ can be bounded independently of i for large enough i. □

Finally, the following theorem provides a method to implement the tuning concept efficiently.

Theorem 6.17 (Implementation of $\mathbb{T}_i$)**.** *Let* $\mathbf{x}^{(i)}$ *be the approximate eigenvector obtained by the ith iteration of inexact inverse iteration. Let* $\mathbf{u}^{(i)} = \mathbf{M}\mathbf{x}^{(i)} - \mathbf{x}^{(i)}$. *Then* $\mathbb{T}_i = \mathbf{I} + \mathbf{u}^{(i)}\dfrac{\mathbf{x}^{(i)H}}{\|\mathbf{x}^{(i)}\|^2}$ *ensures* $\mathbb{T}\mathbf{x}^{(i)} = \mathbf{M}\mathbf{x}^{(i)}$ *and*

$$\mathbb{T}_i^{-1} = \mathbf{I} - \frac{\mathbf{u}^{(i)}\mathbf{x}^{(i)H}}{\mathbf{x}^{(i)H}\mathbf{M}\mathbf{x}^{(i)}}.$$

Proof. Straightforward calculation of $\mathbb{T}_i\mathbf{x}^{(i)}$ and the Sherman-Morrison formula give the required result. □

Before providing a numerical example in the next subsection in order to illustrate the theory we have the following remark on left preconditioning.

Remark 6.18 (Left tuning). *Instead of right tuning as in Theorem 6.14 we may also consider the left tuned system*

$$\mathbb{T}_i^{-1}(\mathbf{A}-\sigma\mathbf{M})\mathbf{y}^{(i)} = \mathbb{T}_i^{-1}\mathbf{M}\mathbf{x}^{(i)}, \tag{6.44}$$

where $\mathbb{T}_i$ is chosen such that (6.33) holds. Using the eigenvalue residual given by (6.4) and the condition (6.33) we obtain

$$\begin{aligned}\mathbb{T}_i^{-1}(\mathbf{A}-\sigma\mathbf{M})\mathbf{x}^{(i)} &= \mathbb{T}_i^{-1}(\rho(\mathbf{x}^{(i)})-\sigma)\mathbf{M}\mathbf{x}^{(i)} + \mathbb{T}_i^{-1}\mathbf{r}^{(i)}\\ &= (\rho(\mathbf{x}^{(i)})-\sigma)\mathbf{x}^{(i)} + \mathbb{T}_i^{-1}\mathbf{r}^{(i)}.\end{aligned}$$

Thus $\mathbf{x}^{(i)}$ is an approximate eigenvector of $\mathbb{T}_i^{-1}(\mathbf{A}-\sigma\mathbf{M})$ with approximate eigenvalue $\rho(\mathbf{x}^{(i)})-\sigma$. A similar analysis as above shows

$$\|\mathcal{P}^{(i)}\mathbb{T}_i^{-1}\mathbf{M}\mathbf{x}^{(i)}\| = \|\mathcal{P}^{(i)}\mathbf{x}^{(i)}\| = \mathcal{O}(\|\mathbf{r}^{(i)}\|),$$

where $\mathcal{P}^{(i)}$ is the appropriate oblique projection.

6.3.3 Numerical example

Here we chose a simple example of a convection-diffusion operator.

Example 6.19. *First, consider the standard eigenvalue problem $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ where $\mathbf{A}$ is the finite difference discretisation (central differences) on a 32×32 grid of the following eigenvalue problem of the convection-diffusion operator*

$$-\Delta u + 5u_x + 5u_y = \lambda u \quad on \quad (0,1)^2, \tag{6.45}$$

with homogeneous Dirichlet boundary conditions. This nonsymmetric eigenvalue problem is also discussed in [50]. The smallest eigenvalue is given by $\lambda_1 \approx 32.18560954$.

Secondly, consider the generalised eigenvalue problem $\mathbf{A}\mathbf{x} = \lambda\mathbf{M}\mathbf{x}$ derived by discretising (6.45) using a Galerkin-FEM on regular triangular elements with piecewise linear functions. We use a 32×32 grid leading to 961 degrees of freedom. Again, we seek the smallest eigenvalue, which in this case is given by $\lambda_1 \approx 32.15825765$.

We apply inexact inverse iteration with fixed shift $\sigma = 20$ to the (a) standard eigenproblem arising from the finite difference discretisation, to the (b) generalised eigenproblem arising from the finite element discretisation and to the (c) tuned generalised eigenproblem. We use left tuning (6.44) as well as right tuning (6.39). For the solve tolerance of GMRES we use $\tau^{(i)} = \min\{0.01, 0.01\,\|\mathbf{r}^{(i)}\|\}$. The overall computation stops once $\|\mathbf{r}^{(i)}\| < 10^{-11}$.

Note that this example is not comparing FE and FD discretisations as the mesh-size $h \to 0$. We are demonstrating the effect of $\mathbf{M}$ on the right hand side of the eigenvalue equation for fixed matrix size n when inverse iteration is used as iterative solver.

Figure 6-1 shows the inner iteration $k^{(i)}$ per outer iteration obtained for Example 6.19. Figure 6-2 shows the improvement in the overall costs of the iteration. By using tuning less than a third of the costs of the untuned method is needed. All methods (a), (b) and (c) converge linearly, since we use a fixed shift and a decreasing tolerance and hence the number of outer iterations of all methods is similar. However the number of inner iterations increases logarithmically for the non-tuned generalised eigenproblem

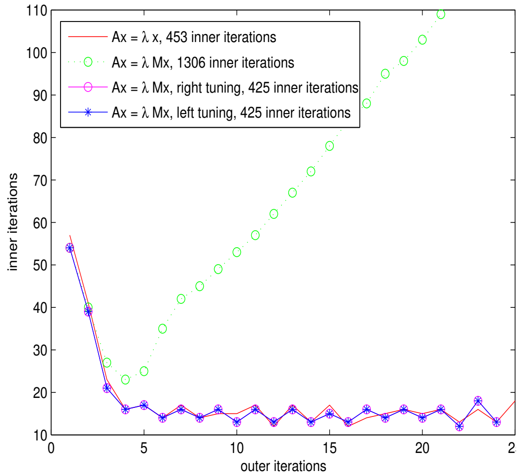


Figure 6-1: *Inner iterations against outer iterations for standard and the generalised eigenproblem with/without tuning (Example 6.19)*

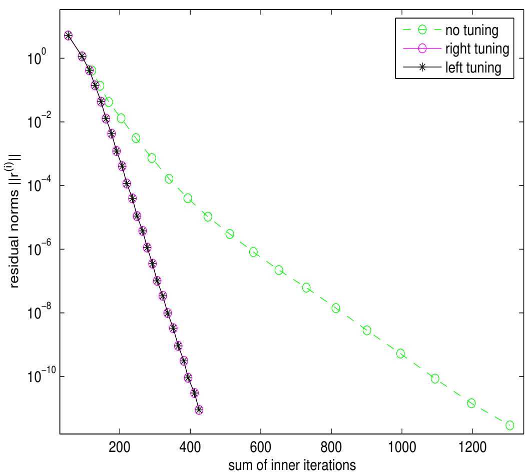


Figure 6-2: *Eigenvalue residual norms against the total number of inner iterations for generalised eigenproblem with/without tuning (Example 6.19)*

as the theory suggests. For the standard eigenproblem condition (6.33) is satisfied trivially with $\mathbb{T}_i = \mathbf{I}$.

Clearly, the tuning strategy (6.33) is only of theoretical nature, since, in practice one would always use a preconditioner to enhance the linear solver. The concept of tuning is applied to preconditioners in the following section.

6.4 Preconditioned GMRES as inner solver for fixed shift case

A good preconditioner accelerates the convergence of the GMRES iteration. To achieve this a matrix $\mathbf{P}$ is constructed such that $\mathbf{A}\mathbf{P}^{-1} \approx \mathbf{I}$ in some sense, so that systems with $\mathbf{A}\mathbf{P}^{-1}$ will become cheaper to solve. We assume here that $\mathbf{P}^{-1}$ is also a good preconditioner for $\mathbf{A} - \sigma\mathbf{M}$.

In this section we also assume throughout that the shifts $\sigma^{(i)}$ are fixed and that the tolerance $\tau^{(i)}$ decreases according to $\tau^{(i)} = \delta\|\mathbf{r}^{(i)}\|$ (as in Theorem 6.2). In Algorithm 4, step 2 is replaced by solving

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

Denote $\tilde{\mathbf{y}}_{k^{(i)}}$ the approximation of $\tilde{\mathbf{y}}^{(i)}$ obtained after $k^{(i)}$ inner iterations. The question is whether $k^{(i)}$ can be bounded independently of i. Since this is not the case for the unpreconditioned problem there is no reason why it should be true for the preconditioned problem. As in Section 6.3.2, where the results are only of theoretical nature, we apply a tuning strategy to the preconditioner, to obtain results similar to those presented in Figure 6-1.

Results for the tuned preconditioner for inexact inverse iteration have been given in Chapter 4 (see also [42]), where tuning applied to the Hermitian positive definite eigenproblem was considered. For the nonsymmetric case, a tuned preconditioner was introduced in Chapter 2 (see also [43]), but with the motivation of a modified Newton method and preservation of an outer convergence rate, not the efficiency of the inner solver. Here we extend these results and give some theoretical justification for the tuned preconditioner.

The main part of this subsection, Subsection 6.4.2 provides analysis and implementation for the tuned preconditioner applied to the nonsymmetric problem. For completeness we consider both left and right preconditioning, although the right preconditioner is easier to treat in terms of the stopping conditions (see, for example [111]).

6.4.1 The ideal preconditioner

In this subsection we will discuss a rather theoretical case. Suppose $\mathbf{x}^{(i)} = \mathbf{x}_1$ (that is, convergence has occurred) and consider the problem of computing

$$(\mathbf{A} - \sigma\mathbf{M})\mathbf{y} = \mathbf{M}\mathbf{x}_1 \tag{6.46}$$

using preconditioned solves. Then we may consider the ideal preconditioner

$$\mathbb{P} = \mathbf{P} + (\mathbf{A} - \mathbf{P})\frac{\mathbf{x}_1}{\|\mathbf{x}_1\|}\frac{\mathbf{x}_1^H}{\|\mathbf{x}_1\|}, \tag{6.47}$$

which satisfies $\mathbb{P}\mathbf{x}_1 = \mathbf{A}\mathbf{x}_1 = \lambda_1\mathbf{M}\mathbf{x}_1$, that is, $\mathbf{x}_1$ is a generalised eigenvector of both $\mathbf{A}$ and $\mathbb{P}$. Furthermore we have $(\mathbf{A}-\sigma\mathbf{M})\mathbb{P}^{-1}\mathbf{M}\mathbf{x}_1 = \dfrac{\lambda_1 - \sigma}{\lambda_1}\mathbf{M}\mathbf{x}_1$, that is $\mathbf{M}\mathbf{x}_1$ (the right hand side of (6.46)) is an exact eigenvector of $(\mathbf{A} - \sigma\mathbf{M})\mathbb{P}^{-1}$ (the iteration matrix for GMRES) for any shift σ. In that case GMRES would converge in just one step. For a zero shift the corresponding eigenvector is one. The next theorem shows that if $\mathbf{P}$ is a good approximation to $\mathbf{A}$, and if we use a zero shift, then the spectrum of $\mathbf{A}\mathbb{P}^{-1}$ should be clustered around one.

Theorem 6.20. *Let $\mathbb{P}$ be given by (6.47) and assume that the block diagonalisation (6.5) is valid. Then the matrix $\mathbf{A}\mathbb{P}^{-1}$ has the same eigenvalues as the matrix*

$$\begin{bmatrix} 1 & \frac{1}{s_{11}}\mathbf{e}_1^T\mathbf{U}^{-1}\mathbf{A}\mathbb{P}^{-1}\mathbf{U}\bar{\mathbf{I}}_{n-1} \\ \mathbf{0} & (\bar{\mathbf{I}}_{n-1}^T\mathbf{U}^{-1}\mathbf{A}\mathbf{X}\bar{\mathbf{I}}_{n-1})(\bar{\mathbf{I}}_{n-1}^T\mathbf{U}^{-1}\mathbf{P}\mathbf{X}\bar{\mathbf{I}}_{n-1})^{-1} \end{bmatrix}.$$

Proof. With $\mathbf{U}^{-1}\mathbf{M}\mathbf{x}_1 = s_{11}\mathbf{e}_1$ and $\mathbf{A}\mathbb{P}^{-1}\mathbf{M}\mathbf{x}_1 = \mathbf{M}\mathbf{x}_1$ we have that $\mathbf{A}\mathbb{P}^{-1}$ has the same eigenvalues as

$$[\mathbf{U}^{-1}\mathbf{M}\mathbf{x}_1\ \bar{\mathbf{I}}_{n-1}]^{-1}\mathbf{U}^{-1}\mathbf{A}\mathbb{P}^{-1}\mathbf{U}[\mathbf{U}^{-1}\mathbf{M}\mathbf{x}_1\ \bar{\mathbf{I}}_{n-1}] = \begin{bmatrix} 1 & \frac{1}{s_{11}}\mathbf{e}_1^T\mathbf{U}^{-1}\mathbf{A}\mathbb{P}^{-1}\mathbf{U}\bar{\mathbf{I}}_{n-1} \\ \mathbf{0} & \bar{\mathbf{I}}_{n-1}^T\mathbf{U}^{-1}\mathbf{A}\mathbb{P}^{-1}\mathbf{U}\bar{\mathbf{I}}_{n-1} \end{bmatrix}.$$

Introducing $\mathbf{U}^{-1}\mathbf{A}\mathbf{X} = \operatorname{diag}(t_{11}, \mathbf{T}_{22})$ we can write

$$[\mathbf{U}^{-1}\mathbf{M}\mathbf{x}_1\ \bar{\mathbf{I}}_{n-1}]^{-1}\mathbf{U}^{-1}\mathbf{A}\mathbb{P}^{-1}\mathbf{U}[\mathbf{U}^{-1}\mathbf{M}\mathbf{x}_1\ \bar{\mathbf{I}}_{n-1}] = \begin{bmatrix} 1 & \frac{1}{s_{11}}\mathbf{e}_1^T\mathbf{U}^{-1}\mathbf{A}\mathbb{P}^{-1}\mathbf{U}\bar{\mathbf{I}}_{n-1} \\ \mathbf{0} & \mathbf{T}_{22}\bar{\mathbf{I}}_{n-1}^T\mathbf{X}^{-1}\mathbb{P}^{-1}\mathbf{U}\bar{\mathbf{I}}_{n-1} \end{bmatrix}.$$

Finally, observe that $\mathbb{P}\mathbf{X}\bar{\mathbf{I}}_{n-1} = \mathbb{P}\mathbf{X}_2 = \mathbf{P}\mathbf{X}_2$ and hence

$$\bar{\mathbf{I}}_{n-1}^T\mathbf{X}^{-1}\mathbb{P}^{-1}\mathbf{U}\bar{\mathbf{I}}_{n-1} = \bar{\mathbf{I}}_{n-1}^T(\mathbf{U}^{-1}\mathbb{P}\mathbf{X})^{-1}\bar{\mathbf{I}}_{n-1}.$$

We have $\mathbf{U}^{-1}\mathbb{P}\mathbf{X} = \mathbf{U}^{-1}\mathbb{P}[\mathbf{x}_1\ \mathbf{X}_2] = \mathbf{U}^{-1}[\lambda_1\mathbf{M}\mathbf{x}_1\ \mathbf{P}\mathbf{X}_2] = [\lambda_1 s_{11}\mathbf{e}_1\ \mathbf{U}^{-1}\mathbf{P}\mathbf{X}_2]$. Taking the inverse using the block structure of $\mathbf{U}^{-1}\mathbb{P}\mathbf{X}$ and using $\mathbf{T}_{22} = \bar{\mathbf{I}}_{n-1}^T\mathbf{U}^{-1}\mathbf{A}\mathbf{X}\bar{\mathbf{I}}_{n-1}$ gives the result. □

Theorem 6.20 shows that one eigenvalue of $\mathbf{A}\mathbb{P}^{-1}$ is equal to one and all the other eigenvalues are equal to eigenvalues of $(\bar{\mathbf{I}}_{n-1}^T\mathbf{U}^{-1}\mathbf{A}\mathbf{X}\bar{\mathbf{I}}_{n-1})(\bar{\mathbf{I}}_{n-1}^T\mathbf{U}^{-1}\mathbf{P}\mathbf{X}\bar{\mathbf{I}}_{n-1})^{-1}$. Therefore if $\mathbf{P}$ is a good preconditioner for $\mathbf{A}$, then $\bar{\mathbf{I}}_{n-1}^T\mathbf{U}^{-1}\mathbf{P}\mathbf{X}\bar{\mathbf{I}}_{n-1}$ will be a good approximation to $\bar{\mathbf{I}}_{n-1}^T\mathbf{U}^{-1}\mathbf{A}\mathbf{X}\bar{\mathbf{I}}_{n-1}$ and hence the eigenvalues of $\mathbf{A}\mathbb{P}^{-1}$ should be clustered around one. Note that for a shifted system $(\mathbf{A}-\sigma\mathbf{M})\mathbb{P}^{-1}$ these eigenvalues are shifted, but, more importantly, also clustered.

Now, since $\mathbf{M}\mathbf{x}_1$ is a simple eigenvector of $(\mathbf{A}-\sigma\mathbf{M})\mathbb{P}^{-1}$, the block-factorisation

$$(\mathbf{A}-\sigma\mathbf{M})\mathbb{P}^{-1} = \begin{bmatrix} \bar{\bar{\mathbf{w}}}_1 & \bar{\bar{\mathbf{W}}}_1^{\perp} \end{bmatrix} \begin{bmatrix} \bar{\bar{\mu}}_1 & \bar{\bar{\mathbf{n}}}_{12}^H \\ \mathbf{0} & \bar{\bar{\mathbf{N}}}_{22} \end{bmatrix} \begin{bmatrix} \bar{\bar{\mathbf{w}}}_1 & \bar{\bar{\mathbf{W}}}_1^{\perp} \end{bmatrix} \tag{6.48}$$

and block-diagonalisation

$$(\mathbf{A}-\sigma\mathbf{M})\mathbb{P}^{-1} = \begin{bmatrix} \bar{\bar{\mathbf{w}}}_1 & \bar{\bar{\mathbf{W}}}_2 \end{bmatrix} \begin{bmatrix} \bar{\bar{\mu}}_1 & \mathbf{0}^H \\ \mathbf{0} & \bar{\bar{\mathbf{C}}} \end{bmatrix} \begin{bmatrix} \bar{\bar{\mathbf{v}}}_1^H \\ \bar{\bar{\mathbf{V}}}_2^H \end{bmatrix}$$

exist, where $\bar{\bar{\mathbf{w}}}_1 = \dfrac{\mathbf{M}\mathbf{x}_1}{\|\mathbf{M}\mathbf{x}_1\|}$ and $\bar{\bar{\mu}}_1 = \dfrac{\lambda_1-\sigma}{\lambda_1}$.

Clearly, the perfect preconditioner introduced in this section is only of theoretical concern. In the next section we will introduce a practical preconditioner, but the ideal preconditioner will be used to prove our main result about the independence of $k^{(i)}$ on i (Theorem 6.23).

6.4.2 The tuned preconditioner

The ideal preconditioner cannot be used in practice since $\mathbf{x}_1$ is unknown. Therefore we propose to use

$$\mathbb{P}_i = \mathbf{P} + (\mathbf{A}-\mathbf{P})\frac{\mathbf{x}^{(i)}}{\|\mathbf{x}^{(i)}\|}\frac{\mathbf{x}^{(i)H}}{\|\mathbf{x}^{(i)}\|}, \tag{6.49}$$

which satisfies

$$\mathbb{P}_i\mathbf{x}^{(i)} = \mathbf{A}\mathbf{x}^{(i)}. \tag{6.50}$$

the same condition as used in Chapters 2 and 4 (see also [42] and [43]). Clearly, as $\mathbf{x}^{(i)} \to \mathbf{x}_1$ the tuned preconditioner $\mathbb{P}_i$ will act like the ideal preconditioner $\mathbb{P}$.

The following lemma is an extension of Lemma 6.12 and is needed to prove the efficiency of the tuned preconditioner given by (6.49).

Lemma 6.21. *Let $\mathcal{E}^{(i)}$ be as in Lemma 6.12 with $\|\mathcal{E}^{(i)}\| \le C_3\varepsilon^{(i)}$. Then*

$$\|(\mathbf{A}-\sigma\mathbf{M})\mathbb{P}^{-1} - (\mathbf{A}-\sigma\mathbf{M})\mathbb{P}_i^{-1}\| \le \beta_2\varepsilon^{(i)},$$

where β_2 is independent of i for large enough i.

Proof. Using (6.36) we have $\mathbb{P}_i = \mathbb{P} + (\mathbf{A}-\mathbf{P})\mathcal{E}^{(i)}$ and therefore we can write

$$\begin{aligned}(\mathbf{A}-\sigma\mathbf{M})\mathbb{P}^{-1} - (\mathbf{A}-\sigma\mathbf{M})\mathbb{P}_i^{-1} &= (\mathbf{A}-\sigma\mathbf{M})\mathbb{P}^{-1}(\mathbb{P}_i-\mathbb{P})\mathbb{P}_i^{-1} \\ &= (\mathbf{A}-\sigma\mathbf{M})\mathbb{P}^{-1}(\mathbf{A}-\mathbf{P})\mathcal{E}^{(i)}(\mathbb{P}+(\mathbf{A}-\mathbf{P})\mathcal{E}^{(i)})^{-1}\end{aligned}$$

and with $\|(\mathbf{A}-\sigma\mathbf{M})\mathbb{P}^{-1}(\mathbf{A}-\mathbf{P})\| \leq C_8$,

$$\begin{aligned}\|(\mathbf{A}-\sigma\mathbf{M})\mathbb{P}^{-1}-(\mathbf{A}-\sigma\mathbf{M})\mathbb{P}_i{}^{-1}\| &\leq C_8\|\mathcal{E}^{(i)}\|\|\mathbb{P}^{-1}\|\|(\mathbf{I}+\mathbb{P}^{-1}(\mathbf{A}-\mathbf{P})\mathcal{E}^{(i)})^{-1}\| \\ &\leq \frac{C_8\|\mathbb{P}^{-1}\|}{1-\|\mathbb{P}^{-1}(\mathbf{A}-\mathbf{P})\|\|\mathcal{E}^{(i)}\|}\|\mathcal{E}^{(i)}\| \\ &\leq \frac{C_8\|\mathbb{P}^{-1}\|}{1-\|\mathbb{P}^{-1}(\mathbf{A}-\mathbf{P})\|C_3\varepsilon^{(i)}}C_3\varepsilon^{(i)}.\end{aligned}$$

Finally, since $\varepsilon^{(i)}$ is decreasing as i increases there exists $\beta_2 > 0$ independent of i such that

$$\frac{C_8\|\mathbb{P}^{-1}\|}{1-\|\mathbb{P}^{-1}(\mathbf{A}-\mathbf{P})\|C_3\varepsilon^{(i)}}C_3 \leq \beta_2,$$

for large enough i which provides the result. □

Now, assume that $\bar{\bar{\mathbf{w}}}_1$ is a simple eigenvector of $(\mathbf{A}-\sigma\mathbf{M})\mathbb{P}^{-1}$. Then for i large enough, $(\mathbf{A}-\sigma\mathbf{M})\mathbb{P}_i^{-1}$ can be block diagonalised as

$$(\mathbf{A}-\sigma\mathbf{M})\mathbb{P}_i^{-1} = \begin{bmatrix}\bar{\bar{\mathbf{w}}}_1^{(i)} & \bar{\bar{\mathbf{W}}}_2^{(i)}\end{bmatrix}\begin{bmatrix}\bar{\bar{\mu}}_1^{(i)} & \mathbf{0}^H \\ \mathbf{0} & \bar{\bar{\mathbf{C}}}^{(i)}\end{bmatrix}\begin{bmatrix}\bar{\bar{\mathbf{v}}}_1^{(i)} & \bar{\bar{\mathbf{V}}}_2^{(i)}\end{bmatrix}^H, \tag{6.51}$$

and similar results hold as in Lemma 6.13, namely

$$\begin{aligned}|\bar{\bar{\mu}}_1-\bar{\bar{\mu}}_1^{(i)}| &\leq d_1\varepsilon^{(i)}, \\ \|\bar{\bar{\mathbf{C}}}-\bar{\bar{\mathbf{C}}}^{(i)}\| &\leq d_2\varepsilon^{(i)}, \\ \|\bar{\bar{\mathbf{V}}}_2-\bar{\bar{\mathbf{V}}}_2^{(i)}\| &\leq d_3\varepsilon^{(i)},\end{aligned}$$

where d_1, d_2 and d_3 are positive constants independent of i. Hence, we have the following theorem.

Theorem 6.22 (Right tuned preconditioner for nonsymmetric eigenproblem)**.** *Let the assumptions of Theorem 6.6 hold and consider the solution of*

$$(\mathbf{A}-\sigma\mathbf{M})\mathbb{P}_i{}^{-1}\tilde{\mathbf{y}}^{(i)} = \mathbf{M}\mathbf{x}^{(i)}, \quad \textit{where} \quad \mathbf{y}^{(i)} = \mathbb{P}_i{}^{-1}\tilde{\mathbf{y}}^{(i)}. \tag{6.52}$$

with $\mathbb{P}_i$ chosen as in (6.50). Let $\bar{\bar{\mu}}_1$ be a simple eigenvalue of $(\mathbf{A}-\sigma\mathbf{M})\mathbb{P}^{-1}$ such that (6.48) holds and let $\varepsilon^{(i)} = \|\mathbf{p}^{(i)}\|$. Furthermore, let $|\rho(\mathbf{x}^{(i)})| > K$, for some positive constant K. Then

$$\|\mathcal{P}^{(i)}\mathbf{M}\mathbf{x}^{(i)}\| \leq C_9\|\mathcal{P}^{(i)}\|\|\mathbf{r}^{(i)}\| \tag{6.53}$$

for some positive constant C_9 independent of i for large enough i, where $\mathcal{P}^{(i)} = \mathbf{I} - \bar{\bar{\mathbf{w}}}_1^{(i)}\bar{\bar{\mathbf{v}}}_1^{(i)H}$ and $\bar{\bar{\mathbf{v}}}_1^{(i)}$ and $\bar{\bar{\mathbf{w}}}_1^{(i)}$ are left and right eigenvectors of $(\mathbf{A}-\sigma\mathbf{M})\mathbb{P}_i^{-1}$.

Proof. Using $\mathcal{P}^{(i)}\bar{\bar{\mathbf{w}}}_1^{(i)} = 0$ we obtain

$$\mathcal{P}^{(i)}\mathbf{M}\mathbf{x}^{(i)} = \mathcal{P}^{(i)}(\mathbf{M}\mathbf{x}^{(i)} - \alpha\bar{\bar{\mathbf{w}}}_1^{(i)}) \quad \forall\alpha\in\mathbb{C}. \tag{6.54}$$

Then, with the eigenvalue residual given by (6.4) and the condition (6.50) as well as $\rho(\mathbf{x}^{(i)}) \neq 0$ we get

$$\begin{aligned}
(\mathbf{A}-\sigma\mathbf{M})\mathbb{P}_i^{-1}\mathbf{M}\mathbf{x}^{(i)} &= (\mathbf{A}-\sigma\mathbf{M})\mathbb{P}_i^{-1}\left(\frac{\mathbf{A}\mathbf{x}^{(i)}-\mathbf{r}^{(i)}}{\rho(\mathbf{x}^{(i)})}\right)\\
&= (\mathbf{A}-\sigma\mathbf{M})\left(\frac{\mathbf{x}^{(i)}-\mathbb{P}_i^{-1}\mathbf{r}^{(i)}}{\rho(\mathbf{x}^{(i)})}\right)\\
&= \left(\frac{\rho(\mathbf{x}^{(i)})-\sigma}{\rho(\mathbf{x}^{(i)})}\right)\mathbf{M}\mathbf{x}^{(i)} + \frac{1}{\rho(\mathbf{x}^{(i)})}(\mathbb{P}_i-(\mathbf{A}-\sigma^{(i)}\mathbf{M}))\mathbb{P}_i^{-1}\mathbf{r}^{(i)}.
\end{aligned}$$

We can rewrite this equation as

$$(\mathbf{A}-\sigma\mathbf{M})\mathbb{P}_i^{-1}\left(\mathbf{M}\mathbf{x}^{(i)}+\frac{\mathbf{r}^{(i)}}{\rho(\mathbf{x}^{(i)})}\right) = \frac{\rho(\mathbf{x}^{(i)})-\sigma}{\rho(\mathbf{x}^{(i)})}\left(\mathbf{M}\mathbf{x}^{(i)}+\frac{\mathbf{r}^{(i)}}{\rho(\mathbf{x}^{(i)})}\right)+\frac{\sigma}{\rho(\mathbf{x}^{(i)})^2}\mathbf{r}^{(i)}. \tag{6.55}$$

Hence, $\mathbf{M}\mathbf{x}^{(i)}+\frac{\mathbf{r}^{(i)}}{\rho(\mathbf{x}^{(i)})}$ is an approximate eigenvector of $(\mathbf{A}-\sigma\mathbf{M})\mathbb{P}_i^{-1}$ with approximate eigenvalue $\frac{\rho(\mathbf{x}^{(i)})-\sigma}{\rho(\mathbf{x}^{(i)})}$. Using (6.4) again we have $\frac{\mathbf{A}\mathbf{x}^{(i)}}{\rho(\mathbf{x}^{(i)})} = \mathbf{M}\mathbf{x}^{(i)}+\frac{\mathbf{r}^{(i)}}{\rho(\mathbf{x}^{(i)})}$. We normalise this approximate eigenvector and obtain

$$(\mathbf{A}-\sigma\mathbf{M})\mathbb{P}_i^{-1}\frac{\mathbf{A}\mathbf{x}^{(i)}}{\|\mathbf{A}\mathbf{x}^{(i)}\|} = \frac{\rho(\mathbf{x}^{(i)})-\sigma}{\rho(\mathbf{x}^{(i)})}\frac{\mathbf{A}\mathbf{x}^{(i)}}{\|\mathbf{A}\mathbf{x}^{(i)}\|}+\frac{\sigma}{\|\mathbf{A}\mathbf{x}^{(i)}\|\rho(\mathbf{x}^{(i)})}\mathbf{r}^{(i)}.$$

For i large enough (and hence $\varepsilon^{(i)}$ as well as $\|\mathbf{r}^{(i)}\|$ small enough) we can apply Proposition 6.5 to $(\mathbf{A}-\sigma\mathbf{M})\mathbb{P}_i^{-1}$ with $\hat{\mathbf{w}} = \frac{\mathbf{A}\mathbf{x}^{(i)}}{\|\mathbf{A}\mathbf{x}^{(i)}\|}$ to get

$$\left\|\frac{\mathbf{A}\mathbf{x}^{(i)}}{\|\mathbf{A}\mathbf{x}^{(i)}\|}-\frac{\bar{\bar{\mathbf{w}}}_1^{(i)}}{\sqrt{1+\bar{\bar{\mathbf{p}}}_i^H\bar{\bar{\mathbf{p}}}_i}}\right\| \leq \frac{2\|\bar{\mathbf{E}}^{(i)}\|}{\operatorname{sep}(\bar{\bar{\mu}}_1^{(i)},\bar{\bar{\mathbf{N}}}_{22}^{(i)})-2\|\bar{\mathbf{E}}^{(i)}\|}, \tag{6.56}$$

where $\bar{\mathbf{E}}^{(i)} = \frac{\sigma\mathbf{r}^{(i)}(\mathbf{A}\mathbf{x}^{(i)})^H}{\|\mathbf{A}\mathbf{x}^{(i)}\|^2\rho(\mathbf{x}^{(i)})}$. We have

$$\|\bar{\mathbf{E}}^{(i)}\| \leq \frac{\sigma}{\|\mathbf{A}\mathbf{x}^{(i)}\||\rho(\mathbf{x}^{(i)})|}\|\mathbf{r}^{(i)}\| \leq C_{10}\|\mathbf{r}^{(i)}\|. \tag{6.57}$$

Multiplying (6.56) by $\frac{\|\mathbf{A}\mathbf{x}^{(i)}\|}{\rho(\mathbf{x}^{(i)})}$ and using (6.57) we have

$$\left\|\mathbf{M}\mathbf{x}^{(i)}+\frac{\mathbf{r}^{(i)}}{\rho(\mathbf{x}^{(i)})}-\frac{\|\mathbf{A}\mathbf{x}^{(i)}\|}{\rho(\mathbf{x}^{(i)})}\frac{\bar{\bar{\mathbf{w}}}_1^{(i)}}{\sqrt{1+\bar{\bar{\mathbf{p}}}_i^H\bar{\bar{\mathbf{p}}}_i}}\right\| \leq \frac{2\|\bar{\mathbf{F}}^{(i)}\|}{\operatorname{sep}(\bar{\bar{\mu}}_1^{(i)},\bar{\bar{\mathbf{N}}}_{22}^{(i)})-2\|\bar{\mathbf{E}}^{(i)}\|}, \tag{6.58}$$

where, with $|\rho(\mathbf{x}^{(i)})| > K$,

$$\|\bar{\mathbf{F}}^{(i)}\| = \frac{\|\mathbf{A}\mathbf{x}^{(i)}\|}{|\rho(\mathbf{x}^{(i)})|}\|\bar{\mathbf{E}}^{(i)}\| \leq \frac{|\sigma|}{|\rho(\mathbf{x}^{(i)})|^2}\|\mathbf{r}^{(i)}\| \leq \frac{|\sigma|}{K^2}\|\mathbf{r}^{(i)}\|. \tag{6.59}$$

Furthermore (6.58) yields

$$\left\| \mathbf{M}\mathbf{x}^{(i)} - \frac{\|\mathbf{A}\mathbf{x}^{(i)}\|}{|\rho(\mathbf{x}^{(i)})|} \frac{\bar{\bar{\mathbf{w}}}_1^{(i)}}{\sqrt{1+\bar{\bar{\mathbf{p}}}_i^H \bar{\bar{\mathbf{p}}}_i}} \right\| - \frac{\|\mathbf{r}^{(i)}\|}{|\rho(\mathbf{x}^{(i)})|} \leq \frac{2\|\bar{\mathbf{F}}^{(i)}\|}{\mathrm{sep}(\bar{\bar{\mu}}_1^{(i)}, \bar{\bar{\mathbf{N}}}_{22}^{(i)}) - 2\|\bar{\mathbf{E}}^{(i)}\|}.$$

Setting $\alpha := \dfrac{\|\mathbf{A}\mathbf{x}^{(i)}\|}{|\rho(\mathbf{x}^{(i)})|\sqrt{1+\bar{\bar{\mathbf{p}}}_i^H \bar{\bar{\mathbf{p}}}_i}}$ in (6.54) we use this bound to obtain

$$\begin{aligned} \|\mathcal{P}^{(i)}\mathbf{M}\mathbf{x}^{(i)}\| &\leq \|\mathcal{P}^{(i)}(\mathbf{M}\mathbf{x}^{(i)} - \alpha\bar{\bar{\mathbf{w}}}_1^{(i)})\| \\ &\leq \|\mathcal{P}^{(i)} \left(\|\frac{2\|\bar{\mathbf{F}}^{(i)}\|}{\mathrm{sep}(\bar{\bar{\mu}}_1^{(i)}, \bar{\bar{\mathbf{N}}}_{22}^{(i)}) - 2\|\bar{\mathbf{E}}^{(i)}\|} + \frac{\|\mathbf{r}^{(i)}\|}{|\rho(\mathbf{x}^{(i)})|} \right). \end{aligned}$$

Finally, with (6.59), (6.57) and $|\rho(\mathbf{x}^{(i)})| > K$ we obtain

$$\|\mathcal{P}^{(i)}\mathbf{M}\mathbf{x}^{(i)}\| \leq \|\mathcal{P}^{(i)}\| \left(\frac{2\frac{|\sigma|}{K^2}\|\mathbf{r}^{(i)}\|}{\mathrm{sep}(\bar{\bar{\mu}}_1^{(i)}, \bar{\bar{\mathbf{N}}}_{22}^{(i)}) - 2C_{10}\|\mathbf{r}^{(i)}\|} + \frac{\|\mathbf{r}^{(i)}\|}{K} \right).$$

Since $\|\mathbf{r}^{(i)}\| \leq C_6\varepsilon^{(i)}$ from (6.8), as well as $|\bar{\bar{\mu}}_1^{(i)} - \bar{\bar{\mu}}_1| \leq d_1\varepsilon^{(i)}$ and $\|\bar{\bar{\mathbf{N}}}_{22}^{(i)} - \bar{\bar{\mathbf{N}}}_{22}\| \leq d_4\varepsilon^{(i)}$ for appropriately chosen constants d_1 and d_4 and $\varepsilon^{(i)}$ small enough (see [137, p. 234] and comments after (6.51)), the term

$$\frac{2}{\mathrm{sep}(\bar{\bar{\mu}}_1^{(i)}, \bar{\bar{\mathbf{N}}}_{22}^{(i)}) - 2C_{10}\|\mathbf{r}^{(i)}\|} \leq \frac{2}{\mathrm{sep}(\bar{\bar{\mu}}_1, \bar{\bar{\mathbf{N}}}_{22}) - d_1\varepsilon^{(i)} - d_4\varepsilon^{(i)} - 2C_6C_{10}\varepsilon^{(i)}},$$

can be bounded by a constant independent of i for large enough i. Hence the result (6.53) is obtained for an appropriately chosen constant C_9. □

We can finally prove the following Theorem which provides the main result of this section.

Theorem 6.23. *Let the assumptions of Theorem 6.22 be satisfied. Then the number $k^{(i)}$ of inner iterations used by preconditioned GMRES to compute $\tilde{\mathbf{y}}_{k^{(i)}}$ satisfying the stopping criterion*

$$\|(\mathbf{A} - \sigma\mathbf{M})\mathbb{P}_i{}^{-1}\tilde{\mathbf{y}}_{k^{(i)}} - \mathbf{M}\mathbf{x}^{(i)}\| \leq \tau^{(i)} = \delta\|\mathbf{r}^{(i)}\|,$$

is bounded independently of i for large enough i.

Proof. Let Ψ and $\mathbf{E}$ be given by Proposition 6.8 applied to $\bar{\bar{\mathbf{C}}}$. For large enough i (and hence small enough $\varepsilon^{(i)}$) decomposition (6.51) exists. By Proposition 6.10 the residual obtained after $k^{(i)}$ iterations of GMRES starting with 0 is less than $\tau^{(i)} = \delta\|\mathbf{r}^{(i)}\|$ if

$$k^{(i)} \geq 1 + \frac{1}{\log|\Psi(0)|} \left(\log \frac{S_\delta\|\bar{\bar{\mu}}_1^{(i)}\mathbf{I} - \bar{\bar{\mathbf{C}}}^{(i)}\|}{|\bar{\bar{\mu}}_1^{(i)}|}\|\bar{\bar{\mathbf{V}}}_2^{(i)}\| + \log \frac{\|\mathcal{P}^{(i)}\mathbf{M}\mathbf{x}^{(i)}\|}{\delta\|\mathbf{r}^{(i)}\|} \right).$$

A similar proof to the one of Theorem 6.16 and using the results of Theorem 6.22 yields bounds for the two terms in brackets by constants independent of i for large enough i and hence gives the result. □

Lemma 6.24 (Implementation of $\mathbb{P}_i^{-1}$). *Let $\mathbf{x}^{(i)}$ be the approximate eigenvector obtained from the ith iteration of Algorithm 4. Set $\mathbf{u}^{(i)} = \mathbf{A}\mathbf{x}^{(i)} - \mathbf{P}\mathbf{x}^{(i)}$, where $\mathbf{P}$ is a standard preconditioner for $\mathbf{A}$. Then $\mathbb{P}_i = \mathbf{P} + \mathbf{u}^{(i)}\dfrac{\mathbf{x}^{(i)H}}{\|\mathbf{x}^{(i)}\|^2}$ ensures (6.50) and we have*

$$\mathbb{P}_i^{-1} = \left(\mathbf{I} - \frac{(\mathbf{P}^{-1}\mathbf{A}\mathbf{x}^{(i)} - \mathbf{x}^{(i)})\mathbf{x}^{(i)H}}{\mathbf{x}^{(i)H}\mathbf{P}^{-1}\mathbf{A}\mathbf{x}^{(i)}}\right)\mathbf{P}^{-1}.$$

Proof. Sherman-Morrison Formula. □

Note that only one extra back solve $\mathbf{P}^{-1}\mathbf{A}\mathbf{x}^{(i)}$ per outer iteration is necessary, which can be computed before the actual inner iteration. All further extra costs are inner products.

We end this subsection by two remarks before giving numerical results in the next subsection.

Remark 6.25 (Left tuned preconditioner). *For left preconditioning, namely*

$$\mathbb{P}_i^{-1}(\mathbf{A} - \sigma^{(i)}\mathbf{M})\mathbf{y}^{(i)} = \mathbb{P}_i^{-1}\mathbf{M}\mathbf{x}^{(i)} \tag{6.60}$$

the tuning works similarly. For $\rho(\mathbf{x}^{(i)}) \neq 0$ we have

$$\mathbb{P}_i^{-1}(\mathbf{A}-\sigma\mathbf{M})\left(\mathbb{P}_i^{-1}\mathbf{M}\mathbf{x}^{(i)} + \frac{\mathbb{P}_i^{-1}\mathbf{r}^{(i)}}{\rho(\mathbf{x}^{(i)})}\right) = \frac{\rho(\mathbf{x}^{(i)}) - \sigma}{\rho(\mathbf{x}^{(i)})}\left(\mathbb{P}_i^{-1}\mathbf{M}\mathbf{x}^{(i)} + \frac{\mathbb{P}_i^{-1}\mathbf{r}^{(i)}}{\rho(\mathbf{x}^{(i)})}\right) + \frac{\sigma\mathbb{P}_i^{-1}\mathbf{r}^{(i)}}{\rho(\mathbf{x}^{(i)})^2},$$

so that $\mathbb{P}_i^{-1}\mathbf{M}\mathbf{x}^{(i)} + \dfrac{\mathbb{P}_i^{-1}\mathbf{r}^{(i)}}{\rho(\mathbf{x}^{(i)})}$ is an approximate eigenvector of $\mathbb{P}_i^{-1}(\mathbf{A} - \sigma\mathbf{M})$. Then, using similar ideas as in the proof of Theorem 6.22 we obtain

$$\|\mathcal{P}^{(i)}\mathbb{P}_i^{-1}\mathbf{M}\mathbf{x}^{(i)}\| = \mathcal{O}(\|\mathbf{r}^{(i)}\|),$$

where $\mathcal{P}^{(i)}$ is the appropriate oblique projection.

Remark 6.26. *Note that as a consequence of (6.50) we have*

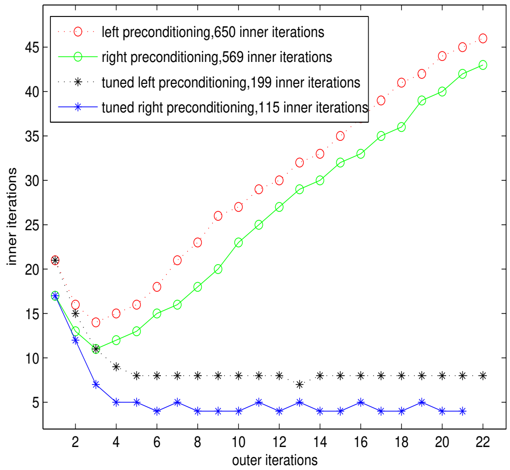
$$(\mathbf{A}\mathbb{P}_i^{-1})\mathbf{A}\mathbf{x}^{(i)} = \mathbf{A}\mathbf{x}^{(i)},$$

that is, $\mathbf{A}\mathbf{x}^{(i)}$ is an eigenvector of $\mathbf{A}\mathbb{P}_i^{-1}$ corresponding to eigenvalue 1.

6.4.3 Numerical examples

In this section we give some numerical examples to illustrate the performance of the tuned preconditioner.

