Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea Magistrale in Matematica

A new convergence model for the GMRES method

Tesi di Laurea in Analisi Numerica

Relatrice:
Chiar.ma Prof.ssa
Valeria Simoncini

Presentata da:
Giulia Sacchi

II Sessione Straordinaria
Anno Accademico 2015/2016

Contents

Notations

$\mathbb{F}^n$	n-dimensional vector space on the generic field $\mathbb{F}$
$\mathbb{R}$	Field of the real numbers
$\mathbb{C}$	Field of the complex numbers
$\langle \cdot, \cdot \rangle$	Natural inner product over $\mathbb{C}^n$
$\| \cdot \|$	2-norm of a vector or a matrix
A^T	Transpose of A
A^H	Conjugate transpose of A
n	Dimension of the space on which the considered linear system $Ax = b$ is defined
m	Dimension of the projected space
$e_1, \ldots, e_n$	Canonical Euclidean basis of $\mathbb{R}^n$
$\mathbb{P}_m$	Vector space of the polynomials of degree smaller or equal to m
$\mathbb{P}_m^*$	Vector subspace of the polynomials in $\mathbb{P}_m$ such that $p(0) = 1$
$\kappa(X)$	2-norm condition number of the matrix X
I_m	$m \times m$ identity matrix
$J_f(x)$	Jacobian matrix of the function f at point x
$\mathcal{H}_f(x)$	Hessian matrix of the function f at point x
blkdiag$(A_1, ..., A_n)$	Block diagonal matrix with matrices $A_1, \ldots, A_n$ on the diagonal

Introduzione

La necessità di risolvere sistemi lineari emerge continuamente da una vasta gamma di campi d'applicazione. Ad esempio, tali sistemi sono ottenuti nella trattazione numerica di equazioni differenziali. Sebbene in alcuni casi l'utilizzo di metodi diretti consente la risoluzione di problemi di dimensione anche notevole, i metodi iterativi risultano spesso la scelta migliore al fine di risolvere numericamente tali sistemi. In questo contesto, di particolare importanza sono i metodi sugli spazi di Krylov. Sebbene le loro potenzialità non siano state colte completamente nelle due decadi successive ai loro primi sviluppi, verso la fine degli anni '70 i metodi sugli spazi di Krylov per sistemi lineari e problemi agli autovalori si diffusero ampiamente e con successo in contesti scientifici ed ingegneristici.
L'idea principale dei metodi sugli spazi di Krylov, ed in generale dei processi di proiezione, è quella di trovare una soluzione approssimata di un sistema $Ax = b$ potenzialmente molto grande risolvendo un sistema di dimensione notevolmente inferiore, ottenuto attraverso la proiezione del sistema originale su un opportuno sottospazio. La quantità di informazione contenuta nel sistema originale che il processo di proiezione riesce a catturare si riflette nell'accuratezza dell'approssimazione della soluzione. I metodi sugli spazi di Krylov costruiscono una successione annidata di sottospazi, dando origine ad una successione di soluzioni approssimate, che tendono alla soluzione esatta. Quando il metodo è ben definito, se un sottospazio della successione dovesse racchiudere tutta l'informazione necessaria alla risoluzione del sistema, allora il processo di proiezione termina fornendo effettivamente la soluzione per $Ax = b$. Pertanto, i metodi sugli spazi di Krylov (ben definiti) sono *processi finiti*. Dal punto di vista computazionale, tuttavia, questi metodi dovrebbero essere considerati *iterativi*, come anche menzionato da Lanczos nel 1952 [9, p.40]:
"Anche se teoricamente parlando l'ultimo vettore (residuo) si annulla esattamente e solo dopo n iterazioni, è piuttosto probabile che esso scenda sotto valori trascurabili già dopo un numero relativamente piccolo di iterazioni."
Ad ogni modo, dal punto di vista matematico, la proprietà di terminazione finita è essenziale, in quanto influisce a livello pratico sul comportamento delle iterate.

Tra i metodi sugli spazi di Krylov, il metodo GMRES (Generalized Minimal RESidual) è adatto per risolvere sistemi non simmetrici. Basato sul processo di ortogonalizzazione di Arnoldi, fu inizialmente proposto da Saad e Schultz nel 1986 come una generalizzazione del metodo MINRES, usato per matrici simmetriche. Questo metodo è caratterizzato dalla proprietà di minimizzare la norma euclidea

del residuo sullo spazio di Krylov affine $x_0 + \mathcal{K}_m(A, r_0)$:

$$\|r_m\| = \min_{z \in x_0 + \mathcal{K}_m(A, r_0)} \|b - Az\|.$$

Uno svantaggio del GMRES è che, a causa delle sue ricorrenze di vettori lunghi, diventa intrattabile per problemi su cui converge troppo lentamente. Comprendere la convergenza di GMRES è dunque molto importante per rendere il suo utilizzo più agevole, ma anche per comprendere il comportamento di altri algoritmi non necessariamente ottimali.

Lo studio del comportamento dei metodi di Krylov porta a problemi non lineari complessi. Nonostante l'intensa attività di ricerca ed i numerosi risultati sia teorici che pratici ottenuti finora, sono ancora tanti i problemi aperti. La complessità di suddetti problemi invita ad approcci risolutivi che coinvolgono ipotesi aggiuntive, al fine di restringersi a determinati casi particolari, come ad esempio avere gli autovalori localizzati in certe aree, oppure avere la possibilità di isolare la parte responsabile del mal condizionamento della matrice del sistema A. Una motivazione a questa seconda ipotesi è la seguente: scrivendo A come $A = XJX^{-1}$, dove J denota la forma canonica di Jordan di A, la norma del residuo di GMRES è limitata da:

$$\|r_m\| \leq \kappa(X)\|r_0\| \min_{p \in \mathcal{P}_m^*} \|p(J)\|.$$

A causa dell'ordine di grandezza potenzialmente elevato, la presenza del termine $\kappa(X)$ spesso rende la stima totalmente priva di ogni utilità. Essere in grado di separare la parte ben condizionata di X da quella mal condizionata è un punto di partenza per fornire una stima più descrittiva.

In questo lavoro di tesi viene analizzato il particolare caso in cui A è diagonalizzabile, e quasi unitaria, e tuttavia è mal condizionata a causa di un piccolo gruppo di autospazi particolarmente vicini tra loro. Tenendo ben presente il carattere polinomiale del residuo di GMRES, ne viene studiata l'evoluzione all'avvicinarsi di tali autospazi. Vengono trattati entrambi i casi in cui i relativi autovalori sono lontani o vicini a loro volta. Quando quest'ultima eventualità si verifica, il polinomio di GMRES si comporta come se il metodo fosse stato applicato al sistema in cui A è sostituita da una matrice a blocchi di Jordan, avente al posto di tali autovalori vicini la loro media aritmetica.
Il principale risultato di questa tesi consiste nella formulazione di un problema di ottimizzazione vincolata, la cui soluzione non si limita a rappresentare una stima per la curva di convergenza di GMRES, ma ne descrive il comportamento, rivelandosi di fatto un nuovo modello di convergenza.

I primi due capitoli sono introduttivi e contestualizzano il metodo GMRES nell'ambito dei metodi di proiezione e ne descrivono le proprietà di convergenza conosciute. I Capitoli 3 e 4 contengono i risultati originali della tesi.
Nel Capitolo 1 vengono introdotti i metodi di proiezione, con particolare attenzione rivolta ai metodi sugli spazi di Krylov. Il metodo GMRES viene presentato nel dettaglio, attraverso la descrizione della sua derivazione ed implementazione.
Il Capitolo 2 raccoglie i risultati ad oggi conosciuti sull'analisi della convergenza

di GMRES. Cominciando dai risultati classici che coinvolgono i polinomi di Chebyshev, l'attenzione si sposta poi verso alcuni degli sviluppi più recenti.
Nel Capitolo 3 sono illustrati diversi risultati minori ed il nuovo modello di convergenza, il quale descrive efficacemente il comportamento di GMRES quando la matrice del sistema mostra mal condizionamento dovuto esclusivamente ad alcuni del suoi autovalori.
Il Capitolo 4 presenta esperimenti numerici a sostegno di quanto affermato nei risultati del Capitolo 3. Sono mostrati esempi su un'ampia gamma di proprietà spettrali delle matrici.

Introduction

The necessity of solving linear algebraic systems continuously arises from a wide range of application fields. For instance, such systems are derived from the discretization of differential equations. While in some cases, direct methods can be used to address fairly large problems, iterative methods are often the best choice for the numerical solution of such systems. In this context, an important role is played by Krylov subspace methods. Although their attractive features were not fully realized in the two decades following the first developments, from the late 1970s Krylov subspace methods for linear algebraic systems and eigenvalue problems became widely and successfully used throughout science and engineering.

The main idea of Krylov subspace methods, and in general of a projection process, is to find an approximate solution of a potentially very large system $Ax = b$ by solving a system of much smaller dimensionality, obtained by projecting the original system onto a suitable subspace. The magnitude of the information contained in the original data that is captured by the projections reflects into the accuracy of the approximate solution. Krylov subspace methods use a sequence of nested subspaces, thus giving a sequence of approximate solutions that converge to the exact solution. When the method is well defined, if some subspace eventually seizes all the information needed to solve the problem then the projection process terminates with the solution of $Ax = b$. Hence well-defined Krylov subspace methods are *finite processes*.
From a computational point of view, however, Krylov subspace methods should rather be considered *iterative*, as Lanczos also mentioned in 1952 [9, p. 40]:
"Even if theoretically speaking the last vector vanishes exactly only after n iterations, it is quite possible that it may drop practically below negligible bounds after a relatively few iterations."
Nevertheless, from a mathematical point of view, the finite termination property is substantial, for it affects the practical behavior of the iterates.

Among Krylov subspace methods, the Generalized Minimal RESidual method (GMRES) is suitable to address non-symmetric systems. Based on the Arnoldi process, it was first proposed by Saad and Schultz in 1986 as a generalization of the MINRES method (used for symmetric matrices). This method is characterized by the optimality property of minimization of the residual Euclidean norm over the affine Krylov subspace $x_0 + \mathcal{K}_m(A, r_0)$:

$$\|r_m\| = \min_{z \in x_0 + \mathcal{K}_m(A, r_0)} \|b - Az\|.$$

One problem with GMRES is that, due to its long vector recurrences, it becomes intractable for problems that converge slowly. Understanding GMRES conver-

gence is thus very important to make its usage easier, but also to understand the behavior of other non necessarily optimal algorithms.