Example 6.27. *Use the generalised eigenproblem of Example 6.19. System (6.11) is solved by preconditioned GMRES with the tolerance $\tau^{(i)} = \min\{0.01, 0.01\,\|\mathbf{r}^{(i)}\|\}$ and the overall computation stops when $\|\mathbf{r}^{(i)}\| < 10^{-11}$. We compare preconditioner $\mathbf{P}$ obtained from an incomplete LU-decomposition of $\mathbf{A}$ with drop tolerance 0.1 and the right tuned preconditioner (6.52) as well as the left preconditioner (6.60) which both satisfy (6.50).*



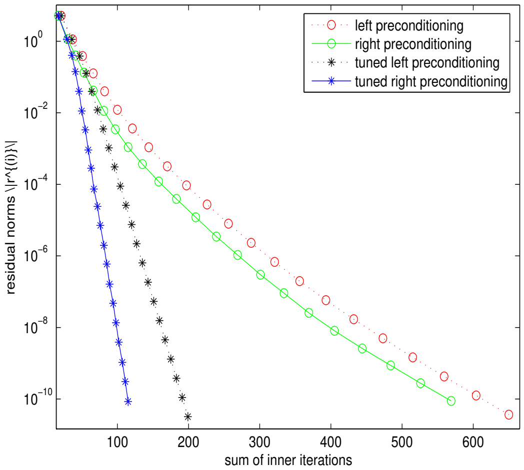


Figure 6-3: *Number of inner iterations against outer iterations with standard and tuned (left and right) preconditioning (Example 6.27)*

Figure 6-4: *Eigenvalue residual norms against the total number of inner iterations with standard and tuned preconditioning (Example 6.27)*

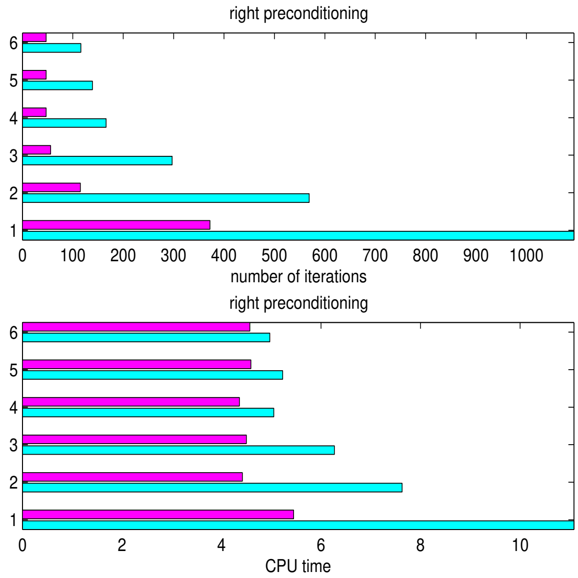


Figure 6-5: *Numerical results for Example 6.27. Comparison of total number of inner iterations and CPU times for different drop tolerances of the preconditioner*

Furthermore, we consider the same problem with only right preconditioning (since it is the more natural preconditioner) but with different drop tolerances. The drop tolerances used are 0.1^i where $i = 1, \ldots, 5$. The total number of iterations and CPU times are compared.

Figure 6-3 displays the number of inner iterations per outer iteration for the left and right preconditioning, both with or without the tuning strategy. We can see the ex-

Table 6.1: *Set of test matrices from the collection [13]*

	Matrix name/s	size n	Description
1	`stiff.mtx/mass.mtx`	961	Convection-Diffusion operator Ex. 6.27
2	`dwa512.mtx/dwb512.mtx`	512	Square Dielectric Waveguide
3	`bcsstk08.mtx/bcsstm08.mtx`	1074	BCS Structural Engineering Matrix
4	`rdb1250l.mtx`	1250	Reaction-Diffusion Brusselator Model $L = 1.0$
5	`cdde1.mtx`	961	Model 2D Convection-Diffusion operator $p_1 = 1$, $p_2 = 2$, $p_3 = 30$
6	`olm2000.mtx`	2000	Olmstead Model

Table 6.2: *Setup for set of test matrices from the collection [13]*

	Matrix name/s	droptol	shift σ	eigenvalue	$\tau^{(0)}$	final $\mathbf{r}^{(i)}$
1	`stiff.mtx/mass.mtx`	1	85	91.6223	0.01	10e-11
2	`dwa512.mtx/dwb512.mtx`	0.001	0.001	1.3957e-3	0.001	10e-8
3	`bcsstk08.mtx/bcsstm08.mtx`	0.01	10	6.90070	0.01	10e-11
4	`rdb1250l.mtx`	0.1	-0.325	-3.20983e-1	0.1	10e-11
5	`cdde1.mtx`	0.1	0.001	-5.17244e-3	0.1	10e-15
6	`olm2000.mtx`	0.1	4.3	4.51010	0.1	10e-9

pected logarithmic increase for the standard preconditioner. Furthermore, Figure 6-4 shows the overall costs of the iteration. We can see that the cost of the tuning is less than a third than the cost for applying the standard preconditioner only.

Furthermore, comparing Figures 6-3 and 6-1 (and Figures 6-4 and 6-2 respectively) we see that that the combination of tuning and preconditioning is more effective than just the tuning strategy alone.

The overall number of iterations which is 1306 without any tuning or preconditioning (see Figures 6-1 and 6-2) can be reduced to only 425 iterations by using tuning (see Figures 6-1 and 6-2), to 569 iterations by using a standard right preconditioner and to only 115 iterations with right preconditioning (see Figures 6-3 and 6-4). This is less than 10 per cent of the initial number of iterations.

Figure 6-5 shows a comparison between different drop tolerances for the preconditioner applied to the the convection-diffusion operator from Example 6.19 used to find the eigenvalue near the shift $\sigma = 20$. We can see that the tuned preconditioner performs always better than the standard preconditioner in terms of CPU time and the total number of inner iterations. Also, we observe that tuning is less effective in cases where the preconditioner is very good, since $\mathbf{P}\mathbf{x}^{(i)} \approx \mathbf{A}\mathbf{x}^{(i)}$ is satisfied because $\mathbf{P} \approx \mathbf{A}$. The worse the standard preconditioner, the larger the gain in the total number of inner iterations by using a tuning approach. The reduction in the iteration numbers is always more than 50 per cent, while the gain in CPU time gets less when the drop tolerance gets smaller, that is then the preconditioner gets a better approximation to $\mathbf{A}$. In that case the extra solve for the preconditioner gets more expensive (due to the fill-in in the incomplete LU-factors.

Example 6.28. *We choose a subset of real $n \times n$ matrices from the Matrix Market Library [13] in addition to the matrices considered in Example 6.27. The details and settings are given in Table 6.1. If only one matrix is given then $\mathbf{M} = \mathbf{I}$.*

The different setups, that is the quality of the preconditioner (i.e. drop tolerance of *the incomplete LU factorisation), the shift and the corresponding sought finite simple eigenvalue,* $\tau^{(0)}$ *for the stopping tolerance of the inner solve* $\tau^{(i)} = \min\{\tau^{(0)}, \tau^{(0)}\|\mathbf{r}^{(i)}\|\}$ *and the starting vector of the six considered problems are given in Table 6.2.*

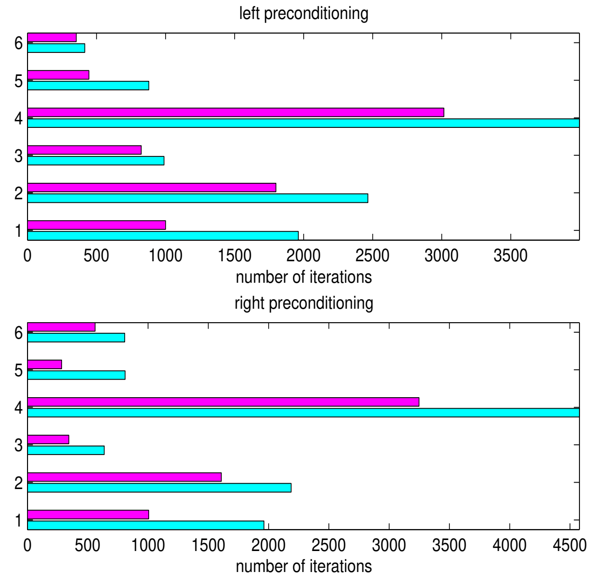


Figure 6-6: *Numerical results for Example 6.28. Total number of inner iterations for left preconditioning with and without tuning (top plot) and for right preconditioning with and without tuning (bottom plot).*

We apply both left and right preconditioning, and also left and right preconditioned tuning to the problems. CPU times and and total number of inner iterations with and without tuning are given in Figures 6-6 and 6-7.

We can see that for both left and right preconditioning tuning is superior to the untuned situation both in terms of CPU time and the total number of inner iterations. Hence tuning in these examples gives always an overall reduction in costs.

6.5 Variable shifts

This section extends the theory of fixed shifts from Section 6.4 to Rayleigh quotient shifts. In fact, many of the results are applicable to other variable shift strategies. Furthermore, both the tuning strategy without preconditioning from Section 6.3.2 and tuned preconditioner (Section 6.4.2) can be used in the variable shift strategy. We only consider preconditioning here, since it is the most practical approach.

6.5.1 The tuned preconditioner applied to systems with variable shift

The idea of a tuning operator $\mathbb{P}$ from (6.47) can be applied to the system with variable shift and with $\rho(\mathbf{x}_1) = \lambda_1$ we have

$$(\mathbf{A} - \rho(\mathbf{x}_1)\mathbf{M})\mathbb{P}^{-1}\mathbf{M}\mathbf{x}_1 = (\mathbf{A} - \rho(\mathbf{x}_1)\mathbf{M})\mathbf{x}_1 = 0,$$

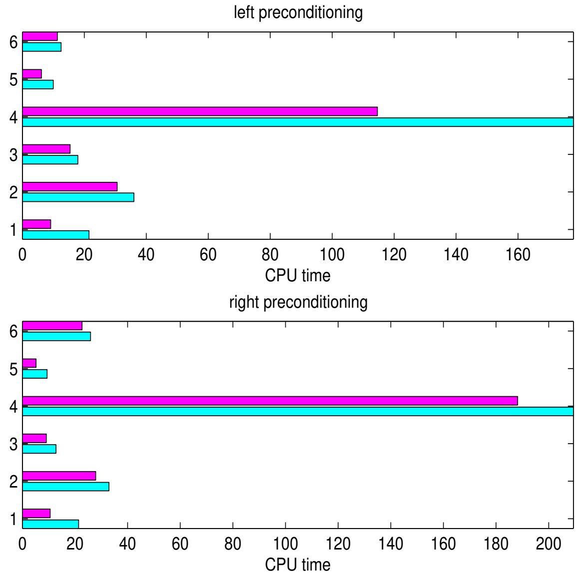


Figure 6-7: *Numerical results for Example 6.28. Total CPU times for left preconditioning with and without tuning (top plot) and for right preconditioning with and without tuning (bottom plot).*

that is, $\mathbf{M}\mathbf{x}_1$ is an exact eigenvector of the singular matrix $(\mathbf{A} - \rho(\mathbf{x}_1)\mathbf{M})\mathbb{P}^{-1}$ corresponding to the eigenvalue 0. The practical tuning operator $\mathbb{P}_i$ from (6.50) can then be applied to the linear system

$$(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbb{P}_i{}^{-1}\tilde{\mathbf{y}}_k = \mathbf{M}\mathbf{x}^{(i)}, \quad \mathbb{P}_i{}^{-1}\tilde{\mathbf{y}}_k = \mathbf{y}_k.$$

The following Lemma is an extension of Lemma 6.12.

Lemma 6.29. *Let $\mathcal{E}^{(i)}$ be as in Lemma 6.12 with $\|\mathcal{E}^{(i)}\| \leq C_3\varepsilon^{(i)}$. Then*

$$\|(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbb{P}_i{}^{-1} - (\mathbf{A} - \rho(\mathbf{x}_1)\mathbf{M})\mathbb{P}^{-1}\| \leq \beta_3\varepsilon^{(i)},$$

where β_3 is independent of i for large enough i.

Proof. Using (6.36) we have $\mathbb{P}_i = \mathbb{P} + (\mathbf{A} - \mathbf{P})\mathcal{E}^{(i)}$ and therefore with $\rho(\mathbf{x}_1) = \lambda_1$ we can write

$$\begin{aligned}(\mathbf{A} - \rho(\mathbf{x}_1)\mathbf{M})\mathbb{P}^{-1} - (\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbb{P}_i{}^{-1} &= (\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})(\mathbb{P}^{-1} - \mathbb{P}_i^{-1}) \\ &\quad + (\rho(\mathbf{x}^{(i)}) - \lambda_1)\mathbf{M}\mathbb{P}^{-1} \\ &= (\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbb{P}^{-1}(\mathbb{P}_i - \mathbb{P})\mathbb{P}_i{}^{-1} \\ &\quad + (\rho(\mathbf{x}^{(i)}) - \lambda_1)\mathbf{M}\mathbb{P}^{-1}\end{aligned}$$

From (6.7) we have $|\rho(\mathbf{x}^{(i)}) - \lambda_1| \leq C_6\varepsilon^{(i)}$ and similar to Lemma 6.21 we have $\|\mathbb{P}^{-1}(\mathbb{P}_i - \mathbb{P})\mathbb{P}_i{}^{-1}\| \leq C_{11}\varepsilon^{(i)}$. Hence, for large enough i, we obtain

$$\|(\mathbf{A} - \rho(\mathbf{x}_1)\mathbf{M})\mathbb{P}^{-1} - (\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbb{P}_i{}^{-1}\| \leq \beta_3\varepsilon^{(i)}$$

for an appropriately chosen constant β_3. $\square$

The results of Theorem 6.22 also hold for system matrices of the form $\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M}$, where equation (6.55) simplifies to

$$(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbb{P}_i^{-1}\left(\mathbf{M}\mathbf{x}^{(i)} + \frac{\mathbf{r}^{(i)}}{\rho(\mathbf{x}^{(i)})}\right) = \frac{1}{\rho(\mathbf{x}^{(i)})}\mathbf{r}^{(i)},$$

and hence $\mathbf{M}\mathbf{x}^{(i)} + \dfrac{\mathbf{r}^{(i)}}{\rho(\mathbf{x}^{(i)})}$ is an approximate eigenvector of $(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbb{P}_i^{-1}$ with approximate eigenvalue 0. Therefore

$$\|\mathcal{P}^{(i)}\mathbf{M}\mathbf{x}^{(i)}\| \leq C_{12}\|\mathcal{P}^{(i)}\|\|\mathbf{r}^{(i)}\|, \tag{6.61}$$

where $\mathcal{P}^{(i)}$ is chosen appropriately and C_{12} is a constant independent of i for large enough i. Finally, the following theorem is an extension of Theorem 6.23 to the solution of the system arising in inexact inverse iteration with variable shifts.

Theorem 6.30. *Let the assumptions of Theorem 6.22 be satisfied. Compute $\tilde{\mathbf{y}}_{k^{(i)}}$ satisfying the stopping criterion*

$$\|(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbb{P}_i{}^{-1}\tilde{\mathbf{y}}_{k^{(i)}} - \mathbf{M}\mathbf{x}^{(i)}\| \leq \tau^{(i)} = \delta\|\mathbf{r}^{(i)}\|^{\zeta} \quad \delta < 1,$$

where $\rho(\mathbf{x}^{(i)})$ is the generalised Rayleigh quotient (6.3) and

(a) $\zeta = 0$ is used for solves with a fixed tolerance and

(b) $\zeta = 1$ is used for solves with a decreasing tolerance.

Then, for large enough i, $k^{(i)}$, the number of inner iterations used by GMRES to compute $\tilde{\mathbf{y}}_{k^{(i)}}$ satisfying this stopping criterion, is

(a) bounded independently of i for $\zeta = 0$,

(b) increasing with order $\log({\varepsilon^{(i)}}^{-1})$ for $\zeta = 1$.

In contrast the number $k^{(i)}$ of inner iterations used by GMRES to compute $\tilde{\mathbf{y}}_{k^{(i)}}$ satisfying the stopping criterion

$$\|(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbf{P}^{-1}\tilde{\mathbf{y}}_{k^{(i)}} - \mathbf{M}\mathbf{x}^{(i)}\| \leq \tau^{(i)} = \delta\|\mathbf{r}^{(i)}\|^{\zeta} \quad \delta < 1,$$

is

(a) increasing with order $\log({\varepsilon^{(i)}}^{-1})$ for $\zeta = 0$,

(b) increasing with order $2\log({\varepsilon^{(i)}}^{-1})$ for $\zeta = 1$.

Proof. $\mathbf{M}\mathbf{x}_1$ is a simple eigenvector of $(\mathbf{A} - \lambda_1\mathbf{M})\mathbb{P}^{-1}$ so that the block-diagonalisation

$$(\mathbf{A} - \lambda_1\mathbf{M})\mathbb{P}^{-1} = \begin{bmatrix} \bar{\bar{\mathbf{w}}}_1 & \bar{\bar{\mathbf{W}}}_2 \end{bmatrix} \begin{bmatrix} 0 & \mathbf{0}^H \\ \mathbf{0} & \bar{\bar{\mathbf{C}}} \end{bmatrix} \begin{bmatrix} \bar{\bar{\mathbf{v}}}_1^H \\ \bar{\bar{\mathbf{V}}}_2^H \end{bmatrix}$$

exist, where $\bar{\bar{\mathbf{w}}}_1 = \dfrac{\mathbf{M}\mathbf{x}_1}{\|\mathbf{M}\mathbf{x}_1\|}$. Since, by Lemma 6.29 $(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbb{P}_i{}^{-1}$ is a perturbation of $(\mathbf{A} - \lambda_1\mathbf{M})\mathbb{P}^{-1}$ we can block diagonalise $(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbb{P}_i^{-1}$ as

$$(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbb{P}_i^{-1} = \begin{bmatrix} \bar{\bar{\mathbf{w}}}_1^{(i)} & \bar{\bar{\mathbf{W}}}_2^{(i)} \end{bmatrix} \begin{bmatrix} \bar{\bar{\mu}}_1^{(i)} & \mathbf{0}^H \\ \mathbf{0} & \bar{\bar{\mathbf{C}}}^{(i)} \end{bmatrix} \begin{bmatrix} \bar{\bar{\mathbf{v}}}_1^{(i)} & \bar{\bar{\mathbf{V}}}_2^{(i)} \end{bmatrix}^H, \tag{6.62}$$

for small enough $\varepsilon^{(i)}$ and similar results hold as in Lemma 6.13, that is

$$\begin{aligned}
|\bar{\bar{\mu}}_1^{(i)}| &\leq f_1\varepsilon^{(i)},\\
\|\bar{\bar{\mathbf{C}}} - \bar{\bar{\mathbf{C}}}^{(i)}\| &\leq f_2\varepsilon^{(i)},\\
\|\bar{\bar{\mathbf{V}}}_2 - \bar{\bar{\mathbf{V}}}_2^{(i)}\| &\leq f_3\varepsilon^{(i)},
\end{aligned}$$

for f_1, f_2 and f_3 independent of i for large enough i. Similar to the proof of Theorem 6.23 the residual obtained after $k^{(i)}$ iterations of GMRES starting with 0 is less than $\tau^{(i)} = \delta\|\mathbf{r}^{(i)}\|^\zeta$ if

$$k^{(i)} \geq 1 + \frac{1}{\log|\Psi(0)|}\left(\log S_\delta\|\bar{\bar{\mu}}_1^{(i)}\mathbf{I} - \bar{\bar{\mathbf{C}}}^{(i)}\|\|\bar{\bar{\mathbf{V}}}_2^{(i)}\| + \log\frac{\|\mathcal{P}^{(i)}\mathbf{M}\mathbf{x}^{(i)}\|}{\delta|\bar{\bar{\mu}}_1^{(i)}|\|\mathbf{r}^{(i)}\|^\zeta}\right).$$

For the first term we have

$$\|\bar{\bar{\mu}}_1^{(i)}\mathbf{I} - \bar{\bar{\mathbf{C}}}^{(i)}\|\|\bar{\bar{\mathbf{V}}}_2^{(i)}\| \leq (\|\bar{\bar{\mathbf{C}}}\| + f_1\varepsilon^{(i)} + f_2\varepsilon^{(i)})(\|\bar{\bar{\mathbf{V}}}\| + f_3\varepsilon^{(i)}),$$

which can be bounded independently of i for large enough i. For the second term, by (6.61) we have $\|\mathcal{P}^{(i)}\mathbf{M}\mathbf{x}^{(i)}\| \leq C_{12}\|\mathcal{P}^{(i)}\|\|\mathbf{r}^{(i)}\|$, and $\|\mathcal{P}^{(i)}\|$ can be bounded by a constant for large enough i (that is, for $\varepsilon^{(i)}$ small enough, see proof of Theorem 6.16). Hence

$$\|\mathcal{P}^{(i)}\mathbf{M}\mathbf{x}^{(i)}\| \leq C_{13}\|\mathbf{r}^{(i)}\|$$

for tuning and

$$\|\mathcal{P}^{(i)}\mathbf{M}\mathbf{x}^{(i)}\| \leq C_{14}$$

if no tuning is applied. Together with the bounds on $|\bar{\bar{\mu}}_1^{(i)}|$ and equivalence of $\varepsilon^{(i)} = \|\mathbf{p}^{(i)}\|$ and $\|\mathbf{r}^{(i)}\|$ (see Lemma 6.1) we obtain the results for $\zeta = 0$ and $\zeta = 1$ respectively. □

6.5.2 Numerical results

Here we give three further examples, where we explore the behavior of the tuned preconditoner for a fixed shift strategy with a decreasing tolerance and for variable shifts with both fixed and decreasing tolerance. In all three examples we take the standard eigenproblem `cdde1.mtx` from the Matrix Market library. [13]. We look for the eigenvalue closest to zero, that is $\lambda_1 = -0.0051724$. We compare the standard preconditioner $\mathbf{P}$ obtained from an incomplete Cholesky factorisation with drop tolerance 0.1 and its tuned version. All iterations stop once $\|\mathbf{r}^{(i)}\| < 10^{-15}$.

Example 6.31. *This example uses a fixed shift $\sigma = 0$ and a decreasing tolerance $\tau^{(i)} = \min\{0.8, 0.8\,\|\mathbf{r}^{(i)}\|\}$. The overall algorithm converges linearly.*

Figures 6-8 and 6-9 show the results for Example 6.31. A fixed shift strategy with a decreasing solve tolerance leads to a logarithmic increase in the number of iterations as the outer iteration proceeds if a standard right preconditioner is applied (see circled dashed line in Figure 6-8) as we have seen in Section 6.4. With the tuned preconditioner this disadvantage is overcome and the number of inner iterations remains approximately constant (see starred solid line in Figure 6-8) as we have proved in Theorem 6.23. Hence the tuning strategy requires only about a fourth of the total number of iterations (177 versus 618 iterations).

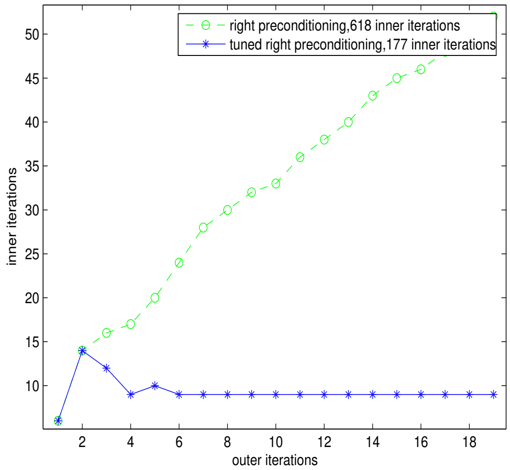


Figure 6-8: *Inner iterations against outer iterations with standard and tuned preconditioning (Example 6.31)*

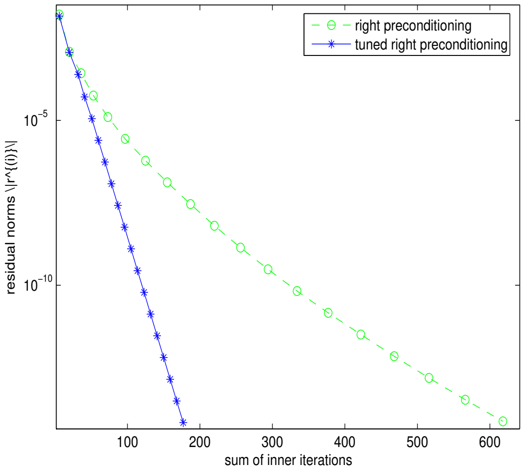


Figure 6-9: *Residual norms vs total number of inner iterations with standard and tuned preconditioning (Example 6.31)*

Example 6.32. *This example uses a variable shift given by* $\rho(\mathbf{x}^{(i)})$ *with starting value* $\sigma^{(0)} = 0$ *and a fixed tolerance* $\tau = 0.6$. *This gives also overall linear convergence.*

The results for Example 6.32 are plotted in Figures 6-10 and 6-11. According to Theorem 6.30 a fixed solve tolerance ($\zeta = 0$) and a Rayleigh quotient shift leads to a logarithmic increase in iteration numbers for the standard preconditioner and to an approximately constant iteration number count for the tuned preconditioner. This is what can indeed be observed from Figure 6-10. The savings in this example are bigger than 80 per cent.

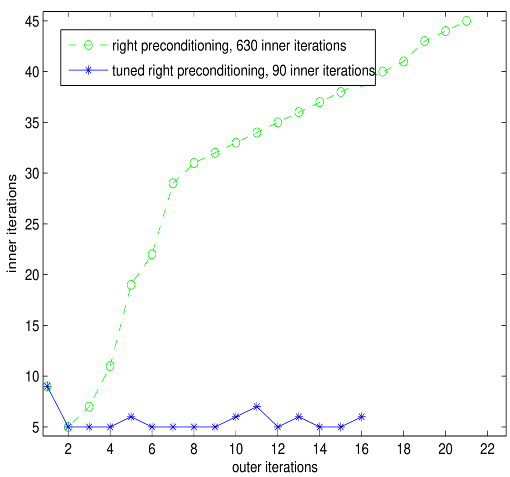


Figure 6-10: *Inner iterations against outer iterations with standard and tuned preconditioning (Example 6.32)*

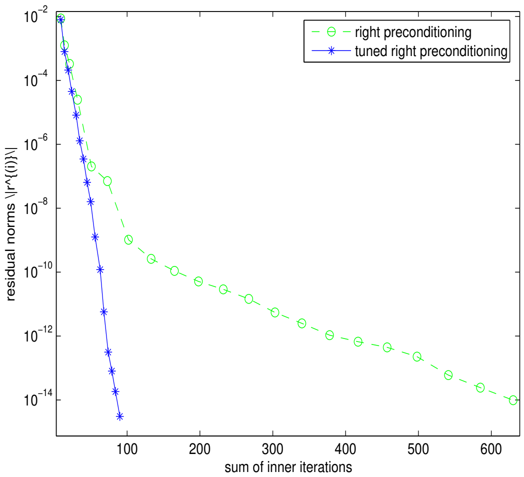


Figure 6-11: *Residual norms vs total number of inner iterations with standard and tuned preconditioning (Example 6.32)*

Example 6.33. *This example uses a variable shift given by* $\rho(\mathbf{x}^{(i)})$ *with starting value* $\sigma^{(0)} = 0$ *and a decreasing tolerance* $\tau^{(i)} = \min\{0.8, 0.8\,\|\mathbf{r}^{(i)}\|\}$. *The overall algorithm converges quadratically.*

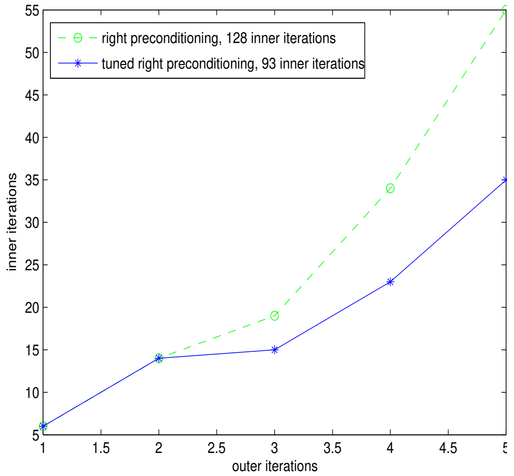


Figure 6-12: *Inner iterations against outer iterations with standard and tuned preconditioning (Example 6.33)*

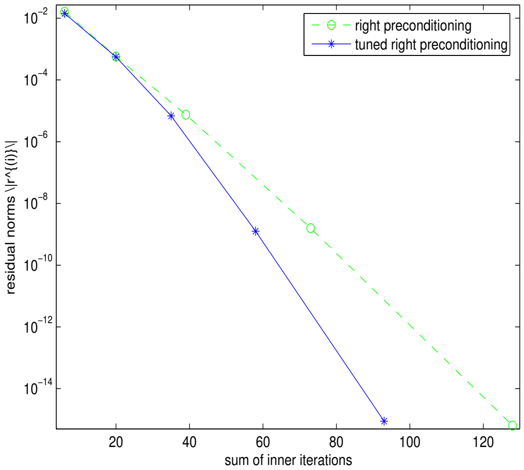


Figure 6-13: *Residual norms vs total number of inner iterations with standard and tuned preconditioning (Example 6.33)*

Figures 6-12 and 6-13 illustrate the iteration numbers and eigenvalue residuals for Example 6.33. Theorem 6.30 states that the number of inner iterations per outer iteration increase with order $2\log({\varepsilon^{(i)}}^{-1})$ if a decreasing solve tolerance $\zeta = 1$, a Rayleigh quotient shift and a standard preconditioner is used, whereas for the tuned preconditioner the increase in the number of inner iterations per outer iterations is only logarithmic, that is half as fast. We can observe this behaviour in Figure 6-12. The savings when using the tuned preconditioner for this particular example is about 25 per cent.

6.6 A comparison of tuned Rayleigh quotient iteration to the Jacobi-Davidson method applied to the generalised eigenproblem

This section contains a brief comparison of the tuning strategy introduced in this chapter with a simplified version of the Jacobi-Davidson method. For the generalised eigenproblem Sleijpen et al. [123] introduced a Jacobi-Davidson type method which we describe briefly. It is a generalised version of the method used in Chapter 3 and discussed in Chapter 5.

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

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

where $\mathbf{u}^H\mathbf{x}^{(i)} \neq 0$ and $\mathbf{w}^H\mathbf{M}\mathbf{x}^{(i)} \neq 0$. Note that in Chapter 3 we used $\mathbf{w} = \mathbf{M}\mathbf{x}^{(i)}$ and

$\mathbf{u} = \mathbf{M}^H\mathbf{M}\mathbf{x}^{(i)}$. With $\mathbf{r}^{(i)}$ defined by (6.4) solve the correction equation

$$\mathbf{P}^{(i)}(\mathbf{A} - \rho(\mathbf{x}^{(i)})\mathbf{M})\mathbf{Q}^{(i)}\mathbf{s}^{(i)} = -\mathbf{r}^{(i)}, \quad \text{where} \quad \mathbf{s}^{(i)} \perp \mathbf{u}, \tag{6.63}$$

for $\mathbf{s}^{(i)}$. This is the Jacobi-Davidson correction equation which maps span$\{\mathbf{u}\}^\perp$ onto span$\{\mathbf{w}\}^\perp$. An improved guess for the eigenvector is given by a suitably normalised $\mathbf{x}^{(i)} + \mathbf{s}^{(i)}$. For a description of the Algorithm of inexact simplified Jacobi-Davidson we refer to Algorithm 5 in Chapter 3.

Several choices for the projectors $\mathbf{P}^{(i)}$ and $\mathbf{Q}^{(i)}$ are possible, depending on the choice of $\mathbf{w}$ and $\mathbf{u}$. In Chapter 5 we have used $\mathbf{w} = \mathbf{M}\mathbf{x}^{(i)}$ and shown that if a tuned preconditioner is used in inexact inverse iteration applied to the generalised eigenproblem then this method is equivalent to the simple Jacobi-Davidson method with correction equation (6.63) and a standard preconditioner.

This generalisation has two practical implications: Firstly, if inexact inverse iteration is used with a tuned preconditioner we obtain the same results as in the inexact simplified Jacobi-Davidson method. Hence, if we use inexact inverse iteration with a tuned preconditioner the the choice of $\mathbf{P}^{(i)}$ and $\mathbf{Q}^{(i)}$ does not have to be taken care of, whereas for the simplified Jacobi-Davidson method we have to think about choices for $\mathbf{P}^{(i)}$ and $\mathbf{Q}^{(i)}$. Another implication is that tuning the preconditioner does not have any effect on the Jacobi-Davidson method.

In this section we show a numerical example where we use GMRES instead of FOM and which supports the result, that it does not matter whether one uses simplified Jacobi-Davidson with a standard preconditioner or inexact inverse iteration with a tuned preconditoner.

Note that by Theorem 5.7, the residuals for FOM and GMRES are related to each other in the sense that the FOM residual norm and the GMRES residual norm will be approximately equal to each other if the GMRES residual norm is reduced at each step. Hence, similar results are expected for FOM and GMRES, although exact equivalence of inexact inverse iteration with a preconditioner which is tuned in a certain way and inexact simplified Jacobi-Davidson method with a standard preconditioner can only be shown for a Galerkin-Krylov method such as FOM.

Example 6.34. *Consider the same generalised eigenproblem arising from the Galerkin-FEM discretisation on a 32×32 grid of the convection-diffusion operator (6.45) as considered in Example 6.19.*

We apply inexact inverse iteration as well as simplified Jacobi-Davidson with Rayleigh quotient shift and a fixed solve tolerance $\tau = 0.2$ to this problem. For both methods we use the same starting guess and the overall computation stops once $\|\mathbf{r}^{(i)}\| < 10^{-12}$. We apply preconditioned GMRES within the inner solve of each method, where the simplified Jacobi-Davidson approach uses the standard preconditioner and the inexact RQ iteration uses a tuned preconditioner.

The results for Example 6.34 are plotted in Figures 6-14 to 6-17. First of all, we can see that for simplified Jacobi-Davidson method tuning the preconditioner has no effect, the results for the standard and the tuned preconditioner are very similar (see Figures 6-16 and 6-17), the total number of iterations is the same.

For inexact Rayleigh quotient iteration tuning the preconditioner reduces the iteration number from 264 to 83 iterations (see Figures 6-14 and 6-15) and the total number of iterations is even smaller than the one for inexact simplified Jacobi-Davidson (83

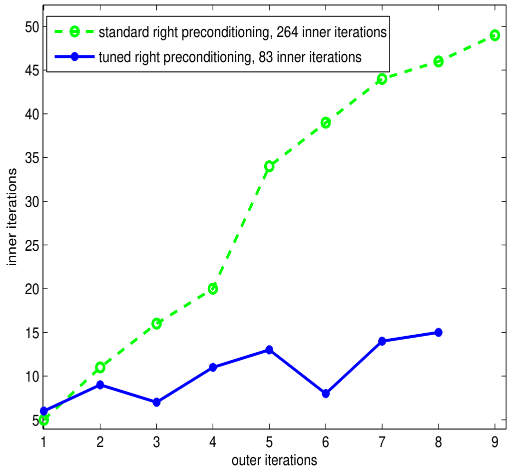


Figure 6-14: *Number of inner iterations against outer iterations with standard and tuned preconditioning (Example 6.34) when using inexact RQ iteration*

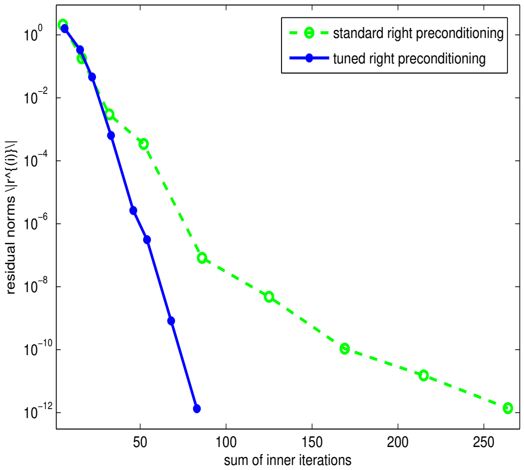


Figure 6-15: *Residual norms vs total number of inner iterations with standard and tuned preconditioning (Example 6.34) when using inexact RQ iteration*

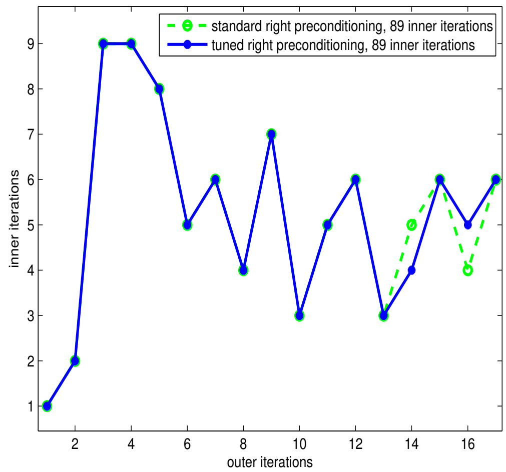


Figure 6-16: *Number of inner iterations against outer iterations with standard and tuned preconditioning (Example 6.34) when using inexact simplified JD*

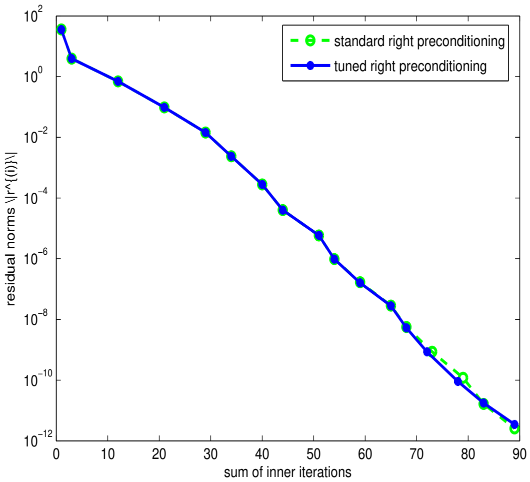


Figure 6-17: *Residual norms vs total number of inner iterations with standard and tuned preconditioning (Example 6.34) when using inexact simplified JD*

versus 89) iterations. In this example we used the same projectors $\mathbf{P}^{(i)}$ and $\mathbf{Q}^{(i)}$ for the simplified Jacobi-Davidson method as in Chapter 3, but choosing the projectors for Jacobi-Davidson for the generalised eigenproblem involves additional analysis in the case of JD, which does not arise if we just use the tuned preconditioner in conjunction with inexact Rayleigh quotient iteration.

6.7 Conclusions

In this chapter we have analysed inexact inverse iteration for the generalised nonsymmetric eigenproblem. We provided a general convergence theory with varying shifts for finding an eigenpair corresponding to a finite simple eigenvalue of the generalised eigenproblem. We presented convergence results on GMRES for the solve of the inner system. Using these findings we showed how the right hand side of the linear system influences the iterative solve of this linear system. With this analysis we derived a new tuned preconditioner in a similar style as presented in [42] for the Hermitian case. We find that tuning may reduce the total number of inner iterations substantially. This approach can be used both for the symmetric and the nonsymmetric problem and it can also be applied to unpreconditioned generalised eigenproblem.

Several numerical examples support our theory and show that tuning yields an improvement over the standard preconditioning with regard to the total number of inner iterations both for fixed and variable shifts. Furthermore, a comparison of the inexact simplified Jacobi-Davidson method with a standard preconditioner to inexact Rayleigh quotient iteration with a tuned preconditioner shows that the later method achieves similar (or even better) results than the Jacobi-Davidson approach.

CHAPTER 7

Inexact preconditioned Arnoldi's method and implicit restarts for eigenvalue computations

7.1 Introduction

In this chapter we consider Arnoldi's method which is one way to extend inverse iteration to a a subspace method. The chapter contains two main results. First we extend the relaxation strategy for Arnoldi's method which was developed in [14] and analysed in [118] to implicitly restarted Arnoldi's method. Secondly, we apply the idea of tuning the preconditioner which was developed in Chapters 4 and 6 to Arnoldi's method and implicitly restarted Arnoldi's method.

Arnoldi's method is an efficient method for the approximation of a few eigenvalues and corresponding eigenvectors of the eigenvalue problem

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}, \quad \|\mathbf{x}\| = 1,$$

where $\mathbf{A} \in \mathbb{C}^{n,n}$ is a large sparse matrix. After $k < n$ steps it produces an upper Hessenberg matrix $\mathbf{H}_k$ of order k, which is the projection of $\mathbf{A}$ onto a Krylov subspace of size k. The eigenvalues of the upper Hessenberg matrix are then approximations to the eigenvalues of $\mathbf{A}$.

Arnoldi's method has two drawbacks: Firstly, a recurrence relation of length $k-1$ is needed to compute kth basis vector and so all previous vectors need to be stored; secondly, the Arnoldi iteration favours the outlying eigenvalues. Both disadvantages can be mitigated. In order to limit the storage requirement, implicitly restarted Arnoldi method (IRA) [130] may be used. To overcome the second disadvantage, interior eigenvalues can be found by applying Arnoldi's method to $\mathcal{A} := (\mathbf{A}-\sigma\mathbf{I})^{-1}$ for an appropriate choice of σ. Hence, within each step of Arnoldi's method (or IRA) a system of the form

$$(\mathbf{A} - \sigma\mathbf{I})\mathbf{y} = \mathbf{q}_k,$$

for a given right hand side $\mathbf{q}_k$ has to be solved. For large systems this solve will probably be done inexactly via an iterative method which leads to a so-called "inner-outer" iterative method. The outer method is the Arnoldi process for the eigencomputation and the inner problem is the iterative solve for the linear system.

A large number of tests carried out in [14] suggested that convergence of Arnoldi's method can be obtained despite the fact that the solve tolerance is relaxed as the outer iteration proceeds. This somewhat surprising different behaviour of Krylov methods (in contrast to Newton-type methods such as inverse iteration and Rayleigh quotient iteration) for eigencomputations was also observed in [51] for symmetric matrices and the case where a projection has to be carried out inexactly. There it was stated that for good approximations to an eigenpair one may start with very accurate inner solve tolerances, but this tolerance may be relaxed as the outer iteration progresses. Simoncini [118] gives a theory for this performance for general inexact matrix-vector multiplications and general nonsymmetric matrices. Similar ideas have been examined in the linear system setting, where inexact matrix vector multiplications were used within Krylov solvers (see [120–122, 148]). Another approach in terms of inexact solves for the arising inner system has been taken in [142]; in this work the spectral transformation is approximated by a fixed-polynomial operator which is computed prior to the Arnoldi iteration.

In this chapter we consider both the inexact solve and relaxation strategies as well as preconditioning for the inner solve. We extend the theory developed in [118] for Arnoldi's method to implicitly restarted Arnoldi's method. In previous chapters a new strategy for preconditioning in eigenproblems was introduced. Chapter 2 (see also [43]) showed that a tuned preconditioner is favourable to a standard preconditioner within the inner solve. In Chapter 4 (see also [42]) we analysed the tuned preconditioner for Hermitian eigenproblems and in Chapter 6 we applied this new preconditioner to non-Hermitian problems and showed how the tuned preconditioner reduces the iteration number for the inner solves. In [104] this strategy was extended to inexact preconditioned subspace iteration and proved to be efficient. In this chapter, we apply this new tuning strategy to Arnoldi iterations and IRA and discuss its efficiency.

The main results of this chapter are the extension of the relaxation result developed in [118] to IRA and the application of the tuned preconditioner to Arnoldi's method and implicitly restarted Arnoldi's method.

The chapter is organised as follows: In Section 7.2 the theory of Arnoldi's method (with and without implicit restarts) for eigencomputations is revised. Section 7.3 contains the main theory on inexact solves in the Arnoldi iteration, including relaxation strategies for Arnoldi and IRA methods. Section 7.4 looks at improvements for preconditioning in shift-invert Arnoldi's method and Section 7.5 considers preconditioning strategies for the shift-invert IRA method. Numerical evidence is given throughout. We use $\|\cdot\| = \|\cdot\|_2$ and $\mathbf{A}^H$ for the conjugate transpose of matrix $\mathbf{A}$.

7.2 Arnoldi's method and implicit restarts

This section gives a short review of Arnoldi's method and implicitly restarted Arnoldi's method discussed in Chapter 1, Section 1.4.3 (see also Appendix A).

Consider the k-dimensional Krylov subspace

$$\mathcal{K}_k(\mathcal{A}, \mathbf{q}_1) = \text{span}\{\mathbf{q}_1, \mathcal{A}\mathbf{q}_1, \mathcal{A}^2\mathbf{q}_1, \ldots, \mathcal{A}^{k-1}\mathbf{q}_1\},$$

The Arnoldi method [3] is used to construct an orthogonal basis $\mathbf{Q}_k$ for $\mathcal{K}_k(\mathcal{A}, \mathbf{q}_1)$. The corresponding k-step Arnoldi factorisation can be written as

$$\mathcal{A}\mathbf{Q}_k = \mathbf{Q}_k\mathbf{H}_k + \mathbf{q}_{k+1}h_{k+1,k}\mathbf{e}_k^H = \mathbf{Q}_{k+1}\begin{bmatrix} \mathbf{H}_k \\ h_{k+1,k}\mathbf{e}_k^H \end{bmatrix}, \quad \mathbf{Q}_{k+1}^H\mathbf{Q}_{k+1} = \mathbf{I}.$$

The matrix $\mathbf{H}_k = \mathbf{Q}_k^H \mathcal{A}\mathbf{Q}_k$ is a $k \times k$ upper Hessenberg matrix, the orthogonal projection of $\mathcal{A}$ to the Krylov subspace. The factorisation can be used to obtain approximate eigenvalues and eigenvectors for $\mathcal{A}$. Using the eigenpairs $(\theta, \mathbf{u})$ of $\mathbf{H}_k$, the vector $\mathbf{x} = \mathbf{Q}_k\mathbf{u}$ satisfies

$$\|\mathbf{r}_k\| = \|\mathcal{A}\mathbf{x} - \theta\mathbf{x}\| = \|(\mathcal{A}\mathbf{Q}_k - \mathbf{Q}_k\mathbf{H}_k)\mathbf{u}\| = |h_{k+1,k}\mathbf{e}_k^H\mathbf{u}|.$$

The values θ and $\mathbf{x}$ are called Ritz value and Ritz vector and are approximate eigenpairs of $\mathcal{A}$ [110]. The residual $\mathbf{r}_k$ provides a bound on the accuracy of the eigenpair approximations, although, in the non-Hermitian case, a small Ritz residual does not necessarily imply an accurate answer [110]. Clearly, $\|\mathbf{r}_k\| = 0$ if and only if $\mathbf{Q}_k$ spans an invariant subspace of $\mathcal{A}$.

Algorithm 8 Implicitly restarted Arnoldi method

Input: $\mathcal{A}$, $\mathbf{q}_1$, ($\|\mathbf{q}_1\|_2 = 1$), wanted eigenvalues k, $m = k + p$ total number of Arnoldi steps. *imax* total number of restarts.
Compute k Arnoldi steps to produce $\mathbf{Q}_k$ and $\mathbf{H}_k$.
for $i = 1, \ldots, i_{\max}$ **do**
  Compute another p Arnoldi steps to produce $\mathbf{Q}_m$ and $\mathbf{H}_m$, $m = k + p$.

$$\mathcal{A}\mathbf{Q}_m = \mathbf{Q}_m\mathbf{H}_m + \mathbf{f}_m\mathbf{e}_m^H, \quad \mathbf{f}_m = h_{m+1,m}\mathbf{q}_{m+1}$$

  if $h_{i+1,i} = 0$ **then**
    span$\{\mathbf{q}_1, \ldots, \mathbf{q}_i\}$ is invariant under $\mathcal{A}$.
  end if
  Compute $\Lambda(\mathbf{H}_m)$ and select p shifts $\nu_1, \ldots, \nu_p$ (unwanted spectrum).
  Set $\mathbf{V} = \mathbf{I}_m$.
  for $j = 1, \ldots, p$ **do**
    Factor $[\mathbf{V}_j, \mathbf{R}_j] = \mathrm{qr}(\mathbf{H}_m - \mu_j\mathbf{I})$;
    $\mathbf{H}_m = \mathbf{V}_j^H\mathbf{H}_m\mathbf{V}_j$;
    $\mathbf{V} = \mathbf{V}\mathbf{V}_j$;
  end for
  Compute $\mathbf{f}_k = \mathbf{q}_{k+1}\mathbf{H}_m(k+1, k) + \mathbf{f}_m\mathbf{V}(m, k)$.
  Compute $\mathbf{Q}_k = \mathbf{Q}_m\mathbf{V}(:, 1:k)$.
  Set $\mathbf{H}_k = \mathbf{H}_m(1:k, 1:k)$.
  Restart with k-step Arnoldi factorisation.
end for
Output: $\mathbf{H}_m$, $\mathbf{Q}_m$.

For large sparse problems, the storage of all basis vectors and the orthogonalisation procedure applied to all of these vectors might not be possible. To limit the storage requirements several acceleration procedures have been developed in order to keep k small. In this chapter we discuss two of them: *restarts* and *spectral transformations*.

The implicitly restarted Arnoldi method (IRA) [130,131] provides a way of limiting the number of basis vectors used in an Arnoldi factorisation by implicitly restarting the iteration with an increasingly better starting vector q_1. We give a short description here. Assume an Arnoldi factorisation of length m is given

$$\mathcal{A}\mathbf{Q}_m = \mathbf{Q}_m\mathbf{H}_m + \mathbf{f}_m\mathbf{e}_m^H, \quad \mathbf{f}_m = h_{m+1,m}\mathbf{q}_{m+1}. \tag{7.1}$$

We choose p shifts $\nu_1, \dots, \nu_p$ and use them to perform p steps of the implicitly shifted QR algorithm on the projected matrix $\mathbf{H}_m$. The overall effect is the generation of a matrix $\mathbf{V}_m$ such that

$$\hat{\mathbf{H}}_m = \mathbf{V}_m^H \mathbf{H}_m \mathbf{V}_m$$

is upper Hessenberg, where

$$q(\mathbf{H}_m) = \mathbf{V}_m \mathbf{R}_m,$$

with $\mathbf{V}_m$ is unitary, $\mathbf{R}_m$ is upper triangular and q is a polynomial of degree p with zeros $\nu_1, \dots, \nu_p$. Then from (7.1) we obtain

$$\mathcal{A}\mathbf{Q}_m \mathbf{V}_m = \mathbf{Q}_m \mathbf{V}_m \mathbf{V}_m^H \mathbf{H}_m \mathbf{V}_m + \mathbf{f}_m \mathbf{e}_m^H \mathbf{V}_m,$$

or with $\hat{\mathbf{Q}}_m = \mathbf{Q}_m \mathbf{V}_m$ and $\hat{\mathbf{H}}_m = \mathbf{V}_m^H \mathbf{H}_m \mathbf{V}_m$

$$\mathcal{A}\hat{\mathbf{Q}}_m = \hat{\mathbf{Q}}_m \hat{\mathbf{H}}_m + \mathbf{f}_m \mathbf{e}_m^H \mathbf{V}_m,$$

The structure of $\mathbf{V}_m$ is such that the first $m-p-1$ components of $\mathbf{e}_m^H \mathbf{V}_m$ are zero (see [130]). Hence with $\hat{\mathbf{H}}_k$ denoting the leading principal submatrix of $\hat{\mathbf{H}}_m$ and setting

$$\hat{\mathbf{f}}_k = \hat{\mathbf{q}}_{k+1} \hat{h}_{k+1,k} + \mathbf{f}_m \mathbf{V}_{m,k} \tag{7.2}$$

we get

$$\mathcal{A}\hat{\mathbf{Q}}_k = \hat{\mathbf{Q}}_k \hat{\mathbf{H}}_k + \hat{\mathbf{f}}_k \mathbf{e}_k^H,$$

an Arnoldi decomposition of order k. Then the Arnoldi process can be restarted from step k rather than starting from the first step. The implicitly restarted Arnoldi Algorithm is stated in Algorithm 8.

With the choice of the shifts $\nu_1, \dots, \nu_p$ we have $q(z) = (z-\nu_1)\cdots(z-\nu_p)$ and obtain

$$\hat{\mathbf{q}}_1 = \frac{q(\mathcal{A})\mathbf{q}_1}{\|q(\mathcal{A})\mathbf{q}_1\|},$$

in other words, the starting vector $\mathbf{q}_1$ has been updated with the polynomial filter $q(z)$ and its roots (or implicit shifts) are chosen to filter out unwanted information from the starting vector, that is small values will be taken near the shifts and large values will be taken away from these points. If we choose the shifts to be the unwanted part of the spectrum then the starting vector $\mathbf{q}_1$ will enhance the wanted part of the spectrum. There are several ways of choosing the shifts $\nu_1, \dots, \nu_p$. In this chapter we chose an exact shift strategy, that is, the unwanted part of the spectrum of $\mathbf{H}_m$ is used as shift. This choice leads to $\hat{h}_{k+1,k} = 0$ in (7.2) (see [130, Lemma 3.10]).

The convergence theory for the Arnoldi process, which leads to a problem in polynomial approximation theory, may be found in [107, 108] and [110] (see also ([68] and [99]). In these papers a bound on the angle between a single eigenvector and a Krylov subspace is given. This bound depends on the clustering and the separation of the eigenvalues of $\mathcal{A}$. It can be shown that Arnoldi's method favours the outer part eigenvalues and associated eigenvectors of $\mathcal{A}$. Convergence to outlying eigenvalues is faster than to other eigenvalues and convergence to those eigenvalues is better the more the rest of the spectrum is clustered.

For the IRA method convergence has been shown for special cases; Sorensen [130] showed convergence in the case of stationary filter polynomials $q(z)$ (that is, the same shifts are used in each restart) for nonsymmetric matrices and in the case of "exact shift" filter polynomials for symmetric matrices. In [6] a convergence analysis using tools from functional analysis, pseudospectra and potential theory is given. Bounds on the gap between the maximal reachable invariant subspace, and a polynomial restarted Krylov subspace are provided. Again this approach amounts to minimising a polynomial that is derived from the roots of the filter polynomial. Linear convergence on a rate depending on the separation of the desired eigenvalues from the remainder of the spectrum is proved. Another approach (see [77]) uses the connection of IRA to nonstationary simultaneous iteration to prove convergence of $\text{span}\{\Psi_i(\mathcal{A})\mathbf{Q}_k\}$, the subspace produced by the restart procedure, to the invariant subspace $\text{span}\{\mathbf{Z}_k\}$ of $\mathcal{A}$ of dimension k. Here, $\Psi_i = q_i \dots q_2 q_1$ is the product of all the restart polynomials. The main result [77, Theorem 5.1], which uses [150, Theorem 5.1] is given by

$$\text{dist}(\text{span}\{\Psi_i(\mathcal{A})\mathbf{Q}_k\}, \text{span}\{\mathbf{Z}_k\}) \leq C\xi_i,$$

where C is a constant and $\Psi_i(\lambda_j) \neq 0$ for $j = 1, \dots, k$

$$\xi_i = \frac{\max_{j=k+1,\dots,n} |\Psi_i(\lambda_j)|}{\min_{j=1,\dots,k} |\Psi_i(\lambda_j)|}.$$

In particular $\text{span}\{\Psi_i(\mathcal{A})\mathbf{Q}_k\} \to \text{span}\{\mathbf{Z}_k\}$ if $\xi_i \to 0$. The latter approximation problem can be solved for special polynomials and special regions (see [77]), however the rigorous convergence theory for exact shift filter polynomials remains an open problem. In practice the exact shift approach is very successful, so for our purposes we assume at least linear convergence, although in practice the convergence might be much faster. Since, for exact shifts we have $\hat{h}_{k+1,k} = 0$ in (7.2) we obtain $\hat{\mathbf{f}}_k = \mathbf{f}_m \mathbf{V}_{m,k}$ and hence

$$\|\hat{\mathbf{f}}_k^{(i)}\| = \|\mathbf{f}_m^{(i)}\| |\mathbf{V}_{m,k}^{(i)}| \leq \eta^{(i)} \|\mathbf{f}_m^{(i)}\|, \quad \text{where} \quad \eta^{(i)} \leq 1, \tag{7.3}$$

using $|\mathbf{V}_{m,k}^{(i)}| \leq 1$ where i denotes the number of restarts. If we assume $\eta^{(i)} < 1$ then, for exact shifts, we obtain $\|\hat{\mathbf{f}}_k^{(i)}\| \to 0$ with at least a linear convergence rate.

7.3 Inexact solves in shift-invert Arnoldi's method with and without implicit restarts

This section contains one of the main results in this chapter. Specifically, we extend the relaxation strategy developed by Bouras and Frayssé [14] and analysed by [118] to the implicitly restarted Arnoldi method.

As noted in the previous section Arnoldi's method favours the outer part of the spectrum. To accelerate convergence to the inner part of the spectrum typically a shift-invert approach is used; if $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ is a given matrix eigenproblem then the spectral transformation

$$(\mathbf{A} - \sigma\mathbf{I})^{-1}\mathbf{x} = \frac{1}{\lambda - \sigma}\mathbf{x}$$

emphasises the eigenvalues of $\mathbf{A}$ close to the shift which then become outlying eigenvalues of $(\mathbf{A} - \sigma\mathbf{I})^{-1}$. Using

$$\mathcal{A} := (\mathbf{A} - \sigma\mathbf{I})^{-1}$$

in Arnoldi's method (7.1) then gives the so-called *shift-invert Arnoldi* method. Note that for a variable shift σ shift-invert Arnoldi becomes a so-called rational Krylov method [105].

Clearly, shift-invert Arnoldi involves the solution of a system

$$(\mathbf{A} - \sigma\mathbf{I})\mathbf{y} = \mathbf{q}_k \tag{7.4}$$

for $\mathbf{y}$ at each step of the Arnoldi (or IRA) process. Since our interest is in using a (preconditioned) iterative method for the solve we only get an inexact solution to that system, that is

$$\mathbf{y} = (\mathbf{A} - \sigma\mathbf{I})^{-1}\mathbf{q}_k + \mathbf{g}_k, \tag{7.5}$$

where $\mathbf{g}_k$ is the error vector produced at step k of the Arnoldi process. Results on inexact solves for (7.5), in particular on the choice of the solve tolerance $\|\mathbf{g}_k\|$ are given in the next subsection.

7.3.1 Bounds for eigenvector and invariant subspace components

Simoncini [118] has provided a relaxation strategy for the solve tolerance $\|\mathbf{g}_k\|$ of (7.5). This idea allows for the solve tolerance to be relaxed as the outer iteration proceeds and still assures convergence of Arnoldi's method to specific eigenpairs closest to σ. The following result on inexact Arnoldi's method was shown, for details and proof we refer to [118, Proposition 2.2]:

Proposition 7.1. *Let $\mathbf{H}_k$ and $\mathbf{H}_m$ be the upper Hessenberg matrices obtained after k and m steps of Arnoldi's method (with $m > k$) applied to $\mathcal{A}$. If $(\mathbf{u}_k, \theta_k)$ is an eigenpair of $\mathbf{H}_k$ (and hence $(\mathbf{Q}_k\mathbf{u}_k, \theta_k)$ an approximate eigenpair of $\mathcal{A}$) where the corresponding norm of the eigenvalue residual (Ritz residual) $\|\mathbf{r}_k\|$ is small enough, then there exists a unit norm eigenvector $\mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}$ of $\mathbf{H}_m$ with $\mathbf{u}_1 \in \mathbb{C}^k$ such that*

$$\|\mathbf{u}_2\| \leq \frac{\tau}{\sqrt{1+\tau^2}}, \quad \textit{with} \quad \tau \in \mathbb{R}, \quad 0 \leq \tau \leq 2\frac{\|\mathbf{r}_k\|}{\delta_{m,k}}.$$

where $\delta_{m,k} = \sigma_{\min}(\mathbf{Y}^H\mathbf{H}_m\mathbf{Y} - \theta_k\mathbf{I})$ with $\mathbf{Y}$ chosen such that $\left[\begin{bmatrix} \mathbf{u}_k \\ \mathbf{0} \end{bmatrix}, \mathbf{Y}\right]$ is unitary.

The value of $\delta_{m,k}$ depends on the separation of θ_k from rest of the spectrum of $\mathbf{H}_m$, for further discussions on this value we refer to [118]. Also, in [118] the idea of Proposition 7.1 has been extended to the approximation of invariant subspaces (see [118, Proposition 2.5]): If a representation of an invariant subspace of $\mathbf{H}_m$ in terms of a unitary matrix $\mathbf{U}$ is given, then the components of this unitary matrix are decreasing in the order of the residual $\mathbf{R}_k$ corresponding to the approximate invariant subspace of dimension k.

The same result for approximate invariant subspaces is shown in the following Theorem, using an alternative idea to Proposition 7.1, which was presented after [118, Proposition 2.5].

Theorem 7.2. *Assume we have carried out $m = k+p$ steps of the Arnoldi factorisation (7.1) and assume $\mathbf{H}_m$ has the Schur decomposition*

$$\mathbf{H}_m = \mathbf{W}\mathbf{T}\mathbf{W}^H, \quad \mathbf{T} = \begin{bmatrix} \mathbf{\Theta} & \star \\ \mathbf{0} & \mathbf{T}_{22} \end{bmatrix}, \tag{7.6}$$

where $\mathbf{W} = \begin{bmatrix} \mathbf{U} & \mathbf{W}_2 \end{bmatrix} \in \mathbb{C}^{m,m}$ *unitary and* $\mathbf{\Theta} \in \mathbb{C}^{k,k}$, $\mathbf{U} \in \mathbb{C}^{m,k}$ *and* $\mathbf{W}_2 \in \mathbb{C}^{m,m-k}$. *Let the columns of* $\mathbf{U}_k \in \mathbb{C}^{k,k}$ *be the orthogonal Schur vectors of* $\mathbf{H}_k \in \mathbb{C}^{k,k}$, *where* $\mathbf{\Theta}_k = \mathbf{U}_k^H \mathbf{H}_k \mathbf{U}_k$ *is the Schur decomposition with the Ritz values being the diagonal entries of* $\mathbf{\Theta}_k$. *Let* $\mathbf{R}_k = h_{k+1,k}\mathbf{q}_{k+1}\mathbf{e}_k^H \mathbf{U}_k$ *be the residual after* k *Arnoldi steps. Then the matrix* $\mathbf{U}$ *can be written as* $\mathbf{U} = \begin{bmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{bmatrix} \in \mathbb{C}^{m,k}$ *with* $\mathbf{U}^H\mathbf{U} = \mathbf{I}$, $\mathbf{U}_1 \in \mathbb{C}^{k,k}$, *and the columns of* $\mathbf{U}$ *span a* k*-dimensional invariant subspace of* $\mathbf{H}_m$, *such that*

$$\|\mathbf{U}_2\| \leq \frac{\|\mathbf{R}_k\|}{sep(\mathbf{T}_{22}, \mathbf{\Theta}_k)}, \tag{7.7}$$

where $sep(\mathbf{T}_{22}, \mathbf{\Theta}_k) := \min_{\|\mathbf{V}\|=1} \|\mathbf{T}_{22}\mathbf{V} - \mathbf{V}\mathbf{\Theta}_k\|$.

Proof. Set $\hat{\mathbf{U}}_k = \begin{bmatrix} \mathbf{U}_k \\ \mathbf{0} \end{bmatrix}$ and let $\mathbf{R}_k$ be the residual after k Arnoldi steps. Then

$$\mathbf{R}_k = \mathcal{A}\mathbf{Q}_k\mathbf{U}_k - \mathbf{Q}_k\mathbf{H}_k\mathbf{U}_k = \mathcal{A}\mathbf{Q}_k\mathbf{U}_k - \mathbf{Q}_k\mathbf{U}_k\mathbf{\Theta}_k. \tag{7.8}$$

We have

$$\begin{aligned} \|\mathbf{R}_k\| \geq \|\mathbf{Q}_k^H\mathbf{R}_k\| &= \|\mathbf{Q}_k^H\mathcal{A}\mathbf{Q}_k\mathbf{U}_k - \mathbf{U}_k\mathbf{\Theta}_k\| \\ &= \|\mathbf{H}_k \begin{bmatrix} \mathbf{U}_k \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{U}_k \\ \mathbf{0} \end{bmatrix} \mathbf{\Theta}_k\| \\ &= \|\mathbf{H}_m \begin{bmatrix} \mathbf{U}_k \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{U}_k \\ \mathbf{0} \end{bmatrix} \mathbf{\Theta}_k\| \\ &= \|\mathbf{W}\mathbf{T}\mathbf{W}^H\hat{\mathbf{U}}_k - \hat{\mathbf{U}}_k\mathbf{\Theta}_k\| \end{aligned}$$

Also with $\mathbf{W}$ unitary

$$\begin{aligned} \|\mathbf{R}_k\| &\geq \|\begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} \mathbf{W}^H\mathbf{Q}_k^H\mathbf{R}_k\| \\ &= \|\begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} \mathbf{T}\mathbf{W}^H\hat{\mathbf{U}}_k - \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} \mathbf{W}^H\hat{\mathbf{U}}_k\mathbf{\Theta}_k\| \\ &= \|\mathbf{T}_{22}\mathbf{W}_2^H\hat{\mathbf{U}}_k - \mathbf{W}_2^H\hat{\mathbf{U}}_k\mathbf{\Theta}_k\| \\ &= \frac{\|\mathbf{T}_{22}\mathbf{W}_2^H\hat{\mathbf{U}}_k - \mathbf{W}_2^H\hat{\mathbf{U}}_k\mathbf{\Theta}_k\|}{\|\mathbf{W}_2^H\hat{\mathbf{U}}_k\|}\|\mathbf{W}_2^H\hat{\mathbf{U}}_k\| \\ &\geq \mathrm{sep}(\mathbf{T}_{22}, \mathbf{\Theta}_k)\|\mathbf{W}_2^H\hat{\mathbf{U}}_k\|. \end{aligned}$$

We can write $\mathbf{U} = \hat{\mathbf{U}}_k\hat{\mathbf{U}}_k^H\mathbf{U} + (\mathbf{I} - \hat{\mathbf{U}}_k\hat{\mathbf{U}}_k^H)\mathbf{U}$. With $\mathbf{U} = \begin{bmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{bmatrix}$ and the definition of $\hat{\mathbf{U}}_k$ we then have $\mathbf{U}_2 = \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} (\mathbf{I} - \hat{\mathbf{U}}_k\hat{\mathbf{U}}_k^H)\mathbf{U}$ and

$$\|\mathbf{U}_2\| \leq \|(\mathbf{I} - \hat{\mathbf{U}}_k\hat{\mathbf{U}}_k^H)\mathbf{U}\|.$$

It can be shown that $\|(\mathbf{I} - \hat{\mathbf{U}}_k\hat{\mathbf{U}}_k^H)\mathbf{U}\| = \|\mathbf{W}_2^H\hat{\mathbf{U}}_k\| = \sin(\mathbf{U}, \hat{\mathbf{U}}_k)$ (see [48]) and hence

$$\|\mathbf{U}_2\| \leq \|\mathbf{W}_2^H\hat{\mathbf{U}}_k\| \leq \frac{\|\mathbf{R}_k\|}{\mathrm{sep}(\mathbf{T}_{22}, \mathbf{\Theta}_k)}.$$

□

Theorem 7.2 states that the norm of the last $m-k$ rows of the matrix $\mathbf{U} \in \mathbb{C}^{m,k}$, representing a basis of the invariant subspace span$\{\mathbf{U}\}$ of $\mathbf{H}_m$ can be bounded by some quantity involving the norm of the residual $\mathbf{R}_k = h_{k+1,k}\mathbf{q}_{k+1}\mathbf{e}_k^H\mathbf{U}_k$ for the approximate invariant subspace of $\mathcal{A}$, where span$\{\mathbf{U}_k\}$ (see (7.8)) is an invariant subspace of the smaller matrix $\mathbf{H}_k$. In particular,

$$\|\mathbf{e}_l^H\mathbf{U}\| \leq \|\mathbf{U}_2\| \leq \frac{\|\mathbf{R}_k\|}{\operatorname{sep}(\mathbf{T}_{22}, \mathbf{\Theta}_k)}, \quad l = k+1, \dots, m. \tag{7.9}$$

for the rows of $\mathbf{U}_2$. This analysis can be carried out for larger submatrices $\mathbf{H}_{\tilde{k}}$ of $\mathbf{H}_m$ with $\tilde{k} > k$, leading to a decrease in the norm of the last $m-k$ rows of the basis of the invariant subspace span$\{\mathbf{U}\}$ where $\mathbf{U} \in \mathbb{C}^{m,k}$.

As noted before, in [118, Proposition 2.2 and Proposition 2.5] a similar result to Theorem 7.2 has been proved. It has been shown that under the condition that $\|\mathbf{R}_k\|$ is small enough we have

$$\|\mathbf{U}_2\| \leq \frac{\tau}{\sqrt{1+\tau^2}}, \quad \text{with} \quad \tau \in \mathbb{R}, \quad 0 \leq \tau < 2\frac{\|\mathbf{R}_k\|}{\operatorname{sep}(\mathbf{\Theta}_k, \mathbf{Y}^H\mathbf{H}_m\mathbf{Y})}, \tag{7.10}$$

with $\mathbf{Y}$ chosen such that $\left[\begin{bmatrix}\mathbf{U}_k \\ \mathbf{0}\end{bmatrix}, \mathbf{Y}\right]$ is unitary. It is not clear which of the bounds (7.7) or (7.10) is sharper (as already noted for the case of an approximate eigenvector in [118]), since they both involve the quantity $\mathbf{H}_m$ which is unknown at step $k < m$. In (7.7) the matrix $\mathbf{T}_{22}$ is completely specified by the spectral properties of the target matrix $\mathbf{H}_m$ whereas (7.10) takes into account the approximate invariant subspace. However, the bound (7.10) requires a condition on the residual, therefore we prefer (7.7) which does not impose a small enough norm of $\mathbf{R}_k$. A precise (theoretical and numerical) comparison of the bounds (7.7) or (7.10) is future research. We will also see, that for implicitly restarted Arnoldi's method good approximations for the unknown quantities depending on $\mathbf{H}_m$ are available after the first restart.

7.3.2 A relaxation strategy for implicitly restarted Arnoldi's method

In [118] the author proposes a relaxation strategy for inexact Arnoldi's method. We prove a result about a relaxation strategy for inexact implicitly restarted Arnoldi's method for finding an invariant subspace. The proof uses Theorem 7.2. We state implications of this proof and give a relaxation strategy for IRA.

If (7.5) holds then clearly $\mathcal{A} = (\mathbf{A} - \sigma\mathbf{I})^{-1}$ is not applied exactly in the Arnoldi method and hence we may write the Arnoldi relation (7.1) after m steps as

$$\mathcal{A}\mathbf{Q}_m + \mathbf{G}_m = \mathbf{Q}_m\mathbf{H}_m + h_{m+1,m}\mathbf{q}_{m+1}\mathbf{e}_m^H, \tag{7.11}$$

where $\mathbf{Q}_m^H\mathbf{Q}_m = \mathbf{I}$ is orthonormal and $\mathbf{G}_m = [\mathbf{g}_1, \dots, \mathbf{g}_m]$. The space originating from this method is no longer a Krylov subspace associated with $\mathcal{A}$, due to the inexact solves. Note that it could be seen as a Krylov subspace of the perturbed matrix $\mathcal{A} + \mathbf{G}_m\mathbf{Q}_m^H$. Clearly $\mathbf{Q}_m$ and the upper Hessenberg matrix $\mathbf{H}_m$ are different from those that would have been obtained from the exact Arnoldi procedure. Similar to [118] we consider the error that has been introduced in the Ritz residual. Let $\mathbf{U} \in \mathbb{C}^{m,k}$ be a unitary matrix forming the basis for a simple invariant subspace of size k of $\mathbf{H}_m$ with matrix representation $\mathbf{\Theta}$. This matrix can be found via the Schur decomposition of $\mathbf{H}_m$. Then with $\mathbf{H}_m\mathbf{U} = \mathbf{U}\mathbf{\Theta}$ and (7.11) we have

$$\begin{aligned} \mathcal{A}\mathbf{Q}_m\mathbf{U} &= \mathbf{Q}_m\mathbf{U}\mathbf{\Theta} + h_{m+1,m}\mathbf{q}_{m+1}\mathbf{e}_m^H\mathbf{U} - \mathbf{G}_m\mathbf{U}, \\ \mathcal{A}\mathbf{Q}_m\mathbf{U} - \mathbf{Q}_m\mathbf{U}\mathbf{\Theta} &= h_{m+1,m}\mathbf{q}_{m+1}\mathbf{e}_m^H\mathbf{U} - \mathbf{G}_m\mathbf{U}, \end{aligned}$$

We adopt the notation by Simoncini [118] and call $\mathcal{A}\mathbf{Q}_m\mathbf{U} - \mathbf{Q}_m\mathbf{U}\mathbf{\Theta}$ the true residual, a quantity which is not available during the computations. The matrix $h_{m+1,m}\mathbf{q}_{m+1}\mathbf{e}_m^H\mathbf{U}$ is called the computed residual, a quantity which in turn is available during the iterations. We are interested in the difference between the true and the computed residual and, in order to achieve accurate results with the inexact methods we want the difference between both quantities to be small. Hence, we consider

$$\|(\mathcal{A}\mathbf{Q}_m\mathbf{U} - \mathbf{Q}_m\mathbf{U}\mathbf{\Theta}) - \mathbf{R}_m\| = \|\mathbf{G}_m\mathbf{U}\|, \tag{7.12}$$

where $\mathbf{R}_m = h_{m+1,m}\mathbf{q}_{m+1}\mathbf{e}_m^H\mathbf{U}$ is the computed residual. We have the following main theorem of this section.

Theorem 7.3. *Assume we have carried out $m = k + p$ steps of Arnoldi's method. Let the Schur decomposition of $\mathbf{H}_m$ be given by (7.6), such that the matrix $\mathbf{U} \in \mathbb{C}^{m,k}$ with orthonormal columns forms a basis for a simple invariant subspace of size k of $\mathbf{H}_m$, and with Ritz values being the diagonal entries of $\mathbf{\Theta} = \mathbf{U}^H\mathbf{H}_m\mathbf{U}$. For any given $\varepsilon \in \mathbb{R}$ with $\varepsilon > 0$ assume that*

$$\|\mathbf{g}_l\| \le \begin{cases} \dfrac{\varepsilon}{2(m-k)} \dfrac{\mathrm{sep}(\mathbf{T}_{22}, \mathbf{\Theta}_k)}{\|\mathbf{R}_k\|} & \text{if} \quad l > k, \\ \dfrac{\varepsilon}{2k} & \text{otherwise.} \end{cases} \tag{7.13}$$

Then

$$\|\mathcal{A}\mathbf{Q}_m\mathbf{U} - \mathbf{Q}_m\mathbf{U}\mathbf{\Theta} - \mathbf{R}_m\| \le \varepsilon. \tag{7.14}$$

Proof. From (7.12) we have

$$\|\mathcal{A}\mathbf{Q}_m\mathbf{U} - \mathbf{Q}_m\mathbf{U}\mathbf{\Theta} - \mathbf{R}_m\| = \|\mathbf{G}_m\mathbf{U}\| = \|[\mathbf{g}_1, \dots, \mathbf{g}_m]\mathbf{U}\|,$$

and using the splitting of $\mathbf{U}$ as $\mathbf{U} = \begin{bmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{bmatrix} \in \mathbb{C}^{m,k}$ with $\mathbf{U}_1 \in \mathbb{C}^{k,k}$ from Theorem 7.2 we get

$$\begin{aligned} \|(\mathcal{A}\mathbf{Q}_m\mathbf{U} - \mathbf{Q}_m\mathbf{U}\mathbf{\Theta}) - \mathbf{R}_m\| &= \|[\mathbf{g}_1, \dots, \mathbf{g}_m]\mathbf{U}\| \\ &\le \|[\mathbf{g}_1, \dots, \mathbf{g}_k]\mathbf{U}_1\| + \|[\mathbf{g}_{k+1}, \dots, \mathbf{g}_m]\mathbf{U}_2\|. \end{aligned}$$

Now using (7.9) and that $\mathbf{U}_1 \in \mathbb{C}^{k,k}$ is unitary we obtain

$$\|(\mathcal{A}\mathbf{Q}_m\mathbf{U} - \mathbf{Q}_m\mathbf{U}\mathbf{\Theta}) - \mathbf{R}_m\| \le \|[\mathbf{g}_1, \dots, \mathbf{g}_k]\| + \|[\mathbf{g}_{k+1}, \dots, \mathbf{g}_m]\| \frac{\|\mathbf{R}_k\|}{\mathrm{sep}(\mathbf{T}_{22}, \mathbf{\Theta}_k)},$$

Then with (7.13) we obtain

$$\begin{aligned} \|(\mathcal{A}\mathbf{Q}_m\mathbf{U} - \mathbf{Q}_m\mathbf{U}\mathbf{\Theta}) - \mathbf{R}_m\| &\le \sum_{l=1}^{k} \|\mathbf{g}_l\| + \frac{\|\mathbf{R}_k\|}{\mathrm{sep}(\mathbf{T}_{22}, \mathbf{\Theta}_k)} \sum_{l=k+1}^{m} \|\mathbf{g}_l\| \\ &\le \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \frac{\|\mathbf{R}_k\|}{\mathrm{sep}(\mathbf{T}_{22}, \mathbf{\Theta}_k)} \frac{\mathrm{sep}(\mathbf{T}_{22}, \mathbf{\Theta}_k)}{\|\mathbf{R}_k\|}, \end{aligned}$$

which gives the required result (7.14) $\square$

Theorem 7.3 provides a relaxation strategy for Arnoldi's method which extends to implicitly restarted Arnoldi's method:

Corollary 7.4 (Relaxation strategy for IRA). *Assume we use implicitly restarted Arnoldi's method to find k eigenvalues using an invariant subspace of maximum size $m = k + p$. Let the Schur decomposition of $\mathbf{H}_m^{(i)}$ be given by (7.6) where the entries depend on i, the number of restarts, such that the unitary matrix $\mathbf{U}^{(i)} \in \mathbb{C}^{m,k}$ forms a basis for a simple invariant subspace of size k of $\mathbf{H}_m^{(i)}$, and with Ritz values being the diagonal entries of $\mathbf{\Theta}^{(i)} = {\mathbf{U}^{(i)}}^H \mathbf{H}_m^{(i)} \mathbf{U}^{(i)}$. For any given $\varepsilon \in \mathbb{R}$ with $\varepsilon > 0$ assume that*

$$\|\mathbf{g}_l^{(i)}\| \leq \begin{cases} \dfrac{\varepsilon}{2(m-k)} \dfrac{sep(\mathbf{T}_{22}^{(i)}, \mathbf{\Theta}_k^{(i)})}{\|\mathbf{R}_k^{(i)}\|} & \text{if} \quad l > k, \\ \dfrac{\varepsilon}{2k} & \text{otherwise} \end{cases} \tag{7.15}$$

holds for each restart i. Then

$$\|\mathcal{A}\mathbf{Q}_m^{(i)}\mathbf{U}^{(i)} - \mathbf{Q}_m^{(i)}\mathbf{U}^{(i)}\mathbf{\Theta}^{(i)} - \mathbf{R}_m^{(i)}\| \leq \varepsilon. \tag{7.16}$$

Omitting the index i for the moment, clearly we have that $\operatorname{sep}(\mathbf{T}_{22}, \mathbf{\Theta}_k)$ is not known at step k, since we do not know $\mathbf{T}_{22}$ from the Schur decomposition of $\mathbf{H}_m$. Hence we propose the following strategy for IRA, which is based on a consequence of the exact shift strategy and a result on the separation function $\operatorname{sep}(\mathbf{T}_{22}, \mathbf{\Theta}_k)$.

We perform $m = k + p$ steps of Arnoldi's method with accuracy $\varepsilon/2(m-k)$. This gives $\mathbf{H}_m$ and hence $\mathbf{\Theta}$ as well as $\mathbf{T}_{22}$. For the implicit restart with exact shifts we then have

$$\hat{\mathbf{H}}_m = \begin{bmatrix} \hat{\mathbf{H}}_k & \star \\ \mathbf{O} & \tilde{\mathbf{T}}_{22} \end{bmatrix},$$

that is $\hat{h}_{k+1,k} = 0$ in (7.2) (see [130, Lemma 3.10]). $\hat{\mathbf{H}}_k$ has the same eigenvalues as $\mathbf{\Theta}$ and $\tilde{\mathbf{T}}_{22}$ has the same eigenvalues as $\mathbf{T}_{22}$. This leads to the observation that, for the restart,

$$\operatorname{sep}(\mathbf{T}_{22}, \mathbf{\Theta}_k) = \operatorname{sep}(\mathbf{T}_{22}, \mathbf{\Theta}),$$

since we restart the iteration with that matrix $\hat{\mathbf{H}}_k$. Therefore $\operatorname{sep}(\mathbf{T}_{22}, \mathbf{\Theta})$ is determined by the separation between the square matrices containing the wanted and the unwanted eigenvalues. With [137, page 233] we have that

$$\operatorname{sep}(\mathbf{T}_{22}, \mathbf{\Theta}) \leq \min |\Lambda(\mathbf{T}_{22}) - \Lambda(\mathbf{\Theta})|.$$

We propose the following relaxation strategy for IRA: Solve the first $m = k+p$ Arnoldi steps exactly; then, from the first restart onwards choose

$$\|\mathbf{g}_l^{(i)}\| = \frac{\varepsilon}{2(m-k)} \frac{\min |\Lambda_W(\mathbf{H}_m^{(i)}) - \Lambda_U(\mathbf{H}_m^{(i)})|}{\|\mathbf{R}_k^{(i)}\|}, \quad l > k,$$

where $\Lambda_W(\mathbf{H}_m^{(i)}) = \{\theta_1^{(i)}, \ldots, \theta_k^{(i)}\}$ represents the wanted eigenvalues whereas $\Lambda_U(\mathbf{H}_m^{(i)}) = \{\nu_1^{(i)}, \ldots, \nu_p^{(i)}\}$ is the unwanted part of the spectrum of $\mathbf{H}_m^{(i)}$ and $i = 1, \ldots, i_{\max}$ denotes the number of the restart. The unwanted part of the spectrum, $\Lambda_U(\mathbf{H}_m^{(i)})$, is also used as shifts. We conclude this section with two remarks.

Remark 7.5. *For standard Arnoldi's method (without restarts), where only one eigenvector and corresponding eigenvalue is sought we use a similar condition to the one proposed in [118], that is*

$$\|\mathbf{g}_k\| = \frac{\delta_{k-1}}{2m\|\mathbf{r}_{k-1}\|}\varepsilon, \quad \delta_{k-1} := \min_{\theta_j \in \Lambda(\mathbf{H}_{k-1})\setminus\{\theta_{k-1}\}} |\theta_{k-1} - \theta_j|,$$

for $k > 1$ and $\|\mathbf{g}_1\| = \frac{\varepsilon}{m}$ for the first solve.

Results for the relaxation strategy for standard Arnoldi's method using Remark 7.5 are shown in part (a) of Example 7.7 (see Figures 7-3 and 7-4)

Remark 7.6. *Note that so far we have used (7.5), that is,*

$$(\mathbf{A} - \sigma\mathbf{I})\mathbf{y} = \mathbf{q}_k + \tilde{\mathbf{g}}_k, \quad \text{where} \quad \tilde{\mathbf{g}}_k = (\mathbf{A} - \sigma\mathbf{I})\mathbf{g}_k,$$

and therefore $\tilde{\mathbf{g}}_k$ is a scaled version of $\mathbf{g}_k$. In the following, for simplicity, we use $\mathbf{g}_k$ instead of $\tilde{\mathbf{g}}_k$, even though a scaled version of $\mathbf{g}_k$ should be used.

7.3.3 Numerical Example

We present three numerical examples supporting the results in this section.

Example 7.7. *Consider the matrix `sherman5.mtx` from the Matrix Market library [13]. This is a real nonsymmetric matrix of size $n = 3312$ which was also used to test relaxation strategies in [118]. The spectrum of this matrix is plotted in Figure 7-1. We use shift-invert Arnoldi's method with a very small fixed solve tolerance, which we call "exact" Arnoldi's method and shift-invert Arnoldi's method with a relaxed solve tolerance, which we call "inexact" Arnoldi's method. Furthermore, we use a starting vector $\mathbf{q}_1$ which is the normalised vector of all ones. Right-preconditioned GMRES with zero starting vector is used for the inner solves, where incomplete LU factorisation with drop tolerance 0.001 is applied.*

We carry out two tests

(a) *Standard Arnoldi method: In order to approximate the eigenvalue closest to zero, which is given by $\lambda = 4.692 \cdot 10^{-2}$. we use standard Arnoldi's method. We carry out a total of $m = 14$ steps (outer iterations) of both "exact" (without relaxation) and "inexact" (with relaxation) Arnoldi's method. We use the following stopping criteria:*

- *For "exact" solves (no relaxation) we use*

$$\|\mathbf{q}_k - \mathbf{A}\mathbf{y}\| \leq \frac{\varepsilon}{m}$$

for each outer iteration.

- *For "inexact" solves (relaxation) we use*

$$\|\mathbf{q}_k - \mathbf{A}\mathbf{y}\| \leq \frac{\delta_{k-1}}{2m\|\mathbf{r}_{k-1}\|}\varepsilon, \quad m = 14, \quad \varepsilon = 10^{-14} \tag{7.17}$$

for $k > 1$, where δ_{k-1} is as in Remark 7.5, and $\frac{\varepsilon}{m}$ for the first outer iteration, which gives a relaxation in the solve tolerance as the outer iteration proceeds.

(b) *Implicitly restarted Arnoldi's method: We use "exact" and "inexact" implicitly restarted Arnoldi's method in order to find the $k = 8$ eigenvalues closest to zero. We use a subspace of total size $m = k + p = 12$ and carry out a maximum of $i_{\max} = 11$ restarts (outer iterations). We use the following stopping criteria:*

- *For "exact" solves (no relaxation) we use*

$$\|\mathbf{q}_k - \mathbf{A}\mathbf{y}\| \leq \frac{\varepsilon}{2k}$$

for each outer iteration.

- *For "inexact" solves (relaxation) we use*

$$\|\mathbf{q}_l - \mathbf{A}\mathbf{y}\| \leq \frac{\varepsilon}{2(m-k)} \frac{\min |\Lambda_W(\mathbf{H}_m) - \Lambda_U(\mathbf{H}_m)|}{\|\mathbf{R}_k\|}, \quad m = 12, \quad \varepsilon = 10^{-13} \tag{7.18}$$

for $l > k$ after the restart and $\dfrac{\varepsilon}{2k}$ for the first m iterations. This achieves a relaxation in the solve tolerance after the restart.

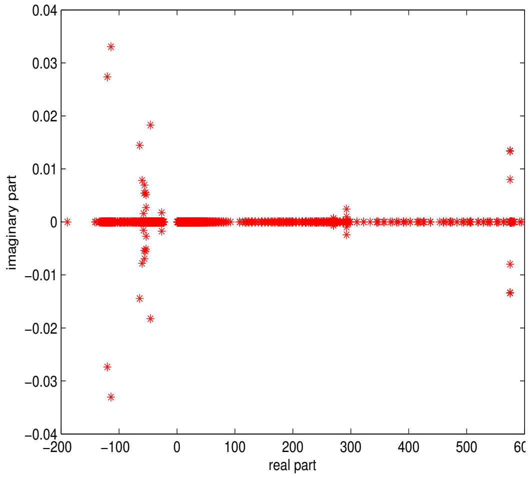


Figure 7-1: *Spectrum of matrix* `sherman5.mtx` *from Example 7.7.*

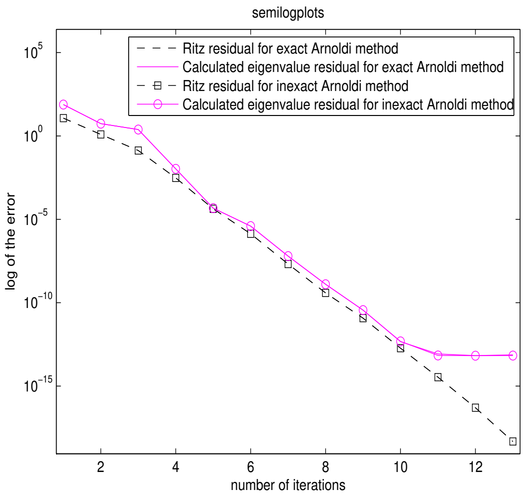


Figure 7-2: *Computed and Ritz residual for exact/inexact Arnoldi's method.*

The results for experiment (a) are plotted in Figures 7-2, 7-3 and 7-4. First note that the computed residual and the Ritz residual are the same for "inexact" and "exact" solves, see Figure 7-2, where plots for "inexact" and "exact" methods overlie each other both for the Ritz residual and the computed residual.

As expected, relaxing the tolerance results in a decreasing number of inner iterations as the outer iteration proceeds (inexact solve in Figure 7-3), whereas "exact" solves keep the number of inner iterations approximately constant. This leads to an improvement of the total number of iterations; the relaxation strategy requires only about 2/3 of the matrix-vector multiplications used in the method with fixed small tolerance solve (see "exact" Arnoldi in Figure 7-4).

Figures 7-5 and 7-6 present the results for experiment (b) using implicit restarts. Again, the number of inner iterations per outer iteration decreases due to the relaxation

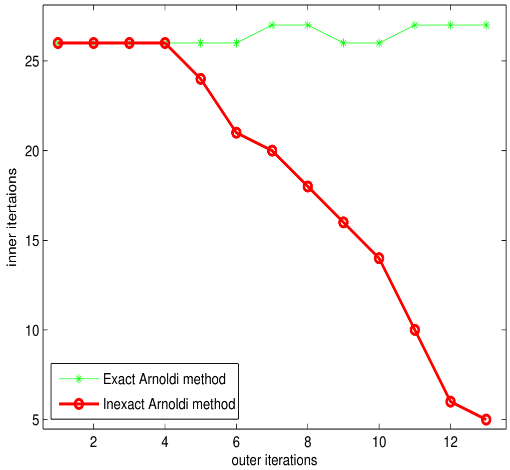


Figure 7-3: *Number of inner iterations against outer iterations for part (a) in Example 7.7.*

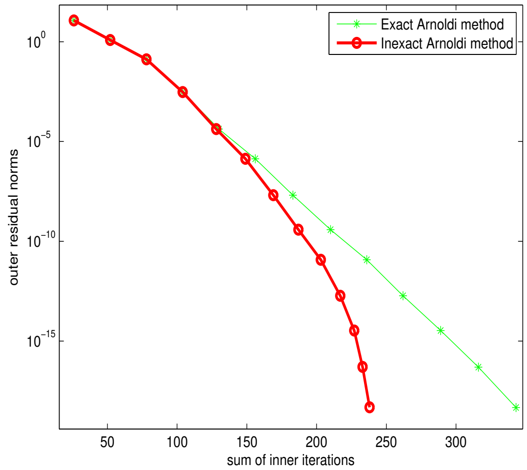


Figure 7-4: *Eigenvalue residual norms against sum of inner iterations for part (a) in Example 7.7.*

of the inner tolerance. The savings in the total number of matrix iterations is about 20 per cent (see Figure 7-6).