The question about convergence behavior of Krylov subspace methods leads to complicated nonlinear problems. Although intense research efforts have been done to study these problems and a variety of theoretical and practical results have been produced, many answers have yet to be discovered. The complexity of the aforementioned problems invites to apply approaches that include additional hypotheses and restrict to relatively peculiar settings, like, for instance, the assumption of having clustered eigenvalues, or the eventuality in which it is somehow possible to isolate the responsible of ill-conditioning of the system matrix A. To motivate this second requirement, it is sufficient to observe that, writing A as $A = XJX^{-1}$, where J denotes the canonical Jordan form of A, GMRES residual can be bounded by

$$\|r_m\| \leq \kappa(X)\|r_0\| \min_{p \in \mathcal{P}_m^*} \|p(J)\|.$$

The main issue with this relation is the presence of the term $\kappa(X)$, whose potentially high magnitude may make the bound totally uninformative. Being able to separate the well-conditioned part of X from the ill-conditioned one has great chance to provide a more descriptive bound.

In this thesis we analyze the particular case in which the system matrix is diagonalizable, almost unitary, and yet it shows ill-conditioning due to the proximity of a small set of eigenspaces.
Keeping in mind the polynomial form of the GMRES residual, we study how it evolves as such eigenspaces become closer. We consider both the cases in which the corresponding eigenvalues are far from each other or close as well. In this last occurrence, the GMRES polynomial behaves like if the system matrix was replaced by a Jordan block matrix in which the close eigenvalues were substituted with their mean value.
The principal result of this thesis is the formulation of a constrained minimization problem whose solution provides a relation that not only bounds the GMRES convergence curve, but also predicts it quite in detail, revealing itself as a new convergence model.

The first two chapters are introductive. They contextualize the GMRES method as a projection method and describe the known convergence properties. Chapter 3 and 4 contain the original results of this thesis.
In Chapter 1 we describe the main features of the projection methods, with a particular focus on the Krylov subspace methods and a detailed presentation of the GMRES method, embracing its derivation and implementation.
Chapter 2 collects the known results about GMRES convergence analysis. We start from the classical results involving Chebyshev polynomials and then move towards some of the most recent developments.
In Chapter 3 we illustrate several minor results along a new convergence model, that efficiently describes GMRES behavior when the system matrix presents ill-conditioning due to just a part of its eigenvalues.

Chapter 4 presents numerical evidence to support the results of Chapter 3. It displays several examples over a wide range of spectral properties of the considered matrices.

Chapter 1

Projection Methods and Krylov Subspace Methods

1.1 Projection Methods

In a projection process that aims to solve a linear algebraic system $Ax = b$, with $A \in \mathbb{F}^{n\times n}$ and $b \in \mathbb{F}^n$, the approximate solution x_m at each step m is sought in the affine space

$$x_0 + \mathcal{S}_m, \tag{1.1}$$

where $x_0 \in \mathbb{F}^n$ is a given initial approximation to x and $\mathcal{S}_m$ in an m-dimensional subspace of $\mathbb{F}^n$, called the *search space*.
If A is non-singular, let x be the solution of the given linear system. The vector $x - x_m$ is called the *m-th error*, and using (1.1) it can be written as

$$x - x_m = x - x_0 - z_m, \qquad \text{for some } z_m \in \mathcal{S}_m.$$

Since $\mathcal{S}_m$ has dimension m, we have m degrees of freedom to construct x_m, therefore we generally need m constraints. These are imposed on the (computable) m-th residual, defined by:

$$\begin{aligned} r_m :=& b - Ax_m \\ =& b - A(x_0 + z_m) = (b - Ax_0) - Az_m \in r_0 + A\mathcal{S}_m, \end{aligned} \tag{1.2}$$

where the quantity $r_0 = b - Ax_0$ is called *initial residual*. In particular, we ask r_m to be orthogonal to a given m-dimensional subspace $\mathcal{C}_m$, the *constraints space*,

$$r_m \perp \mathcal{C}_m \qquad (\text{or equivalently } r_m \in \mathcal{C}_m^{\perp}). \tag{1.3}$$

From (1.2),

$$r_0 = Az_m + r_m, \qquad Az_m \in A\mathcal{S}_m, \quad r_m \in \mathcal{C}_m^{\perp}.$$

If $\mathbb{F}^n$ is the direct sum of $A\mathcal{S}_m$ and $\mathcal{C}_m^{\perp}$ ($\mathbb{F}^n = A\mathcal{S}_m \oplus \mathcal{C}_m^{\perp}$), then the corresponding vectors Az_m and r_m are uniquely determined as the projections of r_0 on the two mentioned subspaces.

We now want to consider a matrix representation of the projection process. Let S_m and C_m be two $n \times m$ matrices whose columns contain (arbitrary) bases for $\mathcal{S}_m$ and $\mathcal{C}_m$ respectively. Then (1.1) becomes

$$x_m = x_0 + S_m t_m$$

for some vector t_m, which is determined by imposing the orthogonality condition (1.3):

$$C_m^H r_m = 0 \Leftrightarrow C_m^H (b - Ax_m) = 0 \Leftrightarrow C_m^H r_0 - C_m^H A S_m t_m = 0,$$

therefore

$$C_m^H A S_m t_m = C_m^H r_0. \tag{1.4}$$

This is called the *projected system*, and the key idea of the projection approach for solving linear systems is to avoid dealing with the (possibly) large system $Ax = b$ by solving at step m of the projection process the projected system, which is of order m. Logically, the aim is to obtain a good approximation $x_m = x_0 + S_m t_m$ for $m \ll n$.

Definition 1.1. A projection process is said to be *well defined at step* m when the solution t_m is uniquely determined, i.e. when $C_m^H A S_m$ is non-singular.

We now give necessary and sufficient conditions in order to have well defined projection processes. These are summed up in the following two theorems (proofs can be found in [10]).

Theorem 1.1. *Let $A \in \mathbb{F}^{n\times n}$ and let $\mathcal{S}_m$ and $\mathcal{C}_m$ be two m-dimensional subspaces of $\mathbb{F}^n$, with bases represented by the columns of the matrices S_m and C_m, as defined above. Then*

$$C_m^H A S_m \text{ is non-singular } \Leftrightarrow \mathbb{F}^n = A\mathcal{S}_m \oplus \mathcal{C}_m^{\perp}.$$

This theorem implies that whether a projection process is well defined at step m depends only on the choices of the subspaces $\mathcal{S}_m$ and $\mathcal{C}_m$. In particular, if $\mathbb{F}^n = A\mathcal{S}_m \oplus \mathcal{C}_m^{\perp}$ then $C_m^H A S_m$ is non-singular for any choice of the bases in S_m and C_m. In this case (1.4) yields

$$t_m = (C_m^H A S_m)^{-1} C_m^H r_0,$$

and therefore

$$x_m = x_0 + S_m t_m = x_0 + S_m (C_m^H A S_m)^{-1} C_m^H r_0,$$

so that

$$\begin{aligned} r_m =& b - Ax_m = (b - Ax_0) - AS_m (C_m^H A S_m)^{-1} C_m^H r_0 \\ =& (I - AS_m (C_m^H A S_m)^{-1} C_m^H) r_0 \\ =& (I - P_m) r_0, \end{aligned} \tag{1.5}$$

where

$$P_m = AS_m(C_m^H AS_m)^{-1}C_m^H$$

is a projector, since $P_m^2 = P_m$. It holds

$$P_m v \in A\mathcal{S}_m \quad \text{and} \quad (I - P_m)v \in \mathcal{C}_m^{\perp} \qquad \forall v \in \mathbb{F}^n,$$

which means that P_m projects onto $A\mathcal{S}_m$ and orthogonally to $\mathcal{C}_m$.
From (1.5) we derive that r_0 can be decomposed as follows:

$$r_0 = P_m r_0 + r_m.$$

Of particular interest for us is the case in which $A\mathcal{S}_m = \mathcal{C}_m$, as it corresponds to the *Krylov subspace methods* known as *Minimal Residual methods*. Under these conditions the projection process is said to be *orthogonal*. Moreover, the orthogonal decomposition of r_0 yields the following result ($\|\cdot\|$ denotes the Euclidean norm):

$$\|r_0\|^2 = \|P_m r_0 + r_m\|^2 = \|P_m r_0\|^2 + \|r_m\|^2 \geq \|r_m\|^2,$$

i.e. for the Minimal Residual methods the sequence of the residual norms is non-increasing.

Remark 1.1. The m-th approximation x_m solves the system $Ax = b$ if and only if $r_m = b - Ax_m = 0$. This happens (when the process is well defined at step m) if ad only if $r_0 = P_m r_0$ (see (1.5)), that is $r_0 \in A\mathcal{S}_m$. A sufficient condition for this is, for instance, to have $r_0 \in \mathcal{S}_m$ and $A\mathcal{S}_m = \mathcal{S}_m$.

A projection process has the *finite termination property* when it finds the solution of the given algebraic linear system in a finite number of steps. From (1.1) we can guess it may be useful to start the projection process with the search space $\mathcal{S}_1 = span\{r_0\}$ and to proceed building up a *nested sequence of search spaces*:

$$\mathcal{S}_1 \subset \mathcal{S}_2 \subset \mathcal{S}_3 \subset \dots$$

such that at some step m the relation $A\mathcal{S}_m = \mathcal{S}_m$ is satisfied. This idea leads to Krylov subspaces, which we will treat in the next section.

1.2 Krylov Subspaces

Krylov subspaces were first studied by Krylov who was interested in finding a method for computing the minimal polynomial of a matrix. Such polynomial, along with the minimal polynomial of a vector with respect to a matrix, play an important role in the convergence analysis of Krylov subspace methods.

Definition 1.2. Given $A \in \mathbb{F}^{n\times n}$ and a non-zero vector $v \in \mathbb{F}^n$, the *Krylov sequence generated by A and b* is

$$v,\ Av,\ A^2v,\ \dots$$

There exists a unique natural number $d = d(A, v)$ such that the vectors v,Av,..., $A^{d-1}v$ are linearly independent and v,Av,..., $A^{d-1}v$, A^dv are linearly dependent. It holds that $1 \leq d \leq n$, since v in non-zero and $n+1$ vectors of $\mathbb{F}^n$ must be linearly dependent.

By construction there exist coefficients $\gamma_0, \ldots, \gamma_{d-1} \in \mathbb{F}$ so that

$$A^d v = \sum_{j=0}^{d-1} \gamma_j A^j v$$

Thus, setting

$$p(z) := z^d - \sum_{j=0}^{d-1} \gamma_j z^j \tag{1.6}$$

we have $p(A)v = 0$.

Definition 1.3. The polynomial $p(z)$ in (1.6) is called the *minimal polynomial of v with respect to A*, and its degree $d(A, v)$ is the *grade of v with respect to A*. The *minimal polynomial of the matrix A* is the non-zero monical polynomial p of lowest degree such that $p(A) = 0$ (i.e. $p(A)v = 0$, $\forall v$).

Remark 1.2. The minimal polynomial of v with respect to A is the non-zero monical polynomial p of lowest degree such that $p(A)v = 0$.

Remark 1.3. It is easy to prove that the grade of any eigenvector v of A with respect to A is 1, while the grade of a generalized eigenvector is at most the dimension of the invariant subspace it belongs to (see [10, Section 4.2]).

Remark 1.4. If v is a basis vector for an invariant subspace under A, its minimal polynomial with respect to A is a divisor of the minimal polynomial of A (again, see [10, Section 4.2]).

Definition 1.4. Given any matrix A and vector v, the *m-th Krylov subspace generated by A and v* ($m = 0, 1, 2, \ldots$) is

$$\mathcal{K}_m(A, v) = span\{v, Av, A^2v, \ldots, A^{m-1}v\}, \qquad m = 1, 2, \ldots.$$

For convention, $\mathcal{K}_0(A, v) = 0$.

It is immediate to see that these spaces form a nested sequence, $\mathcal{K}_m(A, v) \subseteq \mathcal{K}_{m+1}(A, v)$ $\forall m$. Moreover, the existence of the minimal polynomial of v with respect to A (of degree d) implies that

$$A\mathcal{K}_d(A, v) \subseteq \mathcal{K}_d(A, v),$$

namely, such sequence will at some point become invariant under A.

We now present a few basic facts about Krylov subspaces in a lemma:

Lemma 1.2. *Let A be a square matrix and let v be a vector of grade $d \geq 1$ with respect to A. Then*

(i) $\dim \mathcal{K}_m(A, v) = m$ *for* $m = 1, \ldots, d$.

(ii) $d - 1 \leq \dim A\mathcal{K}_d(A, v) \leq d$ *and, if* A *is non-singular, then* $A\mathcal{K}_d(A, v) = \mathcal{K}_d(A, v)$.

(iii) If $A\mathcal{K}_d(A, v) = \mathcal{K}_d(A, v)$ *then* $v \in$*Range*(A).

Proof. (i) By definition of grade of v with respect to A, the vectors $v, Av, \ldots, A^{m-1}v$ are linearly independent for $m = 1, \ldots, d$.
(ii) The $d - 1$ vectors $Av, \ldots, A^{d-1}v \in A\mathcal{K}_d(A, v) \subseteq \mathcal{K}_d(A, v)$ are linearly independent, so $d - 1 \leq \dim A\mathcal{K}_d(A, v) \leq d$.
Now, let A be non-singular. We show that $Av, \ldots, A^d v$ are linearly independent, providing $\dim A\mathcal{K}_d(A, v) = \dim \mathcal{K}_d(A, v)$ and hence $A\mathcal{K}_d(A, v) = \mathcal{K}_d(A, v)$. Suppose

$$\sum_{j=1}^{d} \|r_0\|_j A^j v = 0.$$

Left multiplication by A^{-1} on both sides yields

$$\sum_{j=1}^{d} \|r_0\|_j A^{j-1} v = 0$$

and thus $\|r_0\|_1 = \ldots = \|r_0\|_d = 0$, due to the linear independence of the vectors in this latter sum. But then $Av, \ldots, A^d v$ are linearly independent, too.
(iii) $v \in \mathcal{K}_d(A, v) = A\mathcal{K}_d(A, v) \subseteq$Range$(A)$. □

These facts result in the following theorem:

Theorem 1.3. *Consider the projection process (1.1)-(1.3) and let the search spaces be defined as the Krylov subspaces generated by* A *and* r_0*:*

$$\mathcal{S}_m = \mathcal{K}_m(A, r_0), \qquad m = 1, 2, \ldots$$

If r_0 *is of grade* d *with respect to* A*, then* $r_0 \in \mathcal{S}_1 \subset \mathcal{S}_2 \subset \ldots \subset \mathcal{S}_d = \mathcal{S}_{d+j}$, $\forall j \geq 0$.
Moreover, if A *is non-singular and the projection process is well defined at step* d*, then* $r_d = 0$.

Proof. Only the last statement needs to be proved. If A is non-singular then point (ii) of Lemma 1.2 ensures that $A\mathcal{S}_d = \mathcal{S}_d$, and if in addition the projection process is well defined at step d, we have $r_d = 0$, as already observed in the previous section. □

This theorem gives insight into the importance of invariant subspaces in the context of Krylov subspaces, as it reveals that when the Krylov subspace has become invariant under A, the projection process ceases with a zero residual. Indeed, it is sufficient to note the intrinsic presence of the power method in the building process of a Krylov subspace to intuitively understand that Krylov subspaces tend to contain the dominant information of A with respect to r_0. The main idea is to closely approximate an A-invariant subspace as quickly as possible.

1.2.1 Some Krylov subspace Methods

A common property of all Krylov subspace methods is the *orthogonality* condition (1.3). This property can sometimes be related to a certain *optimality* property, which is a starting point for the investigation of convergence behavior of specific methods. The following theorem gives the mathematical description of several important Krylov subspace methods in term of the search and constraints spaces, and optimality properties. A complete proof can be found in [10].

Theorem 1.4. [10, Theorem 2.3.1] *Consider the projection process (1.1)-(1.3) for solving a linear algebraic system $Ax = b$, with initial approximation x_0. Let the initial residual $r_0 = b - Ax_0$ be of grade $d \geq 1$ with respect to A. Then*

(i) If A is HPD and $\mathcal{S}_m = \mathcal{C}_m = \mathcal{K}_m(A, r_0)$, $m = 1, 2, \ldots$, then the projection is well defined at every step m until it terminates at step d. It is characterized by the orthogonality property

$$x - x_m \perp_A \mathcal{K}_m(A, r_0), \quad \text{or also} \quad x - x_m \in \mathcal{K}_m(A, r_0)^{\perp_A},$$

where

$$\mathcal{K}_m(A, r_0)^{\perp_A} := \{w \in \mathbb{F}^n \ s.t. \ (v, w)_A := w^H A v = 0, \ \forall v \in \mathcal{K}_m(A, r_0)\}.$$

The equivalent optimality property is

$$\|x - x_m\|_A = \min_{z \in x_0 + \mathcal{K}_m(A, r_0)} \|x - z\|_A,$$

where $\|v\|_A := \langle v, v \rangle_A^{1/2}$ is the A-norm of the vector v.
(Mathematical characterization of the Conjugate Gradient (CG) method)

(ii) If A is Hermitian and non-singular, $\mathcal{S}_m = A\mathcal{K}_m(A, r_0)$, and $\mathcal{C}_m = A^{-1}\mathcal{S}_m = \mathcal{K}_m(A, r_0)$, $m = 1, 2, \ldots$, then the projection is well defined at every step m until it terminates at step d. It is characterized by the orthogonality property

$$x - x_m \perp A\mathcal{K}_m(A, r_0).$$

The equivalent optimality property is

$$\|x - x_m\| = \min_{z \in x_0 + \mathcal{K}_m(A, r_0)} \|x - z\|.$$

(Mathematical characterization of the SYMMLQ method)

(iii) If A is non-singular, $\mathcal{S}_m = \mathcal{K}_m(A, r_0)$, and $\mathcal{C}_m = A_m^{\mathcal{S}} = A\mathcal{K}_m(A, r_0)$, $m = 1, 2, \ldots$, then the projection is well defined at every step m until it terminates at step d. It is characterized by the orthogonality property

$$r_m \perp A\mathcal{K}_m(A, r_0), \quad \text{or also} \quad x - x_m \in \mathcal{K}_m(A, r_0)^{\perp_{A^H A}}.$$

The equivalent optimality property is

$$\|r_m\| = \min_{z \in x_0 + \mathcal{K}_m(A, r_0)} \|b - Az\|, \quad \text{or}$$

$$\|x - x_m\|_{A^H A} = \min_{z \in x_0 + \mathcal{K}_m(A, r_0)} \|x - z\|_{A^H A}.$$

(Mathematical characterization of the Minimal Residual (MINRES) method and the Generalized Minimal Residual (GMRES) method)

Remark 1.5. Since the minimization problems in Theorem 1.4 are defined over affine spaces of increasing dimensions, the corresponding sequences of norms, $\|x - x_m\|_A$, $\|x - x_m\|$ and $\|r_m\|$, are *non-increasing* for $m = 0, 1, 2, \ldots, d$.

1.3 Arnoldi Algorithm and the GMRES Method

While from the mathematical point of view each Krylov subspace method is completely determined by the choice of the search and the constraint spaces, for what concerns, instead, the numerical behavior of the methods, the choice of the bases for these spaces is fundamental. By construction, the vectors of the Krylov sequence, $v, Av, \ldots, A^m v, \ldots$, converge towards a dominant eigenvector of A, and thus they will eventually become closer and closer as m grows, leading to loss of information. For this main reason, it is not wise to choose them as a basis for $\mathcal{K}_m(A, v)$. In order to preserve as much information as possible from the original linear system it is advisable to use well-conditioned (and possibly orthonormal) bases for the Krylov subspaces.

Remark 1.6. As shown below, the computation of orthonormal bases for the Krylov subspaces is related to orthogonal (or unitary) transformations of the matrix A. Since in many cases the data in A are affected by errors, these are not amplified by the transformations in any unitarily invariant norm.

From now on, we consider the generic field $\mathbb{F}$ to be $\mathbb{C}$.