Example 7.8. *We use another example from the matrix market library, namely matrix* `qc2534.mtx`*, a real nonsymmetric matrix of size* $n = 2534$ *and spectrum plotted in Figure 7-7. We only test implicitly restarted Arnoldi's method to find the* $k = 6$ *eigenvalues closest to zero. We use a subspace of total size* $m = k + p = 10$ *and carry out a maximum of* $i_{\max} = 11$ *restarts (outer iterations). The starting vector* $\mathbf{q}_1$ *for the Arnoldi iteration is as in Example 7.7. Again, right-preconditioned GMRES with zero starting vector is used for the inner solves, where incomplete LU factorisation with drop tolerance* 0.1 *is applied as a preconditioner. For the stopping criterion we use (7.18) with* $m = 10$ *and* $\varepsilon = 10^{-13}$.

Figures 7-8 to 7-10 show the results for Example 7.8. We plot the computed residual for the first 6 Ritz values and the residual bound after each restart. From Figure 7-8 we can see that there is hardly any difference between the Ritz residual for Arnoldi's method with fixed small tolerance solve ("exact" Arnoldi) and relaxed Arnoldi's method ("inexact" Arnoldi). The computed residuals overlie each other. Figures 7-9 and 7-10 again show the benefits of the relaxation strategy developed in Theorems 7.2 and 7.3 which requires fewer total iterations than the fixed tolerance solves. For this example inexact IRA reduces the total number of matrix-vector products by 20 per cent compared to the use of "exact" solves.

Example 7.9. *Finally, we test a problem generated using the* IFISS *package [32]. The eigenvalue problem generated is of dimension* $n = 834$ *and represents the flow in a lid driven cavity. Taking a regularised cavity with underlying uniform* 16×16 *grid and viscosity parameter* 0.01 *leads to a block-structured eigenvalue problem of the form* $\mathbf{A}\mathbf{x} = \lambda \mathbf{M}\mathbf{x}$*, where*

$$\mathbf{A} = \begin{bmatrix} \mathbf{K} & \mathbf{B}^H \\ \mathbf{B} & \mathbf{0} \end{bmatrix}, \quad \mathbf{M} = \begin{bmatrix} \mathbf{M}_1 & \mathbf{0}^H \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \mathbf{K}, \mathbf{M}_1 \in \mathbb{C}^{578,578} \quad \textit{and} \quad \mathbf{B} \in \mathbb{C}^{256,578},$$

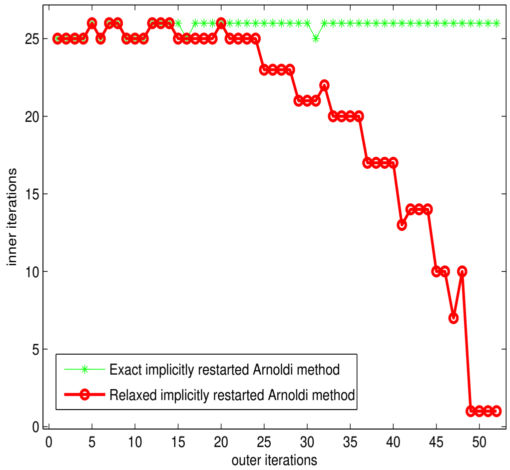


Figure 7-5: *Number of inner iterations against outer iterations for part (b) in Example 7.7.*

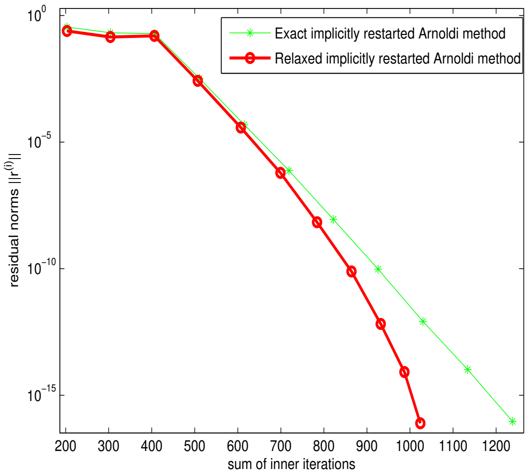


Figure 7-6: *Eigenvalue residual norms against sum of inner iterations for part (b) in Example 7.7.*

after mixed finite element approximation (see [34]). The spectrum of this problem is plotted in Figure 7-11. The eigenvalues closest to zero are found by calculating eigenvalues of largest magnitude of $\mathbf{A}^{-1}\mathbf{M}$*, which is done by means of inexact implicitly restarted shift-invert Arnoldi method. Again, we compare the "exact" and "inexact" strategy to find the* $k = 6$ *eigenvalues closest to zero. The maximum size of the subspace used is* $m = k + p = 16$ *and a maximum of* $i_{\max} = 5$ *restarts is taken. The starting vector* $\mathbf{q}_1$ *for the Arnoldi iteration is as in Example 7.7. Again, right-preconditioned GMRES with zero starting vector is used for the inner solves, where an incomplete LU factorisation of a diagonally perturbed* $\mathbf{A}$ *with drop tolerance* 0.0001 *is applied as a preconditioner. Note that this preconditioner is not optimal. This example is only supposed to show that the relaxation strategy for the outer iterative method (IRA) developed in Subsection 7.3.2 works, where the preconditioner for the inner iteration does not play any role. For the stopping criterion we use (7.18) with* $m = 16$ *and* $\varepsilon = 10^{-11}$

Figures 7-12 to 7-14 present the results for Example 7.9. Observe that there is no difference between the residuals for the first 6 Ritz values no matter if we use the inexact or the exact algorithm. Figures 7-13 and 7-14 again show that the relaxation strategy requires fewer total iterations. For this example inexact IRA improves the total number of matrix-vector products by about 20 per cent compared to the use of exact solves.

7.4 Tuning the preconditioner for shift-invert Arnoldi's method

This section considers the inner iteration for the (inexact) solve of

$$(\mathbf{A} - \sigma\mathbf{I})\mathbf{y} = \mathbf{q}_k$$

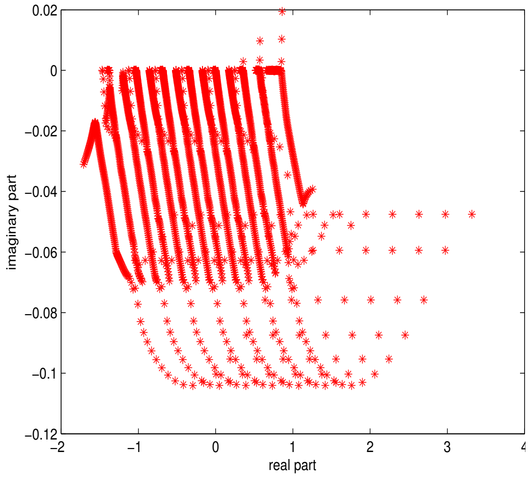


Figure 7-7: *Spectrum of matrix* `qc2534.mtx` *from Example 7.8.*

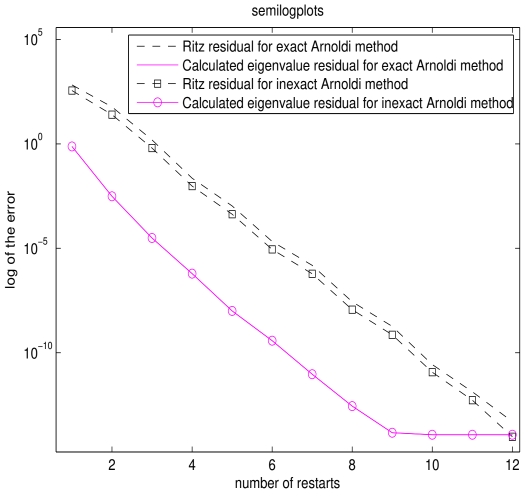


Figure 7-8: *Computed and Ritz residual for exact/inexact IRA method.*

at each step of Arnoldi's method. Since $\mathbf{A}$ is generally nonsymmetric we use preconditioned GMRES for this inner solve. Concentrating on right-preconditioned GMRES, this means at each step of Arnoldi's method we need to solve a system of the form

$$(\mathbf{A}-\sigma\mathbf{I})\mathbf{P}^{-1}\tilde{\mathbf{y}}=\mathbf{q}_k,\quad \mathbf{y}=\mathbf{P}^{-1}\tilde{\mathbf{y}},$$

where $\mathbf{P}$ is a preconditioner for $A-\sigma\mathbf{I}$.

GMRES is an iterative Krylov solver for the general linear system $\mathbf{Bz}=\mathbf{b}$ (here $\mathbf{B}=(\mathbf{A}-\sigma\mathbf{I})\mathbf{P}^{-1}$) whose performance depends on the initial guess $\mathbf{z}_0$, the eigenvalue clustering of $\mathbf{B}$ (see [15]) and the right hand side $\mathbf{b}$, which can be seen by the fact that after j iterations of GMRES, the residual norm $\|\mathbf{s}_j\|=\|\mathbf{b}-\mathbf{Bz}_j\|$ is bounded by (see [111, 112])

$$\|\mathbf{s}_j\|=\kappa(\mathbf{W})\min_{\substack{p\in\Pi_j\\ p(0)=1}}\max_{i=1,\dots,n}|p(\mu_i)|\|\mathbf{s}_0\|, \tag{7.19}$$

where $\mathbf{B}$ is assumed to be diagonalisable $\mathbf{B}=\mathbf{W}\mathrm{diag}\{\mu_1,\dots,\mu_n\}\mathbf{W}^{-1}$ and $\kappa(\mathbf{W})=\|\mathbf{W}\|\|\mathbf{W}^{-1}\|$. Several other bounds are possible, see Appendix B for details. Here, Π_j is the set of polynomials of degree j. The more the eigenvalues μ_i, $i=1,\dots,n$ of $\mathbf{B}$ are clustered the better the convergence bound in (7.19), assuming $\kappa(\mathbf{W})$ is not too large, that is the matrix $\mathbf{B}$ is only mildly non-normal. The dependence on the starting guess and the right hand side are incorporated into $\mathbf{s}_0$. In the following we are interested in the eigenvalue clustering, which clearly improves the approximation problem. For simplicity we consider $\sigma=0$, the results are easily generalised to nonzero values of the shift σ. In the following we assume that $\mathbf{B}=\mathbf{AP}^{-1}$ is only mildly non-normal so that the convergence bound of GMRES in (7.19) mainly depends on the clustering properties of the eigenvalues of $\mathbf{B}$ and on the right hand side $\mathbf{s}_0$ which is equal to $\mathbf{b}$ if the starting guess is given $\mathbf{z}_0=\mathbf{0}$.

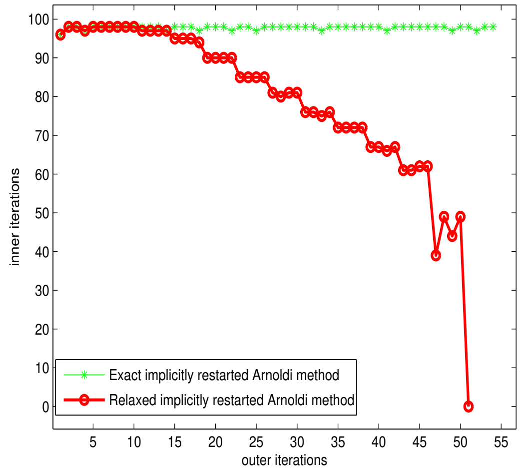


Figure 7-9: *Inner iterations against outer iterations for Example 7.7.*

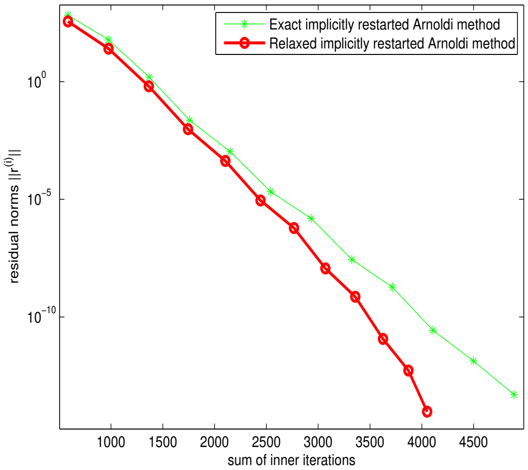


Figure 7-10: *Residual norms against sum of inner iterations for Example 7.7.*

7.4.1 Arnoldi's method applied to $\mathbf{A}^{-1}$ with a tuned preconditioner

In Chapters 4 and 6 (see also [42] and [45]) a tuned version of a preconditioner $\mathbf{P}$ was introduced for inexact inverse iteration for Hermitian and non-Hermitian problems with the aim of reducing the number of inner iterations. In [104] this idea was extended to subspace iteration for nonsymmetric problems. We generalise this tuning strategy to the world of Krylov methods. We prove that the tuned preconditioner amplifies the clustering properties of the eigenvalues of the system matrix $\mathbf{B} = \mathbf{A}\mathbf{P}^{-1}$ (Theorem 7.10) and hence improves the bound (7.19) and reduces the number of inner iterations as is shown in our numerical experiments.

Let the tuned preconditioner $\mathbb{P}_k$ satisfy

$$\mathbb{P}_k\mathbf{Q}_k = \mathbf{A}\mathbf{Q}_k, \quad \mathbf{Q}_k \in \mathbb{C}^{n,k}. \tag{7.20}$$

This condition can be achieved by a simple rank-k change of the usual preconditioner, namely

$$\mathbb{P}_k = \mathbf{P} + (\mathbf{A}-\mathbf{P})\mathbf{Q}_k\mathbf{Q}_k^H. \tag{7.21}$$

Furthermore, assuming $\mathbf{Q}_k^H\mathbf{P}^{-1}\mathbf{A}\mathbf{Q}_k$ is nonsingular, the inverse of $\mathbb{P}_k$ is given by

$$\mathbb{P}_k^{-1} = \mathbf{P}^{-1} - \mathbf{P}^{-1}(\mathbf{A}-\mathbf{P})\mathbf{Q}_k(\mathbf{Q}_k^H\mathbf{P}^{-1}\mathbf{A}\mathbf{Q}_k)^{-1}\mathbf{Q}_k^H\mathbf{P}^{-1}, \tag{7.22}$$

using the Sherman-Morrison-Woodbury formula [48]. We can write this matrix as

$$\mathbb{P}_k^{-1} = \mathbf{P}^{-1}(\mathbf{I}-\mathbf{A}\mathbf{Q}_k(\mathbf{Q}_k^H\mathbf{P}^{-1}\mathbf{A}\mathbf{Q}_k)^{-1}\mathbf{Q}_k^H\mathbf{P}^{-1}) + \mathbf{Q}_k(\mathbf{Q}_k^H\mathbf{P}^{-1}\mathbf{A}\mathbf{Q}_k)^{-1}\mathbf{Q}_k^H\mathbf{P}^{-1}, \tag{7.23}$$

where the first part $\mathbf{I}-\mathbf{A}\mathbf{Q}_k(\mathbf{Q}_k^H\mathbf{P}^{-1}\mathbf{A}\mathbf{Q}_k)^{-1}\mathbf{Q}_k^H\mathbf{P}^{-1}$ is singular and an oblique projector onto $\mathcal{R}(\mathbf{Q}_k^H\mathbf{P}^{-1})^{\perp}$ along $\mathcal{R}(\mathbf{A}\mathbf{Q}_k)$. Note that there is a similarity between

$$\mathbf{P}^{-1}(\mathbf{I}-\mathbf{A}\mathbf{Q}_k(\mathbf{Q}_k^H\mathbf{P}^{-1}\mathbf{A}\mathbf{Q}_k)^{-1}\mathbf{Q}_k^H\mathbf{P}^{-1})$$

and the deflation-based preconditioner introduced for Hermitian systems in [40], namely $\mathbf{I}-\mathbf{A}\mathbf{Z}(\mathbf{Z}^H\mathbf{A}\mathbf{Z})^{-1}\mathbf{Z}^H$, where the deflation subspace is $\mathcal{R}(\mathbf{Z})$. For $\mathbf{P} = \mathbf{I}$ and $\mathbf{Z} =$

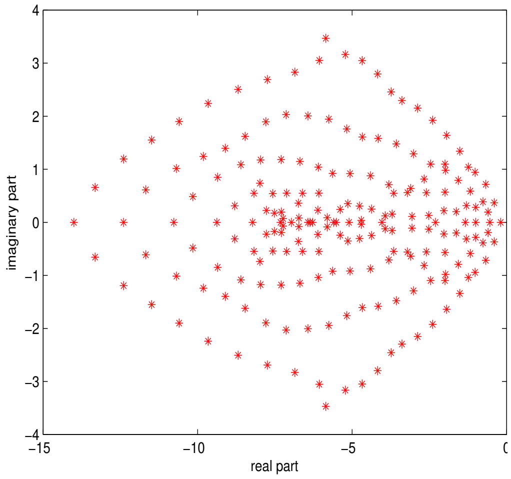


Figure 7-11: *Spectrum of matrix $\mathbf{A}$ from Example 7.9.*

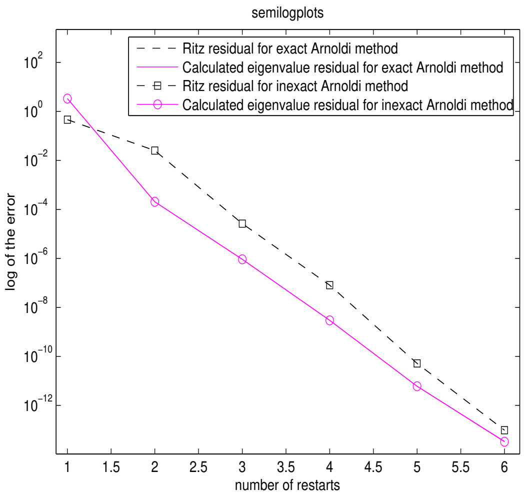


Figure 7-12: *Computed and Ritz residual for exact/inexact IRA method.*

$\mathbf{Q}_k$ both projections are equal. Also note that the tuned preconditioner $\mathbb{P}_k^{-1}$ bears a resemblance to the balanced preconditioner (see [37]) given by

$$(\mathbf{I} - \mathbf{Z}(\mathbf{Y}^H\mathbf{A}\mathbf{Z})^{-1}\mathbf{Y}^H\mathbf{A})\mathbf{P}^{-1}(\mathbf{I} - \mathbf{A}\mathbf{Z}(\mathbf{Y}^H\mathbf{A}\mathbf{Z})^{-1}\mathbf{Y}^H) + \mathbf{Z}(\mathbf{Y}^H\mathbf{A}\mathbf{Z})^{-1}\mathbf{Y}^H.$$

Equivalence between both hold for $\mathbf{I} - \mathbf{Z}(\mathbf{Y}^H\mathbf{A}\mathbf{Z})^{-1}\mathbf{Y}^H\mathbf{A} = \mathbf{I}$ and $\mathbf{Z} = \mathbf{Y} = \mathbf{Q}_k$ as well as $\mathbf{P} = \mathbf{I}$. Condition $\mathbf{I} - \mathbf{Z}(\mathbf{Y}^H\mathbf{A}\mathbf{Z})^{-1}\mathbf{Y}^H\mathbf{A} = \mathbf{I}$ with $\mathbf{Z} = \mathbf{Y} = \mathbf{Q}_k$ holds if $\mathbf{Q}_k$ is a left invariant subspace of $\mathbf{A}$. Note that the deflation-based preconditioner and the balanced preconditioner are projections and hence singular, whereas the tuned preconditioner defined by (7.21) is nonsingular. We concentrate on the tuned preconditioner (7.21).

It is well-known that convergence of GMRES improves if eigenvalues are significantly clustered around a point away from zero [15] and the following theorem shows that the tuned preconditioner achieves such a clustering.

Theorem 7.10. *Let $\mathbf{P}$ be a preconditioner for $\mathbf{A}$ with $\mathbf{A} = \mathbf{P} + \mathbf{E}$. Assume $\mathbb{P}_k^{-1}$ given by (7.22) exists and we have carried out k steps of Arnoldi method applied to $\mathbf{A}^{-1}$, with the inner iteration being solved with preconditioner $\mathbb{P}_k^{-1}$. Then the matrix $\mathbf{A}\mathbb{P}_k^{-1}$ has at least k eigenvalues equal to one and $n-k$ eigenvalues that are close to one, in the sense that they are eigenvalues of $\mathbf{L}_2 \in \mathbb{C}^{n-k,n-k}$, which satisfies*

$$\|\mathbf{L}_2 - \mathbf{I}\| \leq C\|\mathbf{E}\|,$$

where $\mathbf{I}$ is the identity matrix of size $(n-k)\times(n-k)$ and C is a constant of the order of $\|\mathbf{A}\|$.

Proof. From (7.20) we have

$$\mathbf{A}\mathbb{P}_k^{-1}\mathbf{A}\mathbf{Q}_k = \mathbf{A}\mathbf{Q}_k,$$

so that $\mathbf{A}\mathbf{Q}_k$ is a k-dimensional invariant subspace of $\mathbf{A}\mathbb{P}_k^{-1}$ corresponding to the eigenvalue 1. Furthermore, with span$\{\mathbf{Q}_k^\perp\}$ being the $n-k$-dimensional subspace orthogonal to span$\{\mathbf{Q}_k\}$ we have from (7.23)

$$\mathbf{A}\mathbb{P}_k^{-1}\mathbf{P}\mathbf{Q}_k^\perp = \mathbf{A}\mathbf{Q}_k^\perp = \mathbf{P}\mathbf{Q}_k^\perp + \mathbf{E}\mathbf{Q}_k^\perp,$$

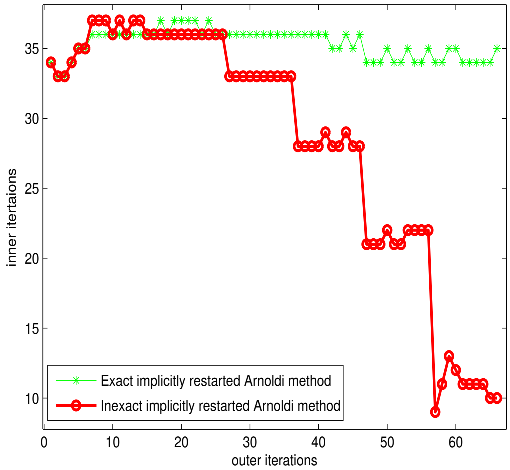


Figure 7-13: *Inner iterations per outer iteration for Example 7.9.*

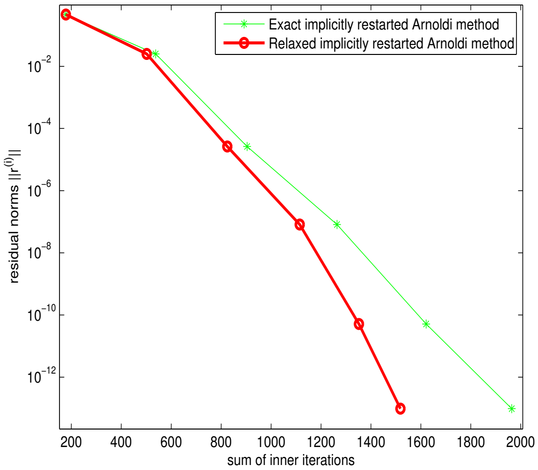


Figure 7-14: *Residual norms against sum of inner iterations for Example 7.9.*

and hence

$$(\mathbf{A}\mathbb{P}_k^{-1} - \mathbf{E}\mathbf{Q}_k^{\perp}(\mathbf{P}\mathbf{Q}_k^{\perp})^{\perp})\mathbf{P}\mathbf{Q}_k^{\perp} = \mathbf{P}\mathbf{Q}_k^{\perp}, \tag{7.24}$$

where $(\mathbf{P}\mathbf{Q}_k^{\perp})^{\perp} \perp \mathbf{P}\mathbf{Q}_k^{\perp}$. Then, $\mathbf{Z}_1 := \mathbf{P}\mathbf{Q}_k^{\perp}$ forms a basis for a $n-k$-dimensional invariant subspace of

$$\tilde{\mathbf{A}} := (\mathbf{A}\mathbb{P}_k^{-1} - \mathbf{E}\mathbf{Q}_k^{\perp}(\mathbf{P}\mathbf{Q}_k^{\perp})^{\perp})$$

with corresponding eigenvalues 1 ($n-k$ times). Omitting the index k and taking the QR-factorisation of $\mathbf{Z}_1 = \tilde{\mathbf{Z}}_1\mathbf{R}$, where $\tilde{\mathbf{Z}}_1 \in \mathbb{C}^{n,n-k}$ unitary we can write (7.24) as

$$\tilde{\mathbf{A}}\tilde{\mathbf{Z}}_1 = \tilde{\mathbf{Z}}_1, \quad \tilde{\mathbf{Z}}_1 \in \mathbb{C}^{n,n-k},$$

which, assuming $\mathcal{R}(\tilde{\mathbf{Z}}_1)$ is a simple invariant subspace of $\tilde{\mathbf{A}}$ leads to the decomposition

$$\begin{bmatrix} \tilde{\mathbf{Z}}_1 & \tilde{\mathbf{Z}}_1^{\perp} \end{bmatrix}^H \tilde{\mathbf{A}} \begin{bmatrix} \tilde{\mathbf{Z}}_1 & \tilde{\mathbf{Z}}_1^{\perp} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{H} \\ \mathbf{0} & \mathbf{L}_1 \end{bmatrix},$$

where $\mathbf{L}_1 \in \mathbb{C}^{k,k}$. Note that we are interested in the $n-k$-dimensional invariant subspace for the perturbation result of the second part of the theorem, hence the unusual blocks. If 1 is not an eigenvalue of $\mathbf{L}_1$ then $\tilde{\mathbf{A}}$ can be block-diagonalised [48], which gives

$$\begin{bmatrix} \tilde{\mathbf{Z}}_1 & \tilde{\mathbf{Z}}_2 \end{bmatrix}^{-1} \tilde{\mathbf{A}} \begin{bmatrix} \tilde{\mathbf{Z}}_1 & \tilde{\mathbf{Z}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{L}}_1 \end{bmatrix},$$

where $\tilde{\mathbf{L}}_1 \in \mathbb{C}^{k,k}$ has the same spectrum as $\mathbf{L}_1 \in \mathbb{C}^{k,k}$. Letting $\mathbf{F} := \mathbf{E}\mathbf{Q}_k^{\perp}(\mathbf{P}\mathbf{Q}_k^{\perp})^{\perp})$ and partitioning $\mathbf{F}$ in the same way as $\tilde{\mathbf{A}}$

$$\begin{bmatrix} \tilde{\mathbf{Z}}_1 & \tilde{\mathbf{Z}}_2 \end{bmatrix}^{-1} \mathbf{F} \begin{bmatrix} \tilde{\mathbf{Z}}_1 & \tilde{\mathbf{Z}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{bmatrix},$$

perturbation theory of simple invariant subspaces (see [137]) proves the existence of a simple right-invariant subspace of $\tilde{\mathbf{A}} + \mathbf{F} = \mathbf{A}\mathbb{P}_k^{-1}$ with representation matrix $\mathbf{L}_2 =$

$\mathbf{I} + \mathbf{F}_{11} + \mathbf{F}_{12}\mathbf{J}$, where

$$\|\mathbf{J}\| \leq 2\frac{\|\mathbf{F}_{21}\|}{\text{sep}(\mathbf{I}, \tilde{\mathbf{L}}_1) - \|\mathbf{F}_{11}\| - \|\mathbf{F}_{22}\|}.$$

Combining the results and using $\|\mathbf{Q}_k\| = 1$ gives $\|\mathbf{L}_2 - \mathbf{I}\| \leq C\|\mathbf{E}\|$. □

Note that from the proof of Theorem 7.10 it follows that $C = \mathcal{O}(\|\mathbf{F}\|) \approx \mathcal{O}(\|\mathbf{P}\|) \approx \mathcal{O}(\|\mathbf{A}\|)$ which is not necessarily small.

Theorem 7.10 shows that the tuned preconditioner has the nice property of clustering at least part of the spectrum of $\mathbf{A}\mathbb{P}_k^{-1}$ around 1. Theorem 7.10 is general, in the sense that it holds for any matrix $\mathbf{A}$ and orthogonal matrices $\mathbf{Q}_k$ which satisfy (7.20). The following theorem shows how the eigenvalues of $\mathbf{A}\mathbb{P}_k^{-1}$ behave, if Arnoldi's method applied to $\mathbf{A}^{-1}$ is used to compute the orthogonal basis $\mathbf{Q}_k$ of the Krylov subspace $\mathcal{K}_k(\mathbf{A}^{-1}, \mathbf{q}_1)$.

Theorem 7.11. *Let $\mathbb{P}_k$ be given by (7.21) and assume Arnoldi's method is applied to $\mathbf{A}^{-1}$ without errors, that is*

$$\mathbf{A}^{-1}\mathbf{Q}_k = \mathbf{Q}_k\mathbf{H}_k + \mathbf{q}_{k+1}h_{k+1,k}\mathbf{e}_k^H, \tag{7.25}$$

and hence $\mathbf{A}^{-1}$ has the upper Hessenberg form

$$\begin{bmatrix} \mathbf{Q}_k & \mathbf{Q}_k^{\perp} \end{bmatrix}^H \mathbf{A}^{-1} \begin{bmatrix} \mathbf{Q}_k & \mathbf{Q}_k^{\perp} \end{bmatrix} = \begin{bmatrix} \mathbf{H}_k & \mathbf{T}_{12} \\ h_{k+1,k}\mathbf{e}_1\mathbf{e}_k^H & \mathbf{T}_{22} \end{bmatrix}, \tag{7.26}$$

where $\begin{bmatrix} \mathbf{Q}_k & \mathbf{Q}_k^{\perp} \end{bmatrix}$ is unitary and $\mathbf{H}_k \in \mathbb{C}^{k,k}$ and $\mathbf{T}_{22} \in \mathbb{C}^{n-k,n-k}$ are upper Hessenberg. Assume $\mathbb{P}_k^{-1}$ given by (7.22) exists. Then $\mathbf{A}\mathbb{P}_k^{-1}$ has the same eigenvalues as the matrix

$$\begin{bmatrix} \mathbf{I} + \mathbf{Q}_k^H(\mathbf{I} - \mathbf{A}\mathbb{P}_k^{-1})\mathbf{A}\mathbf{q}_{k+1}h_{k+1,k}\mathbf{e}_k^H & \mathbf{Q}_k^H\mathbf{A}\mathbb{P}_k^{-1}\mathbf{Q}_k^{\perp} \\ {\mathbf{Q}_k^{\perp}}^H(\mathbf{I} - \mathbf{A}\mathbb{P}_k^{-1})\mathbf{A}\mathbf{q}_{k+1}h_{k+1,k}\mathbf{e}_k^H & \mathbf{K}_{22} \end{bmatrix}, \tag{7.27}$$

where

$$\begin{aligned}\mathbf{K}_{22} &= \mathbf{T}_{22}^{-1}({\mathbf{Q}_k^{\perp}}^H\mathbf{P}\mathbf{Q}_k^{\perp})^{-1} \\ &\quad + \mathbf{T}_{22}^{-1}({\mathbf{Q}_k^{\perp}}^H\mathbb{P}_k^{-1}\mathbf{A}\mathbf{q}_{k+1}h_{k+1,k}\mathbf{e}_k^H\mathbf{Q}_k^H\mathbf{P}\mathbf{Q}_k^{\perp}({\mathbf{Q}_k^{\perp}}^H\mathbf{P}\mathbf{Q}_k^{\perp})^{-1} - h_{k+1,k}\mathbf{q}_k^H\mathbf{A}\mathbb{P}_k^{-1}\mathbf{Q}_k^{\perp}).\end{aligned}$$

Proof. With (7.25) and (7.26) we have

$$\mathbf{Q}_k^H\mathbf{A}^{-1}\mathbf{Q}_k = \mathbf{H}_k \qquad {\mathbf{Q}_k^{\perp}}^H\mathbf{A}^{-1}\mathbf{Q}_k^{\perp} = \mathbf{T}_{22} \tag{7.28}$$

$${\mathbf{Q}_k^{\perp}}^H\mathbf{A}^{-1}\mathbf{Q}_k = h_{k+1,k}\mathbf{e}_1\mathbf{e}_k^H \qquad \mathbf{Q}_k^H\mathbf{A}^{-1}\mathbf{Q}_k^{\perp} = \mathbf{T}_{12}. \tag{7.29}$$

Then we have

$$\mathbf{A}\mathbf{Q}_k = \mathbb{P}_k\mathbf{Q}_k = \mathbf{Q}_k\mathbf{H}_k^{-1} + \mathbf{A}\mathbf{q}_{k+1}h_{k+1,k}\mathbf{e}_k^H\mathbf{H}_k^{-1} \tag{7.30}$$

Multiplying by $\mathbf{Q}_k^H\mathbf{A}\mathbb{P}_k^{-1}$ from the left and by $\mathbf{H}_k$ from the right leads to

$$\mathbf{Q}_k^H\mathbf{A}\mathbf{Q}_k\mathbf{H}_k = \mathbf{Q}_k^H\mathbf{A}\mathbb{P}_k^{-1}\mathbf{Q}_k + \mathbf{Q}_k^H\mathbf{A}\mathbb{P}_k^{-1}\mathbf{A}\mathbf{q}_{k+1}h_{k+1,k}\mathbf{e}_k^H.$$

Rewriting this equality and also using the fact that from (7.30)

$$\mathbf{Q}_k^H\mathbf{A}\mathbf{Q}_k = \mathbf{H}_k^{-1} + \mathbf{Q}_k^H\mathbf{A}\mathbf{q}_{k+1}h_{k+1,k}\mathbf{e}_k^H\mathbf{H}_k^{-1},$$

we obtain

$$\mathbf{Q}_k^H\mathbf{A}\mathbb{P}_k^{-1}\mathbf{Q}_k = \mathbf{I} + \mathbf{Q}_k^H\mathbf{A}\mathbf{q}_{k+1}h_{k+1,k}\mathbf{e}_k^H - \mathbf{Q}_k^H\mathbf{A}\mathbb{P}_k^{-1}\mathbf{A}\mathbf{q}_{k+1}h_{k+1,k}\mathbf{e}_k^H,$$

and hence we obtain the left upper block of $\begin{bmatrix} \mathbf{Q}_k & \mathbf{Q}_k^\perp \end{bmatrix}^H \mathbf{A}\mathbb{P}_k^{-1} \begin{bmatrix} \mathbf{Q}_k & \mathbf{Q}_k^\perp \end{bmatrix}$. Then, multiplying (7.30) by ${\mathbf{Q}_k^\perp}^H\mathbf{A}\mathbb{P}_k^{-1}$ we obtain

$${\mathbf{Q}_k^\perp}^H\mathbf{A}\mathbf{Q}_k = {\mathbf{Q}_k^\perp}^H\mathbf{A}\mathbb{P}_k^{-1}\mathbf{Q}_k\mathbf{H}_k^{-1} + {\mathbf{Q}_k^\perp}^H\mathbf{A}\mathbb{P}_k^{-1}\mathbf{A}\mathbf{q}_{k+1}h_{k+1,k}\mathbf{e}_k^H\mathbf{H}_k^{-1},$$

and also from (7.30) ${\mathbf{Q}_k^\perp}^H\mathbf{A}\mathbf{Q}_k = {\mathbf{Q}_k^\perp}^H\mathbf{A}\mathbf{q}_{k+1}h_{k+1,k}\mathbf{e}_k^H\mathbf{H}_k^{-1}$ holds, which gives

$${\mathbf{Q}_k^\perp}^H\mathbf{A}\mathbb{P}_k^{-1}\mathbf{Q}_k = {\mathbf{Q}_k^\perp}^H\mathbf{A}\mathbf{q}_{k+1}h_{k+1,k}\mathbf{e}_k^H - {\mathbf{Q}_k^\perp}^H\mathbf{A}\mathbb{P}_k^{-1}\mathbf{A}\mathbf{q}_{k+1}h_{k+1,k}\mathbf{e}_k^H,$$

and hence the lower left block of $\begin{bmatrix} \mathbf{Q}_k & \mathbf{Q}_k^\perp \end{bmatrix}^H \mathbf{A}\mathbb{P}_k^{-1} \begin{bmatrix} \mathbf{Q}_k & \mathbf{Q}_k^\perp \end{bmatrix}$. To obtain the lower right block use (7.26) and observe that

$${\mathbf{Q}_k^\perp}^H\mathbf{A}^{-1} = h_{k+1}\mathbf{e}_k^H\mathbf{Q}_k^H + \mathbf{T}_{22}{\mathbf{Q}_k^\perp}^H.$$

Multiplying by $\mathbf{A}\mathbb{P}_k^{-1}\mathbf{Q}_k^\perp$ from the right and $\mathbf{T}_{22}^{-1}$ from the left we get

$$\mathbf{T}_{22}^{-1}({\mathbf{Q}_k^\perp}^H\mathbb{P}_k^{-1}\mathbf{Q}_k^\perp) = \mathbf{T}_{22}^{-1}h_{k+1}\mathbf{q}_k^H\mathbf{A}\mathbb{P}_k^{-1}\mathbf{Q}_k^\perp + {\mathbf{Q}_k^\perp}^H\mathbf{A}\mathbb{P}_k^{-1}\mathbf{Q}_k^\perp,$$

and hence, after reordering a formula for the lower right block in (7.27). Finally observing that $\mathbb{P}_k\mathbf{Q}_k^\perp = \mathbf{P}\mathbf{Q}_k^\perp$ and we obtain

$$\begin{aligned}
({\mathbf{Q}_k^\perp}^H\mathbb{P}_k^{-1}\mathbf{Q}_k^\perp)({\mathbf{Q}_k^\perp}^H\mathbf{P}\mathbf{Q}_k^\perp) &= ({\mathbf{Q}_k^\perp}^H\mathbb{P}_k^{-1}\mathbf{Q}_k^\perp)({\mathbf{Q}_k^\perp}^H\mathbb{P}_k\mathbf{Q}_k^\perp) \\
&= {\mathbf{Q}_k^\perp}^H\mathbb{P}_k^{-1}(\mathbf{I} - \mathbf{Q}_k\mathbf{Q}_k^H)\mathbb{P}_k\mathbf{Q}_k^\perp \\
&= \mathbf{I} + {\mathbf{Q}_k^\perp}^H\mathbb{P}_k^{-1}\mathbf{A}\mathbf{q}_{k+1}h_{k+1}\mathbf{e}_k^H\mathbf{Q}_k^H\mathbb{P}_k\mathbf{Q}_k^\perp,
\end{aligned}$$

which leads to $\mathbf{K}_{22}$ after some modifications. Hence, $\mathbf{A}\mathbb{P}_k^{-1}$ has the same eigenvalues as (7.27). □

Both Theorems 7.10 and 7.11 show that the tuned preconditioner has the advantageous property of clustering parts of the spectrum of $\mathbf{A}\mathbb{P}_k^{-1}$ around 1, and hence improving the convergence bound of GMRES given in (7.19), where $\mathbf{B} = \mathbf{A}\mathbb{P}_k^{-1}$. The next section gives an example of this behaviour.

7.4.2 Numerical examples

We state one example for the tuning strategy applied to the inexact Arnoldi method.

Example 7.12. *Consider Example 7.7 again. We carry out Arnoldi's method (part (a)) with and without the tuned preconditioner in order to approximate the eigenvalue closest to zero, which is given by $\lambda = 4.692 \cdot 10^{-2}$. We use $m = 14$ steps (outer iterations) of (in)exact Arnoldi's method.*

Figures 7-15 to 7-17 illustrate the results for Example 7.12. Figure 7-15 shows the inner iterations per outer iteration for the exact Arnoldi method with the standard and the tuned preconditioner as well as the Arnoldi method with the relaxation strategy as

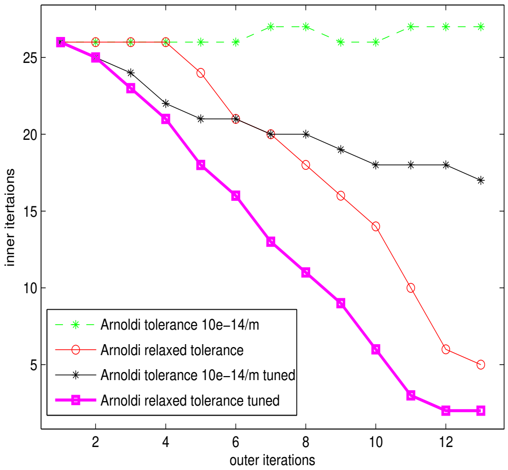


Figure 7-15: *Inner iterations against outer iterations in Example 7.12.*

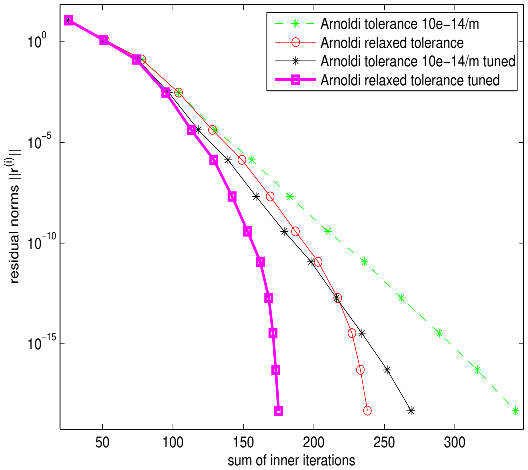


Figure 7-16: *Residual norms against sum of inner iterations in Example 7.12.*

introduced in Section 7.3 with both standard and tuned preconditioners, Figure 7-16 shows the behaviour of the eigenvalue residual during the outer iterations compared with the number of total iterations. Clearly, the combined tuning and relaxation strategy proves to be very efficient, using only about half the total number of inner iterations of the exact Arnoldi method with standard preconditioning.

Table 7.1 shows the CPU times used for all four methods and indicates the actual costs of using the tuned preconditioner. From this table we can see that the costs of applying the tuned preconditioner is slightly higher, since the modification of $\mathbf{P}$ is done with tall matrices and not just vectors. Hence an extra solve has to be performed per outer iteration, but this time with a matrix in the right hand side (instead of just a vector for tuning in inexact inverse iteration). These slightly increased costs are reflected in Table 7.1. For example, the combined tuning and relaxation strategy gives a total saving in CPU time of about 38 per cent compared to the "exact" method, whilst the saving in the total number of iterations is almost 50 per cent.

Table 7.1: *CPU times for exact and relaxed Arnoldi with standard and tuned preconditioner for Example 7.12.*

Method	"Exact" Arnoldi	"Relaxed" Arnoldi	"Tuned" Arnoldi	"Relaxed" and "tuned" Arnoldi
CPU time	6.17	4.41	5.61	3.90

Figure 7-17 shows the ratio of the maximum absolute value over the minimum absolute value of the eigenvalues of $\mathbf{AP}^{-1}$ and $\mathbf{A}\mathbb{P}_k^{-1}$ as the outer iteration progresses. This ratio is taken as a measure for the clustering of the eigenvalues of $\mathbf{AP}^{-1}$ and $\mathbf{A}\mathbb{P}_k^{-1}$ respectively. If the standard preconditioner is used this ratio is about 33 and clearly, the relaxation strategy makes no difference to this ratio. We see that the tuned preconditioner reduces this ratio and hence improves the clustering properties of $\mathbf{A}\mathbb{P}_k^{-1}$ over $\mathbf{AP}^{-1}$ (as indicated by Theorem 7.10) by about one order of magnitude (to about

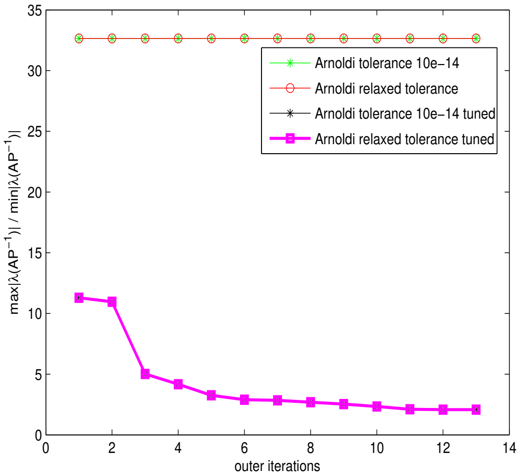


Figure 7-17: *Ratio of the maximum absolute value over the minimum absolute value of the eigenvalues of* $\mathbf{AP}^{-1}$ *and* $\mathbf{A}\mathbb{P}_k^{-1}$ *as the outer iteration proceeds for Example 7.12.*

3) and hence improves the convergence of GMRES within the inner iterations.