1.3.1 Arnoldi's Orthogonalization Algorithm

The Arnoldi algorithm can be seen as a variant of the Gram-Schmidt orthogonalization method applied to the Krylov sequence in order to generate an orthonormal basis for the Krylov subspace. It has been first introduced in 1951 as a means to reduce a dense matrix in Hessenberg form.
The algorithm produces the orthonormal basis of $\mathcal{K}_m(A, v)$, $v_1, \ldots, v_d$ (where d is the grade of v with respect to A), applying, at each step m, the Gram-Schmidt orthogonalization to Av_{m-1} instead of $A^{m-1}v$. Due to numerical instability reasons, the classical Gram-Schmidt implementation is rarely used. A very common implementation in practical computations is the *modified Gram-Schmidt orthogonalization*, which is mathematically equivalent, that is, in absence of rounding errors the two procedures build identical vectors.

Algorithm 1.1. *Arnoldi's algorithm, modified Gram-Schmidt implementation*

1. Define $v_1 = v/\|v\|$.
2. For $j = 1, 2, \ldots$
3. $\quad w_j := Av_j$
4. $\quad$ For $i = 1, \ldots, j$
5. $\quad\quad h_{i,j} = \langle w_j, v_i \rangle$

6. $w_j = w_j - h_{i,j} v_i$
7. End
8. $h_{j+1,j} = \|w_j\|_2$. If $h_{j+1,j} = 0$ then stop.
9. $v_{j+1} = w_j / h_{j+1,j}$
10. End

Proposition 1.5. *Arnoldi's algorithm stops at step j (i.e. $h_{j+1,j} = 0$)* if and only if $j = d = d(A, v)$ *(grade of v with respect to A).*
Furthermore, $\forall m \leq d$ the computed vectors $v_1, \ldots, v_m$ form an orthonormal basis of the space $\mathcal{K}_m(A, v_1) = span\{v_1, Av_1, \ldots, A^{m-1}v_1\}$.

Proof. Let us first consider a given $m \leq d$, and show that the vectors $v_1, \ldots, v_m$ are an orthonormal basis of $\mathcal{K}_m(A, v_1)$. Such vectors are orthonormal by construction, so we only need to show they belong to $\mathcal{K}_m(A, v_1)$. In order to prove this, it is sufficient to point out that each v_j, $j = 1, \ldots, m$ is of the form $v_j = q_{j-1}(A)v_1$, where q_{j-1} is a polynomial of degree $j - 1$. By induction on j:
If $j = 1$: $v_1 = q_0(A)v_1$, with $q_0(z) = 1$;
If $1 < j < m$:

$$v_{j+1} = \frac{Av_j - \sum_{i=1}^{j} h_{i,j} v_i}{\left\| Av_j - \sum_{i=1}^{j} h_{i,j} v_i \right\|}.$$

Call c the denominator and, by inductive hypothesis, write v_j in polynomial form, $v_j = q_{j-1}(A)v_1$:

$$v_{j+1} = \frac{1}{c} \left(Aq_{j-1}(A)v_1 - \sum_{i=1}^{j} h_{i,j} q_{i-1}(A)v_1 \right) = \frac{1}{c} q_j(A)v_1.$$

We now prove the first statement. If $j = d$, w_d must be 0 after all the subtractions in line 6 of Algorithm 1.1, otherwise it would be possible to define a vector v_{d+1}, linearly independent of the previously generated ones. But this would mean $\dim \mathcal{K}_{d+1}(A, v) = d + 1$, that contradicts the hypothesis of d to be grade of v with respect to A. □

The next proposition describes the Arnoldi orthogonalization process in a matrix form (note that the notation here introduced will be adopted also in the rest of this thesis):

Proposition 1.6. *Consider $m \in \{1, \ldots, d\}$, and define:*

- *V_m: the $n \times m$ matrix whose columns are $v_1, \ldots, v_m$, i.e. the orthonormal basis of $\mathcal{K}_m(A, v_1)$:*

$$V_m := [v_1,\, v_2,\, \ldots,\, v_m],$$

- $\underline{H}_m$*: the* $(m+1) \times m$ *upper Hessenberg matrix defined by the elements* $h_{i,j}$ *in Algorithm 1.1,*

$$\underline{H}_m := \begin{pmatrix} h_{1,1} & h_{1,2} & \dots & \dots & h_{1,m} \\ h_{2,1} & h_{2,2} & & & h_{2,m} \\ 0 & h_{3,2} & \ddots & & \vdots \\ & & \ddots & \ddots & \vdots \\ & & & h_{m,m-1} & h_{m,m} \\ 0 & & & 0 & h_{m+1,m} \end{pmatrix}$$

$$= \begin{pmatrix} \langle Av_1, v_1\rangle & \langle Av_2, v_1\rangle & \dots & \dots & \langle Av_m, v_1\rangle \\ \|w_1\| & \langle Av_2, v_2\rangle & & & \langle Av_m, v_2\rangle \\ 0 & \langle Av_2, v_3\rangle & \ddots & & \vdots \\ & & \ddots & \ddots & \vdots \\ & & & \|w_{m-1}\| & h\langle Av_m, v_m\rangle \\ 0 & & & 0 & \|w_m\| \end{pmatrix}$$

- H_m*: the* $m \times m$ *upper Hessenberg matrix obtained from* $\underline{H}_m$ *deleting the last row.*

Then it holds:

$$(i) \qquad AV_m = V_m H_m + h_{m+1,m} v_{m+1} e_m^H = V_{m+1} \underline{H}_m \,, \tag{1.7}$$

$$(ii) \qquad V_m^H A V_m = H_m. \tag{1.8}$$

Proof. (i): It follows from the matrix reformulation of Algorithm 1.1, lines $4, 5, 7$. For $j = 1, \dots, m-1$:

$$v_{j+1} = \frac{1}{h_{j+1,j}} \left(Av_j - \sum_{i=1}^{j} h_{i,j} v_i \right),$$

thus, isolating Av_j:

$$Av_j = h_{j+1,j} v_{j+1} + \sum_{i=1}^{j} h_{i,j} v_i = \sum_{i=1}^{j+1} h_{i,j} v_i, \qquad j = 1, \dots, m-1.$$

For $j = m$:

$$Av_m = \sum_{i=1}^{m+1} h_{i,m} v_i = \sum_{i=1}^{m} h_{i,m} v_i + h_{m+1,m} v_{m+1}.$$

Therefore $AV_m = V_m H_m + h_{m+1,m} v_{m+1} e_m^H = V_{m+1} \underline{H}_m$.
(ii): From before, with a left multiplication by V^H applied to

$$AV_m = V_m H_m + h_{m+1,m} v_{m+1} e_m^H,$$

we obtain (remember that $V_{m+1} = [v_1, \dots, v_{m+1}]$ is an orthogonal matrix):

$$V_m^H A V_m = V_m^H V_m H_m + h_{m+1,m} V_m^H v_{m+1} e_m^H = H_m.$$

□

Remark 1.7. Note that, when $h_{m+1,m} = 0$, from (i) we have $AV_m = V_m H_m$, which means that the space spanned by the orthonormal columns of V_m is invariant under A.

1.3.2 Deriving the GMRES Method

We will now focus on the GMRES method. Let $A \in \mathbb{C}^{n\times n}$ be a general non-singular matrix. As already stated in Theorem 1.4, this method implements the projection process characterized by:

$$x_m \in x_0 + \mathcal{K}_m(A, r_0) \quad \text{and} \quad r_m \perp A\mathcal{K}_m(A, r_0), \tag{1.9}$$

with optimality property

$$\|r_m\| = \min_{z\in x_0+\mathcal{K}_m(A,r_0)} \|b - Az\|.$$

The method can be implemented through the application of the Arnoldi algorithm to A, with initial vector $v = r_0$ and the use of the residual minimization property. More precisely, naming $V_m := [v_1, \ldots, v_m]$ the orthonormal basis of $K_m(A, r_0)$ provided by the orthogonalization process, we have that each vector of $x_0 + \mathcal{K}_m(A, r_0)$ can be written as:

$$x = x_0 + V_m y, \qquad y \in \mathbb{C}^m.$$

Here m is always smaller than d. The approximate solution $x_m \in x_0 + \mathcal{K}_m(A, r_0)$ is then determined by imposing the minimization of the residual 2-norm. We have

$$\begin{aligned} b - Ax =& b - Ax_0 - AV_m y \qquad (\text{using } (1.7)) \\ =& r_0 - V_{m+1}\underline{H}_m y = \|r_0\|v_1 - V_{m+1}\underline{H}_m y \\ =& \|r_0\|V_{m+1}e_1 - V_{m+1}\underline{H}_m y \\ =& V_{m+1}(\|r_0\|e_1 - \underline{H}_m y). \end{aligned}$$

Therefore

$$\|b - Ax\| = \|V_{m+1}(\|r_0\|e_1 - \underline{H}_m y)\| = \big\|\|r_0\|e_1 - \underline{H}_m y\big\|;$$

and it will be

$$x_m = x_0 + V_m y_m, \qquad y_m = \text{argmin}_{y\in\mathbb{C}^m}\big\|\|r_0\|e_1 - \underline{H}_m y\big\|. \tag{1.10}$$

Thanks to the fact that the matrix $\underline{H}_m \in \mathbb{C}^{(m+1)\times m}$ has full rank m, y_m (and hence the solution x_m) is uniquely determined by

$$y_m = (\underline{H}_m^H \underline{H}_m)^{-1}\underline{H}_m^H(\|r_0\|e_1),$$

where the matrix $(\underline{H}_m^H \underline{H}_m)^{-1}\underline{H}_m^H$ is the *Moore-Penrose pseudoinverse* of $\underline{H}_m$.

In practice y_m is obtained by solving the least squares problem via the QR factorization, as described in the next section.

1.3.3 GMRES practical implementation issues

Being GMRES a Krylov subspace method, its basic idea is to try to solve a small projected system instead of the (possibly) large given one. At each step m, the considered Krylov subspace is expanded by the addition of a new vector in the basis, and the least squares problem in (1.10) is built. The algorithm terminates at step m when the stopping criterion is matched by the corresponding residual norm $\|r_m\|$. An interesting fact is that it is possible to determine how good the m-th approximation is without the need to explicitly compute x_m, nor r_m, at each step.
Keeping this in mind, here we discuss how to address the least squares problem $\min_y \left\|\|r_0\|e_1 - \underline{H}_m y\right\|$ in order to obtain the m-th residual norm while avoiding unnecessary computations.

To individuate the minimizer y_m, it is natural to transform the upper Hessenberg matrix $\underline{H}_m$ into its upper triangular form, through a QR factorization. Due to its structure, $\underline{H}_m$ can be efficiently treated by the application of Givens planar rotations. These operations are represented by $(m+1)\times(m+1)$ matrices

$$\Omega_i = \begin{bmatrix} I_{i-1} & & & \\ & c_i & s_i & \\ & -s_i & c_i & \\ & & & I_{m-i} \end{bmatrix}, \qquad c_i^2 + s_i^2 = 1, \quad i = 1, \dots, m.$$

Coefficients c_i and s_i are determined sequentially, in order to annihilate the non-zero elements under the diagonal, i.e. imposing $(\Omega_i \underline{H}_m^{(i-1)})_{i+1,i} = 0$, for each $i = 1, \dots, m$. Here $\underline{H}_m^{(i-1)} := \Omega_{i-1} \dots \Omega_1 \underline{H}_m$. It is then easy to prove that:

$$s_i = \frac{h_{i+1,i}}{\sqrt{\left(h_{i,i}^{(i-1)}\right)^2 + h_{i+1,i}^2}}, \quad c_i = \frac{h_{i+1,i}^{(i-1)}}{\sqrt{\left(h_{i,i}^{(i-1)}\right)^2 + h_{i+1,i}^2}}. \tag{1.11}$$

Therefore, defining

$$\begin{aligned} Q_m &:= \Omega_m \Omega_{m-1} \dots \Omega_1 \\ \underline{R}_m &:= \underline{H}_m^{(m)} = Q_m \underline{H}_m \\ \underline{g}_m &= (\gamma_1, \dots, \gamma_m)^T := Q_m(\|r_0\|e_1) \end{aligned}$$

we have:

$$\min_y \left\|\|r_0\|e_1 - \underline{H}_m y\right\| = \min_y \left\|Q_m(\|r_0\|e_1 - \underline{H}_m y)\right\| = \min_y \|\underline{g}_m - \underline{R}_m y\|.$$

The solution y_m is thus given by solving the upper triangular system $R_m y = g_m$, where R_m and g_m are obtained by removing the last row of $\underline{R}_m$ (which is a row of zeros) and the last entry of $\underline{g}_m$ respectively. In particular, since

$$\|\underline{g}_m - \underline{R}_m y\|^2 = |\gamma_{m+1}|^2 + \|g_m - R_m y\|^2,$$

it holds

$$\min_{y \in \mathbb{C}^m} \|\underline{g}_m - \underline{R}_m y\| = |\gamma_{m+1}|.$$

Remark 1.8. Even though $\underline{H}_m$ grows one row and one column each step, from the discussion above in the previous paragraph we deduce that the QR factorization of $\underline{H}_{m+1}$ does not need to be recomputed from scratch every time. Indeed, it is sufficient to just update the QR factorization of $\underline{H}_m$ by expanding it with the last column of $\underline{H}_{m+1}$ and subsequently applying the Given rotations $\Omega_1, \Omega_2, \ldots, \Omega_{m+1}$ to the last two columns. This means that the cost of the QR factorization process on the fly is very limited.

Next proposition summarizes the observations above.

Proposition 1.7. *Consider Ω_i, $i = 1, \ldots, m$, $\underline{H}_m$, $\underline{R}_m$, R_m, $\underline{R}_m$, g_m, defined as before. It holds:*

(i) rank(AV_m) =rank(R_m). In particular, if $(R_m)_{m,m} = 0$ then A is singular.

(ii) The minimizing vector y_m of $\|\|r_0\|e_1 - \underline{H}_m y\|$ is given by $y_m = R_m^{-1} g_m$.

(iii) The GMRES *m-th residual vector satisfies:*

$$r_m = b - Ax_m = V_{m+1}(\|r_0\|e_1 - \underline{H}_m y_m) = V_{m+1} Q_m^H(\gamma_{m+1} e_1)$$

and hence

$$\|r_m\| = |\gamma_{m+1}| = \|r_0\| \left| e_{m+1}^H Q_m e_1 \right|.$$

Proof. (i). From (1.7):

$$\begin{aligned} AV_m =& V_{m+1}\underline{H}_m \\ =& V_{m+1} Q_m^H (Q_m \underline{H}_m) \\ =& V_{m+1} Q_m^H \underline{R}_m \end{aligned}$$

and since $V_{m+1}Q_m^H$, being orthogonal, has full rank, then

$$\text{rank}(AV_m) = \text{rank}(\underline{R}_m) = \text{rank}(R_m).$$

Now, if $(R_m)_{m,m} = 0$ then rank(AV_m) =rank$(R_m) \leq m-1$, but V_m has full rank (equal to m), thus A must be singular.
(ii). The result has already been proved in the discussion that anticipates the proposition. Note that all the steps are well posed if A is non-singular: from (i) $(R_m)_{m,m} \neq 0$ and R_m^{-1} exists.
(iii). If $x_m = x_0 + V_m y_m$, $y_m = R_m^{-1} g_m$, then

$$\begin{aligned} r_m =& b - Ax_m = r_0 - AV_m y_m = \|r_0\| v_1 - V_{m+1}\underline{H}_m y_m \\ =& V_{m+1}(\|r_0\|e_1 - \underline{H}_m y_m) = V_{m+1} Q_m^H Q_m (\|r_0\|e_1 - \underline{H}_m y_m) \\ =& V_{m+1} Q_m^H (\underline{g}_m - \underline{R}_m y_m) = V_{m+1} Q_m^H \left(\begin{bmatrix} g_m \\ \gamma_{m+1} \end{bmatrix} - \begin{bmatrix} R_m y_m \\ 0 \end{bmatrix} \right) \\ =& V_{m+1} Q_m^H (\gamma_{m+1} e_{m+1}), \end{aligned}$$

and therefore $\|r_m\| = |\gamma_{m+1}|$. Moreover, since $\underline{g}_m = Q_m(\|r_0\|e_1)$, then $\gamma_{m+1} = \|r_0\| e_{m+1}^H Q_m e_1$. □

Remark 1.9. Thanks to this approach, the m-th residual norm is implicitly computed at each step ($\|r_m\| = |\gamma_{m+1}|$), without the need to solve the system $R_m y_m = g_m$, which will be addressed only when $\|r_m\|$ is small enough.

We are now ready to give the algorithm.

Algorithm 1.2. *Basic* GMRES *Algorithm*

1. Define $r_0 = b - Ax_0$, $\beta = \|r_0\|$, $v_1 = r_0/\beta$, $m = 0$.
2. While ($\|r_m\| > tol$ & $m \leq n$):
3. $m = m + 1$
4. $w_m := Av_m$
5. For $i = 1, \ldots, m$
6. $h_{i,m} = \langle w_m, v_i \rangle$
7. $w_m = w_m - h_{i,m} v_i$
8. End
9. $h_{m+1,m} = \|w_m\|$. If $h_{m+1,m} = 0$, go to *line 14*.
10. $v_{m+1} = w_m / h_{m+1,m}$
11. Compute the Givens rotation coefficients s_m and c_m.
12. Update R_m and Q_m in QR factorization.
13. Compute γ_{m+1} and set $\|r_m\| := |\gamma_{m+1}|$
14. End
15. Compute $y_m = \text{argmin}_y \|\beta e_1 - \underline{H}_m y\|$
16. $x_m = x_0 + V_m y_m$

Observe that, in exact arithmetic, GMRES may break down in *line 9* of Algorithm 1.2, that is when $h_{m+1,m} = 0$. However (in exact arithmetic) this is called a *lucky breakdown*, because in this case the exact solution is found, as stated in the next proposition.

Proposition 1.8 (Breakdown of GMRES)**.** *Let A be a non-singular matrix. Then*

GMRES *stops at step m (i.e. $h_{m+1,m} = 0$) $\Leftrightarrow$ Approximate solution x_m is exact.*

Proof. ($\Rightarrow$) Remember that $\gamma_{m+1}^{(m)}$ (the superscript indicates "at step m") is built using Givens rotations over $\|r_0\|e_1 \in F^m$, more precisely

$$\underline{g}_m^{(m)} = \begin{pmatrix} \gamma_1^{(m)} \\ \vdots \\ \gamma_m^{(m)} \\ \gamma_{m+1}^{(m)} \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \Omega_m \dots \Omega_1(\|r_0\|e_1) = \Omega_m \begin{pmatrix} \gamma_1^{(m-1)} \\ \vdots \\ \gamma_m^{(m-1)} \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} \gamma_1^{(m-1)} \\ \vdots \\ c_m\gamma_m^{(m-1)} \\ -s_m\gamma_m^{(m-1)} \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

therefore, using item (*iii*) of Theorem Proposition 1.7 and (1.11)

$$\|r_m\| = |\gamma_{m+1}| = |s_m\gamma_m^{(m-1)}| = \frac{|h_{m+1,m}|}{\sqrt{\left(h_{m,m}^{(m-1)}\right)^2 + h_{m+1,m}^2}}|\gamma_m| = 0.$$

($\Leftarrow$) Conversely, if x_m is exact (and x_{m-1} is not), $r_m = 0$ and thus, from the previous chain of equalities, $h_{m+1,m}$ must be zero. □

1.4 GMRES residual polynomial

As already stated in (2.4), r_m has a polynomial form involving the matrix A and the initial residual r_0. An explicit expression for such polynomial is known, and this also results in a way to compute its roots. Due to their clear connection with the projected system solved ad each step of GMRES, the GMRES residual polynomial and its roots, known as *harmonic Ritz values*, have been deeply studied (see, for instance, survey [16, Section 6 and its references]).

Theorem 1.9. *Consider the linear algebraic system $Ax = b$, and suppose the initial residual $r_0 = b - Ax_0$ to be of grade d with respect to A. For any $m \leq d$ consider the m-th* GMRES *approximation x_m to the solution x, as in (1.9):*

$$x_m = x_0 + V_m y_m, \qquad \text{where} \quad y_m = (\underline{H}_m^H \underline{H}_m)^{-1} H_m^H(\|r_0 e_1\|).$$

Then the corresponding m-th GMRES *residual polynomial is given by*

$$\varphi_m(z) = \frac{det(zH_m^H - \underline{H}_m^H \underline{H}_m)}{(-1)^m det(\underline{H}_m^H \underline{H}_m)},$$

and hence the zeros of $\varphi_m(z)$ are the eigenvalues of the generalized eigenvalue problem

$$(\underline{H}_m^H \underline{H}_m)v = zH_m^H v. \tag{1.12}$$

Remember that the $(m+1) \times m$ upper Hessenberg matrix $\underline{H}_m$ has full rank m, so $\underline{H}_m^H \underline{H}_m$ is non-singular.