Table 7.2: *Ritz values of exact Arnoldi's method and inexact Arnoldi's method with the tuning strategy compared to exact eigenvalues closest to zero after* 14 *shift-invert Arnoldi steps.*

Exact eigenvalues	Ritz values (exact Arnoldi)	Ritz values (inexact Arnoldi with tuning)
+4.69249563e-02	+$\underline{4.69249563}$e-02	+$\underline{4.69249563}$e-02
+1.25445378e-01	+$\underline{1.25445378}$e-01	+$\underline{1.25445378}$e-01
+4.02658363e-01	+$\underline{4.02658}$347e-01	+$\underline{4.02658}$244e-01
+5.79574381e-01	+$\underline{5.79}$625498e-01	+$\underline{5.79}$817301e-01
+6.18836405e-01	+$\underline{6.18}$798666e-01	+$\underline{6.18}$650849e-01

In Table 7.2 the Ritz values for the inexact Arnoldi method and those obtained by the exact Arnoldi method are shown. The exact digits in both methods are underlined. One would expect that the relaxed solves ("inexact" Arnoldi) creates errors in the Ritz values. However, as one can see from Table 7.2, even the 5th smallest Ritz values is correct to 3 digits, and the difference between carrying out "exact" Arnoldi (Shift-invert Arnoldi with a small fixed solve tolerance) and "inexact" Arnoldi is marginal.

7.5 Preconditioners for implicitly restarted shift-invert Arnoldi's method

As in Arnoldi's method implicitly restarted Arnoldi method requires a linear system solve of the form $\mathbf{Ay} = \mathbf{q}_k$ for $\mathbf{y}$ at each step, which is usually done via an iterative method such as preconditioned GMRES.

7.5.1 Implicitly restarted Arnoldi's method applied to A^{-1} with a tuned preconditioner

As in Section 7.4 we introduce a tuned version of a preconditioner $\mathbf{P}$ given by $\mathbb{P}_k$ and satisfying (7.20).

For the implicitly restarted Arnoldi method (IRA) the tuned preconditioner has two advantages: Firstly, the eigenvalue clustering improves as we have shown in Theorems 7.10 and 7.11. Secondly, as we will see in Theorem 7.17, the preconditioner is an improvement of the iteration matrix for the inner solve in the sense that the right hand side $\mathbf{q}_k$ of the system will become an increasingly better approximation to an eigenvector of the the system matrix $\mathbf{A}\mathbb{P}_k^{-1}$, which improves the convergence of GMRES. This second property of a tuned preconditioner has been used in Chapters 4 and 6 for inexact inverse iteration.

First, we have the following Corollary from Theorem 7.11

Corollary 7.13. *Let the assumptions of Theorem 7.11 hold. Assume that $h_{k+1,k} = 0$, that is an invariant subspace span$\{\mathbf{Q}_k\}$ has been found and $\mathbf{A}^{-1}\mathbf{Q}_k = \mathbf{Q}_k\mathbf{H}_k$. Then*

$$\begin{bmatrix} \mathbf{Q}_k & \mathbf{Q}_k^{\perp} \end{bmatrix}^H \mathbf{A}\mathbb{P}_k^{-1} \begin{bmatrix} \mathbf{Q}_k & \mathbf{Q}_k^{\perp} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{Q}^H\mathbf{A}\mathbb{P}_k^{-1}\mathbf{Q}_k^{\perp} \\ \mathbf{0} & \mathbf{T}_{22}^{-1}({\mathbf{Q}_k^{\perp}}^H\mathbf{P}\mathbf{Q}_k^{\perp})^{-1} \end{bmatrix}. \tag{7.31}$$

Proof. Setting $h_{k+1,k} = 0$ in Theorem 7.11 gives the result. □

Corollary 7.13 shows, that if $\mathbf{P}$ is a good preconditioner to $\mathbf{A}$, then $({\mathbf{Q}_k^{\perp}}^H\mathbf{P}\mathbf{Q}_k^{\perp})^{-1}$ is a good approximation to $({\mathbf{Q}_k^{\perp}}^H\mathbf{A}^{-1}\mathbf{Q}_k^{\perp}) = \mathbf{T}_{22}$. Therefore the eigenvalues of $\mathbf{A}\mathbb{P}_k^{-1}$ should be either located at 1 or clustered around 1 and hence the clustering of the eigenvalues of $\mathbf{A}\mathbb{P}_k^{-1}$ is much improved over the clustering of the eigenvalues of $A\mathbf{P}^{-1}$ (see Figure (7-20)).

In addition to improving the eigenvalue clustering properties of $A\mathbf{P}^{-1}$, the tuned preconditioner reduces the total number of iterations by having a positive effect on the right hand side. We shall see that, as the outer iteration progresses, the right hand side becomes a good approximation to the eigenvector of the system matrix. This property is shown in the following proposition.

Proposition 7.14. *Assume we have found an invariant subspace via Arnoldi's method applied to $\mathbf{A}^{-1}$, that is*

$$\mathbf{A}^{-1}\mathbf{Q}_k = \mathbf{Q}_k\mathbf{H}_k. \tag{7.32}$$

Then the right hand side $\mathbf{q}_k$ of the linear system with the tuned preconditioner $\mathbf{A}\mathbb{P}_k^{-1}$ is an eigenvector of the system matrix $\mathbf{A}\mathbb{P}_k^{-1}$ corresponding to the eigenvalue 1.

Proof. The system to be solved at each step of Arnoldi's method is

$$\mathbf{A}\mathbb{P}_k^{-1}\tilde{\mathbf{y}} = \mathbf{q}_k, \quad \mathbf{y} = \mathbb{P}_k^{-1}\tilde{\mathbf{y}}. \tag{7.33}$$

Using (7.31) from Corollary 7.13 we have

$$\mathbf{A}\mathbb{P}_k^{-1}\mathbf{Q}_k = \mathbf{Q}_k,$$

and multiplying by $\mathbf{e}_k$ from the right gives the result. □

Proposition 7.14 shows a nice property of tuning. The right hand side $\mathbf{q}_k$ is an eigenvector of $\mathbf{A}\mathbb{P}_k^{-1}$ corresponding to eigenvalue 1. In Chapters 4 and 6 (see also [42] and [45]) a similar result has been shown for a tuned preconditioner in inexact inverse iteration.

A Krylov method like GMRES with zero starting vector applied to (7.33) requires only one iteration to converge, since the right hand side is an eigenvector of the system matrix. We give a more detailed account since Proposition 7.14 only holds for the case where an invariant subspace has been found. Within the IRA iteration condition (7.32) only holds approximately. We therefore have the following Theorem, which is a generalisation of Theorem 6.6 in Chapter 6.

Theorem 7.15. *Assume $\mathbb{P}_k^{-1}$ given by (7.22) exists. Suppose the nonsymmetric matrix $\mathbf{A}\mathbb{P}_k^{-1} \in \mathbb{C}^{n,n}$ has an invariant subspace span$\{\mathbf{W}_1^{(k)}\}$ and hence a block-diagonalisation as follows*

$$\begin{aligned}\mathbf{A}\mathbb{P}_k^{-1} &= \begin{bmatrix} \mathbf{W}_1^{(k)} & \mathbf{W}_2^{(k)} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{22}^{(k)} \end{bmatrix} \begin{bmatrix} \mathbf{W}_1^{(k)} & \mathbf{W}_2^{(k)} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} \mathbf{W}_1^{(k)} & \mathbf{W}_2^{(k)} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{22}^{(k)} \end{bmatrix} \begin{bmatrix} {\mathbf{V}_1^{(k)}}^H \\ {\mathbf{V}_2^{(k)}}^H \end{bmatrix},\end{aligned}$$

where span$\{\mathbf{V}_1^{(k)}\}$ is the left invariant subspace of $\mathbf{A}\mathbb{P}_k^{-1} \in \mathbb{C}^{n,n}$. Let $\mathcal{P}_k = \mathbf{I} - \mathbf{W}_1^{(k)}{\mathbf{V}_1^{(k)}}^H$ be the oblique projector onto $\mathcal{R}(\mathbf{W}_2^{(k)})$. Let $\mathbf{y}_j^{(k)}$ be the result of applying GMRES to $\mathbf{A}\mathbb{P}_k^{-1}\mathbf{y}^{(k)} = \mathbf{q}_k$ with starting value $\mathbf{y}_0^{(k)} = \mathbf{0}$. Then

$$\|\mathbf{q}_k - \mathbf{A}\mathbb{P}_k^{-1}\mathbf{y}_j^{(k)}\| \leq \min_{p_{j-1}\in\Pi_{j-1}} \|p_{j-1}(\mathbf{K}_{22}^{(k)})\|\|\mathbf{I} - \mathbf{K}_{22}^{(k)}\|\|\mathbf{V}_2^{(k)}\|\|\mathcal{P}_k\mathbf{q}_k\|. \tag{7.34}$$

Proof. Following the proof of Theorem 6.6 in Chapter 6 we have

$$\|\mathbf{q}_k - \mathbf{A}\mathbb{P}_k^{-1}\mathbf{y}_j^{(k)}\| = \min_{p_j\in\Pi_j} \|p_j(\mathbf{A}\mathbb{P}_k^{-1})\mathbf{q}_k\|,$$

for the GMRES residual (see [55]), where Π_j is the set of polynomials of degree j with $p(0) = 1$. Introduce special polynomials $\hat{p}_j \in \Pi_j$, given by

$$\hat{p}_j(z) = p_{j-1}(z)(1 - z),$$

where $p_{j-1} \in \Pi_{j-1}$. Then we can write

$$\begin{aligned}\|\mathbf{q}_k - \mathbf{A}\mathbb{P}_k^{-1}\mathbf{y}_j^{(k)}\| &= \min_{p_j\in\Pi_j} \|p_j(\mathbf{A}\mathbb{P}_k^{-1})\mathcal{P}_k\mathbf{q}_k + p_j(\mathbf{A}\mathbb{P}_k^{-1})(\mathbf{I} - \mathcal{P}_k)\mathbf{q}_k\| \\ &\leq \min_{\hat{p}_j\in\Pi_j} \|\hat{p}_j(\mathbf{A}\mathbb{P}_k^{-1})\mathcal{P}_k\mathbf{q}_k + \hat{p}_j(\mathbf{A}\mathbb{P}_k^{-1})(\mathbf{I} - \mathcal{P}_k)\mathbf{q}_k\| \\ &\leq \min_{p_{j-1}\in\Pi_{j-1}} \|p_{j-1}(\mathbf{A}\mathbb{P}_k^{-1})(\mathbf{I} - \mathbf{A}\mathbb{P}_k^{-1})\mathcal{P}_k\mathbf{q}_k \\ &\qquad\qquad + p_{j-1}(\mathbf{A}\mathbb{P}_k^{-1})(\mathbf{I} - \mathbf{A}\mathbb{P}_k^{-1})(\mathbf{I} - \mathcal{P}_k)\mathbf{q}_k\|.\end{aligned}$$

For the second term we have

$$(\mathbf{I} - \mathbf{A}\mathbb{P}_k^{-1})(\mathbf{I} - \mathcal{P}_k)\mathbf{q}_k = (\mathbf{I} - \mathbf{A}\mathbb{P}_k^{-1})\mathbf{W}_1^{(k)}{\mathbf{V}_1^{(k)}}^H\mathbf{q}_k = \mathbf{0},$$

using $\mathbf{A}\mathbb{P}_k^{-1}\mathbf{W}_1^{(k)} = \mathbf{W}_1^{(k)}$. Therefore

$$\|\mathbf{q}_k - \mathbf{A}\mathbb{P}_k^{-1}\mathbf{y}_j^{(k)}\| \leq \min_{p_{j-1}\in\Pi_{j-1}} \|p_{j-1}(\mathbf{A}\mathbb{P}_k^{-1})(\mathbf{I} - \mathbf{A}\mathbb{P}_k^{-1})\mathcal{P}_k\mathbf{q}_k\|. \tag{7.35}$$

With $\mathcal{P}_k^2 = \mathcal{P}_k$ and $\mathcal{P}_k\mathbf{A}\mathbb{P}_k^{-1} = \mathbf{A}\mathbb{P}_k^{-1}\mathcal{P}_k$ we have

$$\begin{aligned} p_{j-1}(\mathbf{A}\mathbb{P}_k^{-1})(\mathbf{I} - \mathbf{A}\mathbb{P}_k^{-1})\mathcal{P}_k\mathbf{q}_k &= p_{j-1}(\mathbf{W}_2^{(k)}\mathbf{K}_{22}^{(k)}\mathbf{V}_2^{(k)H})(\mathbf{I} - \mathbf{A}\mathbb{P}_k^{-1})\mathcal{P}_k\mathbf{q}_k \\ &= \mathbf{W}_2^{(k)}p_{j-1}(\mathbf{K}_{22}^{(k)})(\mathbf{I} - \mathbf{K}_{22}^{(k)})\mathbf{V}_2^{(k)H}\mathcal{P}_k\mathbf{q}_k, \end{aligned}$$

and hence

$$\|\mathbf{q}_k - \mathbf{A}\mathbb{P}_k^{-1}\mathbf{y}_j^{(k)}\| \leq \min_{p_{j-1}\in\Pi_{j-1}} \|p_{j-1}(\mathbf{K}_{22}^{(k)})\|\|\mathbf{I} - \mathbf{K}_{22}^{(k)}\|\|\mathbf{V}_2^{(k)}\|\|\mathcal{P}_k\mathbf{q}_k\|, \tag{7.36}$$

since $\mathbf{W}_2^{(k)}$ can be chosen to have orthonormal columns, see Theorem 6.6 in Chapter 6. □

We are particularly interested in the term $\|\mathcal{P}_k\mathbf{q}_k\|$. Note that if $\mathbf{q}_k$ is an exact eigenvector of $\mathbf{A}\mathbb{P}_k^{-1}$ (see also Proposition 7.14), then $\|\mathcal{P}_k\mathbf{q}_k\| = 0$ and convergence is immediate. Hence, we investigate how close $\mathbf{q}_k$ is to an exact eigenvector of $\mathbf{A}\mathbb{P}_k^{-1}$.

The following theorem states that under the condition that $|h_{k+1,k}|$ is small enough then $\mathbf{q}_k$ is an approximate eigenvector of $\mathbf{A}\mathbb{P}_k^{-1}$, which is an extension of the exact case in Proposition 7.14.

Theorem 7.16. *Let $\mathbb{P}_k$ be given by (7.21) and assume that $\mathbf{Q}_k^H\mathbf{P}^{-1}\mathbf{A}\mathbf{Q}_k$ is nonsingular so that $\mathbb{P}_k^{-1}$ given by (7.22) exists. Further assume $\mathbf{Q}_k^H\mathbf{P}^{-1}\mathbf{Q}_k$ is nonsingular and $|h_{k+1,k}|$ is small enough. At each outer step of shift-invert Arnoldi method we have*

$$\|\mathbf{A}\mathbb{P}_k^{-1}\mathbf{q}_k - \mathbf{q}_k\| = C_0|h_{k+1,k}|, \tag{7.37}$$

where C_0 depends on the norm of $\mathbf{A}$ and $\mathbf{P}^{-1}$.

Proof. Using the definition of $\mathbb{P}_k^{-1}$ we have

$$\mathbf{A}\mathbb{P}_k^{-1}\mathbf{q}_k = \mathbf{A}\mathbf{P}^{-1}\mathbf{q}_k - \mathbf{A}(\mathbf{P}^{-1}\mathbf{A} - \mathbf{I})\mathbf{Q}_k(\mathbf{Q}_k^H\mathbf{P}^{-1}\mathbf{A}\mathbf{Q}_k)^{-1}\mathbf{Q}_k^H\mathbf{P}^{-1}\mathbf{Q}_k\mathbf{e}_k \tag{7.38}$$

and with the Arnoldi relation (7.30) we get

$$\begin{aligned} \mathbf{Q}_k^H\mathbf{P}^{-1}\mathbf{A}\mathbf{Q}_k &= \mathbf{Q}_k^H\mathbf{P}^{-1}\mathbf{Q}_k\mathbf{H}_k^{-1} + \mathbf{Q}_k^H\mathbf{P}^{-1}\mathbf{A}\mathbf{q}_{k+1}h_{k+1,k}\mathbf{e}_k^H\mathbf{H}_k^{-1} \\ &= \mathbf{Q}_k^H\mathbf{P}^{-1}\mathbf{Q}_k(\mathbf{I} + (\mathbf{Q}_k^H\mathbf{P}^{-1}\mathbf{Q}_k)^{-1}\mathbf{Q}_k^H\mathbf{P}^{-1}\mathbf{A}\mathbf{q}_{k+1}h_{k+1,k}\mathbf{e}_k^H)\mathbf{H}_k^{-1} \\ &= \mathbf{Q}_k^H\mathbf{P}^{-1}\mathbf{Q}_k(\mathbf{I} + \mathbf{E}_k)\mathbf{H}_k^{-1}, \end{aligned}$$

where $\mathbf{E}_k^{(1)} := (\mathbf{Q}_k^H\mathbf{P}^{-1}\mathbf{Q}_k)^{-1}\mathbf{Q}_k^H\mathbf{P}^{-1}\mathbf{A}\mathbf{q}_{k+1}h_{k+1,k}\mathbf{e}_k^H$. Hence from (7.38) we obtain

$$\begin{aligned} \mathbf{A}\mathbb{P}_k^{-1}\mathbf{q}_k &= \mathbf{A}\mathbf{P}^{-1}\mathbf{q}_k - \mathbf{A}(\mathbf{P}^{-1}\mathbf{A} - \mathbf{I})\mathbf{Q}_k(\mathbf{Q}_k^H\mathbf{P}^{-1}\mathbf{Q}_k(\mathbf{I} + \mathbf{E}_k^{(1)})\mathbf{H}_k^{-1})^{-1}\mathbf{Q}_k^H\mathbf{P}^{-1}\mathbf{Q}_k\mathbf{e}_k \\ &= \mathbf{A}\mathbf{P}^{-1}\mathbf{q}_k - (\mathbf{A}\mathbf{P}^{-1} - \mathbf{I})\mathbf{A}\mathbf{Q}_k\mathbf{H}_k(\mathbf{I} + \mathbf{E}_k^{(1)})^{-1}\mathbf{e}_k \\ &= \mathbf{A}\mathbf{P}^{-1}\mathbf{q}_k - (\mathbf{A}\mathbf{P}^{-1} - \mathbf{I})(\mathbf{Q}_k + \mathbf{A}\mathbf{q}_{k+1}h_{k+1,k}\mathbf{e}_k^H)(\mathbf{I} + \mathbf{E}_k^{(1)})^{-1}\mathbf{e}_k \\ &= \mathbf{A}\mathbf{P}^{-1}\mathbf{q}_k - (\mathbf{A}\mathbf{P}^{-1} - \mathbf{I})(\mathbf{Q}_k + \mathbf{E}_k^{(2)})(\mathbf{I} + \mathbf{E}_k^{(1)})^{-1}\mathbf{e}_k \end{aligned}$$

where $\mathbf{E}_k^{(2)} := \mathbf{A}\mathbf{q}_{k+1}h_{k+1,k}\mathbf{e}_k^H$. Assuming that $\|\mathbf{E}_k^{(1)}\| < 1$ (that is, the $\|\mathbf{P}^{-1}\|$ is moderate and $|h_{k+1,k}|$ is small enough), using Neumann series (see [48]) we have

$$(\mathbf{I} + \mathbf{E}_k^{(1)})^{-1} = \mathbf{I} + \sum_{i=1}^{\infty}(-\mathbf{E}_k^{(1)})^i,$$

and hence

$$\mathbf{A}\mathbb{P}_k^{-1}\mathbf{q}_k = \mathbf{q}_k - (\mathbf{A}\mathbf{P}^{-1} - \mathbf{I})\left(\mathbf{E}_k^{(2)}\sum_{i=0}^{\infty}(-\mathbf{E}_k^{(1)})^i\mathbf{e}_k + \mathbf{Q}_k\sum_{i=1}^{\infty}(-\mathbf{E}_k^{(1)})^i\mathbf{e}_k\right).$$

With the definitions of the Neumann series and using $\mathbf{Q}_k$ unitary we find that

$$\|\mathbf{A}\mathbb{P}_k^{-1}\mathbf{q}_k - \mathbf{q}_k\| \leq \|\mathbf{A}\mathbf{P}^{-1} - \mathbf{I}\|\left(\|\mathbf{E}_k^{(2)}\|\frac{1}{1 - \|\mathbf{E}_k^{(1)}\|} + \frac{1}{1 - \|\mathbf{E}_k^{(1)}\|} - 1\right),$$

and hence, using the definitions of $\mathbf{E}_k^{(1)}$ and $\mathbf{E}_k^{(2)}$ we obtain

$$\|\mathbf{A}\mathbb{P}_k^{-1}\mathbf{q}_k - \mathbf{q}_k\| \leq \|\mathbf{A}\mathbf{P}^{-1} - \mathbf{I}\|\frac{\|\mathbf{A}\||h_{k+1,k}| + \|(\mathbf{Q}_k^H\mathbf{P}^{-1}\mathbf{Q}_k)^{-1}\|\|\mathbf{P}^{-1}\mathbf{A}\||h_{k+1,k}|}{1 - \|(\mathbf{Q}_k^H\mathbf{P}^{-1}\mathbf{Q}_k)^{-1}\|\|\mathbf{P}^{-1}\mathbf{A}\||h_{k+1,k}|},$$

and Taylor series expansion around $|h_{k+1,k}|$ gives

$$\|\mathbf{A}\mathbb{P}_k^{-1}\mathbf{q}_k - \mathbf{q}_k\| \leq \|\mathbf{A}\mathbf{P}^{-1} - \mathbf{I}\|(\|\mathbf{A}\| + \|(\mathbf{Q}_k^H\mathbf{P}^{-1}\mathbf{Q}_k)^{-1}\|\|\mathbf{P}^{-1}\mathbf{A}\|)|h_{k+1,k}|,$$

for small enough $|h_{k+1,k}|$ and hence a constant C_0 independent of k, since $\mathbf{Q}_k$ is unitary. $\square$

We know from (7.3) that for convergence in IRA with exact shifts we have that

$$\|\hat{\mathbf{f}}_k^{(i)}\| = |h_{k+1,k}^{(i)}| \to 0,$$

where i denotes the number of restarts. Hence, we expect that span$\{\mathbf{Q}_k\}$ becomes closer to an invariant subspace as the iteration proceeds and in the limit the right hand side of the linear system given by $\mathbf{q}_k$ is an eigenvector of $\mathbf{A}\mathbb{P}_k^{-1}$. Hence we have the following theorem for $\|\mathcal{P}_k\mathbf{q}_k\|$.

Theorem 7.17. *Let the assumptions of Theorem 7.15 hold and let $\mathcal{P}_k$ be given as in Theorem 7.15. Assume shift-inverted IRA converges, that is $|h_{k+1,k}| \to 0$. Then $\|\mathcal{P}_k\mathbf{q}_k\| \to 0$.*

Proof. By Theorem (7.16) we have that $\mathbf{q}_k$ is an approximate eigenvalue of $\mathbf{A}\mathbb{P}_k^{-1}$, and hence, with $\mathbf{w}_k$ being the exact eigenvector we have that

$$\|\mathbf{q}_k - \mathbf{w}_k\| \leq C_1|h_{k+1,k}|$$

for some constant C_1. With $\mathcal{P}_k\mathbf{w}_k = \mathbf{0}$ we get

$$\|\mathcal{P}_k\mathbf{q}_k\| = \|\mathcal{P}_k(\mathbf{q}_k - \mathbf{w}_k)\| \leq C_1\|\mathcal{P}_k\||h_{k+1,k}|.$$

With $|h_{k+1,k}| \to 0$ and $\|\mathcal{P}_k\| < C_2$ for small enough $h_{k+1,k}$ (see [104]) we obtain the result. $\square$

Hence, with Theorem 7.17 and the GMRES bound (7.34), we expect the number of inner iterations per outer iteration to decrease. This is indeed observed in the examples in the following section.

7.5.2 Numerical examples

We present three numerical examples and use the the same matrices as Section 7.3.

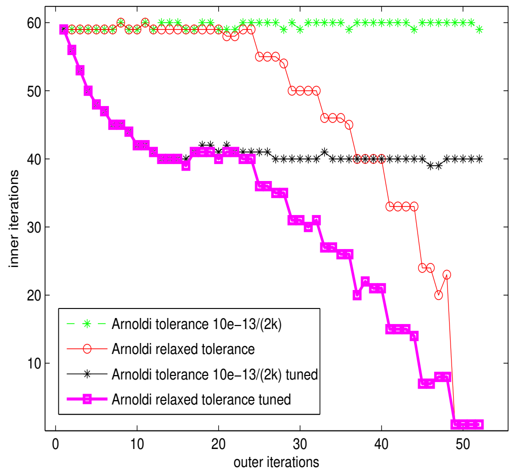


Figure 7-18: *Inner iterations per outer iteration in Example 7.18.*

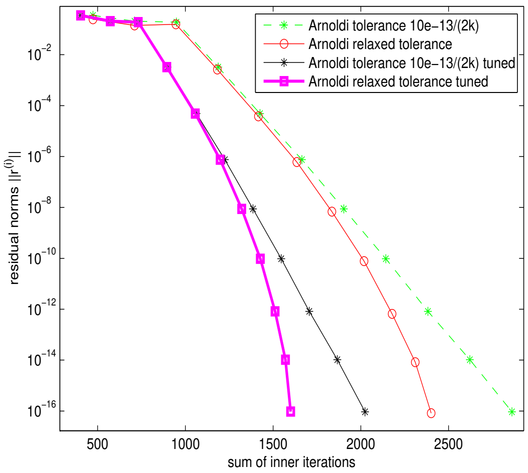


Figure 7-19: *Residual norms against sum of inner iterations in Example 7.18.*

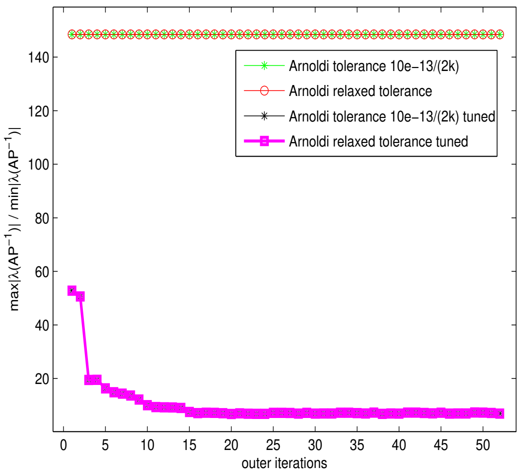


Figure 7-20: *Ratio of the maximum over the minimum absolute value of the eigenvalues of* $\mathbf{A}\mathbf{P}^{-1}$ *and* $\mathbf{A}\mathbb{P}_k^{-1}$ *vs outer iterations for Example 7.18.*

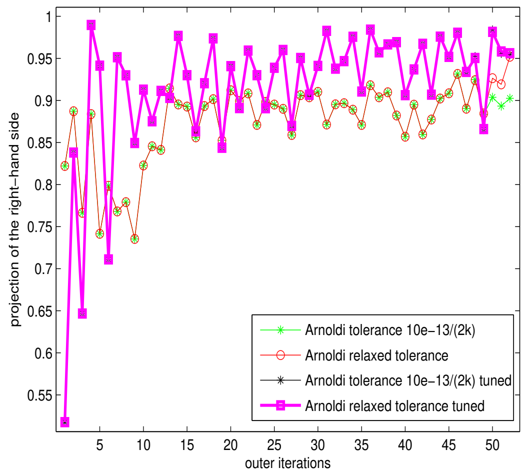


Figure 7-21: *Cosine of the angle between the right hand side vector* $\mathbf{q}_k$ *and the vector* $\mathbf{A}\mathbb{P}_k^{-1}\mathbf{q}_k$ *as the outer iteration proceeds for Example 7.18.*

Example 7.18. *Consider the same matrix and setup as in Example 7.7. However this time we use an incomplete LU factorisation with a larger drop tolerance* 0.008 *(instead of* 0.001*) as standard preconditioner. We apply implicitly restarted Arnoldi's method with exact shifts for finding the* $k = 8$ *eigenvalues closest to zero. We use a subspace of total size* $m = k + p = 12$ *with* $i_{\max} = 10$ *restarts (outer iterations). We apply the same relaxation strategies as in Example 7.7, part (b)and compare 4 methods:*

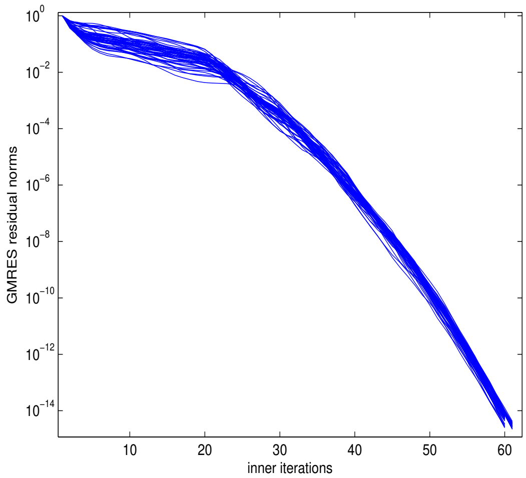


Figure 7-22: *Relative GMRES residual norms for Example 7.18 with standard preconditioner.*

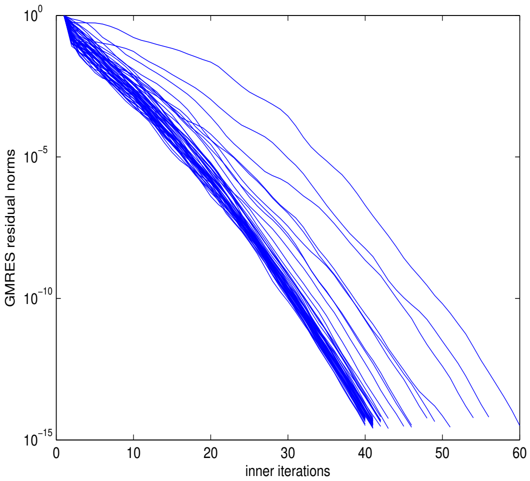


Figure 7-23: *Relative GMRES residual norms for Example 7.18 with tuned preconditioner.*

(a) exact IRA method with standard preconditioner $\mathbf{P}$*,*

(b) IRA method with relaxation strategy as done in Example 7.7 with standard preconditioner $\mathbf{P}$*,*

(c) exact IRA method with tuned preconditioner $\mathbb{P}_k$ *at each outer step* k*,*

(d) IRA method with relaxation strategy as done in Example 7.7 and *with tuned preconditioner* $\mathbb{P}_k$ *at each outer step* k*.*

The Results for Example 7.18 are presented in Figures 7-18 to 7-23. From Figure 7-18 we can see how the tuned preconditioner reduces the number of inner iterations per outer iteration both for the case when exact solves are used (method (c) compared with method (a)) and when inexact solves are used with a relaxation strategy (method (d) compared with method (b)). In Figure 7-19 we see that the tuning strategy gives an improvement of about 30 per cent (in terms of the number of inner iterations) both for exact solves (compare the starred line with the dashed line) and the relaxation strategy (compare the squared line with the circled line). If the relaxation strategy is combined with a tuned preconditioner in the inner solves, the reduction in total iterations is about 46 per cent.

Table 7.3: *CPU times for exact and relaxed Arnoldi with standard and tuned preconditioner for Example 7.18.*

Method	"Exact" Arnoldi (a)	"Relaxed" Arnoldi (b)	"Tuned" Arnoldi (c)	"Relaxed" and "tuned" Arnoldi (d)
CPU time	56.83	44.02	42.41	33.25

As noted before in Example 7.12, the actual cost of the tuned preconditioner is slightly higher, due to extra solves with $\mathbf{P}$. The actual savings in the costs, given by the CPU times for all four methods, is presented in Table 7.3. From there, comparing the second with the last column, we see that the saving of using the relaxation strategy together with a tuned preconditioner in the inner solves is about 42 per cent.

In Figure 7-20 the ratio of the maximum absolute value over the minimum absolute value of the eigenvalues of $\mathbf{AP}^{-1}$ and $\mathbf{A}\mathbb{P}_k^{-1}$ as the outer iteration proceeds is shown. This ratio is taken as a measure for the clustering of the eigenvalues of $\mathbf{AP}^{-1}$ and $\mathbf{A}\mathbb{P}_k^{-1}$ respectively. For $\mathbf{A}\mathbb{P}_k^{-1}$ this ratio is about 10 per cent of the ratio for $\mathbf{AP}^{-1}$, indicating improvements in the eigenvector clustering. This supports the result in Theorem 7.11. Figure 7-21 the cosine of the angle between $\mathbf{q}_k$ and $\mathbf{APq}_k$ ($\mathbf{A}\mathbb{P}_k^{-1}\mathbf{q}_k$ respectively) is shown for each outer iteration step k. For the tuned preconditioner the angle between $\mathbf{q}_k$ and $\mathbf{A}\mathbb{P}_k^{-1}\mathbf{q}_k$ is smaller as indicated by Theorem 7.17. The plots in Figures 7-22 and 7-23 show how the relative GMRES residual norms behave as the outer iteration proceeds. The progress of the outer iteration is read from the upper right corner to the lower left corner. Observe that for the tuned preconditioner (Figure 7-23) the GMRES residual is decreasing faster as the outer iteration proceeds, whilst for the standard preconditioner (Figure 7-22) the reduction rate remains approximately constant.

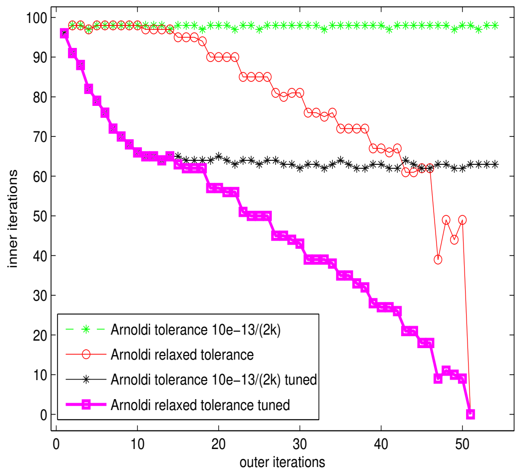


Figure 7-24: *Inner iterations per outer iteration in Example 7.19.*

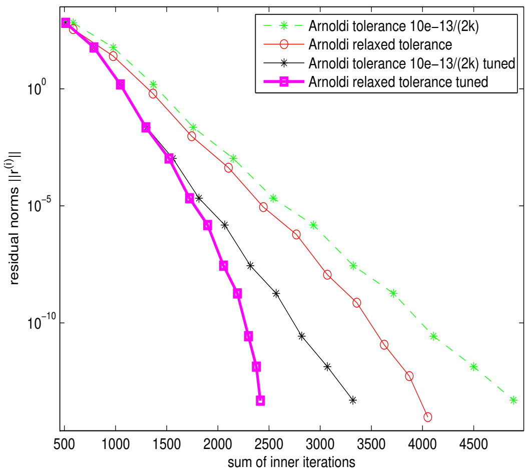


Figure 7-25: *Residual norms against sum of inner iterations in Example 7.19.*

Example 7.19. *Use the same setup and the same matrix as in Example 7.8, where the* 6 *smallest eigenvalues of matrix* `qc2534.mtx` *were to be found via inexact IRA and a maximum size of the subspace of* 10. *We apply the same relaxation strategies and compare the same for methods as in the example above.*

The Results for Example 7.19 are presented in Figures 7-24 and 7-25. Again we see that the use of the tuned preconditioner instead of the standard preconditioner in each inner iteration reduces the inner iteration count significantly. In this example, the combination of the relaxation strategy and tuning reduces the total iteration number by about 52 percent, as it can be observed in 7-25. In order to obtain the same residual norm of about 10^{-14} of the invariant subspace only about 2400 versus about 5000 inner iterations are needed, if the relaxation strategy together with the tuned preconditioner is applied in each iteration step.

Table 7.4: *CPU times for exact and relaxed Arnoldi with standard and tuned preconditioner for Example 7.19.*

Method	"Exact" Arnoldi (a)	"Relaxed" Arnoldi (b)	"Tuned" Arnoldi (c)	"Relaxed" and "tuned" Arnoldi (d)
CPU time	233.27	164.82	147.41	98.02

The CPU times for all four methods in Example 7.19 are given in Table 7.4. We observe that the CPU time is reduced by over 50 per cent in this example if tuning the preconditioner and the relaxation strategy are combined.

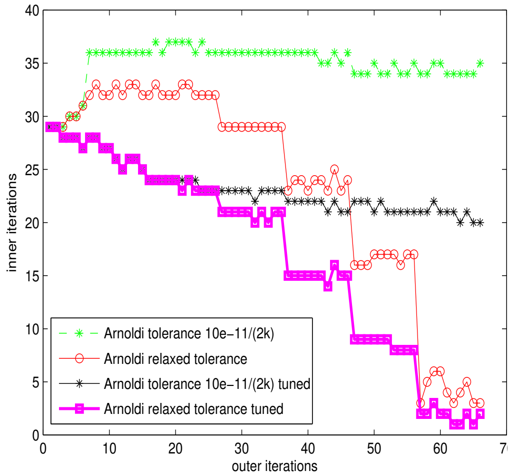


Figure 7-26: *Inner iterations per outer iteration in Example 7.20.*

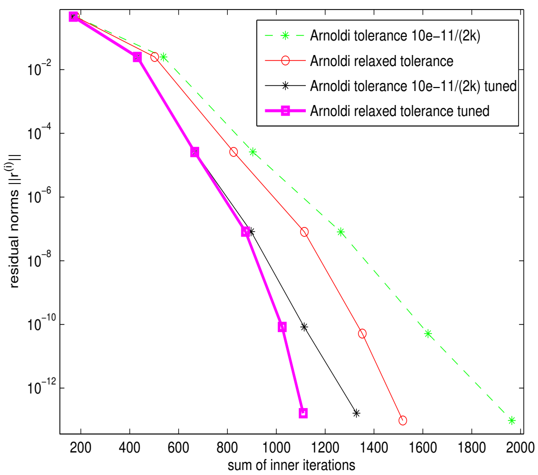


Figure 7-27: *Residual norms against sum of inner iterations in Example 7.20.*

Example 7.20. *Consider Example 7.9 again, where the $k = 6$ eigenvalues closest to zero of a generalised eigenproblem from the* IFISS *package [32] are sought using exact and inexact IRA and a maximum size of the subspace of* 16*. We apply the same relaxation strategies and compare the same for methods as in the the two examples above.*

Figures 7-26 and 7-27 illustrate the results for Example 7.20. The total reduction of inner iterations in this example is about 45 per cent (if relaxation and tuning are combined) as it can be observed from the plots in 7-27. The reduction in CPU time for this example is of similar size, see Table 7.5.

Table 7.5: *CPU times for exact and relaxed Arnoldi with standard and tuned preconditioner for Example 7.20.*

Method	"Exact" Arnoldi (a)	"Relaxed" Arnoldi (b)	"Tuned" Arnoldi (c)	"Relaxed" and "tuned" Arnoldi (d)
CPU time	136.03	89.82	93.41	70.96

7.6 Conclusions

In this chapter, we extended the relaxation strategy for shift-invert Arnoldi method to implicitly restarted Arnoldi method. In addition, we extended the tuning strategy developed for preconditioned inexact inverse iteration (see Chapters 4 and 6) to shift-invert Arnoldi method with inexact inner solves. We showed that the tuned preconditioner, a modified version of the standard preconditioner improves the behaviour of the inner iterative solve. We also point out that the tuned preconditioner in combination with the relaxation strategy for the inexact inner solves within shift-invert Arnoldi's method (and shift-invert IRA) significantly improves the costs of the methods by up to 50 per cent.

CHAPTER 8

Conclusions and further work

This thesis is concerned with the important topic of iterative methods for large, sparse eigenvalue problems. In particular it deals with inner-outer iterative methods, where the outer iteration is a subspace method for eigencomputations and the inner part is an inexact solve of the linear system arising within each step of the outer iteration.

We have contributed both to

1. the development of the convergence theory of the outer iterative methods when the inner shifted linear system is solved to a prescribed tolerance only and
2. the efficiency of the inner iterative solves of the linear systems through the construction of new and improved preconditioners for the inner solvers.

A list of future work includes:

- a comparison of the tuned preconditioner in shift-and-invert Arnoldi's method to the full Jacobi-Davidson method with subspace expansion,
- an analysis of shift-and-invert Arnoldi's method for the generalised eigenproblem with the tuned preconditioner,
- an analysis of the rational Krylov method with tuned preconditioner and comparison to preconditioned full Jacobi-Davidson method,
- the extension of the idea of the tuned preconditioner to block Arnoldi and block Lanczos methods,
- an analysis of deflation strategies in combination with the tuned preconditioner,
- the exploitation of stopping criteria for inexact inverse iteration and subspace iteration with fixed and increasing dimension,
- an analysis of further applications of preconditioners for linear systems with low rank modifications.

APPENDIX A

A list of basic iterative methods

A.1 A list of basic iterative methods for eigenvalue problems and numerical examples

We are going to discuss some basic iterative algorithms here, which we will use in the following chapters. We will give some examples for vector iterations (subspace methods with $\dim(\mathcal{S}_i) = 1\ \forall i$), subspace iteration (subspace methods with fixed dimension $\dim(\mathcal{S}_i) = p\ \forall i$) and subspace algorithms with increasing dimension ($\dim(\mathcal{S}_i) = i\ \forall i$ and $\mathcal{S}_{i-1} \subset \mathcal{S}_i$). Note that as an introduction we consider the exact algorithms, in the following chapters various aspects and new developments of the inexact algorithms are studied. Note that the algorithms are written out in pseudocode.

A.1.1 Single vector iterations

Single vector iterations and their convergence theory are discussed very detailed in Parlett [101]. The power method is the basic iterative method which takes a starting vector $\mathbf{x}^{(0)}$ and lets the matrix $\mathbf{A}$ operate on it until we get a vector close to the largest eigenvector. Suppose that $\mathbf{A}$ is nondefective, so it has n eigenvectors that span $\mathbb{C}^n$,

Algorithm 9 Power Method

Input: $\mathbf{A}$, $\mathbf{x}^{(0)}$ ($\|\mathbf{x}^{(0)}\| = 1$), i_{max}.
  for $i = 1, \ldots, i_{max}$ **do**
    $\mathbf{y}^{(i)} = \mathbf{A}\mathbf{x}^{(i-1)}$
    $\mathbf{x}^{(i)} = \dfrac{\mathbf{y}^{(i)}}{\|\mathbf{y}^{(i)}\|}$
    $\theta^{(i)} = \mathbf{x}^{(i)H}\mathbf{A}\mathbf{x}^{(i)}$
  end for
Output: $\theta^{(i_{max})}$, $\mathbf{x}^{(i_{max})}$.

they satisfy $\mathbf{A}\mathbf{v}_j = \lambda_j \mathbf{v}_j$ and the starting vector $\mathbf{x}^{(0)}$ can be decomposed as

$$\mathbf{x}^{(0)} = \sum_{j=1}^{n} \alpha_j \mathbf{v}_j \quad \text{for some} \quad \alpha_j.$$

Hence, if $\mathbf{A}$ is i times applied to the vector $\mathbf{x}^{(0)}$ and let's say λ_1 is the eigenvalue with largest absolute value, that is

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|,$$

we get

$$\mathbf{A}^i \mathbf{x}^{(0)} = \sum_{j=1}^{n} \lambda_j^i \alpha_j \mathbf{v}_j = \lambda_1^i \sum_{j=1}^{n} \left(\frac{\lambda_j}{\lambda_1}\right)^i \alpha_j \mathbf{v}_j$$

and $\mathbf{v}_1$ will dominate the iteration completely:

$$\lim_{i\to\infty} \lambda_1^{-i} \mathbf{A}^i \mathbf{x}^{(0)} = \alpha_1 \mathbf{v}_1 + \lim_{i\to\infty} \sum_{j=2}^{n} \left(\frac{\lambda_j}{\lambda_1}\right)^i \alpha_j \mathbf{v}_j = \alpha_1 \mathbf{v}_1,$$

and hence the sequence $\lambda_1^{-i} \mathbf{A}^i \mathbf{x}^{(0)}$ convergences to an eigenvector, if $\alpha_1 \neq 0$. The convergence factor is $\mathcal{O}\left(\left|\frac{\lambda_1}{\lambda_2}\right|^i\right)$, where λ_2 is the second largest eigenvalue. Thus the power method approximates the eigenvector corresponding to the largest eigenvalue with a linear convergence rate. The corresponding approximate eigenvalue can be recovered via the Rayleigh-Ritz procedure, which we will discuss in the next Section A.1.2, and which in this special one-dimensional case is just the Rayleigh quotient $\rho(\mathbf{x}^{(i)}) = {\mathbf{x}^{(i)}}^H \mathbf{A} \mathbf{x}^{(i)}$.

The vector sequence produced by the power method converges to the eigenvector belonging to the largest eigenvalue. If we want to find the eigenvalue close to some $\sigma \in \mathbb{C}$ then we can apply the power method to $(\mathbf{A} - \sigma \mathbf{I})^{-1}$ which has the eigenvalues $(\lambda_j - \sigma)^{-1}$ for $j = 1, \ldots, n$. Then the power method is called inverse iteration [151].. If

Algorithm 10 Inverse Iteration

Input: $\mathbf{A}$, $\mathbf{x}^{(0)}$ $(\|\mathbf{x}^{(0)}\| = 1)$,$\sigma$, i_{max}.
  for $i = 1, \ldots, i_{max}$ **do**
    $\mathbf{y}^{(i)} = (\mathbf{A} - \sigma \mathbf{I})^{-1} \mathbf{x}^{(i-1)}$
    $\mathbf{x}^{(i)} = \dfrac{\mathbf{y}^{(i)}}{\|\mathbf{y}^{(i)}\|}$
    $\theta^{(i)} = {\mathbf{x}^{(i)}}^H \mathbf{A} \mathbf{x}^{(i)}$
  end for
Output: $\theta^{(i_{max})}$, $\mathbf{x}^{(i_{max})}$.

$|\lambda_s - \sigma| < |\lambda_t - \sigma| \leq |\lambda_j - \sigma|$, $\forall j \neq s, t$, then $\lambda_s - \sigma$ is the largest eigenvalue of $(\mathbf{A} - \sigma \mathbf{I})^{-1}$ and the iteration vectors of inverse iteration converge to $\mathbf{v}_s$ with a linear convergence rate $\mathcal{O}\left(\left|\frac{\lambda_s - \sigma}{\lambda_t - \sigma}\right|^i\right)$. The closer the shift σ is to λ_s the faster is the convergence. Again the eigenvalues are found with the Rayleigh quotient. We remark that it is necessary to compute the inverse of the matrix for this method. As we will note later, this might be too expensive for large matrices both in storage requirements and computation time and is therefore done inexactly.

If in inverse iteration a variable shift is used instead of a fixed one, and, if this shift is chosen to be the Rayleigh quotient $\theta^{(i)} = {\mathbf{x}^{(i)}}^H \mathbf{A} \mathbf{x}^{(i)}$, then inverse iteration is called

Algorithm 11 Rayleigh Quotient Iteration

Input: $\mathbf{A}$, $\mathbf{x}^{(0)}$ ($\|\mathbf{x}^{(0)}\| = 1$), i_{max}.
  $\theta^{(0)} = \mathbf{x}^{(0)H}\mathbf{A}\mathbf{x}^{(0)}$
  for $i = 1, \ldots, i_{max}$ **do**
    $\mathbf{y}^{(i)} = (\mathbf{A} - \theta^{(i-1)}\mathbf{I})^{-1}\mathbf{x}^{(i-1)}$
    $\mathbf{x}^{(i)} = \dfrac{\mathbf{y}^{(i)}}{\|\mathbf{y}^{(i)}\|}$
    $\theta^{(i)} = \mathbf{x}^{(i)H}\mathbf{A}\mathbf{x}^{(i)}$
  end for
Output: $\theta^{(i_{max})}$, $\mathbf{x}^{(i_{max})}$.

Rayleigh quotient iteration. This accelerates the rate of convergence, because after a few steps the shift $\theta^{(i)}$ will be very close to the sought eigenvalue. Indeed, the rate of convergence of this algorithm is at least quadratic, if $\mathbf{A} = \mathbf{A}^H$ then the convergence is even cubic, for details we refer to [101]. A disadvantage of this algorithm is that the approximate eigenvalue is not necessarily the closest to $\theta^{(0)}$. Therefore usually a combined inverse iteration with fixed and variable shift is used [139].

We want to illustrate the various algorithms and their convergence behaviour in the exact case with a few numerical tests. We chose simple test matrices, since the only purpose is to illustrate some convergence results of the described algorithms. The computations are carried out in MATLAB.

We applied the algorithms to symmetric matrices $\mathbf{A}_{sym}$ and nonsymmetric matrices $\mathbf{A}_{unsym}$ of size 100×100. Their eigenvalues are $\lambda_j = j, \forall j$, which is achieved by $\mathbf{A}_{sym} = \text{diag}(1, 2, \ldots, 100)$ and $\mathbf{A}_{unsym} = \mathbf{W}\text{diag}(1, 2, \ldots, 100)\mathbf{W}^{-1}$, where $\mathbf{W}$ is a random matrix. First we want to consider the algorithms using single vector iterations, that is the power method, inverse iteration and Rayleigh quotient iteration.

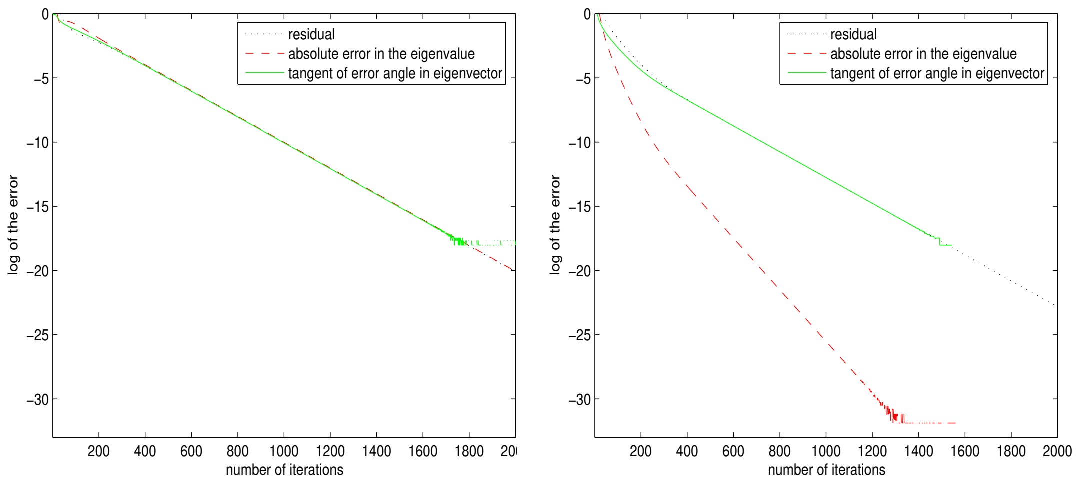


Figure A-1: *Power method with* $\mathbf{A}_{unsym}$ **Figure A-2:** *Power method with* $\mathbf{A}_{sym}$

For the power method and inverse iteration we chose random starting guesses and for Rayleigh quotient iteration we chose a starting vector whose component in the eigen-

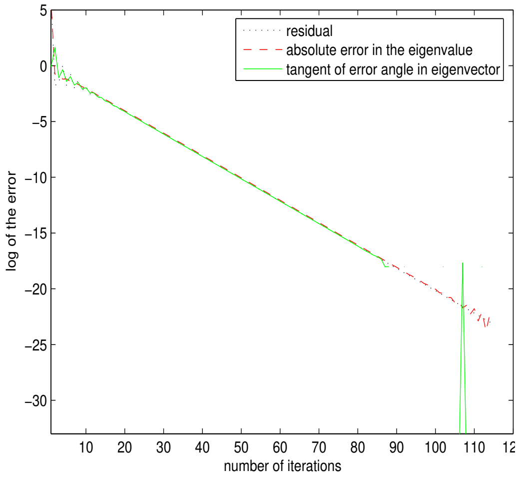


Figure A-3: *Inverse iteration with* $\mathbf{A}_{unsym}$ *and shift* $\sigma = 30.45$

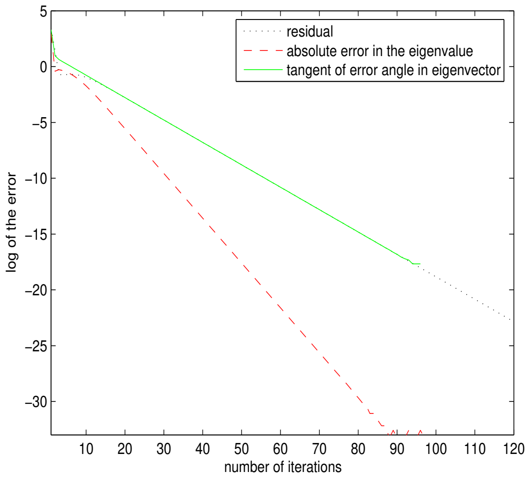


Figure A-4: *Inverse iteration with* $\mathbf{A}_{sym}$ *and shift* $\sigma = 30.45$

vector direction corresponding to λ_{30} was large enough. In each experiment the errors in the eigenvector and eigenvalue approximations as well as the eigenvalue residual have been calculated during the iterations. We plotted their log's in the figures. The dotted line represents the eigenvalue residuals, the solid line represents the tangent of the angle between the wanted eigenvector and its current approximation and the dashed line shows the absolute value of the error in the eigenvalue approximation. The algorithms have been stopped once the 2-norm of the eigenvalue residual was smaller than 10^{-10}.

In the Figures A-1 and A-2 the linear convergence of the power method can be observed. The convergence is slow in this case, because the convergence factor is 0.99 leading to a relatively high number of iterations. There is a difference between $\mathbf{A}_{unsym}$ and $\mathbf{A}_{sym}$ in the eigenvalue approximation. In the nonsymmetric case the errors are about the same as for the eigenvectors, whereas in the symmetric case they are about the square of those errors. This is due to the bound

$$|\theta - \lambda_1| \leq \rho_{\max} \sin^2 \Phi,$$

where Φ is the angle between the eigenvector approximation $\mathbf{x}$ and the eigenvector $\mathbf{v}_1$ and $\rho_{\max} = \max_i |\lambda_i - \lambda_1|$, see [129].

For inverse iteration we chose two different shifts, $\sigma = 30.45$ and $\sigma = 30.1$, to see how the shift influences the rate of convergence. Inverse iteration will then converge to the eigenvector corresponding to λ_{30} and since it is just a modified power method, the Figures A-3 to A-6 are similar to the Figures A-1 and A-2 from the power method in the symmetric and nonsymmetric cases and again they show linear convergence. However the shift gives a difference in the convergence rate: For the shift $\sigma = 30.45$ we get a convergence rate of $(30 - 30.45)/(31 - 30.45) \approx 0.82$, whereas the shift $\sigma = 30.1$ gives a convergence rate of $(30 - 30.1)/(31 - 30.1) \approx 0.11$, and hence the last choice needs only 12 iterations to obtain the required accuracy while the first choice needed about 110 iterations.

In Figures A-7 and A-8 the results for Rayleigh quotient iteration in the symmetric and nonsymmetric case are shown for the approximation of λ_{30} and the corresponding

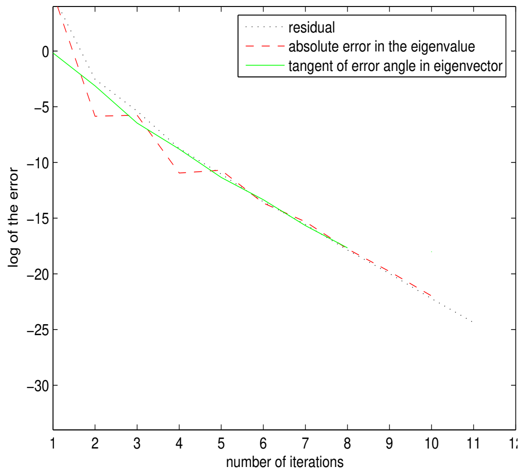


Figure A-5: *Inverse iteration with* $\mathbf{A}_{unsym}$ *and shift* $\sigma = 30.1$

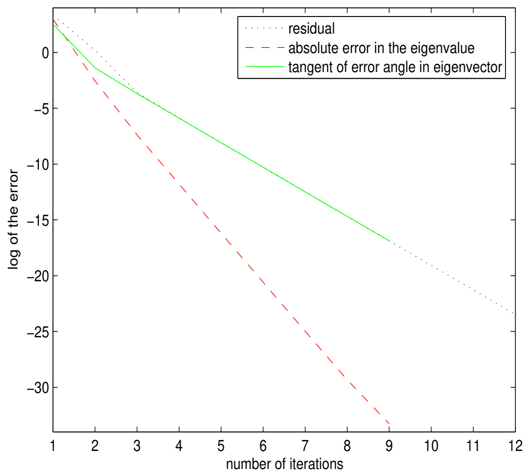


Figure A-6: *Inverse iteration with* $\mathbf{A}_{sym}$ *and shift* $\sigma = 30.1$

eigenvector are shown. The convergence is quadratic in the nonsymmetric case (see Figure A-7) and the algorithms converges in only 5 iterations. For the symmetric matrix (see Figure A-8) the convergence is cubic.

A.1.2 Subspace iteration - fixed dimension

Subspace or simultaneous iteration lets the matrix operate on a set of vectors simultaneously, until the iterated vectors span the invariant subspace of the leading eigenvalues. It is therefore a generalisation of the power method. Details on subspace iteration can be found in Parlett [101]. The subspaces of dimension p are spanned by the columns of

Algorithm 12 Subspace Iteration

Input: $\mathbf{A}$, $\mathbf{X}^{(0)}$ $({\mathbf{X}^{(0)}}^H\mathbf{X}^{(0)} = \mathbf{I})$, i_{max}.
  for $i = 1, \ldots, i_{max}$ **do**
    $\mathbf{Y}^{(i)} = \mathbf{A}\mathbf{X}^{(i-1)}$
    $\mathbf{X}^{(i)} = orth(\mathbf{Y}^{(i)})$
    $\mathbf{H}^{(i)} = {\mathbf{X}^{(i)}}^H\mathbf{A}\mathbf{X}^{(i)}$
  end for
Output: $\mathbf{H}^{(i_{max})}$, $\mathbf{X}^{(i_{max})}$.

the matrices $\mathbf{X}^{(i)}$, which are kept orthonormal by using the Gram-Schmidt orthonormalisation. We apply the Rayleigh-Ritz procedure onto these subspaces in order to determine the projected matrix $\mathbf{H}^{(i)} = {\mathbf{X}^{(i)}}^H\mathbf{A}\mathbf{X}^{(i)}$. Then eigenvalues and eigenvectors of the smaller matrix $\mathbf{H}^{(i)} \in \mathbb{R}^{p\times p}$ can be calculated. Thus ${\mathbf{X}^{(i)}}^H\mathbf{A}\mathbf{X}^{(i)}\mathbf{y}_j^{(i)} = \theta_j\mathbf{y}_j^{(i)}$ for $j = 1, \ldots, p$ and we have the following definition.

Definition A.1. *The eigenvalues* θ_j, $j = 1, \ldots, p$ *of the projected matrix* $\mathbf{H} = \mathbf{X}^H\mathbf{A}\mathbf{X}$, *where* $\mathbf{H}$ *is the projection of* $\mathbf{A}$ *onto the subspace span*$\{\mathbf{X}\}$ *of dimension* p *are the Ritz values, the corresponding vectors* $\mathbf{X}\mathbf{y}_j$ *are the Ritz vectors.*

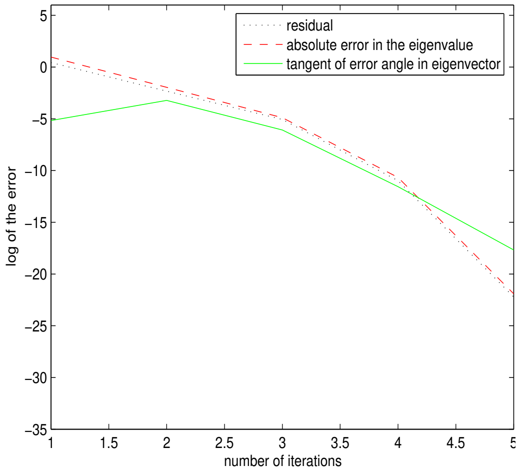


Figure A-7: *Rayleigh quotient iteration with* $\mathbf{A}_{unsym}$

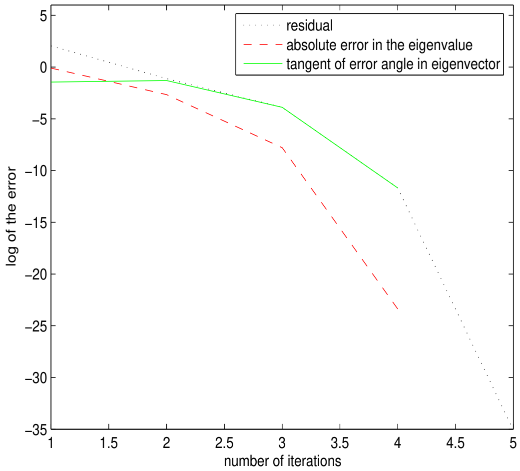


Figure A-8: *Rayleigh quotient iteration with* $\mathbf{A}_{sym}$

Algorithm 13 Rayleigh-Ritz Procedure

Input: $\mathbf{A}$, $\mathbf{X}$, $(\mathbf{X}^H\mathbf{X} = \mathbf{I})$.
$\mathbf{H} = \mathbf{X}^H\mathbf{A}\mathbf{X}$
calculate $\mathbf{Y}, \mathbf{\Theta} \in \mathbb{R}^{\mathbf{p}\times\mathbf{p}}$ with $\mathbf{Y}$ invertible and $\mathbf{\Theta}$ diagonal matrix of eigenvalues such that
$\mathbf{H}\mathbf{Y} = \mathbf{Y}\mathbf{\Theta}$
$\mathbf{Z} = \mathbf{X}\mathbf{Y}$
Output: $\mathbf{\Theta}$, $\mathbf{Z}$.

For $p = 1$ there is only one Ritz value which is equal to the well-known Rayleigh-quotient $\theta^{(i)} = \dfrac{\mathbf{x}^{(i)H}\mathbf{A}\mathbf{x}^{(i)}}{\mathbf{x}^{(i)H}\mathbf{x}^{(i)}} = \mathbf{x}^{(i)H}\mathbf{A}\mathbf{x}^{(i)}$ since $\|\mathbf{x}^{(i)}\| = 1$.

The p Ritz values converge linearly to the p eigenvalues to which they correspond to. The analysis is similar to the one in the power method, generalised to dimension p.

Consider the same simple matrix example as in the previous subsection. Now, we use subspace iteration to approximate the eigenvectors belonging to λ_{100}, λ_{99} and λ_{98}. As initial subspace we chose a random 3-dimensional orthonormalised subspace.

In each experiment the errors in the eigenvector and eigenvalue approximations as well as the eigenvalue residual have been calculated during the iterations. We plotted their log's in the figures. The dotted line represents the 2-norm of the generalised eigenvalue residuals, the solid line represents the tangent of the angle between the wanted and the current invariant subspace and the dashed line shows the 2-norm of the vector containing the errors in the Ritz values. The algorithms have been stopped once the 2-norm of the generalised eigenvalue residual $\mathbf{AX} - \mathbf{XT}$ (where $\mathbf{T} = \mathbf{X}^H\mathbf{A}\mathbf{X}$) was smaller than 10^{-7}.

The results of the subspace iteration are shown in Figures A-9 and A-10. The convergence is linear with a convergence rate of 97/98 (this result follows similar to the power method) and with the same phenomenon of faster eigenvalue convergence

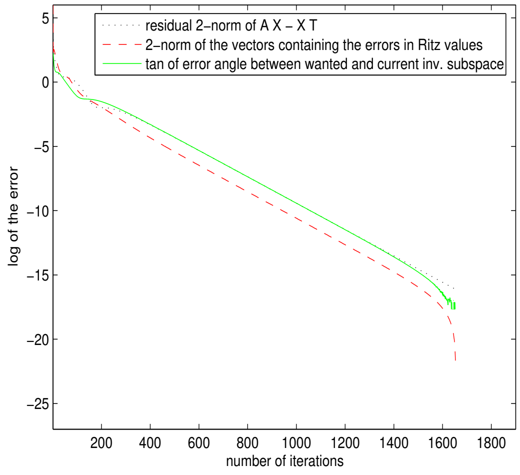


Figure A-9: *Subspace iteration with* $\mathbf{A}_{unsym}$

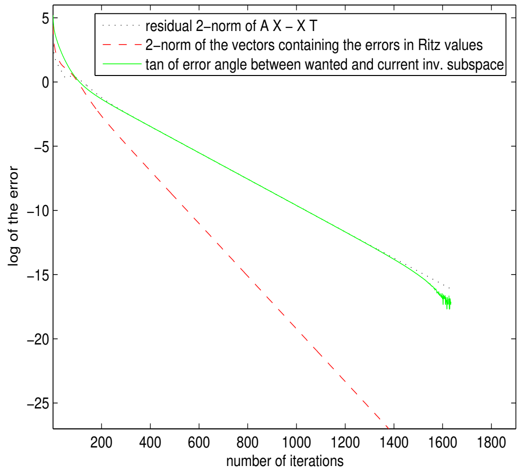


Figure A-10: *Subspace iteration with* $\mathbf{A}_{sym}$

for symmetric matrices as for the single vector iterations.

A.1.3 Subspace iteration - increasing dimension

So far we have only considered subspaces algorithms, where the dimension of the subspaces was fixed during the algorithm. Now we describe a few methods with increasing subspace dimension. The algorithms usually start off with a single vector and in each iteration step the subspace is expanded by one vector which is orthogonal to the previous ones and the subspace dimension is incremented.

Arnoldi's method or Lanczos' method (in the special case of Hermitian matrices) are closely related to subspace (or simultaneous) iteration, but they make much better use of the information obtained by remembering all the directions computed so far and orthonormalising them using some form of the Gram-Schmidt process. Also, the algorithm always lets the matrix operate on a vector orthogonal to all those previously tried. They build up an orthogonal basis of the Krylov sequence with respect to $\mathbf{A}$ and $\mathbf{x}$, which is given by the iteration vectors of the simple power method: $\mathbf{x}, \mathbf{A}\mathbf{x}, \mathbf{A}^2\mathbf{x}, \ldots$. Subspaces with this structure are called Krylov subspaces:

Definition A.2 (Krylov subspace)**.** *Krylov subspaces of dimension i and are defined as*

$$\mathcal{K}_i(\mathbf{A}, \mathbf{x}) = span\{\mathbf{x}, \mathbf{A}\mathbf{x}, \mathbf{A}^2\mathbf{x}, \ldots, \mathbf{A}^{i-1}\mathbf{x}\}.$$

For completeness we denote some interesting properties of Krylov spaces here.

Definition A.3 (Minimal polynomial)**.** *The minimal polynomial p of a vector $\mathbf{x}$ w.r.t. $\mathbf{A}$ is the nonzero polynomial of minimal degree such that*

$$p(\mathbf{A})\mathbf{x} = 0.$$

The degree of the minimal polynomial is called grade of $\mathbf{x}$ with respect to $\mathbf{A}$.

Corollary A.4. $\mathcal{K}_i(\mathbf{A}, \mathbf{x}) = span\{\mathbf{x}, \mathbf{A}\mathbf{x}, \mathbf{A}^2\mathbf{x}, \ldots, \mathbf{A}^{i-1}\mathbf{x}\}$ *has dimension i if and only if the grade of $\mathbf{x}$ is larger than $i - 1$.*

Algorithm 14 Arnoldi Algorithm

Input: $\mathbf{A}$, $\mathbf{x}^{(1)}$, ($\|\mathbf{x}^{(1)}\|_2 = 1$), i_{max}.
  for $i = 1, \ldots, i_{max}$ **do**
    $\mathbf{x}^{(i+1)} = \mathbf{A}\mathbf{x}^{(i)}$
    for $j = 1$ to i **do**
      $\mathbf{h}_{ji} = \mathbf{x}_j^H \mathbf{x}^{(i+1)}$
      $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i+1)} - \mathbf{x}^{(j)}\mathbf{h}_{ji}$
    end for
    reorthogonalise
    $\mathbf{h}_{i+1,i} = \|\mathbf{x}^{(i+1)}\|_2$
    if $\mathbf{h}_{i+1,i} = 0$ **then**
      span$\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)}\}$ is invariant under $\mathbf{A}$
    end if
    $\mathbf{x}^{(i+1)} = \dfrac{\mathbf{x}^{(i+1)}}{\mathbf{h}_{i+1,i}}$
    $\mathbf{H}^{(i)} = (\mathbf{h}_{kj})_{1 \le k,j \le i}$
    $\mathbf{X}^{(i)} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)})$
  end for
Output: $\mathbf{H}^{(i_{max})}$, $\mathbf{X}^{(i_{max})}$.

Proof. The vectors $\mathbf{x}, \mathbf{A}\mathbf{x}, \mathbf{A}^2\mathbf{x}, \ldots, \mathbf{A}^{i-1}\mathbf{x}$ form a basis of the Krylov subspace $\mathcal{K}_i$ if and only if from $\sum_{j=0}^{i-1} \alpha_j \mathbf{A}^j \mathbf{x}$ it follows that $\alpha_j = 0$ for all α_j. But this is equivalent to the condition that there exists no polynomial p of maximum degree $i-1$ such that $p(\mathbf{A})\mathbf{x} = 0$ and hence the grade of $\mathbf{x}$ is larger than $i-1$. □

Within the Arnoldi process the subspace is orthonormalised and in this orthonormal basis, the matrix operator is represented by an upper Hessenberg matrix $\mathbf{H}^{(i)}$ whose eigenvalues yield Ritz approximations to several eigenvalues. Hence, let $\mathbf{x}^{(1)}$ be the starting vector and let $\mathbf{X}^{(i)} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)})$ be an orthonormal basis of $\mathcal{K}_i(\mathbf{A}, \mathbf{x}^{(1)})$. Then the subspace in step i is expanded by calculating $\mathbf{A}\mathbf{x}^{(i)}$ and orthonormalising the result against span$\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)}\}$. It can be shown that this procedure is the same as calculating $\mathbf{A}^i\mathbf{x}^{(1)}$ and orthonormalising this vector (see, for example [149]).

Now, the Ritz pairs are calculated from the upper Hessenberg matrix ${\mathbf{X}^{(i)}}^H \mathbf{A}\mathbf{X}^{(i)} = \mathbf{H}^{(i)}$, whose elements are generated during the orthonormalisation process in the Arnoldi algorithm. The Arnoldi process can be written in the form

$$\mathbf{A}\mathbf{X}^{(i)} = \mathbf{X}^{(i)}\mathbf{H}^{(i)} + \mathbf{x}^{(i+1)}\mathbf{h}_{i+1,i}\mathbf{e}_i^H. \tag{A.1}$$

If the Arnoldi process breaks down, that is $\mathbf{h}_{i+1,i} = 0$, we have found an invariant subspace and with $\mathbf{A}\mathbf{X}^{(i)} = \mathbf{X}^{(i)}\mathbf{H}^{(i)}$ the eigenvalues of $\mathbf{H}^{(i)}$ are eigenvalues of $\mathbf{A}$ by a similarity transform. If this is not true we get at least the following error estimate for the Ritz values: If $(\theta, \mathbf{y})$ ($\|\mathbf{y}\|_2 = 1$) is an eigenpair of $\mathbf{H}^{(i)}$ obtained by the Rayleigh-Ritz procedure applied to $\mathbf{A}$ and $\mathbf{X}^{(i)}$ then, with $\mathbf{z} = \mathbf{X}^{(i)}\mathbf{y}$ we have

$$\|\mathbf{A}\mathbf{z} - \theta\mathbf{z}\|_2 = |\mathbf{h}_{i+1,i}||\mathbf{y}_i|$$

where $\mathbf{y}_i$ denotes the last component of $\mathbf{y}$. Typically, loss of numerical orthogonality of the vectors in $\mathbf{X}^{(i)}$ takes place, requiring reorthogonalisation (see [23]). Typically the correction suggested in [21] is used.

Several variants of Arnoldi's method try to reduce $|\mathbf{h}_{i+1,i}|$, for example by restarting techniques (see [130]). For more details on Arnoldi's method we refer to the original paper [3] and [110].

If $\mathbf{A}^H = \mathbf{A}$ then also ${\mathbf{H}^{(i)}}^H = \mathbf{H}^{(i)}$ and so $\mathbf{H}^{(i)}$ is tridiagonal and then often called $\mathbf{T}^{(i)}$. It simplifies the Arnoldi algorithm, which is then called Lanczos algorithm and significantly reduces the amount of storage, since, due to a tree-term recurrence, only three vectors are stored at each Lanczos step. For more details on Lanczos' method

Algorithm 15 Lanczos Algorithm

Input: $\mathbf{A}$, $\mathbf{x}^{(1)}$, $(\|\mathbf{x}^{(1)}\|_2 = 1)$, i_{max}.
  for $i = 1, \ldots, i_{max}$ **do**
    $\mathbf{x}^{(i+1)} = \mathbf{A}\mathbf{x}^{(i)}$
    $\alpha_i = {\mathbf{x}^{(i)}}^H \mathbf{x}^{(i+1)}$
    $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i+1)} - \alpha_i \mathbf{x}^{(i)}$
    if $i > 1$ **then**
      $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i+1)} - \beta_{i-1} \mathbf{x}^{(i-1)}$
    end if
    reorthogonalise
    $\beta_i = \|\mathbf{x}^{(i+1)}\|_2$
    if $\beta_i = 0$ **then**
      $\text{span}\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)}\}$ is invariant under $\mathbf{A}$
    end if
    $\mathbf{x}^{(i+1)} = \dfrac{\mathbf{x}^{(i+1)}}{\beta_i}$
    $\mathbf{T}^{(i)} = \text{tridiag}(\alpha_j, \beta_k)_{1 \le j \le i, 1 \le k \le i}$
    $\mathbf{X}^{(i)} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)})$
  end for
Output: $\mathbf{T}^{(i_{max})}$, $\mathbf{X}^{(i_{max})}$.

we refer to the original paper [76] and Saad's book [110], which also contains some convergence theory.

Again, use the example from the previous subsections. Now, Arnoldi's method (or Lanczos' method for $\mathbf{A}_{sym}$) was used to approximate the largest eigenvalue λ_{100} and the eigenvalue λ_{30}. We took a random initial guess and again plotted the log's of the errors. The dotted line represents the eigenvalue residual, the solid line represents the tangent of the angle between the wanted eigenvector and its current approximation and the dashed line shows the absolute value of the error in the eigenvalue approximation. The algorithms have been stopped once the 2-norm of the eigenvalue residual was smaller than 10^{-10}.

Figures A-11 to A-14 show the results for Arnoldi's and Lanczos' method. The Arnoldi algorithm used to approximate the extreme eigenvalue converges in about 70 iterations to the desired accuracy, the convergence is superlinear. In the symmetric case the convergence is smooth and monotonically decreasing (see Figure A-12), which can be explained using convergence theory by Kaniel and Paige (see [48]), whereas in the nonsymmetric case there are irregularities (see Figure A-11).

For the interior eigenvalue (Figures A-13 and A-14) we have a long stagnation period before convergence, the number of iterations needed for convergence to the

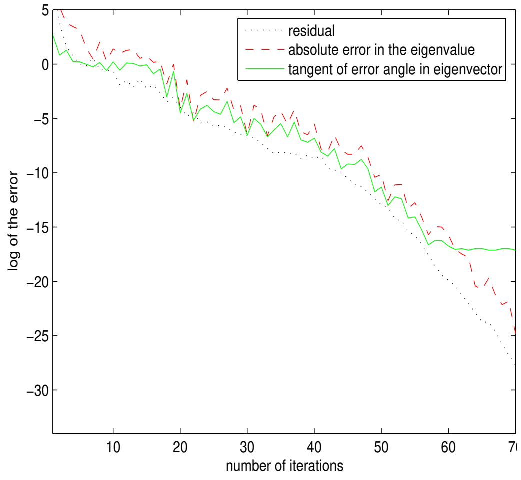


Figure A-11: *Arnoldi method with* $\mathbf{A}_{unsym}$ *(extreme eigenvalue)*

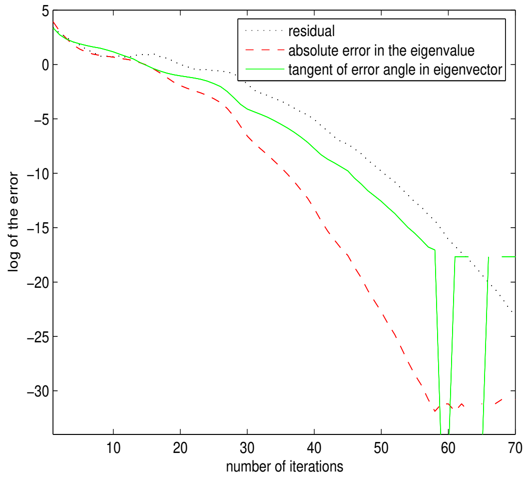


Figure A-12: *Lanczos method with* $\mathbf{A}_{sym}$ *(extreme eigenvalue)*

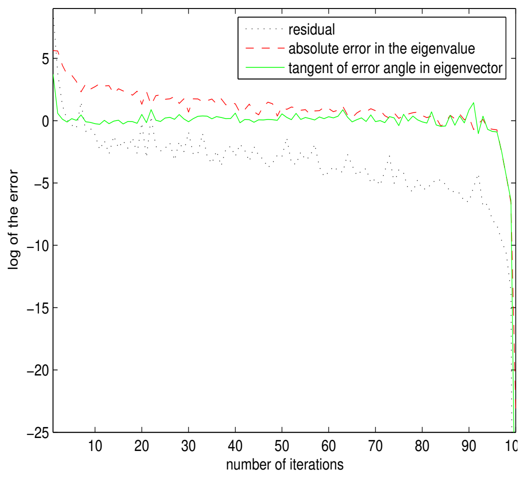


Figure A-13: *Arnoldi method with* $\mathbf{A}_{unsym}$ *(interior eigenvalue)*

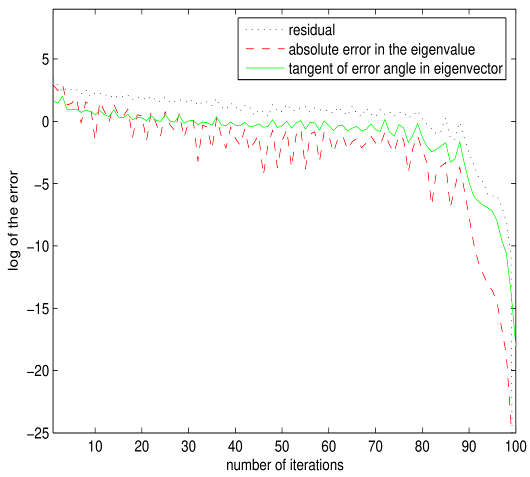


Figure A-14: *Lanczos method with* $\mathbf{A}_{sym}$ *(interior eigenvalue)*

desired accuracy is almost 100. This illustrates that Arnoldi and Lanczos approximate extreme eigenvalues better than the inner ones. This observation suggests a shift-invert procedure to be used for interior eigenvalues.

Finally we want to consider the Jacobi-Davidson method which is also an subspace algorithm with increasing dimension of the subspace. It was introduced by Sleijpen and van der Vorst [124] (see also [63] for an overview). It combines ideas from algorithms by Jacobi and Davidson. Again, the Rayleigh-Ritz procedure is applied to a sequence of subspaces of increasing dimension. But, in the Jacobi-Davidson method the constructed subspaces are no longer Krylov subspaces. Instead, in each step, the subspace is expanded with an orthogonal correction to a Ritz vector in order to obtain a better approximation of an eigenvector. We give a short derivation of the algorithm.

Algorithm 16 Jacobi-Davidson Algorithm

Input: $\mathbf{A}$, σ, $\mathbf{x}^{(1)}$, $(\|\mathbf{x}^{(1)}\|_2 = 1)$, i_{max}.
$\mathbf{z}^{(1)} = \mathbf{x}^{(1)}$
$\theta^{(1)} = \mathbf{x}^{(1)H}\mathbf{A}\mathbf{x}^{(1)}$
$\mathbf{r}^{(1)} = (\mathbf{A} - \theta^{(1)}\mathbf{I})\mathbf{z}^{(1)}$
for $i = 2, \ldots, i_{max}$ **do**
calculate $\mathbf{y}^{(i)} \perp \mathbf{z}^{(i-1)}$ that satisfies (approximately)

$$(\mathbf{I} - \mathbf{z}^{(i-1)}\mathbf{z}^{(i-1)H})(\mathbf{A} - \theta^{(i-1)}\mathbf{I})(\mathbf{I} - \mathbf{z}^{(i-1)}\mathbf{z}^{(i-1)H})\mathbf{y}^{(i)} = -\mathbf{r}^{(i-1)}$$

$\mathbf{y}^{(i)} = \mathbf{y}^{(i)} - \sum_{j=1}^{i-1} \mathbf{x}_j\mathbf{x}_j{}^H\mathbf{y}^{(i)}$
$\mathbf{x}^{(i)} = \dfrac{\mathbf{y}^{(i)}}{\|\mathbf{y}^{(i)}\|_2}$
$\mathbf{X}^{(i)} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)})$
$\mathbf{H}^{(i)} = \mathbf{X}^{(i)H}\mathbf{A}\mathbf{X}^{(i)}$
calculate the Ritz vector $\mathbf{z}^{(i)}$ whose Ritz value $\theta^{(i)}$ is closest to σ
$\mathbf{r}^{(i)} = (\mathbf{A} - \theta^{(i)}\mathbf{I})\mathbf{z}^{(i)}$
end for
Output: $\mathbf{H}^{(i_{max})}$, $\mathbf{X}^{(i_{max})}$.

Let θ be the Ritz value closest to a $\sigma \in \mathbb{C}$ and $\mathbf{z}$ the corresponding Ritz vector of norm one. Then the residual is given by $\mathbf{r} = \mathbf{A}\mathbf{z} - \theta\mathbf{z}$. Further, let λ be an eigenvalue of $\mathbf{A}$, which is also closest to σ. Then the orthogonal correction $\mathbf{y}$ should ideally be equal to $\hat{\mathbf{y}}$ satisfying

$$\mathbf{A}(\mathbf{z} + \hat{\mathbf{y}}) = \lambda(\mathbf{z} + \hat{\mathbf{y}}), \quad \text{and} \quad \hat{\mathbf{y}} \perp \mathbf{z},$$

which is equivalent to

$$(\mathbf{A} - \lambda\mathbf{I})\hat{\mathbf{y}} = \lambda\mathbf{z} - \mathbf{A}\mathbf{z} = (\lambda - \theta)\mathbf{z} - (\mathbf{A}\mathbf{z} - \theta\mathbf{z}) = (\lambda - \theta)\mathbf{z} - \mathbf{r} \quad \text{and} \quad \hat{\mathbf{y}} \perp \mathbf{z}.$$

Multiplying this equation by $\mathbf{z}\mathbf{z}^H$ we get

$$\mathbf{z}\mathbf{z}^H(\mathbf{A} - \lambda\mathbf{I})\hat{\mathbf{y}} = (\lambda - \theta)\mathbf{z}\mathbf{z}^H\mathbf{z} - \mathbf{z}\mathbf{z}^H\mathbf{r} = (\lambda - \theta)\mathbf{z},$$

since $\mathbf{z}^H\mathbf{r} = 0$. Hence $\hat{\mathbf{y}}$ satisfies

$$(\mathbf{I} - \mathbf{z}\mathbf{z}^H)(\mathbf{A} - \lambda\mathbf{I})\hat{\mathbf{y}} = -\mathbf{r} \quad \text{and} \quad \hat{\mathbf{y}} \perp \mathbf{z}.$$

Because λ is unknown it is substituted by the Ritz value θ, which, together with the orthogonality condition $\hat{\mathbf{y}} \perp \mathbf{z}$ leads to the essential correction equation for $\mathbf{y}$ for the Jacobi-Davidson method:

$$(\mathbf{I} - \mathbf{z}\mathbf{z}^H)(\mathbf{A} - \theta\mathbf{I})(\mathbf{I} - \mathbf{z}\mathbf{z}^H)\mathbf{y} = -\mathbf{r}.$$

We also want to note here, that one step of the Jacobi-Davidson method can be seen as a Newton-method. The correction equation can be written as

$$(\mathbf{A} - \theta\mathbf{I})\mathbf{y} = -\mathbf{r} + \beta\mathbf{z}$$

where β is chosen such that $\mathbf{z} \perp \mathbf{y}$. This can be written as a Newton system applied to $\mathbf{F}(\mathbf{z}, \theta) = [(\mathbf{A} - \theta\mathbf{I})\mathbf{z}, (1 - \mathbf{z}^H\mathbf{z})/2]^T$,

$$\begin{bmatrix} \mathbf{A} - \theta\mathbf{I} & -\mathbf{z} \\ -\mathbf{z}^H & 0 \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \beta \end{bmatrix} = \begin{bmatrix} -\mathbf{r} \\ 0 \end{bmatrix},$$

which gives generally quadratic convergence. For Hermitian matrices even cubic convergence can be expected for exact solves of the correction equation, since it can be seen as a subspace accelerated Rayleigh quotient iteration. Note that instead of θ we can use a fixed value σ as a shift in the Jacobi-Davidson method. Then the algorithm can be seen as a subspace accelerated inverse iteration. Hence the Jacobi-Davidson method acts like inexact inverse iteration or RQI in which the use of the iterative solver is made easier, because the arising linear systems use a projected matrix which is better conditioned than the shifted matrix arising in classical inverse iteration and RQI. Also a key difference between inexact inverse iteration and Jacobi-Davidson arises in the use of preconditioners, since in the preconditioner for the Jacobi-Davidson method also has to be restricted orthogonal to the current approximation. We will give more detailed analysis of the Jacobi-Davidson method including equivalence results with Rayleigh quotient iteration in Chapter 3.

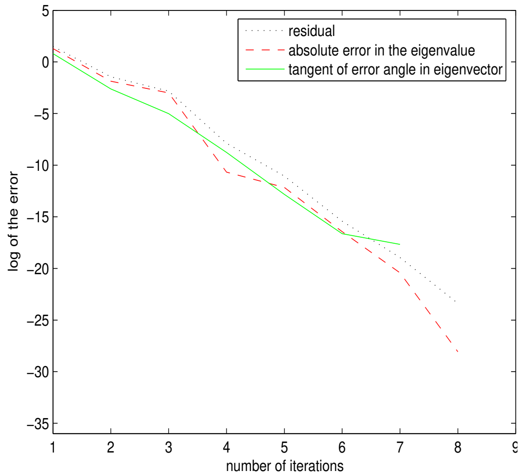


Figure A-15: *Jacobi-Davidson method with* $\mathbf{A}_{unsym}$ *(fixed shift)*

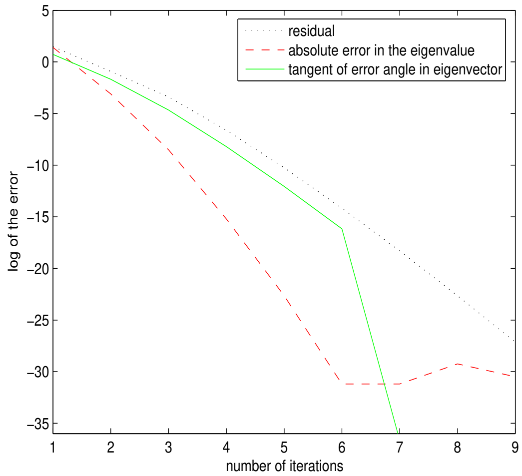


Figure A-16: *Jacobi-Davidson method with* $\mathbf{A}_{sym}$ *(fixed shift)*

Finally, the Jacobi-Davidson method with fixed and Rayleigh quotient shift is used to approximate the extreme eigenvalue λ_{100} of the simple example matrix. We chose a starting vector of all ones and carry out a few steps of the power method in order to get a good initial guess in all cases. We plotted the log's of the error, where the lines in the plot are as in the power method. The algorithms have been stopped once the 2-norm of the eigenvalue residual was smaller than 10^{-10}.