Remark 1.10. If, additionally, the matrix $H_m = V_m^H A V_m$ is non-singular, then the generalized eigenvalue problem is equivalent to

$$H_m^{-H}(\underline{H}_m^H \underline{H}_m)v = zv. \tag{1.13}$$

Furthermore, in the same hypotheses, we have another available form for such problem, that is

$$(H_m + h_m^2 H_m^{-H} e_m e_m^H)v = zv, \tag{1.14}$$

where h_m denotes the entry of indices $(m+1, m)$ of $\underline{H}_m$.
A few passages transform (1.13) into (1.14), namely:

$$\underline{H}_m^H \underline{H}_m = \begin{bmatrix} H_m \\ h_m e_m^H \end{bmatrix} \begin{bmatrix} H_m^H & h_m e_m \end{bmatrix} = H_m H_m^H + h_m^2 e_m e_m^H.$$

Then

$$H_m^{-H}(\underline{H}_m^H \underline{H}_m) = H_m^{-H}(H_m H_m^H + h_m^2 e_m e_m^H) = H_m + h_m^2 H_m^{-H} e_m e_m^H.$$

From the computational point of view, for stability reasons, it is preferable to consider version (1.14), as it allows to avoid potentially dangerous computations like, for instance, the multiplication $\underline{H}_m^H \underline{H}_m$.

1.5 Restarted GMRES

Before starting a more detailed discussion about convergence, which will be the main topic of next chapter, we shortly present an important variant of the GMRES algorithm, *restarted* GMRES, with its main pros and cons.

Considering the algorithm from a practical point of view, GMRES becomes infeasible when m grows too much. This is due to the high cost required for the Gram-Schmidt orthogonalization process and the storage of all the basis vectors in V_m. Indeed, as m increases, the computational cost goes up at least as $\mathcal{O}(m^2)n$, while the memory cost as $\mathcal{O}(mn)$. For large n this limits the maximum usable value for m, thus the scheme might not be continued for an m large enough to meet the requested stopping criterion.
Therefore, the method is usually stopped for a reasonably small m and then restarted by constructing the subspace $\mathcal{K}_m(A, r_m)$; $r_m = b - Ax_m$, the current approximation x_m becomes the starting approximation for the next phase. However, this leads to the loss of the GMRES global optimality property, for the minimization now occurs only with respect to the last "partial" basis which is built. Moreover, for the same reason, the restarted process is not always ensured to converge, as stagnation may take place. A detailed discussion about these issues can be found in [15].

Chapter 2

Convergence analysis of GMRES

In exact arithmetic, well defined Krylov subspace methods terminate in a finite number of steps. Therefore no limit can be formed, and terms like "*convergence*" or "*rate of convergence*" cannot be interpreted in the classical way. The conceptual difference between having a *convergence bound* and having a description of the *convergence behavior* has always to be kept in mind. More precisely, a convergence bound basically represents an area in which the convergence curve will lay, but may not give any other qualitative information about the trend of such curve. On the other hand, the convergence behavior is an asymptotic description of the effective trend of the convergence curve. In the context of this thesis it will refer to the residual history "after a certain number of iterations", still within the dimension n of the addressed system.

The goal of the convergence analysis of Krylov subspace methods is to describe the convergence of this process in terms of input data of the given problem, i.e. in terms of properties of the system matrix, the right-hand side and the initial guess.

2.1 Chebyshev polynomials and classical results

In this section we briefly recall Chebyshev polynomials, along with their optimality properties, which are exploited in some known GMRES convergence bounds.

Definition 2.1. The Chebyshev polynomials of the first type are defined recursively as

$$\begin{cases} C_0(z) := 1 \\ C_1(z) := t \\ C_{k+1}(z) := 2zC_k(z) - C_{k-1}(z), \quad k \geq 2, \end{cases} \tag{2.1}$$

with $z \in \mathbb{C}$.

Remark 2.1. It is easy to see that $C_k(z)$ is a polynomial of degree k in z.

Remark 2.2. In the real case, $t \in \mathbb{R}$, Definition 2.1 can be written as

$$C_k(t) = \begin{cases} \cos(k \arccos t) & \text{if } |t| \leq 1 \\ \cosh(k \text{ arccosh } t) & \text{if } t \geq 1 \\ (-1)^k \cosh(k \text{ arccosh } (-t)) & \text{if } t \leq -1 \end{cases} . \tag{2.2}$$

Furthermore, for $k \gg 1$

$$C_k(t) \approx \frac{1}{2}\left(t + \sqrt{t^2-1}\right)^k .$$

Theorem 2.1 (Optimality of Chebyshev polynomials in a real interval)**.** *Let $[\alpha, \beta] \subset \mathbb{R}$, and let $\gamma \in \mathbb{R} \setminus [\alpha, \beta]$. Then the minimum*

$$\min_{\substack{p \in \mathbb{P}_k \\ p(\gamma)=1}} \max_{t \in [\alpha,\beta]} |p(t)|,$$

where $\mathbb{P}_k$ denotes the vector space of the polynomials of degree less or equal to k, with coefficients in $\mathbb{R}$, is reached by the polynomial

$$\widehat{C}_k(t) = \frac{C_k\left(1 + 2\frac{t-\beta}{\beta-\alpha}\right)}{C_k\left(1 + 2\frac{\gamma-\beta}{\beta-\alpha}\right)}.$$

From now on, in this section we consider Chebyshev polynomials in the more general complex case.

Remark 2.3. For $z \in \mathbb{C}$, it holds:

$$C_k(z) = \frac{1}{2}\left(w^k + w^{-k}\right) , \qquad \text{where } w \text{ is s.t. } z = \frac{1}{2}\left(w + w^{-1}\right).$$

The function

$$J(w) = \frac{1}{2}\left(w + w^{-1}\right)$$

is called *Joukowski's mapping*. It transforms a circle $\mathcal{C}(0, \rho)$, with center in the origin and radius ρ in an ellipse E_ρ, again centered at the origin, with foci ± 1, major semi-axis $\frac{1}{2}(\rho + \rho^{-1})$ and minor semi-axis $\frac{1}{2}|\rho - \rho^{-1}|$. Note that $J(\mathcal{C}(0, \rho)) = J(\mathcal{C}(0, 1/\rho))$, and this is the reason why in the following we will consider only $\rho \geq 1$.

As we will now see, in the complex case Chebyshev polynomials are only asymptotically optimal.

Lemma 2.2 (Zarantonello)**.** *Consider $\mathcal{C}(0, \rho)$ the circle of center 0 and radius $\rho \geq 1$, and choose $\gamma \in \mathbb{C}$, $|\gamma| > \rho$. Then*

$$\min_{\substack{p \in \mathbb{P}_k \\ p(\gamma)=1}} \max_{z \in \mathcal{C}(0,\rho)} |p(z)| = \left(\frac{\rho}{|\gamma|}\right)^k ,$$

that is, the minimum is reached by the polynomial $p(z) = \left(\frac{z}{\gamma}\right)^k$.

Remark 2.4. Scaling and translating, the result is valid also for a circle $\mathcal{C}(c,\rho)$, and $\gamma \in \mathcal{C}$, $|\gamma| > \rho$:

$$\min_{\substack{p\in\mathbb{P}_k \\ p(\gamma)=1}} \max_{z\in\mathcal{C}(c,\rho)} |p(z)| = \left(\frac{\rho}{|\gamma - c|}\right)^k.$$

Let us now consider an ellipse E_ρ centered at the origin, with foci ± 1 and semi-major axis a, $E_\rho = J(\mathcal{C}(0,\rho))$.

Theorem 2.3. *Let $E_\rho = J(\mathcal{C}(0,\rho))$, $\rho \geq 1$ and let $\gamma \in \mathbb{C}$ be external to E_ρ. Then*

$$\frac{\rho^k}{|w_\gamma|^k} \overset{(i)}{\leq} \min_{\substack{p\in\mathbb{P}_k \\ p(\gamma)=1}} \max_{z\in E_\rho} |p(z)| \overset{(ii)}{\leq} \frac{\rho^k + \rho^{-k}}{|w_\gamma^k + w_\gamma^{-k}|},$$

where w_γ is the dominant root of the equation $J(w) = \gamma$.

Proof. (ii). Each polynomial p of degree k such that $p(\gamma) = 1$ can be written as

$$p(z) = \frac{\sum_{j=0}^k \xi_j z^j}{\sum_{j=0}^k \xi_j \gamma^j}.$$

If $z \in E_\rho$, $z = J(w)$, with $w \in \mathcal{C}(0,\rho)$, and w_γ is the element of maximum norm of $J^{-1}(\gamma)$, $p(z)$ can be rewritten:

$$p(z) = \frac{\displaystyle\sum_{j=0}^k \xi_j (w^j + w^{-j})}{\displaystyle\sum_{j=0}^k \xi_j (w_\gamma^j + w_\gamma^{-j})}.$$

Consider now the particular polynomial with $\xi_k = 1$, $\xi_j = 0$, $\forall j \neq k$:

$$p^*(z) = \frac{w^k + w^{-k}}{w_\gamma^k + w_\gamma^{-k}} \qquad (= 2C_k(z))$$

$|p^*(z)|$ is maximum when $w = \rho e^{i\theta} \in \mathbb{R}$, i.e when $w = \rho$. Therefore

$$\max_{z\in E_\rho} |p^*(z)| = \frac{\rho^k + \rho^{-k}}{|w_\gamma^k + w_\gamma^{-k}|} \geq \min_{\substack{p\in\mathbb{P}_k \\ p(\gamma)=1}} \max_{z\in E_\rho} |p(z)|.$$

(i). We have

$$p(z) = \frac{\displaystyle\sum_{j=0}^k \xi_j (w^j + w^{-j})}{\displaystyle\sum_{j=0}^k \xi_j (w_\gamma^j + w_\gamma^{-j})} = \frac{\displaystyle w^{-k}\sum_{j=0}^k \xi_j (w^{k+j} + w^{k-j})}{\displaystyle w_\gamma^{-k}\sum_{j=0}^k \xi_j (w_\gamma^{k+j} + w_\gamma^{k-j})}.$$

Keeping in mind that $w = \rho e^{i\theta}$,

$$|p(z)| = \frac{\rho^{-k}}{|w_\gamma|^{-k}} \left| \frac{\displaystyle\sum_{j=0}^{k} \xi_j (w^{k+j} + w^{k-j})}{\displaystyle\sum_{j=0}^{k} \xi_j (w_\gamma^{k+j} + w_\gamma^{k-j})} \right|.$$

The argument of the absolute value is a $2k$ degree polynomial in w, taking the value 1 when evaluated at w_γ. From Lemma 2.2, its maximum absolute value is greater or equal to $\frac{\rho^{2k}}{|w_\gamma|^{2k}}$. Hence

$$\max_{z \in E_\rho} |p(z)| \geq \frac{\rho^{-k}}{|w_\gamma|^{-k}} \frac{\rho^{2k}}{|w_\gamma|^{2k}} = \frac{\rho^k}{|w_\gamma|^k} \qquad \forall p \in \mathbb{P}_k \text{ s.t. } p(\gamma) = 1,$$

and therefore

$$\min_{\substack{p \in \mathbb{P}_k \\ p(\gamma)=1}} \max_{z \in E_\rho} |p(z)| \geq \frac{\rho^k}{|w_\gamma|^k}.$$

□

Remark 2.5 (Asymptotical optimality of Chebyshev polynomials). The difference between the two bounds in the theorem goes to 0 as k goes to infinity, therefore we deduce that, for $k \gg 1$, the Chebyshev polynomial

$$p^*(z) = \frac{w^k + w^{-k}}{w_\gamma^k + w_\gamma^{-k}}, \qquad \text{with } z = \frac{w + w^{-1}}{2},$$

is close to the optimal polynomial. Chebyshev polynomials are therefore *asymptotically optimal*.

2.2 Known results on GMRES convergence

Before proceeding, it is important to recall that both the error and the residual of GMRES at each step m can be written in a polynomial form. More precisely, since $x_m \in x_0 + \mathcal{K}_m(A, r_0)$,

$$x - x_m = (x - x_0) + \underbrace{(x_0 - x_m)}_{= q_{m-1}(A) r_0} = x - x_0 + q_{m-1}(A) A (x - x_0) = p_m(A)(x - x_0),$$

where $q_{m-1} \in \mathbb{P}_{m-1}$ and $p_m \in \mathbb{P}_m$. From the definition of r_m, we have

$$r_m = b - A x_m = A \underbrace{(x - x_m)}_{= p_m(A)(x - x_0)} = p_m(A) A (x - x_0) = p_m(A) r_0.$$

Moreover, considering that $r_m \in r_0 + A\mathcal{K}_m(A, r_0)$, there exists another $m-1$ degree polynomial p_{m-1} such that

$$r_m = r_0 + A p_{m-1}(A) r_0 = (I - A p_{m-1}(A)) r_0.$$

This means the previously cited polynomial p_m satisfies $p_m(z) = 1 - zp_{m-1}(z)$, so $p_m(0) = 1$.
To sum up, there exists a polynomial $p_m \in \mathbb{P}_m^* := \{p \in \mathbb{P}_m \text{ s.t. } p(0) = 1\}$, uniquely determined by the orthogonality conditions (1.3) in the GMRES case, such that:

$$x - x_m = p_m(A)(x - x_0), \tag{2.3}$$
$$r_m = p_m(A)r_0. \tag{2.4}$$

Remark 2.6. For later notation observe that, without loss of generality, it is possible to assume $r_0 = b$. Indeed, for any initial guess x_0 for GMRES applied to the system $Ax = b$, we have:

$$Ax = b \Leftrightarrow Ax - Ax_0 = b - Ax_0 \Leftrightarrow A(x - x_0) = r_0.$$

Namely, considering $x - x_0$ as the unknown, it is always possible to return to the case in which the initial guess is 0 and $b = r_0$.

By construction the m-th residual has minimal 2-norm. In term of polynomials, this optimality property can be expressed as

$$\|r_m\| = \min_{p \in \mathbb{P}_m^*} \|p(A)r_0\|. \tag{2.5}$$

Several studies have been made in order to find a good bound for the right-hand side in (2.5), however obtaining results that are useful in practice is a real challenge. The main reason is because the bounds are often tied to the (whole) spectrum of A, which is usually not available. Another important issue is the presence of possibly large multiplicative constants in the relations, which may result in the loss of the quantitative - and sometimes even any qualitative - information the bound could carry.

In this section we summarize the principal results in the literature. As we will see, many expedients have been considered, attempting to bypass the cited problems.

A is diagonalizable.

We start with the case of A diagonalizable (with complex eigenvalues, in general), $A = X\Lambda X^{-1}$, with invertible X and $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$. In this case:

$$\begin{aligned} \|r_m\| &= \min_{p \in \mathbb{P}_m^*} \|p(A)r_0\| = \min_{p \in \mathbb{P}_m^*} \|Xp(\Lambda)X^{-1}r_0\| \\ &\leq \min_{p \in \mathbb{P}_m^*} \|p(\Lambda)\|\|X\|\|X^{-1}\|\|r_0\| = \kappa(X)\|r_0\| \min_{p \in \mathbb{P}_m^*} \|p(\Lambda)\|, \end{aligned} \tag{2.6}$$

where $\kappa(X)$ is the 2-norm condition number of the matrix X, $\kappa(X) := \|X\|\|X^{-1}\|$. Therefore we have the following bound on the relative residual norm

$$\frac{\|r_m\|}{\|r_0\|} \leq \kappa(X) \min_{p \in \mathbb{P}_m^*} \max_{k=1,\ldots,n} |p(\lambda_k)|. \tag{2.7}$$

If A is normal the spectral theorem assures it is *unitary* diagonalizable, i.e. X is unitary, thus $X^{-1} = X^H$ and $\|X\| = \|X\|^{-1} = 1$. In this lucky case, $\kappa(X) = 1$ and (2.7) reduces to

$$\frac{\|r_m\|}{\|r_0\|} \leq \min_{p \in \mathbb{P}_m^*} \max_{k=1,\dots,n} |p(\lambda_k)|. \tag{2.8}$$

As anticipated earlier, results in this form have limited practical interest, as they involve all the eigenvalues of A, which are not explicitly known. In order to estimate that *min-max quantity* we can move from the discrete set of the eigenvalues of A, let it be $\sigma(A)$, to compact continuous subsets of $\mathbb{C}$ containing $\sigma(A)$, on which the values of (quasi) optimal polynomials (e.g. Chebyshev polynomials) are explicitly known. Due to the co-normality property required to the polynomials (i.e. $p(0) = 1$), we will consider sets that do not contain the origin, otherwise the maximum of $|p(z)|$ over such a set would be not smaller than 1, loosing any possibility to be an informative bound.
We can use the results in section 2.1 to further estimate the residual in (2.8):

- If $\sigma(A)$ is contained in a disk $\mathcal{C}(c, r)$ ($c \in \mathbb{C}$, $r > 0$, $0 \notin \mathcal{C}(c, r)$), from Remark 2.4:

  $$\min_{p \in \mathbb{P}_m^*} \max_{k=1,\dots,n} |p(\lambda_k)| \leq \left(\frac{r}{|c|}\right)^m.$$

  In particular, a disk with a small radius, that is far from the origin guarantees fast convergence of GMRES residual norms.

- If $\sigma(A)$ is enclosed in an ellipse $\mathcal{E}(c, d, a)$, (center $c \in \mathbb{R}$, focal distance $d > 0$, semi-major axis $a > 0$)

  $$\min_{p \in \mathbb{P}_m^*} \max_{k=1,\dots,n} |p(\lambda_k)| \leq \frac{C_m\left(\frac{a}{d}\right)}{\left|C_m\left(\frac{c}{d}\right)\right|} \approx \left(\frac{a + \sqrt{a^2 - d^2}}{c + \sqrt{c^2 - d^2}}\right)^m,$$

  where $C_m(z)$ is the Chebyshev polynomial of degree m. We remark that the polynomials $C_m(z)/C_m(z)$ are in general not the optimal min-max polynomials on $\mathcal{E}$, as shown by Fisher and Freund [6]. However, their asymptotical optimality allows to correctly predict the rate of convergence of the min-max approximation problem.

- Another option is to consider the *numerical range* (or *field of values*) of A,

  $$\mathcal{F}(A) := \{w = x^H A x, x \in \mathbb{C}^n, \|x\| = 1\}.$$

  Denoting with $\nu_{\mathcal{F}(A)}$ the distance of $\mathcal{F}(A)$ from the origin,

  $$\nu_{\mathcal{F}(A)} := \min_{z \in \mathcal{F}(A)} |z|,$$

  it holds, when $0 \notin \mathcal{F}(A)$,

  $$\min_{p \in \mathbb{P}_m} \|p(A)\| \leq \left(1 - \nu_{\mathcal{F}(A)} \nu_{\mathcal{F}(A^{-1})}\right)^{n/2}.$$

  But again, it is generally difficult to compute the numerical range of a matrix, moreover, this kind of bound presents slackness especially when the

spectrum has some outlayer eigenvalues, which make significantly increase the value $\nu_{\mathcal{F}(A)}$.
If A is normal, $\mathcal{F}(A)$ coincides with the convex hull of $\sigma(A)$, that is the smallest convex set containing the spectrum. Hence, in the normal case, replacing $\mathcal{E}(c,d,a)$ with $\mathcal{F}(A)$ can improve the bound in the previous item.