We see linear convergence with the same phenomenon as for the power method and inverse iteration for the fixed shift Jacobi-Davidson method (see Figures A-15 and A-16), but with fewer iterations, since the Jacobi-Davidson method can be seen as a subspace accelerated inverse iteration. For the Rayleigh quotient shift we observe similar behaviour, since then the Jacobi-Davidson method can be interpreted as a

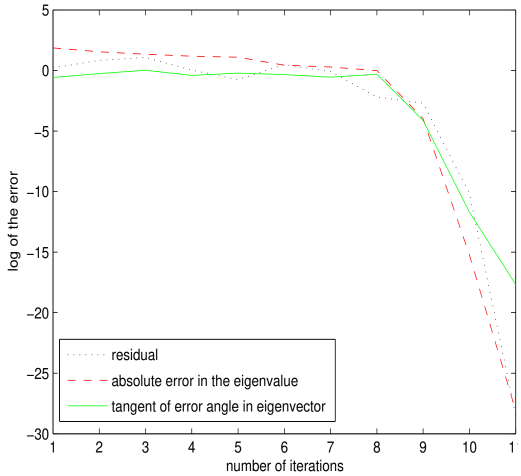


Figure A-17: *Jacobi-Davidson method with $\mathbf{A}_{unsym}$ (Ritz value shift)*

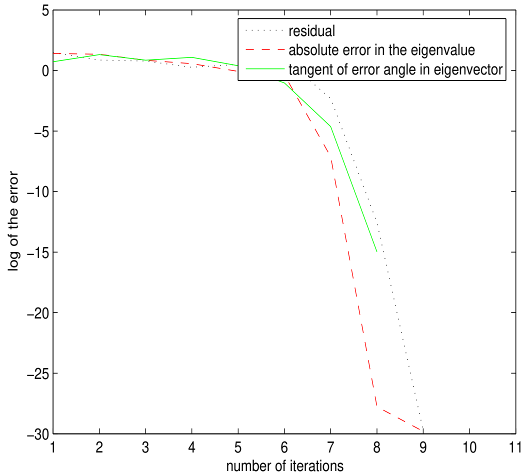


Figure A-18: *Jacobi-Davidson method with $\mathbf{A}_{sym}$ (Ritz value shift)*

subspace accelerated Rayleigh-quotient iteration. We have fast (quadratic) convergence for $\mathbf{A}_{unsym}$ (see Figure A-17) and even faster (cubic) convergence for $\mathbf{A}_{sym}$ (see Figure A-18) after an initial stagnation stage.

A.2 Iterative solvers for linear systems

In this subsection we will briefly describe CG, MINRES and GMRES, three important iterative solvers for linear systems given by

$$\mathbf{Bz} = \mathbf{b}, \tag{A.2}$$

where $\mathbf{B}$ is a square n by n matrix, $\mathbf{b}$ is a column vector and $\mathbf{z}$ is the sought solution. For our inner-outer iterative methods for eigenvalue problems we usually have $\mathbf{B} = \mathbf{A} - \sigma\mathbf{I}$ for some shift σ. CG, MINRES and GMRES belong to the important class of Krylov subspace methods (see, for example [67]).

Detailed discussion on those methods along with other methods can be found in [111], [55] and [5].

A.2.1 CG for Hermitian positive definite systems

The Conjugate Gradient (CG) algorithm is one of the best known algorithms for solving sparse Hermitian positive definite systems. Let $\mathbf{z}^*$ be the exact solution of (A.2) and let $\mathbf{z}_k$ be the kth iterate of some iterative solution technique. The error $\mathbf{e}_k$ and the residual $\mathbf{r}_k$ at step k are given by

$$\mathbf{e}_k = \mathbf{z}^* - \mathbf{z}_k$$

and

$$\mathbf{r}_k = \mathbf{b} - \mathbf{Bz}_k.$$

respectively. The CG algorithm aims to minimise the $\mathbf{B}$-norm of the error $\|\mathbf{e}_k\|_{\mathbf{B}} = \sqrt{\mathbf{e}_k^H \mathbf{B} \mathbf{e}_k}$ (note that this is only defined for positive definite $\mathbf{B}$) over the affine space $\mathbf{z}_0 + \mathcal{K}_k(\mathbf{B}, \mathbf{r}_0)$ where the kth Krylov subspace $\mathcal{K}_k(\mathbf{B}, \mathbf{r}_0)$ is given by

$$\mathcal{K}_k(\mathbf{B}, \mathbf{r}_0) = \text{span}\{\mathbf{r}_0, \mathbf{B}\mathbf{r}_0, \ldots, \mathbf{B}^{k-1}\mathbf{r}_0\}.$$

In one sentence the algorithm is an orthogonal projection technique for the error $\mathbf{z}^* - \mathbf{z}_k$ onto the Krylov subspace $\mathcal{K}_k(\mathbf{B}, \mathbf{r}_0)$ and satisfies the Galerkin condition $\mathbf{r}_k \perp \mathcal{K}_k(\mathbf{B}, \mathbf{r}_0)$. The following algorithm achieves this projection [59].

Algorithm 17 CG

Input: $\mathbf{B}$, $\mathbf{b}$ $\mathbf{z}_0$, k_{max}, ε.
  $\mathbf{r}_0 = \mathbf{b} - \mathbf{B}\mathbf{z}_0$, $\mathbf{p}_0 = \mathbf{r}_0$.
  for $k = 1, \ldots, k_{max}$ **do**
    if $\mathbf{r}_{k-1} < \varepsilon \|\mathbf{b}\|_2$ **then**
      algorithm converged after $k-1$ iterations.
    end if
    Compute $\mathbf{B}\mathbf{p}_{k-1}$.
    $\alpha_{k-1} = \dfrac{< \mathbf{r}_{k-1}, \mathbf{r}_{k-1} >}{< \mathbf{p}_{k-1}, \mathbf{B}\mathbf{p}_{k-1} >}$.
    $\mathbf{z}_k = \mathbf{z}_{k-1} + \alpha_{k-1}\mathbf{p}_{k-1}$.
    $\mathbf{r}_k = \mathbf{r}_{k-1} - \alpha_{k-1}\mathbf{B}\mathbf{p}_{k-1}$.
    $\beta_{k-1} = \dfrac{< \mathbf{r}_k, \mathbf{r}_k >}{< \mathbf{r}_{k-1}, \mathbf{r}_{k-1} >}$
    $\mathbf{p}_k = \mathbf{r}_k + \beta_{k-1}\mathbf{p}_{k-1}$
  end for
Output: $\mathbf{z}_k$.

A result of the algorithm is

$$\mathbf{r}_i^H \mathbf{r}_j = 0, \quad \mathbf{p}_i^H \mathbf{B} \mathbf{p}_j = 0, \quad \text{for} \quad 0 \leq i, j \leq k-1, \, i \neq j,$$

that is the residuals $\mathbf{r}_j$ form an orthogonal basis of $\mathcal{K}_k$ and the search directions $\mathbf{p}_j$ are a conjugate basis. Note that only one matrix-vector multiplication is needed in the algorithm during each iteration, the remaining of the method just consists of inner products.

Since $\mathbf{z}_k$ minimizes $\|\mathbf{z}^* - \mathbf{z}_k\|_{\mathbf{B}}$ over $\mathbf{z}_0 + \mathcal{K}_k(\mathbf{B}, \mathbf{r}_0)$,

$$\|\mathbf{z}^* - \mathbf{z}_k\|_{\mathbf{B}} \leq \|\mathbf{z}^* - \mathbf{w}\|_{\mathbf{B}}$$

holds and with $\mathbf{w} \in \mathbf{z}_0 + \mathcal{K}_k$ we can write $\mathbf{w} = \sum_{j=0}^{k-1} \gamma_j \mathbf{B}^j \mathbf{r}_0 + \mathbf{z}_0$ and hence

$$\|\mathbf{z}^* - \mathbf{z}_k\|_{\mathbf{B}} = \min_{p \in \Pi_k, p(0)=1} \|p(\mathbf{B})(\mathbf{z}^* - \mathbf{z}_0)\|_{\mathbf{B}},$$

where Π_k denotes denotes the set of polynomials of degree k. Using the fact that any Hermitian matrix has a complete set of orthonormal eigenvectors we obtain the following bound:

$$\|\mathbf{e}_k\|_{\mathbf{B}} \leq \min_{p \in P_k, p(0)=1} \max_{i=1,\ldots,n} |p(\lambda_i)| \|\mathbf{e}_0\|_{\mathbf{B}}$$

where λ_i denote the eigenvalues of the system matrix $\mathbf{B}$. A well-known result from approximation theory, using Chebychev polynomials (see, for example Saad [111]) yields

$$\|\mathbf{e}_k\|_{\mathbf{B}} \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k \|\mathbf{e}_0\|_{\mathbf{B}},$$

where $\kappa = \dfrac{\lambda_{max}}{\lambda_{min}}$. Also note that we can express the 2-norm of the residual in terms of the A-norm of the error to get (see, for example Kelley [71])

$$\frac{\|\mathbf{r}_k\|_2}{\|\mathbf{r}_0\|_2} \leq \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}\frac{\|\mathbf{e}_k\|_{\mathbf{B}}}{\|\mathbf{e}_0\|_{\mathbf{B}}}.$$

Several versions of CG exist, such as PCG (preconditioned conjugate Gradient method). Detailed theory on the CG algorithm can be found in Kelley [71], Saad [111] or Greenbaum [55].

A.2.2 GMRES for general systems

The Generalised Minimum Residual (GMRES) method was proposed in [112] and it aims to minimise the 2-norm of the residual $\mathbf{r}_k$. Hence the kth iterate of GMRES is the solution to the least squares problem

$$\min_{\mathbf{z}_k \in \mathbf{z}_0 + \mathcal{K}_k(\mathbf{B},\mathbf{r}_0)} \|\mathbf{b} - \mathbf{B}\mathbf{z}_k\|_2.$$

With $\mathbf{z}_k = \mathbf{z}_0 + \sum_{j=0}^{k-1} \gamma_j \mathbf{B}^j \mathbf{r}_0$ and assuming $\mathbf{B}$ is diagonalisable $\mathbf{B} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-\mathbf{1}}$ a similar analysis to the one for the CG algorithm yields

$$\|\mathbf{r}_k\|_2 \leq \kappa_2(\mathbf{V}) \min_{p\in P_k, p(0)=1} \max_{i=1,\dots,n} |p(\lambda_i)| \|\mathbf{r}_0\|_2,$$

where $\kappa_2(\mathbf{V}) = \|\mathbf{V}\|_2\|\mathbf{V}^{-1}\|_2$ is the condition number of the eigenvector matrix $\mathbf{V}$. Again, describing the convergence of GMRES reduces to a problem in approximation theory. We only state the result for the case where the (possibly complex) eigenvalues λ_i are located in a circle $C(\mathbf{c},\rho)$ with centre $\mathbf{c}$ and radius ρ not containing the origin. Then

$$\min_{p\in P_k, p(0)=1} \max_{\lambda_i \in C(\mathbf{c},\rho)} |p(\lambda_i)| = \left(\frac{\rho}{|\mathbf{c}|}\right)^k$$

and hence

$$\|\mathbf{r}_k\|_2 \leq \kappa_2(\mathbf{V})\left(\frac{\rho}{|\mathbf{c}|}\right)^k \|\mathbf{r}_0\|_2.$$

For further analysis we refer to Saad [111] and Appendix B .
In order to implement the algorithm we remark that $\mathbf{z}_k$ has to be of the form

$$\mathbf{z}_k = \mathbf{z}_0 + \mathbf{Q}_k\mathbf{y},$$

where the second term is a linear combination of the orthonormal basis vectors for the Krylov subspace. The Krylov subspace is orthonormalised using the Arnoldi algorithm, which we can write as (see (A.1))

$$\mathbf{B}\mathbf{Q}_k = \mathbf{Q}_k\mathbf{H}_k + \mathbf{h}_{k+1,k}\mathbf{q}_{k+1}\mathbf{e}_k^H = \mathbf{Q}_{k+1}\mathbf{H}_{k+1}.$$

Algorithm 18 GMRES

Input: $\mathbf{B}$, $\mathbf{b}$ $\mathbf{z}_0$, k_{max}, ε.
  $\mathbf{r}_0 = \mathbf{b} - \mathbf{B}\mathbf{z}_0$, $\rho = \beta = \|\mathbf{r}_0\|_2$, $\mathbf{q}_1 = \frac{\mathbf{r}_0}{\beta}$.
  for $k = 1, \ldots, k_{max}$ **do**
    if $\rho < \varepsilon \|\mathbf{b}\|_2$ **then**
      algorithm converged after k iterations.
    end if
    $\mathbf{q}_{k+1} = \mathbf{B}\mathbf{q}_k$.
    for $j = 1, \ldots, k$ **do**
      $\mathbf{h}_{jk} = \mathbf{q}_j^H \mathbf{q}_{k+1}$
      $\mathbf{q}_{k+1} = \mathbf{q}_{k+1} - \mathbf{h}_{jk}\mathbf{q}_j$
    end for
    $\mathbf{h}_{k+1,k} = \|\mathbf{q}_{k+1}\|_2$
    test for orthogonality and reorthogonalise, if necessary
    $\mathbf{q}_{k+1} = \dfrac{\mathbf{q}_{k+1}}{\mathbf{h}_{k+1,k}}$
    $\mathbf{e}_1 = [1, 0, \ldots, 0]^T \in \mathbb{R}^{k+1}$
    minimise $\|\beta\mathbf{e}_1 - \mathbf{H}_{k+1}\mathbf{y}^k\|_2$ to obtain $\mathbf{y}^k \in \mathbb{R}^k$
    $\rho = \|\beta\mathbf{e}_1 - \mathbf{H}_{k+1}\mathbf{y}^k\|_2$
  end for
  $\mathbf{z}_k = \mathbf{z}_0 + \mathbf{Q}_k\mathbf{y}^k$.
Output: $\mathbf{z}_k$.

Hence we obtain a least squares problem in $\mathbb{R}^k$

$$\begin{aligned}\min_{\mathbf{z}_k \in \mathbf{z}_0 + \mathcal{K}_k(\mathbf{B}, \mathbf{r}_0)} \|\mathbf{b} - \mathbf{B}\mathbf{z}_k\|_2 &= \min_{\mathbf{y} \in \mathbb{R}^k} \|\mathbf{b} - \mathbf{B}(\mathbf{z}_0 + \mathbf{Q}_k\mathbf{y})\|_2 \\ &= \min_{\mathbf{y} \in \mathbb{R}^k} \|\mathbf{r}_0 - \mathbf{B}\mathbf{Q}_k\mathbf{y}\|_2.\end{aligned}$$

Finally we can write

$$\begin{aligned}\min_{\mathbf{y} \in \mathbb{R}^k} \|\mathbf{r}_0 - \mathbf{B}\mathbf{Q}_k\mathbf{y}\|_2 &= \min_{\mathbf{y} \in \mathbb{R}^k} \|\mathbf{r}_0 - \mathbf{Q}_{k+1}\mathbf{H}_{k+1}\mathbf{y}\|_2 \\ &= \min_{\mathbf{y} \in \mathbb{R}^k} \|\mathbf{Q}_{k+1}(\beta\mathbf{e}_1 - \mathbf{H}_{k+1}\mathbf{y})\|_2 \\ &= \min_{\mathbf{y} \in \mathbb{R}^k} \|\beta\mathbf{e}_1 - \mathbf{H}_{k+1}\mathbf{y}\|_2,\end{aligned}$$

where $\beta = \|\mathbf{r}_0\|_2$. The problem reduces to solving a least squares problem for $\mathbf{y}$ with upper Hessenberg matrix $\mathbf{H}_{k+1}$, where the solution vector can be calculated by $\mathbf{z}_k = \mathbf{z}_0 + \mathbf{Q}_k\mathbf{y}$. The least squares problem is solved using a QR-factorisation of $\mathbf{H}_{k+1}$ which can be implemented in an efficient way using Givens transformations, that reduces the costs for calculation and storage.

There exist several variants of GMRES in order to improve the convergence rate, such as restarted GMRES or preconditioned GMRES.

For details we refer to Kelley [71], Saad [111] or Greenbaum [55].

A.2.3 MINRES for symmetric systems

The Minimum Residual (MINRES) method is basically GMRES applied to Hermitian systems. The algorithms simplifies in this case. Instead of the Arnoldi process we can use the Lanczos process in order to find an orthonormal basis of the Krylov subspace. Then the orthonormalisation of the Krylov subspace is reduced to a three-term recurrence. The upper Hessenberg matrix $\mathbf{H}_{k+1}$ reduces to a tridiagonal matrix and hence $\mathbf{R}$ in its QR-decomposition has only three nonzero diagonals. Therefore $\mathbf{z}_k$ can be updated from $\mathbf{z}_{k-1}$ and the GMRES algorithm simplifies significantly, reducing storage and computation costs. For more details on the MINRES algorithm we refer to Greenbaum [55].

We only state the convergence result for the MINRES algorithm here. MINRES, as GMRES, minimises the 2-norm of the residual $\mathbf{r}_k$ over $\mathbf{z}_0 + \mathcal{K}_k(\mathbf{B}, \mathbf{r}_0)$. If the MINRES algorithm is applied to Hermitian positive definite systems we get an analogous result to the one for the CG algorithm, namely

$$\|\mathbf{r}_k\|_2 \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k \|\mathbf{r}_0\|_2.$$

Bounds can also be derived if one eigenvalue of $\mathbf{B}$ is much larger than we others. We refer to [55] for further analysis.

For MINRES applied to Hermitian indefinite systems we obtain a different error bound. If zero is not an eigenvalue of $\mathbf{B}$ we get

$$\|\mathbf{r}_k\|_2 \leq 2\left(\frac{\kappa-1}{\kappa+1}\right)^{\left[\frac{k}{2}\right]} \|\mathbf{r}_0\|_2 \leq 2\left(\frac{\kappa-1}{\kappa+1}\right)^{k-1} \|\mathbf{r}_0\|_2,$$

where $\kappa = \dfrac{\max_{i=1,\ldots,n}|\lambda_i|}{\min_{i=1,\ldots,n}|\lambda_i|}$ and $[\ \cdot\ ]$ denotes the integer part. Relations between CG and MINRES approximations and in particular several different basis for the associated Krylov subspace are given in [98]. Further comments and results on iterative methods of linear systems, in particular GMRES will be presented in Appendix B. A survey on Krylov subspace methods for linear systems and and new developments has recently been published in [122].

We illustrate the performance of the various algorithms and their convergence behaviour through some numerical examples.

We implemented the above iterative methods and applied them to matrices of size 100×100. We chose a symmetric positive definite matrix $\mathbf{B} = \mathbf{Q}\mathrm{diag}(1, 2, \ldots, 100)\mathbf{Q}^T$, where $\mathbf{Q}$ is a random orthogonal matrix and $\mathbf{b} = (1, \ldots, 1)^T$. Figure A-19 shows the convergence curves for this example, where CG, MINRES and GMRES all converge after 57 iterations. Then we use $\mathbf{B} = \mathbf{Q}\mathrm{diag}(-100, \ldots, -51, 51, \ldots, 100)\mathbf{Q}^T$ and $\mathbf{B} = \mathbf{Q}\mathrm{diag}(-50, \ldots, -1, 1, \ldots, 50)\mathbf{Q}^T$ as system matrices, which are symmetric, but not positive definite. The results, which show that MINRES and GMRES converge well for these matrices are shown in figures A-20 and A-21. CG, which was designed for positive definite matrices, gives oscillations or fails to converge completely. Finally we chose a random nonsymmetric matrix, where we have to use GMRES to solve the system. The convergence curve for this method is shown in Figure A-22. Typically, one would use preconditioners to accelerate the convergence of iterative methods, especially to avoid the initial plateau in the GMRES convergence curve. More details on special preconditioners can be found in [111] and [55].

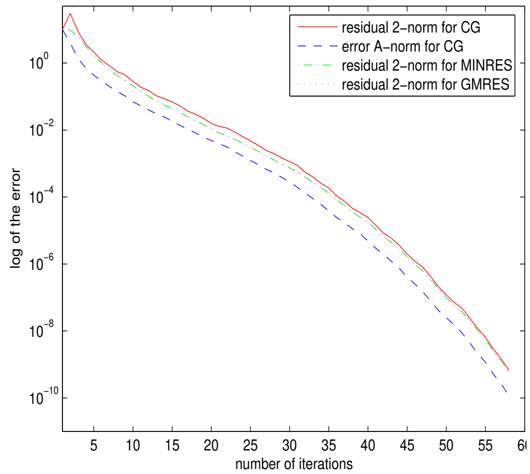


Figure A-19: *CG, MINRES and GMRES convergence for a symmetric positive definite matrix*

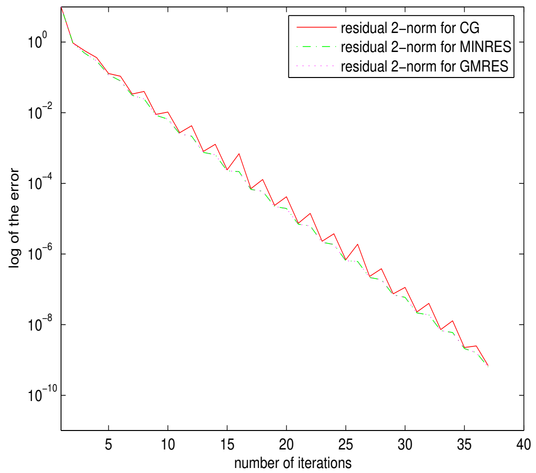


Figure A-20: *CG, MINRES and GMRES convergence for a symmetric matrix with* $\kappa(\mathbf{B}) \approx 2$

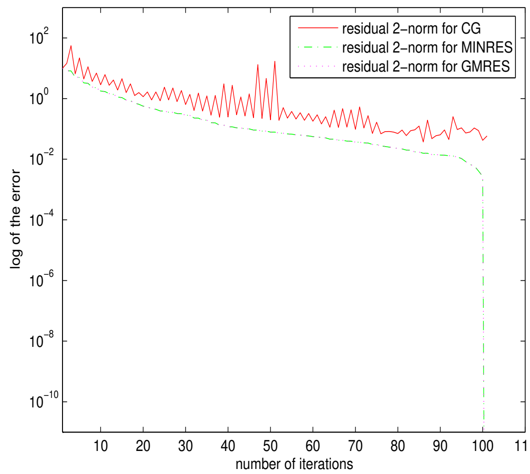


Figure A-21: *Convergence curves for CG, MINRES and GMRES for a symmetric matrix with* $\kappa(\mathbf{B}) = 50$

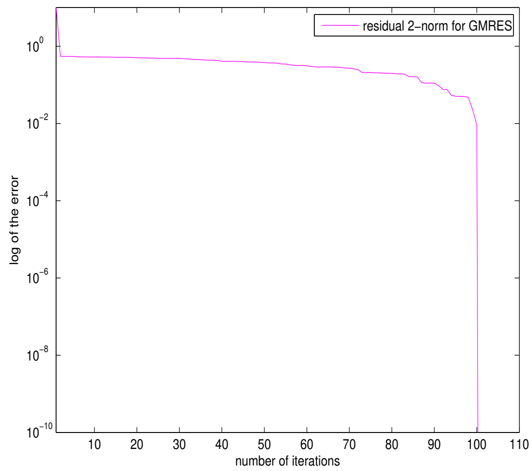


Figure A-22: *Convergence curves for GMRES for an nonsymmetric matrix with* $\kappa(\mathbf{B}) = 8.7e + 03$

APPENDIX B

Convergence theory for GMRES

B.1 Introduction

We have seen that in inexact methods for the computation of interior eigenvalues systems of the form $\mathbf{Bz} = \mathbf{b}$, where $\mathbf{B} \in \mathbb{C}^{n\times n}$ and $\mathbf{b} \in \mathbb{C}^n$, have to be solved. Typically $\mathbf{B} := \mathbf{A} - \sigma\mathbf{M}$ or $\mathbf{B} := \mathbf{A} - \sigma\mathbf{I}$ for some shift σ. It is important to know how the solution of these systems behaves. We use GMRES as a linear solver for these systems and therefore we need to deal with the performance of GMRES (see [112]). We give a short introduction to GMRES and some more details on its convergence theory.

GMRES is an iterative Krylov subspace method which computes approximate solutions $\mathbf{z}_k$ to the system $\mathbf{Bz} = \mathbf{b}$ of the form

$$\mathbf{z}_k = \mathbf{z}_0 + \mathcal{K}_k(\mathbf{B}, \mathbf{r}_0) \quad \text{s.t.} \quad \mathbf{b} - \mathbf{Bz}_k \perp \mathbf{B}\mathcal{K}_k(\mathbf{B}, \mathbf{r}_0) \tag{B.1}$$

This particular choice of the constraint space $\mathbf{B}\mathcal{K}_k(\mathbf{B}, \mathbf{r}_0)$ results in the minimisation of the residual norm for all approximate solutions in the search space $\mathbf{z}_0 + \mathcal{K}_k(\mathbf{B}, \mathbf{r}_0)$. Condition (B.1) is also commonly known as Petrov-Galerkin condition. Recall that the Krylov subspace is given by

$$\mathcal{K}_k(\mathbf{B}, \mathbf{r}_0) := \operatorname{span}\{\mathbf{r}_0, \mathbf{Br}_0, \mathbf{B}^2\mathbf{r}_0, \ldots, \mathbf{B}^{k-1}\mathbf{r}_0\},$$

which is used as a search space. There are many other methods where the search space and the constraint space are either equal (orthogonal projection methods) or not equal (oblique projection methods). One famous example for an orthogonal projection method is CG. We stick to GMRES here.

A measure for the quality of the approximate solution $\mathbf{z}_k$ is the residual

$$\mathbf{r}_k = \mathbf{b} - \mathbf{Bz}_k.$$

Since the vector $\mathbf{z}_k$ is the vector in $\mathbf{z}_0 + \mathcal{K}_k(\mathbf{B}, \mathbf{r}_0)$ which has the smallest residual, we have

$$\|\mathbf{b} - \mathbf{Bz}_k\|_2 = \min_{\mathbf{z}\in\mathbf{z}_0+\mathcal{K}_k(\mathbf{B},\mathbf{r}_0)} \|\mathbf{b} - \mathbf{Bz}\|_2.$$

The approximate solution can be written as $\mathbf{z}_k = \mathbf{z}_0 + q(\mathbf{B})\mathbf{r}_0$, where $q \in \Pi_{k-1}$ is called the iteration polynomial. Using this result, the residual at step k can be written as

$$\mathbf{r}_k = \mathbf{b} - \mathbf{Bz}_k = \mathbf{b} - \mathbf{Bz}_0 - \mathbf{B}q(\mathbf{B})\mathbf{r}_0 = (\mathbf{I} - \mathbf{B}q(\mathbf{B}))\mathbf{r}_0 = p(\mathbf{B})\mathbf{r}_0,$$

where $p \in \Pi_k$ with $p(0) = 1$ is the so called residual polynomial. The minimisation problem can then be written as

$$\|\mathbf{r}_k\|_2 = \min_{\substack{p \in \Pi_k \\ p(0)=1}} \|p(\mathbf{B})\mathbf{r}_0\|_2,$$

and this is the standard way of examining GMRES convergence. Embree [35] summarised three approaches for the convergence bounds on GMRES, which we want to state here. All three approaches apply the inequality

$$\|\mathbf{r}_k\|_2 \leq \min_{\substack{p \in \Pi_k \\ p(0)=1}} \|p(\mathbf{B})\|_2 \|\mathbf{r}_0\|_2,$$

first, which might already be misleading, because it leads to upper bounds for worst case GMRES convergence and also does not consider any specialties of the right hand side. We still use this inequality, since our GMRES bounds deal with special right hand side approximations and upper bounds for worst-case GMRES are sufficient for our analysis. Many classical GMRES convergence bounds are given in [55] and [111].

B.2 Three convergence bounds

Eigenvector conditioning If $\mathbf{B}$ is diagonalisable, that is $\mathbf{B} = \mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^{-1}$ and $\mathbf{B}$ has a complete set of eigenvectors which are the columns of $\mathbf{X}$ and $\boldsymbol{\Lambda}$ is a diagonal matrix containing the eigenvalues $\Lambda(\mathbf{B})$ of $\mathbf{B}$, then we have (see, for example [31])

$$\|\mathbf{r}_k\|_2 \leq \min_{\substack{p \in \Pi_k \\ p(0)=1}} \|\mathbf{X}p(\boldsymbol{\Lambda})\mathbf{X}^{-1}\|_2 \|\mathbf{r}_0\|_2.$$

giving the bound

$$\frac{\|\mathbf{r}_k\|_2}{\|\mathbf{r}_0\|_2} \leq \kappa(\mathbf{X}) \min_{\substack{p \in \Pi_k \\ p(0)=1}} \max_{\lambda \in \Lambda(\mathbf{B})} |p(\lambda)|, \quad \text{where} \quad \kappa(\mathbf{X}) = \|\mathbf{X}\|_2 \|\mathbf{X}^{-1}\|_2. \tag{B.2}$$

For normal matrices, where $\kappa(\mathbf{X}) = 1$, this bound is very satisfying. It is sharp and describes the worst-case behaviour of GMRES (see, [82]). Whereas, for nonnormal matrices, $\kappa(\mathbf{X})$ can be very large and this bound does not work so well. In addition, if $\mathbf{B}$ is not diagonalisable, the bound is not applicable at all. To obtain the actual convergence bound a (complex) polynomial approximation problem over $\Lambda(\mathbf{B})$ has to be solved.

Field of Values The field of values (or numerical range), which is given by the set of Rayleigh quotients

$$\mathcal{F}(\mathbf{B}) = \left\{ \frac{\mathbf{z}^*\mathbf{B}\mathbf{z}}{\mathbf{z}^*\mathbf{z}},\, \mathbf{z} \in \mathbb{C}^n,\, \mathbf{z} \neq 0 \right\},$$

is an alternative for bounding the GMRES polynomial, provided that $0 \notin \mathcal{F}(\mathbf{B})$. The largest absolute value of a point in $\mathcal{F}(\mathbf{B})$ is given by the numerical radius, $\nu(\mathbf{B}) := \max_{z \in \mathcal{F}(\mathbf{B})} |\mathbf{z}|$ (see [65]). Several different bounds have been developed using the field of values. We give the simplest one, that was already given in [55] (see also [35]). It uses the fact that for the 2-norm of a matrix $\|\mathbf{B}\|_2 \leq 2\nu(\mathbf{B})$ (see [55, 21]) and hence also for the polynomial $\|p(\mathbf{B})\|_2 \leq 2\nu(p(\mathbf{B}))$. Also, the numerical radius satisfies a power inequality $\nu(\mathbf{B}^m) \leq (\nu(\mathbf{B}))^m$ (see [102]) for a proof), which applies to polynomials as well, that is $\nu(p(\mathbf{B})) \leq \max_{z\in\mathcal{F}(\mathbf{B})} |p(\mathbf{z})|$, so we obtain the bound

$$\frac{\|\mathbf{r}_k\|_2}{\|\mathbf{r}_0\|_2} \leq 2 \min_{\substack{p\in\Pi_k\\ p(0)=1}} \max_{z\in\mathcal{F}(\mathbf{B})} |p(z)|. \tag{B.3}$$

Again, in order to obtain the actual convergence bound a (complex) polynomial approximation problem over $\mathcal{F}(\mathbf{B})$ has to be solved.

The field of values bound can be useful especially if the problem comes from the discretisation of elliptic PDE's. However, since $\mathcal{F}(\mathbf{B})$ is a convex set that contains the convex hull of the eigenvalues of $\mathbf{B}$, the requirement of $0 \notin \mathcal{F}(\mathbf{B})$ makes the bound useless in many situations, in particular for indefinite problems.

Pseudospectra An alternative way to provide GMRES convergence bounds are pseudospectra. The ε-pseudospectrum (see, for example [145], [36]) of a matrix $\mathbf{B}$ is defined by

$$\Lambda_\varepsilon(\mathbf{B}) := \{z \in \mathbb{C} \,|\, \|(z\mathbf{I}-\mathbf{B})^{-1}\|_2 > \varepsilon^{-1}\}.$$

GMRES bounds can then be obtained using the Dunford-Taylor integral as done in [143], which writes any polynomial $p(\mathbf{B})$ in terms of an integral

$$p(\mathbf{B}) = \frac{1}{2\pi i}\int_\Gamma p(z)(z\mathbf{I}-\mathbf{B})^{-1}dz,$$

where Γ is any simple closed curve or a union of simple closed curves containing $\Lambda(\mathbf{B})$ in its interior. If, for a fixed ε, we choose Γ_ε to be the boundary of the pseudospectrum $\Lambda_\varepsilon(\mathbf{B})$ and take norms in the previous expression, we get

$$\|p(\mathbf{B})\|_2 \leq \frac{1}{2\pi}\int_{\Gamma_\varepsilon} |p(z)|\|(z\mathbf{I}-\mathbf{B})^{-1}\|_2 d|z|.$$

Since the resolvent norm $\|(z\mathbf{I}-\mathbf{B})^{-1}\|_2$ is constant with value ε on the boundary, we get

$$\|p(\mathbf{B})\|_2 \leq \frac{\mathcal{L}(\Gamma_\varepsilon)}{2\pi\varepsilon} \max_{z\in\Lambda_\varepsilon(\mathbf{B})} |p(z)|. \tag{B.4}$$

Hence, the GMRES approximation problem gives

$$\frac{\|\mathbf{r}_k\|_2}{\|\mathbf{r}_0\|_2} \leq \frac{\mathcal{L}(\Gamma_\varepsilon)}{2\pi\varepsilon} \min_{\substack{p\in\Pi_k\\ p(0)=1}} \max_{z\in\Lambda_\varepsilon(\mathbf{B})} |p(z)|.$$

A nice property of the pseudospectral bound is that it applies to different values of ε, and so the bound may also be applied to different stages of the iteration (see, for example [35] for details). Also, as it was noted in [35], pseudospectral bounds inherit properties of both the field of values bound and the eigenvector bound. If we write $\operatorname{conv}(S)$ for the convex hull of a set $S \subseteq \mathbb{C}$ we know that

$$\operatorname{conv}(\Lambda(\mathbf{B})) \subseteq \mathcal{F}(\mathbf{B}),$$

with equality when $\mathbf{B}$ is normal (see [65], [55]). Furthermore, for the ε-pseudospectrum we have

$$\Lambda_\varepsilon(\mathbf{B}) \subseteq \mathcal{F}(\mathbf{B}) + \Delta_\varepsilon,$$

where Δ_ε is the closed disk of radius ε. For details of this result we refer to the book [145]. Hence, from the previous two inclusions, we may remark the following, which is also noted in [145]: If $\varepsilon \to 0$, then the spectrum $\Lambda(\mathbf{B})$ is determined by the ε-pseudospectrum $\Lambda_\varepsilon(\mathbf{B})$. On the other hand, if ε becomes large, that is $\varepsilon \to \infty$, then the numerical range $\mathcal{F}(\mathbf{B})$ is determined by $\Lambda_\varepsilon(\mathbf{B})$. We finally note that the pseudospectral bound is more useful than the field of values bound in the case of indefinite problems. However, in most cases the bound is of theoretic nature, since the pseudospectrum might be hard to calculate.

B.3 The actual convergence bound

All three bounds noted in the previous section are associated with a complex approximation problem over one of the sets $\Lambda(\mathbf{B})$, $\mathcal{F}(\mathbf{B})$ and $\Lambda_\varepsilon(\mathbf{B})$. This approximation problem is not trivial and obtaining optimal bounds is still an active area of research.

Most bounds require the set that is approximated over to be at least simply connected, closed and bounded, and ideally convex. This is not always the case, since especially for the pseudospectra bound we may have to deal with disconnected sets.

To this end, consider general complex optimisation problem

$$s_k(E) = \min_{\substack{p\in\Pi_k\\ p(0)=1}} \max_{z\in E} |p(z)|, \tag{B.5}$$

where E is a non-empty and compact subset of $\mathbb{C}$, which contains either of the sets $\Lambda(\mathbf{B})$, $\mathcal{F}(\mathbf{B})$ or $\Lambda_\varepsilon(\mathbf{B})$.

B.3.1 Convex simply connected compact sets

In order to prepare the setting and make things easier, let E be a simply connected compact set which is also convex. Further let Θ be a conformal mapping from the exterior of the convex set E, that is $\bar{\mathbb{C}} \setminus E$ onto the exterior of the unit disk, $\{|w| > 1\}$, with $\Theta(\infty) = \infty$ and assume $0 \notin E$. This conformal mapping exists, since E is a simply connected domain and we can apply the Riemann mapping Theorem (see, for example [90, page 72]). We then have the following Theorem which is a combination of [8, Lemma 2.2] and the results used in [61], [62]. In these papers Faber polynomials are used in order solve the complex approximation problem. Faber polynomials have also been used in [30] and [90] for the analysis of iterative methods. The proof follows the one in [8, Lemma 2.2].

Theorem B.1. *Let $s_k(E)$ be given by (B.5) and let E be convex closed and bounded with $0 \notin E$. Then*

$$s_k(E) \leq \min\{2+\gamma, \frac{2}{1-\gamma^{k+1}}\}\gamma^k, \quad \textit{where} \quad \gamma = \frac{1}{|\Theta(0)|}. \tag{B.6}$$

Proof. Consider Faber polynomials F_k of degree k associated with E. They are given by the polynomial part of the Laurent expansion of $\Theta(z)^k$, that is $\Theta(z)^k = F_k(z) + \mathcal{O}(z^{-1})$ as $z \to \infty$. Then we can define the polynomial $\tilde{p}_\delta(z)$ of degree k depending on some parameter δ by

$$\tilde{p}_\delta(z) := F_k(z) + \delta(\Theta(0)^k - F_k(0)).$$

and then $p_\delta(z) := \dfrac{\tilde{p}_\delta(z)}{\tilde{p}_\delta(0)}$ is of degree k with $p_\delta(0) = 1$.

Using a result on Faber polynomials shown in [74, Theorem 2], which holds for general *convex* sets E we have

$$\nu_k := |F_k(z) - \Theta(z)^k)| \leq 1, \quad \text{for} \quad z \in \bar{C} \setminus E, \tag{B.7}$$

which is especially satisfied for $z = 0$ and $z \in \partial E$. First, we apply the maximum principle to $\Theta(z)F_k(z) - \Theta(z)^{k+1}$, which gives $|\Theta(0)||F_k(0) - \Theta(0)^k| < \max_{z \in \partial E} |\Theta(z)||F_k(z) - \Theta(z)^k| \leq \nu_k$, since $|\Theta(z)| \geq 1$. Hence, the maximum principle applied to $\tilde{p}_\delta(z)$ yields

$$\begin{aligned}
\max_{z \in E} |\tilde{p}_\delta(z)| &= \max_{z \in \partial E} |\tilde{p}_\delta(z)| \\
&\leq \max_{z \in \partial E} |\Theta(z)^k| + \max_{z \in \partial E} |\Theta(z)^k - F_k(z)| + \delta |\Theta(0)^k - F_k(0))| \\
&\leq 1 + \nu_k + \delta \frac{\nu_k}{|\Theta(0)|}
\end{aligned}$$

For $\tilde{p}_\delta(0)$ we get

$$\begin{aligned}
|\tilde{p}_\delta(0)| &= |\Theta(0)^k - (\Theta(0)^k - F_k(0)) + \delta(\Theta(0)^k - F_k(0))| \\
&\geq |\Theta(0)|^k - |\Theta(0)^k - F_k(0)| + \delta |\Theta(0)^k - F_k(0)| \\
&\geq |\Theta(0)|^k - (1 - \delta) \frac{\nu_k}{|\Theta(0)|} \\
&\geq \frac{1}{\gamma^k} - (1 - \delta) \frac{\nu_k \gamma^{k+1}}{\gamma^k}.
\end{aligned}$$

Thus, finally with $\gamma = \dfrac{1}{|\Theta(0)|}$ we obtain

$$s_k(E) \leq \min_{\delta \in [0,1]} \max_{z \in E} |p_\delta(z)| \leq \min_{\delta \in [0,1]} \gamma^k \frac{1 + \nu_k(1 + \delta\gamma)}{1 - (1 - \delta)\nu_k \gamma^{k+1}},$$

leading to (B.6) by substituting $\delta = 0$ or $\delta = 1$ and using $0 \leq \nu_k \leq 1$. $\square$

Note that the approach in the proof is a different one from the one taken in [61] where instead of (B.6)

$$s_k(E) \leq 3\gamma^k, \quad \text{where} \quad \gamma = \frac{1}{|\Theta(0)|},$$

a slightly worse bound is obtained. In a very recent result by Beckermann [7] this bound was improved even further. Using profound techniques of complex analysis he showed that for a convex compact set E, where the field of values $\mathcal{F}(\mathbf{B}) \subset E$ the Faber polynomials of degree $k \geq 1$ on E satisfy $\|F_k(\mathbf{B})\| \leq 2$. Using this result he uses the normalised Faber polynomials itself as suitable polynomials of degree k. Then he applies the GMRES property

$$\frac{\|\mathbf{r}_k\|_2}{\|\mathbf{r}_0\|_2} = \min_{p\in\Pi_k} \frac{\|p(\mathbf{B})\mathbf{r}_0\|_2}{|p(0)|\|\mathbf{r}_0\|_2} \leq \frac{\|F_k(\mathbf{B})\|_2}{|F_k(0)|} \leq \frac{2}{|F_k(0)|} \leq \frac{2\gamma^k}{1-\gamma^{k+1}}$$

where $\gamma = \dfrac{1}{|\Theta(0)|}$ and with $\min\{1, 2\gamma^k/(1-\gamma^{k+1})\} \leq (2+\gamma)\gamma^k$ we obtain the result in (B.6). Note that this bound is better than the one in (B.3), since the constant 2 can be replaced by 1. Further, Beckermann et al. [8] improved the so-called Elman estimate for GMRES: By constructing a special conformal mapping he determines γ to be

$$\gamma_\beta := 2\sin\left(\frac{\beta}{4-2\beta/\pi}\right) < \sin(\beta),$$

where $0 \notin \mathcal{F}(\mathbf{B})$ and $\beta \in (0, \pi/2)$ is given by

$$\cos(\beta) = \frac{\operatorname{dist}(0, \mathcal{F}(\mathbf{B}))}{\|\mathbf{B}\|_2}.$$

For a proof of this result we refer to [8]. This bound improves one of the earliest GMRES bounds given by Elman et al. (see [31] for a proof): If $\mathbf{B}$ has positive definite Hermitian part $(\mathbf{B}+\mathbf{B}^*)/2$ the following upper bound on the GMRES residual $\mathbf{r}_k$ holds:

$$\frac{\|\mathbf{r}_k\|_2}{\|\mathbf{r}_0\|_2} < \sin^k(\beta) \quad \text{where} \quad \cos(\beta) = \frac{\lambda_{\min}(\mathbf{B}+\mathbf{B}^*)/2}{\|\mathbf{B}\|_2}.$$

In the following we give some special cases for E and hence for the conformal mapping Θ and therefore obtain well known convergence bounds.

Disks Let $E = D(z_0, r)$ be a closed disk with radius r and center z_0 which does not include the origin of the form

$$D(z_0, r) := \{z \,|\, |z - z_0| \leq r\} \quad \text{with} \quad 0 < r < |z_0|, \quad z_0 \in \mathbb{C}.$$

In order to map the exterior of $D(z_0, r)$ onto the exterior of the unit disk a suitable conformal mapping is given by $\Theta(z) = \dfrac{z - z_0}{r}$. Then

$$\gamma = \frac{1}{|\Theta(0)|} = \frac{r}{|z_0|},$$

which coincides with the general error bounds given in [55, page 56] and [111, page 189].

Intervals The simplest example for a set E is an interval, so if we choose $E = I = [\lambda_{\min}, \lambda_{\max}]$ then the map of its exterior onto the exterior of a unit disk is the inverse of a Joukowski map , which generally carries circles to ellipses in the complex plane (see, for example [111, page 88]). Since an interval (more precisely an interval travelled through twice) may also be seen as a degenerate ellipse (i.e. an ellipse with minor semi-axis 0) the Joukowski transformation is the right choice to use. First note the Joukowski mapping

$$J(w) = \frac{1}{2}(w + w^{-1})$$

transforms circles of radius r with center at the origin $D(0,r)$ into an ellipse of center at the origin, with foci $-1, 1$, major semi-axis $\frac{1}{2}(r+r^{-1})$ and minor semi-axis $\frac{1}{2}(r-r^{-1})$, the inverse of the Joukowski mapping is not unique, there are two circles, one with radius r, one with radius r^{-1} which have the same image under $J(w)$. Since we consider maps onto the exterior of the unit disk, we have to consider $r > 1$ and hence choose the inverse of the Joukowski map $H(z)$ to be

$$H(z) = z \pm \sqrt{z^2 - 1},$$

where the square root is chosen such that $|H(z)| > 1$. For $r = 1$ we have that $H(z)$ maps the interval $[-1, 1]$ onto the unit disk. In addition we need to transform the interval $I = [\lambda_{\min}, \lambda_{\max}]$ into $[-1, 1]$, which requires a translation and a scalar multiplication. This is given by

$$z(t) = \frac{2t - (\lambda_{\max} + \lambda_{\min})}{\lambda_{\max} - \lambda_{\min}}.$$

Therefore $\Theta(t) := H(z(t))$ with $|H(z)| > 1$ maps $I = [\lambda_{\min}, \lambda_{\max}]$ onto the unit disk and hence

$$\Theta(0) = H(z(0)) = -\frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}} \pm \sqrt{\left(\frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}}\right)^2 - 1}$$

which gives, after some elementary calculations

$$\gamma = \frac{1}{|\Theta(0)|} = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}, \quad \text{where} \quad \kappa = \frac{\lambda_{\max}}{\lambda_{\min}}.$$

This is the well known convergence bound associated with MINRES (for positive definite systems) and CG.

Ellipses Finally we consider E to be an ellipse $E = E(z_0, d, a)$ with center z_0, focal distance d and major semi axis a. The two foci then lie on $(z_0 \pm d)$. We follow a similar approach taken in [38]. In the case of an ellipse both the convergence factor $\gamma = 1/|\Theta(0)|$ and the Faber polynomials are known explicitly. It turns out that the Faber polynomials are just suitably scaled Chebychev polynomials and we can then describe $F_k(z)/F_k(0)$. An ellipse can be characterised by the set

$$E(z_0, d, a) := \{z \in \mathbb{C} : \quad |z - (z_0 + d)| + |z - (z_0 - d)| \le 2a\}.$$

Instead of this general ellipse, we initially consider the standard ellipse with foci at ± 1 and center at the origin which is given by

$$E(0, 1, b) := \{z \in \mathbb{C} : \quad |z - 1| + |z + 1| \le 2b\}.$$

A given ellipse $E(z_0, d, a)$ can be mapped onto $E(0, 1, b)$ by the linear transformation

$$z(t) = \frac{z_0 - t}{d}$$

and $b = a/|d|$, where $E(z_0, d, a)$ and $E(0, 1, b)$ have the same eccentricity $|d|/a$. Introduce the complex Chebychev polynomials of degree k, which are given by (see, for example [111, page 188])

$$T_k(z) = \cosh(k \cosh^{-1}(z)), \quad k = 1, 2, \ldots,$$

and satisfy the recursion

$$\begin{aligned}
T_0(z) &= 1,\\
T_1(z) &= z,\\
T_{k+1}(z) &= 2zT_k(z) - T_{k-1}(z), \quad k > 1.
\end{aligned}$$

One can also show that the Chebychev polynomials satisfy

$$T_k(z) = \frac{1}{2}(w^k + w^{-k}), \quad \text{where} \quad z = \frac{1}{2}(w + w^{-1}),$$

with $w \neq 0$. Note that $z = \frac{1}{2}(w + w^{-1})$ has two solutions $w = z \pm \sqrt{z^2 - 1}$ that are inverse of each other and therefore comparing to the computation of $T_k(z)$ the actual value of $T_k(z)$ does not depend on which solution is chosen.

Our problem is now to find the conformal mapping Θ that maps the exterior of the standard ellipse $E(0, 1, b)$ onto the exterior of the unit circle. Using the given linear transformation $z(t)$ we can then map the exterior of $E(z_0, d, a)$ onto the exterior of the unit disk. Again, we can make use of the Joukowski mapping $J(w)$ which maps the exterior $|w| > 1$ of a unit disk onto $\bar{C} \setminus [-1, 1]$, in particular it maps each circle $|w| = r > 1$ onto $E(0, 1, b)$ with $b = \frac{r + r^{-1}}{2}$. The inverse of J is again given by

$$H(z) = z \pm \sqrt{z^2 - 1}, \quad \text{with} \quad |H(z)| > 1.$$

Using the results about Chebychev polynomials we can then write

$$T_k(z) = \frac{1}{2}(w^k + w^{-k}) = \frac{1}{2}(H(z)^k + H(z)^{-k}) \quad z \notin [-1, 1].$$

Finally using the transformation $z(t)$ we can give the scaled kth Faber polynomial for $E(z_0, d, a)$ which is given by

$$p_k(t) = \frac{F_k(t)}{F_k(0)} = \frac{T_k(z(t))}{T_k(z(0))} = \frac{T_k\left(\frac{z_0 - t}{d}\right)}{T_k\left(\frac{z_0}{d}\right)}.$$

This is the same bound that was obtained in [111, page 191]. Note that, since $0 \notin E(z_0, d, a)$ and hence $z(0) = z_0/d \notin E(0, 1, b)$ and the zeros of T_k all lying in $(0, 1)$ implies $T_k(z(0)) \neq 0$. By examining the expression $\frac{1}{2}(w^k + w^{-k})$ for $w = re^{i\theta}$ we see that

$$|T_k(z)| \leq \frac{1}{2}(r^k + r^{-k})$$

and hence

$$|p_k(t)| \leq \frac{1}{2}(r^k + r^{-k}) \frac{1}{|T_k\left(\frac{z_0}{d}\right)|}.$$

Finally we need to examine $|T_k\left(\frac{z_0}{d}\right)|$. Letting $\tilde{\gamma} := H(z(0))$ which lies outside the unit circle, its image under J which is given by $z(0) = \frac{z_0}{d}$, lies on the boundary of the ellipse

$$E(0, 1, b) = E\left(0, 1, \frac{H(z(0)) + H(z(0)^{-1}}{2}\right) = E\left(0, 1, \frac{\tilde{\gamma}^{-1} + \tilde{\gamma}}{2}\right),$$

which gives

$$\frac{1}{2}(\tilde{\gamma}^{-k} - \tilde{\gamma}^{k}) \leq |T_k\left(\frac{z_0}{d}\right)| \leq \frac{1}{2}(\tilde{\gamma}^{-k} + \tilde{\gamma}^{k})$$

and hence

$$|p_k(t)| \leq \frac{r^k + r^{-k}}{\tilde{\gamma}^{-k} - \tilde{\gamma}^{k}}.$$

From this bound we can see that the asymptotic convergence rate γ is given by

$$\gamma = \frac{r}{\tilde{\gamma}}.$$

This rate can also be obtained using the conformal mapping Θ. Noting that $H(z)$, the inverse of the Joukowski mapping that maps $E(0,1,b)$ with $b = \dfrac{r + r^{-1}}{2}$ onto the circle $|w| = r > 1$, we consider the mapping $\tilde{H}(z) = H(z)/r$, then this mapping transforms the exterior of $E(0,1,b)$ with $b = \dfrac{r + r^{-1}}{2}$ onto the exterior of the unit circle sharply. Hence

$$\gamma = \frac{1}{|\Theta(0)|} = \frac{1}{|\tilde{H}(z(0))|} = \frac{r}{|H\left(\frac{z_0}{d}\right)|} = \frac{r}{\tilde{\gamma}},$$

and since $z(0) \notin E(0,1,b)$ we have $|H(\frac{z_0}{d})| > r$ and so $\gamma < 1$. Using

$$\tilde{\gamma} = H(z(0)) = \frac{z_0}{d} \pm \sqrt{\left(\frac{z_0}{d}\right)^2 - 1},$$

where the root is chosen such that $\tilde{\gamma} > r$ and with $b = \dfrac{a}{|d|} = \dfrac{r + r^{-1}}{2}$:

$$r = \frac{a}{|d|} \pm \sqrt{\left(\frac{a}{|d|}\right)^2 - 1},$$

where the root is chosen such that $r > 1$ we obtain the same approximate convergence rate as obtained in [111, page 196], but there the results were only stated for the case when the foci and the origin are colinear.

B.3.2 Simply connected compact sets

So far we have considered convex sets E only. Since we always have to assume $0 \notin E$, this is a severe drawback, since the bounds generally cannot be applied to indefinite systems. The main result in our bound (B.6) which can be generally written as $s_k(E) \leq 3\gamma^k$ where $\gamma = \dfrac{1}{|\Theta(0)|}$ uses a property of Faber polynomials on *convex* sets E, which was proved in [74, Theorem 2]. If the set is not convex we still get a convergence factor of $\gamma = \dfrac{1}{|\Theta(0)|}$, but the constant C in $s_k(E) \leq C\gamma^k$ might be much larger. We may define the asymptotic convergence factor (see [27]), which will also be used in the next section.

Definition B.2. *Let the quantity $s_k(E)$ be given by*

$$s_k(E) = \min_{\substack{p \in \Pi_k \\ p(0)=1}} \max_{z \in E} |p(z)|.$$

The limit of the decreasing sequence

$$\lim_{k\to\infty} (s_k(E))^{\frac{1}{k}} =: \rho_E < 1$$

is called asymptotic convergence factor and for large k

$$s_k(E) = \min_{\substack{p\in\Pi_k\\ p(0)=1}} \max_{z\in E} \approx \rho_E^k$$

is called estimated asymptotic convergence factor.

The asymptotic convergence factor is given by

$$\rho_E = \frac{1}{|\Theta(0)|} = \gamma,$$

and it holds $\gamma < 1$ if $0 \notin E$.

B.3.3 Compact sets which are not simply connected

The setting gets even harder when we are dealing with disconnected sets. A potential theory approach that treats this setting is given in [27]. We summarise this theory here. Let $E \subseteq \mathbb{C}$ be a compact set with $0 \notin E$ and no isolated points and let ρ_E be the asymptotic convergence factor. Let $p(z) = \prod_{i=1}^k (z - z_i)$ be a polynomial of degree k. Taking the absolute value and the logarithm of this polynomial gives

$$\log|p(z)| = \sum_{i=1}^k \log|z - z_i|.$$

The aim is to minimise $\dfrac{|p(z)|}{|p(0)|}$ on E and by the maximum principle this is equivalent to minimising the same property on ∂E, that is minimise

$$\log|p(z)| - \log|p(0)| = \log\prod_{i=1}^k \left|1 - \frac{z}{z_i}\right|.$$

on ∂E. We rescale the problem to

$$g(z) = \frac{1}{k}\sum_{i=1}^k \log|z - z_i| + C,$$

and want to minimise $\max_{z\in E} g(z) - g(0)$. The function $g(z)$ then is the Green's function associated with E, a unique function defined in the exterior of E satisfying $\nabla^2 g = 0$ outside E, $g(z) \to 0$ for $z \to \partial E$ and $g(z) - \log|z| \to C$ as $|z| \to \infty$. The asymptotic convergence factor is then given by

$$\rho_E = \exp(-g(0)).$$

To actually find ρ_E is generally not trivial, since the computation of Green's function in the complex plane is hard.

Finally we would like to add that if $0 \in E$ then the minimising polynomial is $p(z) = 1$ and then $s_k(E) = 1$ for all k.

APPENDIX C

Results on Eigenvector Perturbation

This short section contains some error bounds for eigenvalues and eigenvectors of a perturbed matrix. These results were proved for the more general case of invariant subspaces and their representation by Stewart (see [137] and [134]).

He sensitivity of an invariant subspace (or eigenvector) depends on the distance between the eigenvalues representing the invariant subspaces. A measure for this distance is given by the separation sep between two operators $\mathbf{B}_{11} \in \mathbb{C}^{l \times l}$ and $\mathbf{B}_{22} \in \mathbb{C}^{n-l \times n-l}$. It is defined by

$$\operatorname{sep}(\mathbf{B}_{11}, \mathbf{B}_{22}) = \min_{\mathbf{X} \neq \mathbf{0}} \frac{\|\mathbf{B}_{11}\mathbf{X} - \mathbf{X}\mathbf{B}_{22}\|}{\|\mathbf{X}\|}, \quad \mathbf{X} \in \mathbb{C}^{l \times n-l},$$

which with the operator $\mathbf{T} : \mathbb{C}^{l \times n-l} \to \mathbb{C}^{l \times n-l}$ given by

$$\mathbf{T}\mathbf{X} = \mathbf{B}_{11}\mathbf{X} - \mathbf{X}\mathbf{B}_{22}$$

is equivalent to

$$\operatorname{sep}(\mathbf{B}_{11}, \mathbf{B}_{22}) = \begin{cases} \|\mathbf{T}^{-1}\|^{-1}, & 0 \notin \lambda(\mathbf{T}) \\ 0, & 0 \in \lambda(\mathbf{T}) \end{cases}.$$

Here, we only use the case $l = 1$, where

$$\operatorname{sep}(b_{11}, \mathbf{B}_{22}) := \begin{cases} \|(b_{11}\mathbf{I} - \mathbf{B}_{22})^{-1}\|^{-1}, & b_{11} \notin \Lambda(\mathbf{B}_{22}) \\ 0, & b_{11} \in \Lambda(\mathbf{B}_{22}) \end{cases}.$$

Now we can describe the behaviour of an eigenvector (and corresponding eigenvalue) of a matrix $\mathbf{B}$ under a perturbation $\mathbf{E}$.

Theorem C.1 (Eigenvector perturbation theory). *Let μ_1 be a simple eigenvalue of $\mathbf{B}$ with corresponding right eigenvector $\mathbf{w}_1$ and let $\mathbf{W} = [\mathbf{w}_1, \mathbf{W}_1^{\perp}]$ be unitary such that*

$$\begin{bmatrix} \mathbf{w}_1 & \mathbf{W}_1^{\perp} \end{bmatrix}^H \mathbf{B} \begin{bmatrix} \mathbf{w}_1 & \mathbf{W}_1^{\perp} \end{bmatrix} = \begin{bmatrix} \mu_1 & \mathbf{n}_{12}^H \\ \mathbf{0} & \mathbf{N}_{22} \end{bmatrix}.$$

Given a perturbation matrix $\mathbf{E}$, partition $\mathbf{W}^H\mathbf{E}\mathbf{W}$ conformally such that

$$\begin{bmatrix} \mathbf{w}_1 & \mathbf{W}_1^{\perp} \end{bmatrix}^H \mathbf{E} \begin{bmatrix} \mathbf{w}_1 & \mathbf{W}_1^{\perp} \end{bmatrix} = \begin{bmatrix} e_{11} & \mathbf{e}_{12}^H \\ \mathbf{e}_{21} & \mathbf{E}_{22} \end{bmatrix}.$$

Let

$$\zeta = sep(\mu_1, \mathbf{N}_{22}) - |e_{11}| - \|\mathbf{E}_{22}\|.$$

If $\zeta > 0$ *and*

$$\frac{\|\mathbf{e}_{21}\|(\|\mathbf{n}_{12}\| + \|\mathbf{e}_{12}\|)}{\zeta^2} < \frac{1}{4}, \tag{C.1}$$

then there exists a unique vector $\mathbf{p} \in \mathbb{C}^{n-1,1}$ *satisfying*

$$\|\mathbf{p}\| \leq 2\frac{\|\mathbf{e}_{21}\|}{\zeta} \tag{C.2}$$

such that

$$\hat{\mathbf{w}} = \frac{\mathbf{w}_1 + \mathbf{W}_1^{\perp}\mathbf{p}}{\sqrt{1 + \mathbf{p}^H\mathbf{p}}}$$

is a simple right eigenvector of $\mathbf{B} + \mathbf{E}$*. The representation of* $\mathbf{B} + \mathbf{E}$ *with respect to the perturbed invariant subspaces is*

$$\hat{\mu}_1 = \mu_1 + e_{11} + (\mathbf{n}_{12}^H + \mathbf{e}_{12}^H)\mathbf{p}.$$

Proof. See [137] and [134]. ☐

Remark C.2. *Condition (C.1) in Theorem C.1 is satisfied if the eigenvalue separation is large enough and if the perturbation is small enough. Also we have*

$$\begin{aligned}
\left\| \hat{\mathbf{w}} - \frac{\mathbf{w}_1}{\sqrt{1 + \mathbf{p}^H\mathbf{p}}} \right\| &= \frac{\|\mathbf{W}_1^{\perp}\mathbf{p}\|}{\sqrt{1 + \mathbf{p}^H\mathbf{p}}} \leq \|\mathbf{p}\| \leq 2\frac{\|\mathbf{e}_{21}\|}{\zeta} \\
&\leq 2\frac{\|\mathbf{E}\|}{sep(\mu_1, \mathbf{N}_{22}) - 2\|\mathbf{E}\|} = \frac{2}{sep(\mu_1, \mathbf{N}_{22})}\|\mathbf{E}\| + \mathcal{O}(\|\mathbf{E}\|^2),
\end{aligned}$$

where the last equality holds for small enough $\|\mathbf{E}\|$ *using Taylor expansion.*

Another perturbation result, which generalises C.1, is given in terms of the spectral resolution of $\mathbf{B}$. We will use both theorems in our applications.

Theorem C.3 (Eigenvector perturbation theory)**.** *Let* μ_1 *be a simple eigenvalue of* $\mathbf{B}$ *with corresponding right eigenvector* $\mathbf{w}_1$ *and let* $[\mathbf{w}_1, \mathbf{W}_2]^{-1} = [\mathbf{v}_1, \mathbf{V}_2]^H$

$$\begin{bmatrix} \mathbf{w}_1 & \mathbf{W}_2 \end{bmatrix}^{-1} \mathbf{B} \begin{bmatrix} \mathbf{w}_1 & \mathbf{W}_2 \end{bmatrix} = \begin{bmatrix} \mu_1 & \mathbf{0}^H \\ \mathbf{0} & \mathbf{C} \end{bmatrix}.$$

Given a perturbation matrix $\mathbf{E}$*, partition conformally such that*

$$\begin{bmatrix} \mathbf{w}_1 & \mathbf{W}_2 \end{bmatrix}^{-1} \mathbf{E} \begin{bmatrix} \mathbf{w}_1 & \mathbf{W}_2 \end{bmatrix} = \begin{bmatrix} e_{11} & \mathbf{e}_{12}^H \\ \mathbf{e}_{21} & \mathbf{E}_{22} \end{bmatrix}.$$

Let

$$\zeta = sep(\mu_1, \mathbf{C}) - |e_{11}| - \|\mathbf{E}_{22}\|.$$

If $\zeta > 0$ *and*

$$\frac{\|\mathbf{e}_{21}\|\|\mathbf{e}_{12}\|}{\zeta^2} < \frac{1}{4}, \tag{C.3}$$

then there exists a unique vector $\mathbf{p} \in \mathbb{C}^{n-1,1}$ *satisfying*

$$\|\mathbf{p}\| \leq 2\frac{\|\mathbf{e}_{21}\|}{\zeta} \tag{C.4}$$

such that

$$\hat{\mathbf{w}} = \mathbf{w}_1 + \mathbf{W}_2\mathbf{p} \quad \textit{and} \quad \hat{\mathbf{V}} = \mathbf{V}_2 - \mathbf{v}_1\mathbf{p}^H$$

are simple right eigenvector and left invariant subspace of $\mathbf{B} + \mathbf{E}$*. The representation of* $\mathbf{B} + \mathbf{E}$ *with respect to* $\hat{\mathbf{w}}$ *is*

$$\hat{\mu}_1 = \mu_1 + e_{11} + \mathbf{e}_{12}^H\mathbf{p}$$

and with respect to $\hat{\mathbf{V}}$ *it is*

$$\hat{\mathbf{C}} = \mathbf{C} + \mathbf{E}_{22} - \mathbf{p}\mathbf{e}_{12}^H.$$

Proof. See [137] and [134]. □

BIBLIOGRAPHY

[1] E. L. Allgower and H. Schwetlick, *A general view of minimally extended systems for simple bifurcation points*, ZAMM Z. angew. Math. Mech, 77 (1997), pp. 83–98.

[2] P. Arbenz and G. H. Golub, *On the spectral decomposition of Hermitian matrices modified by low rank perturbations with applications*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 40–58.

[3] W. E. Arnoldi, *The principle of minimized iteration in the solution of the matrix eigenvalue problem*, Quarterly of Applied Mathematics, 9 (1951), pp. 17–29.

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

[5] R. Barrett, M. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. A. van der Vorst, *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods, 2nd Edition*, SIAM, Philadelphia, PA, 1994.

[6] C. A. Beattie, M. Embree, and D. C. Sorensen, *Convergence of polynomial restart Krylov methods for eigenvalue computations*, SIAM Rev., 47 (2005), pp. 492–515.

[7] B. Beckermann, *Image numérique, GMRES et polynômes de Faber*, C. R. Acad. Sci. Paris, Ser. I, 340 (2005), pp. 855–860.

[8] B. Beckermann, S. A. Goreinov, and E. E. Tyrtyshnikov, *Some remarks on the Elman estimate for GMRES*, SIAM J. Matrix Anal. Appl., 27 (2006), pp. 772–778.

[9] J. Berns-Müller, *Inexact Inverse Iteration using Galerkin Krylov Solvers*, PhD thesis, University of Bath, Department of Mathematical Sciences, 2003.

[10] J. Berns-Müller, I. G. Graham, and A. Spence, *Inexact inverse iteration for symmetric matrices*, Linear Algebra Appl., 416 (2006), pp. 389–413.

[11] J. Berns-Müller and A. Spence, *Inexact inverse iteration and GMRES*, 2006. In preparation.

[12] ——, *Inexact inverse iteration with variable shift for nonsymmetric generalized eigenvalue problems*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 1069–1082.

[13] B. Boisvert, R. Pozo, K. Remington, B. Miller, and R. Lipman, *Matrix market*. available online at `http://math.nist.gov/MatrixMarket/`, 2004.

[14] A. Bouras and V. Frayssé, *A relaxation strategy for the Arnoldi method in eigenproblems*, technical report 16, CERFACS, Toulouse, France, 2000.

[15] S. Campbell, I. Ipsen, C. Kelley, and C. Meyer, *GMRES and the minimal polynomial*, BIT, 36 (1996), pp. 664–675.

[16] F. Chatelin, *Eigenvalues of matrices*, John Wiley & Sons Ltd, Chichester, West Sussex, England, 1993. Originally published in two separate volumes by Masson, Paris: Valeurs propres de matrices (1988) and Exercises de valeurs propres de matrices (1989).

[17] K. A. Cliffe, T. J. Garratt, and A. Spence, *Eigenvalues of the discretized navier-stokes equation with application to the detection of Hopf bifurcations*, Advances in Computational Mathematics, 1 (1993), pp. 337–356.

[18] ——, *Eigenvalues of block matrices arising from problems in fluid mechanics*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1310–1318.

[19] L. Collatz, *Functional analysis and numerical mathematics*, Academic Press, New York and London, 1966.

[20] J. Cullum and A. Greenbaum, *Relations between Galerkin and norm-minimizing iterative methods for solving linear systems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 223–247.

[21] J. Daniel, W. B. Gragg, L. Kaufman, and G. W. Stewart, *Reorthogonalization and stable algorithms for updating the Gram-Schmidt QR factorization*, Math. Comp., 30 (1976), pp. 772–795.

[22] R. S. Dembo, S. C. Eisenstat, and T. Steihaug, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408.

[23] J. W. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, PA, 1997.

[24] J. E. Dennis and J. J. Moré, *Quasi-Newton methods, motivation and theory*, SIAM Rev., 19 (1977), pp. 46–89.

[25] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, vol. 16 of Classics in Applied Mathematics, SIAM, Philadelphia, PA, 1996. Unabridged, corrected reprint of the 1983 original.

[26] P. Deuflhard, *Newton Methods for Nonlinear Problems*, vol. 35 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin Heidelberg, 2004.

[27] T. A. Driscoll, K.-C. Toh, and L. N. Trefethen, *From potential theory to matrix iterations in six steps*, SIAM Rev., 40 (1998), pp. 547–578.

[28] Z. Drmač, M. Omladič, and K. Veselić, *On the perturbation of the Cholesky factorization*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1319–1332.

[29] A. Edelman, T. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 303–353.

[30] M. Eiermann, *On semiiterative methods generated by Faber polynomials*, Numer. Math., 56 (1989), pp. 357–375.

[31] S. C. Eisenstat, H. C. Elman, and M. H. Schultz, *Variational iterative methods for nonsymmetric systems of linear equations*, SIAM J. Numer. Anal., 20 (1983), pp. 345–357.

[32] H. C. Elman, A. Ramage, and D. J. Silvester, *Algorithm 866: IFISS, a Matlab toolbox for modelling incompressible flow*, ACM Trans. Math. Softw., 33 (2007), p. 14.

[33] H. C. Elman and D. Silvester, *Fast nonsymmetric iterations and preconditioning for Navier-Stokes equations*, SIAM J. Sci. Comput., 17 (1996), pp. 33–46.

[34] H. C. Elman, D. J. Silvester, and A. J. Wathen, *Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics*, Oxford University Press, Oxford, UK, 2005.

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

[36] M. Embree and L. N. Trefethen, *Generalizing eigenvalue theorems to pseudospectra theorems*, SIAM J. Sci. Comput., 23 (2002), pp. 583–590.

[37] Y. A. Erlangga and R. Nabben, *Deflation and balancing preconditioners for Krylov subspace methods applied to nonsymmetric matrices*, 2006. Submitted.

[38] O. G. Ernst, *Residual-minimizing Krylov subspace methods for stabilized discretizations of convection-diffusion equations*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1079–1101.

[39] D. R. Fokkema, G. L. G. Sleijpen, and H. A. Van der Vorst, *Jacobi-Davidson style QR and QZ algorithms for the reduction of matrix pencils*, SIAM J. Sci. Comput., 20 (1999), pp. 94–125.

[40] J. Frank and C. Vuik, *On the construction of deflation-based preconditioners*, SIAM J. Sci. Comput., 23 (2001), pp. 442–462.

[41] M. A. Freitag and A. Spence, *A tuned preconditioner for inexact inverse iteration applied to Hermitian eigenvalue problems*, technical report, Dept. of Mathematical Sciences, University of Bath, 2005. available from `ftp://ftp.maths.bath.ac.uk/pub/preprints/maths0506.ps.gz`.

[42] ——, *A tuned preconditioner for inexact inverse iteration applied to Hermitian eigenvalue problems*, 2005. to appear in IMA J. Numer. Anal.

[43] ——, *Convergence rates for inexact inverse iteration with application to preconditioned iterative solves*, BIT, 47 (2007), pp. 27–44.

[44] ——, *Convergence theory for inexact inverse iteration applied to the generalised nonsymmetric eigenproblem*, Electron. Trans. Numer. Anal., 28 (2007), pp. 40–64.

[45] M. A. Freitag, A. Spence, and E. Vainikko, *Tuning for inexact inverse iteration applied to the generalised nonsymmetric eigenproblem*, 2007. In preparation.

[46] S. K. Godunov, *Modern Aspects of Linear Algebra*, Providence, RI, 1998. Translation of Mathematical Monographs, 175. Amer. Math. Soc.

[47] G. H. Golub and C. F. Van Loan, *Matrix Computations*, John Hopkins University Press, Baltimore, 2nd ed., 1989.

[48] ——, *Matrix Computations*, John Hopkins University Press, Baltimore, 3rd ed., 1996.

[49] G. H. Golub and J. Wilkinson, *Ill-conditioned eigensystems and the computation of the Jordan canonical form*, SIAM Review, 18 (1976), pp. 578–619.

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

[51] G. H. Golub, Z. Zhang, and H. Zha, *Large sparse symmetric eigenvalue problems with homogeneous linear constraints: the Lanczos process with inner-outer iterations*, Linear Algebra Appl., 309 (2000), pp. 289–306.

[52] *Google*. `http://www.google.com`, 2005.

[53] I. G. Graham and A. Spence, *Numerical methods for bifurcation problems*, in The Graduate Student's Guide to Numerical Analysis '98, M. Ainsworth, J. Levesley, and M. Marletta, eds., Springer-Verlag, Berlin, 1999, pp. 177–216.

[54] I. G. Graham, A. Spence, and E. Vainikko, *Parallel iterative methods for Navier-Stokes equations and application to eigenvalue computation*, Concurrency and Computation: Practice and Experience, 15 (2003), pp. 1151–1168.

[55] A. Greenbaum, *Iterative methods for solving linear systems*, vol. 17 of Frontiers in Applied Mathematics, SIAM, Philadelphia, PA, 1997.

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

[57] S. C. Hawkins, *The Computation of Eigenvalues of Large Sparse Matrices*, PhD thesis, University of Bath, Department of Mathematical Sciences, 1999.

[58] V. Hernández, J. Román, A. Tomás, and V. Vidal, *A survey of software for sparse eigenvalue problems*, slepc technical report str-6, Universidad Politecnica de Valencia, 2005. Available at http://www.grycap.upv.es/slepc.

[59] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Stand., 49 (1952), pp. 409–436.

[60] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, PA, 2nd ed., 2002.

[61] M. Hochbruck and C. Lubich, *On Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 34 (1997), pp. 1911–1925.

[62] ——, *Error analysis of Krylov methods in a nutshell*, SIAM J. Sci. Comput., 19 (1998), pp. 695–701.

[63] M. E. Hochstenbach and Y. Notay, *The Jacobi-Davidson method*, GAMM Mitt., (2004).

[64] M. E. Hochstenbach and G. L. G. Sleijpen, *Two-sided and alternating Jacobi-Davidson*, Linear Algebra Appl., 358 (2003), pp. 145–172. Special Issue on accurate solution of eigenvalue problems (Hagen, 2000).

[65] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, 1991.

[66] I. C. F. Ipsen, *Computing an eigenvector with inverse iteration*, SIAM Review, 39 (1997), pp. 254–291.

[67] I. C. F. Ipsen and C. D. Meyer, *The idea behind Krylov methods*, Amer. Math. Monthly, 105 (1998), pp. 889–899.

[68] S. Kaniel, *Estimates for some computational techniques in linear algebra*, Math. Comp., 20 (1966), pp. 369–378.

[69] T. Kato, *Perturbation Theory for Linear Operators*, Springer Verlag, 1966.

[70] H. B. Keller, *Numerical solution of bifurcation and nonlinear eigenvalue problems*, in Applications of Bifurcation Theory, P. H. Rabinowitz, ed., Academic Press, New York, 1977, pp. 359–384.

[71] C. T. Kelley, *Iterative Methods for Linear and Nonlinear Equations*, vol. 16 of Frontiers in Applied Mathematics, SIAM, Philadelphia, PA, 1995.

[72] A. V. Knyazev, *Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method*, SIAM Journal on Scientific Computing, 23 (2001), pp. 517–541.

[73] A. V. Knyazev and K. Neymeyr, *A geometric theory for preconditioned inverse iteration. III: a short and sharp convergence estimate for generalized eigenvalue problems*, Linear Algebra Appl., 358 (2003), pp. 95–114.

[74] T. Kövari and C. Pommerenke, *On Faber polynomials and Faber expansions*, Math. Zeitschr., 99 (1967), pp. 193–206.

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

[76] C. Lanczos, *An iterative method for the solution of the eigenvalue problem of linear differential and integral operators*, Journal of Research of the National Bureau of Standards, 45 (1950), pp. 255–282.

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

[78] R. B. Lehoucq and K. Meerbergen, *Using generalized Cayley transformations within an inexact rational Krylov sequence method*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 131–148.

[79] R. B. Lehoucq and J. A. Scott, *Implicitly restarted Arnoldi methods and eigenvalues of the discretized Navier-Stokes equations*, 1997. Technical Report SAND97-2712J, Sandia National Laboratories, Albuquerque, New Mexico.

[80] R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK Users' Guide, Solution of Large-Scale Eigenvalue Problems by Implicitly Restarted Arnoldi Methods*, SIAM, Philadelphia, PA, 1998.

[81] J. Liesen and Z. Strakoš, *GMRES convergence analysis for a convection-diffusion model problem*, SIAM J. Sci. Comput., 26 (2005), pp. 1989–2009.

[82] J. Liesen and P. Tichý, *Convergence analysis of Krylov subspace methods*, GAMM Mitt. Ges. Angew. Math. Mech., 27 (2004), pp. 153–173 (2005).

[83] E. Lindström and L. Eldén, *Adaptive eigenvalue computations using Newton's method on the Grassmann manifold*, SIAM J. Matrix Anal. Appl., 23 (2002), pp. 819–839.

[84] R. Lösche, H. Schwetlick, and G. Timmermann, *A modified block Newton iteration for approximating an invariant subspace of a symmetric matrix*, Linear Algebra Appl., 275-276 (1998), pp. 381–400.

[85] K. Meerbergen and A. Spence, *Implicitly restarted Arnoldi with purification for the shift-invert transformation*, Math. Comp., 66 (1997), pp. 667–689.

[86] G. Meurant and Z. Strakŏs, *The Lanczos and conjugate gradient algorithms in finite precision arithmetic*, Acta Numerica, 15 (2006), pp. 471–542.

[87] R. B. Morgan and D. S. Scott, *Preconditioning the Lanczos algorithm for sparse symmetric eigenvalue problems*, SIAM J. Sci. Comput., 14 (1993), pp. 585–593.

[88] R. Nabben and C. Vuik, *A comparison of deflation and the balancing preconditioner*, SIAM J. Sci. Comput., 27 (2006), pp. 1742–1759.

[89] ——, *Domain Decomposition methods and deflated Krylov subspace iterations*, in European Conference on Computational Fluid Dynamics ECCOMAS CFD 2006, P. Wesseling, E. Onate, and J. Periaux, eds., TU Delft, Delft, 2006.

[90] O. Nevanlinna, *Convergence of Iterations for Linear Equations*, Birkhäuser Verlag, Basel, 1993.

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

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

[93] Y. Notay, *Combination of Jacobi-Davidson and conjugate gradients for the partial symmetric eigenproblem*, Numer. Linear Algebra Appl., 9 (2002), pp. 21–44.

[94] ——, *Convergence analysis of inexact Rayleigh quotient iteration*, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 627–644.

[95] ——, *Inner iterations in eigenvalue solvers*, 2005. Report GANMN 05-01, Université Libre de Bruxelles, Brussels, Belgium, `http://homepages.ulb.ac.be/ ynotay`.

[96] B. Nour-Omid, B. N. Parlett, T. Ericsson, and P. S. Jensen, *How to implement the sprectral transformation*, Math. Comp., 48 (1987), pp. 663–673.

[97] A. 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 (1957), pp. 233–241.

[98] C. Paige, B. N. Parlett, and H. van der Vorst, *Approximate solutions and eigenvalue bounds for Krylov subspaces*, Numer. Linear Algebra Appl., 2 (1995), pp. 115–133.

[99] C. C. Paige, *The computation of eigenvalues and eigenvectors of very large sparse matrices*, PhD thesis, University of London, 1971.

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

[101] ——, *The Symmetric Eigenvalue Problem*, vol. 20 of Classics in Applied Mathematics, SIAM, Philadelphia, PA, 1998. Corrected reprint of the 1980 original.

[102] C. Pearcy, *An elementary proof of the power inequality for the numerical radius*, Michigan Math. J., 13 (1966), pp. 289–291.

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

[104] M. Robbé, M. Sadkane, and A. Spence, *Inexact inverse subspace iteration with preconditioning applied to non-Hermitian eigenvalue problems*, 2006. Submitted to SIMAX.

[105] A. Ruhe, *Rational Krylov sequence methods for eigenvalue computations*, Lin. Alg. Appl., 58 (1984), pp. 391–405.

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

[107] Y. Saad, *On the rates of convergence of the Lanczos and Block-Lanczos methods*, SIAM J. Numer. Anal., 17 (1980), pp. 687–706.

[108] ——, *Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices*, Linear Alg. Appl., 34 (1980), pp. 269–295.

[109] ——, *Chebyshev acceleration techniques for solving nonsymmetric eigenvalue problems*, Math. Comp., 42 (1984), pp. 567–588.

[110] ——, *Numerical Methods for Large Eigenvalue Problems*, Halsted Press, New York, 1992.

[111] ——, *Iterative Methods for Sparse Linear Systems*, SIAM, Philadelphia, PA, 2nd ed., 2003.

[112] Y. Saad and M. Schultz, *GMRES a generalised minimum residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[113] R. Scheichl, *Parallel Solution of the Transient Multigroup Neutron Diffusion Equations with Multi-Grid and Preconditioned Krylov-Subspace Methods*, Master's thesis, Johannes Kepler Universität Linz, Austria, 1997.

[114] H. Schwetlick and R. Lösche, *A generalized Rayleigh quotient iteration for computing simple eigenvalues of nonnormal matrices*, ZAMM Z. angew. Math. Mech, 80 (2000), pp. 9–25.

[115] H. Schwetlick and K. Schreiber, *A primal-dual jacobi-davidson-like method for nonlinear eigenvalue problems*, preprint zih-ir-0613, Technische Universität Dresden, 2006.

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

[117] V. Simoncini, *Algebraic formulations for the solution of the nullspace-free eigenvalue problem using the inexact shift-and-invert Lanczos method*, Numer. Linear Algebra Appl., 10 (2003), pp. 357–375.

[118] ——, *Variable accuracy of matrix-vector products in projection methods for eigencomputation*, SIAM J. Numer. Anal., 43 (2005), pp. 1155–1174.

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

[120] V. Simoncini and D. B. Szyld, *Theory of inexact Krylov subspace methods and applications to scientific computing*, SIAM J. on Sci. Comput., 25 (2003), pp. 454–477.

[121] ——, *Relaxed Krylov subspace approximation*, PAMM, 5 (2005), pp. 797–800.

[122] ——, *Recent computational developments in Krylov subspace methods for linear systems*, Numer. Linear Algebra Appl., 14 (2007), pp. 1–59.

[123] G. L. G. Sleijpen, A. G. L. Booten, D. R. Fokkema, and H. A. van der Vorst, *Jacobi-Davidson type methods for generalized eigenproblems and polynomial eigenproblems*, BIT, 36 (1996), pp. 595–633. International Linear Algebra Year (Toulouse, 1995).

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

[125] ——, *The Jacobi-Davidson method for eigenvalue problems and its relation with accelerated inexact Newton schemes*, in Iterative methods in Linear Algebra, II., S. D. Margenov and P. S. Vassilevski, eds., vol. 3 of IMACS Series in Computational and Applied Mathematics, New Brunswick, NJ, U.S.A., 1996, IMACS, pp. 377–389. Proceedings of the Second IMACS International Symposium on Iterative Methods in Linear Algebra, June 17-20, 1995, Blagoevgrad.

[126] ——, *A Jacobi-Davidson Iteration Method for Linear Eigenvalue Problems*, SIAM Rev., 42 (2000), pp. 267–293.

[127] G. L. G. Sleijpen, H. A. van der Vorst, and E. Meijerink, *Efficient expansion of subspaces in the Jacobi-Davidson method for standard and generalized eigenproblems*, Electron. Trans. Numer. Anal., 7 (1998), pp. 75–89.

[128] P. Smit, *Numerical Analysis of Eigenvalue Algorithms based on Subspace Iterations*, PhD thesis, Katholieke Universiteit Brabant, 1997.

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

[130] D. C. Sorensen, *Implicit application of polynomial filters in a k-step Arnoldi method*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 357–385.

[131] ——, *Implicitly restarted Arnoldi/Lanczos methods for large scale eigenvalue calculations*, 1995. in Parallel Numerical Algorithms: Proceedings of an ICASE/LaRC Workshop, May 23-25, 1994, Hampton, VA, D. E. Keyes, A. Sameh, and V. Venkatakrishnan, eds., Kluwer.

[132] A. Stathopoulos and Y. Saad, *Restarting techniques for the (Jacobi-)Davidson symmetric eigenvalue methods*, Electron. Trans. Numer. Anal., 7 (1998), pp. 163–181.

[133] G. W. Stewart, *On the Sensitivity of the Eigenvalue Problem $Ax = \lambda Bx$*, SIAM J. Numer. Anal., 9 (1972), pp. 669–686.

[134] ——, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764.

[135] ——, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

[136] ——, *Matrix Algorithms II: Eigensystems*, SIAM, Philadelphia, PA, 2001.

[137] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.

[138] H. Symm and J. Wilkinson, *Realistic error bounds for a simple eigenvalue and its associated eigenvector*, Numer. Math., 35 (1980), pp. 113–126.

[139] D. B. Szyld, *Criteria for combining inverse and Rayleigh quotient iteration*, SIAM J. Numer. Anal., 25 (1988), pp. 1369–1375.

[140] ——, *On recurring theorems on the norm of oblique projections*, 2005. Research Report 05-12-30, Department of Mathematics, Temple University, December 2005.

[141] R. T. Tapia and D. L. Whitley, *The projected Newton method has order* $1+\sqrt{2}$ *for the symmetric eigenvalue problem*, SIAM J. Numer. Anal., 25 (1988), pp. 1376–1382.

[142] H. K. Thornquist, *Fixed-Polynomial Approximate Spectral Transformations for Preconditioning the Eigenvalue Problem*, PhD thesis, Rice University, Houston, Texas, 2006.

[143] L. N. Trefethen, *Approximation theory and numerical linear algebra*, in Algorithms for Approximation II, J. C. Mason and M. G. Cox, eds., Chapman and Hall, London, 1990.

[144] L. N. Trefethen and D. I. Bau, *Numerical Linear Algebra*, SIAM, Philadelphia, PA, 1997.

[145] L. N. Trefethen and M. Embree, *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*, Princeton University Press, 2005.

[146] H. Unger, *Nichtlineare Behandlung von Eigenwertaufgaben*, ZAMM Z. Angew. Math. Mech., 30 (1950), pp. 281–282.

[147] J. van den Eshof, *The convergence of Jacobi-Davidson for Hermitian eigenproblems*, Numer. Linear Algebra Appl., 9 (2002), pp. 163–179.

[148] J. van den Eshof and G. L. G. Sleijpen, *Inexact Krylov subspace methods for linear systems*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 125–153.

[149] D. S. Watkins, *Fundamentals of Matrix Computations*, Pure and Applied Mathematics. Wiley-Interscience [John Wiley & Sons], New York, 2nd ed., 2002.

[150] D. S. Watkins and L. Elsner, *Convergence of algorithms of decomposition type for the eigenvalue problem*, Linear Algebra Appl., 143 (1991), pp. 19–47.

[151] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, UK, 1965.

[152] K. Wu, A. Stathopoulos, and Y. Saad, *Inexact Newton preconditioning techniques for large symmetric eigenvalue problems*, Electron. Trans. Numer. Anal., 7 (1998), pp. 202–214.

[153] K.-B. Yu, *Recursive updating the eigenvalue decomposition of a covariance matrix*, IEEE Transactions on Signal Processing, 39 (1991), pp. 1136–1145.

INDEX

CONTRIBUTIONS OF THIS THESIS

Chapter 2 In this chapter we have

1. provided a convergence theory for inexact inverse iteration applied to a generalised eigenproblem $\mathbf{Ax} = \lambda\mathbf{Mx}$ with a special variable shift by showing its equivalence to modified Newton's method,
2. compared inexact Newton's method to a simplified version of Jacobi-Davidson method and obtained convergence results,
3. introduced a new preconditioner which yields both a fast (quadratic) outer convergence rate and small iteration numbers.

Chapter 3 In this chapter we have

1. provided a full convergence theory for inexact inverse iteration for fixed and variable shifts applied to the generalised eigenproblem $\mathbf{Ax} = \lambda\mathbf{Mx}$ with minimal assumptions on $\mathbf{A}$ and $\mathbf{M}$ by introducing a new convergence measure,
2. shown that convergence of inexact inverse iteration leads to an increase of the norm of the solution and hence no projection is necessary for inexact inverse iteration applied to a constraint eigenproblem,
3. compared inexact Rayleigh quotient iteration to a simplified version of Jacobi-Davidson method with Rayleigh quotient shift and inexact solves, shown that both methods are equivalent in a certain sense and hence provided convergence results for Jacobi-Davidson method.

Chapter 4 In this chapter we have

1. shown that inexact inverse iteration with a fixed shift applied to the Hermitian positive definite eigenproblem with iterative inner solves using MINRES/CG gives no growth in the number of inner iterations whereas the use of preconditioned MINRES leads to an increase of the iteration number,
2. introduced an ideal and a practical tuned preconditioner, which is a rank-one change of the standard preconditioner and leads to no increase in the number of inner iterations,

3. given interlacing and perturbation results for the eigenvalues of the system matrices arising from the tuned and the standard preconditioners,

4. provided numerical experiments for Rayleigh quotient iteration to show that the tuned preconditioner is also superior to both the standard preconditioner and a modified right hand side approach.

Chapter 5 In this chapter we have

1. described an equivalence result between the inexact simplified Jacobi-Davidson method using a standard preconditioner and inexact inverse iteration (with fixed and variable shifts) using a tuned version of the preconditioner when Galerkin-Krylov methods are used,

2. given numerical results for Hermitian and non-Hermitian case using preconditioned FOM and Lanczos (CG),

3. showed numerically that even for norm-minimising methods such as preconditioned GMRES and MINRES and for using the usual tuned preconditioner the approximate solutions are very close to each other.

Chapter 6 In this chapter we have

1. shown that for the generalised non-Hermitian eigenproblem inexact inverse iteration with a fixed shift and iterative inner solves with GMRES gives growth in the number of inner iterations no matter if preconditioned solves are used or not,

2. introduced a tuning operator, which is a rank-one change of the identity preconditioner and leads to no increase in the number of inner iterations,

3. investigated the tuned preconditioner for the generalised eigenproblem and proved that the number of inner iterations per outer iteration does not grow,

4. analysed the inner solve if the tuned preconditioner is used within inexact inverse iteration with variable shifts

5. gave a comparison to simplified Jacobi-Davidson method for generalised eigenproblem and numerical results

Chapter 7 In this chapter we have

1. extended the idea of a tuned preconditioner as a preconditioner with low rank change to Arnoldi's method with and without restarts,

2. supplied analysis of the tuned preconditioner and shown how tuning reduces the conditioner number of the system matrix,

3. extended the relaxation strategy for Arnoldi's method to implicitly restarted Arnoldi method,

4. provided numerical experiments which support the theory.