- If A is *hermitian* then the eigenvalues are all real and the convex hull of $\sigma(A)$ reduces to an interval $I := [\lambda_{\min}, \lambda_{\max}]$. If, moreover, A is HPD or HND, I does not intersect zero. Hence

$$\max_{k=1,\dots,n} |p(\lambda_k)| \leq \max_{t\in I} |p(t)|.$$

However, when A is not definite, the presence of 0 in such interval I can be avoided by splitting it into two subintervals. Let λ_s be the largest negative eigenvalue, and λ_{s+1} the smallest positive one. The two intervals

$$I^- := [\lambda_{\min}, \lambda_s] \quad \text{and} \quad I^+ := [\lambda_{s+1}, \lambda_{\max}]$$

include the whole spectrum. In [11], Liesen and Tichý report that when $\lambda_{\max} - \lambda_{s+1} = \lambda_s - \lambda_{\min}$, the following relation holds:

$$\begin{aligned}\frac{\|r_m\|}{\|r_0\|} &\leq \min_{p\in\mathbb{P}^*_m} \max_{k=1,\dots,n} |p(\lambda_k)| \leq \min_{p\in\mathbb{P}^*_m} \max_{z\in I^-\cup I^+} |p(t)| \\ &\leq 2\left(\frac{\sqrt{|\lambda_{min}\lambda_{max}|} - \sqrt{|\lambda_s\lambda_{s+1}|}}{\sqrt{|\lambda_{min}\lambda_{max}|} + \sqrt{|\lambda_s\lambda_{s+1}|}}\right)^{\left[\frac{n}{2}\right]},\end{aligned}$$

where $[n/2]$ is the integer part of $n/2$.

Another aspect to keep in mind is the dependence of (2.7) on $\kappa(X)$. If A is far from normal, then $\kappa(X)$ may have large magnitude and this utterly inficiates the bound, making it possibly too pessimistic, since the relative residual norm on the left-hand side remains smaller than 1. However, we will return to this argument later, after having discussed the non-diagonalizable case as well.

An approach to understand the worst case GMRES convergence in the generic non-normal case is to replace the minimization problem (2.5) by another one, which in some sense approximates it, and that is easier to analyze. These natural bounds arise when the influence of $\|r_0\|$ is excluded:

$$\begin{aligned}\frac{\|r_m\|}{\|r_0\|} &= \min_{p\in\mathbb{P}_m*} \frac{\|p(A)r_0\|}{\|r_0\|} && \text{(GMRES)} \\ &\leq \max_{\|v\|=1} \min_{p\in\mathbb{P}_m*} \|p(A)v\| && \text{(worst-case GMRES)} && (2.9) \\ &\leq \min_{p\in\mathbb{P}_m*} \|p(A)\| && \text{(ideal GMRES)} && (2.10)\end{aligned}$$

The bound (2.9) corresponds to the *worst-case* GMRES behavior and represents a sharp upper bound, i.e. a bound that is attainable by the GMRES residual norm. In this sense, it is the best bound on the relative residual norm that is

independent of r_0.
The bound (2.10) represents instead a matrix approximation problem. A possible way to deal with it is to determine sets $\Omega_1, \Omega_2 \subset \mathbb{C}$ that are somehow associated with A, and that provide lower and upper bounds on (2.10):

$$c_1 \min_{p\in\mathbb{P}_m^*} \max_{z\in\Omega_1} |p(z)| \leq \min_{p\in\mathbb{P}_m^*} \|p(A)\| \leq c_2 \min_{p\in\mathbb{P}_m^*} \max_{z\in\Omega_2} |p(z)|.$$

Here c_1 and c_2 should be some moderate size constants depending on A and possibly on m. This generalizes the idea of taking the spectrum of A as an appropriate set. A possible choice for Ω_2 is the *ε-pseudospectrum* of A (as was first suggested by Trefethen, see [17])

$$\Lambda_\varepsilon(A) := \{z \in \mathbb{C} : \|(zI - A)^{-1}\| \geq \varepsilon^{-1}\}.$$

Denoting by L the arc length of the boundary of $\Lambda_\varepsilon(A)$, the following bound can be derived

$$\min_{p\in\mathbb{P}_m} \|p(A)\| \leq \frac{L}{2\pi\varepsilon} \min_{p\in\mathbb{P}_m} \max_{z\in\Lambda_\varepsilon(A)} |p(z)|.$$

The parameter ε gives flexibility, but choosing a good value can be hard. In fact, in order to make the right-hand side reasonably small, ε should be large enough to make the constant $L/(2\pi\varepsilon)$ small, but also small enough to make the pseudospectrum not too large.

A is not diagonalizable.

When A is not diagonalizable, a relation similar to (2.7), involving the Jordan form of A, can be derived. The idea is exactly the same Freund presented in [7] for the Quasi-Minimal residual method. We write the relation as a theorem, for it generalizes the previously showed bounds, and reduces to each of them in the corresponding particular cases.

Theorem 2.4. *Let A be a non-singular $n \times n$ matrix, with eigenvalues $\lambda_1, \ldots, \lambda_r$ of algebraic multiplicity $\mu_1, \ldots, \mu_r$ respectively. Consider the Jordan form of A, $A = X^{-1}JX$, and denote with $J(\lambda_k)$ the Jordan block which corresponds to the eigenvalue λ_k; let $\ell_k \leq \mu_k$ be its order. Then it holds*

$$\begin{aligned}\frac{\|r_m\|}{\|r_0\|} &\leq \kappa(X) \min_{p\in\mathbb{P}_m^*} \max_{k=1,\ldots,r} \|p(J(\lambda_k))\| \\ &\leq \kappa(X) \min_{p\in\mathbb{P}_m^*} \max_{k=1,\ldots,r} \left(\sum_{j=0}^{\ell_k - 1} \frac{1}{j!} \left| p^{(j)}(\lambda_k) \right| \right). \end{aligned} \tag{2.11}$$

Proof. With the same passages used in (2.6) we can write

$$\|r_m\| \leq \kappa(X) \min_{p\in\mathbb{P}_m *} \|p(J)\| \|r_0\|,$$

hence we just need to estimate $\|p(J)\|$. Since $p(J)$ is block diagonal, its norm is the maximum of the norms of its blocks:

$$\|p(J)\| = \max \left\{ \max_{k=1,\ldots,r} \|p(J(\lambda_k))\|, \max_{\substack{k=1,\ldots,r \\ \mu_k > \ell_k}} |p(\lambda_k)| \right\} = \max_{k=1,\ldots,r} \|p(J(\lambda_k))\|.$$

The result is an immediate consequence of the standard relation

$$p(J(\lambda_k)) = \sum_{j=0}^{\ell_k - 1} \frac{1}{j!} p^{(j)}(\lambda_k) N_k^j,$$

where N_k is the $\ell_k \times \ell_k$ nilpotent matrix

$$N_k = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix}$$

and thus has norm 1. □

A particular case: A has clustered eigenvalues

Several studies have been performed in case of clustered eigenvalues. Here we briefly report the main results presented in [2]. The authors' bound gives a qualitative description of the GMRES behavior. The relation they show involves the clusters radii, the relative distance of the outliers from the clusters and between the clusters.

First we introduce some notations and the setting. A is again a non-singular $n \times n$ matrix, with distinct eigenvalues $\lambda_1, \ldots, \lambda_r$ of algebraic multiplicities $\mu_1, \ldots, \mu_r$, indices[1] $\ell_1, \ldots, \ell_r$, and spectral projectors

$$X_j := \frac{1}{2\pi i} \int_{\Gamma_j} (zI - A)^{-1} dz, \qquad j = 1, \ldots, r,$$

where $i^2 = -1$ and Γ_j is any circle around λ_j containing no other eigenvalue of A. Note that the spectral projectors satisfy

$$\begin{aligned} & X_j A = A X_j, \\ & X_j X_j = X_j \quad \text{and} \quad X_i X_j = 0 \ \text{ for } i \neq j, \\ & \text{Range}(X_j) = \text{Ker}(\lambda_j I - A)^{\ell_j}. \end{aligned}$$

Now suppose A has M_1 outlying eigenvalues and P non-intersecting clusters of eigenvalues, and consider

$$d_{out} := \sum_{j=1}^{M_1} \ell_j$$

[1] The index of an eigenvalue λ_j is the smallest positive integer ℓ_j such that $Ker(\lambda_j I - A)^{\ell_j} = Ker(\lambda_j I - A)^{\ell_j + 1}$, i.e. ℓ_j is the dimension of the Jordan block $J(\lambda_j)$.

the degree of the minimal polynomial associated to the outliers $\{\lambda_j\}_{j=1}^{M_1}$, and Z_1 their spectral projector:

$$Z_1 := \sum_{j=1}^{M_1} X_j.$$

The (relative) clusters[2] are centered at distinct non-zero points $\gamma_2, \ldots, \gamma_{P+1} \in \mathbb{C}$, and are given by

$$\{\lambda_j\}_{j=M_{h-1}+1}^{M_h} \subset \{z \in \mathbb{C} : |z - \gamma_h| < \rho|\gamma_h|\}, \qquad h = 2, \ldots, P+1, \quad \rho > 0.$$

The associated spectral projectors are

$$Z_h := \sum_{q=M_{h-1}+1}^{M_h} X_q, \qquad h = 2, \ldots, P+1.$$

Let us separate the clusters and the outliers by decomposing

$$A = A_1 + \sum_{h=1}^{P+1} A_h, \qquad \text{where } A_1 = Z_1 A,\ A_h = Z_h A.$$

Finally, we can state the following proposition.

Proposition 2.5. *Consider $\rho > 0$, the distinct non-zero complex numbers $\{\gamma_h\}_{h=2,\ldots,P+1}$, and the integers*

$$0 \le M_1 \le M_2 \le \ldots \le M_{P+1} = r$$

so that, for $h = 2, \ldots, P+1$, the non-intersecting sets

$$\{\lambda_j\}_{j=M_{h-1}+1}^{M_h} \subset \{z \in \mathbb{C} : |z - \gamma_h| < \rho|\gamma_h|\}$$

are clusters, while

$$\{\lambda_j\}_{j=1}^{M_1} \subset \{z \in \mathbb{C} : |z - \gamma_h| > \rho|\gamma_h| \ \ \textit{for } h = 2, \ldots, P+1\}$$

are the outliers.
Define the distance of the outliers from the clusters as

$$\delta := \max_{2 \le h \le P+1} \max_{|z-\gamma_h| = \rho|\gamma_h|} \max_{1 \le j \le M_1} \frac{|\lambda_j - z|}{|\lambda_j|}$$

and the maximal distance between clusters as

$$\sigma := \max_{2 \le h \le P+1} \max_{|z-\gamma_h| = \rho|\gamma_h|} \max_{q \ne p} \frac{|\gamma_q - z|}{|\gamma_q|}.$$

Then, for any b, x_0 and k, it holds

$$\|r_{d_{out}+kP}\| \le C(\sigma^{P-1}\rho)^k \|r_0\|, \tag{2.12}$$

where the constant C is independent on k and given by

$$C := P\rho\delta^{d_{out}} \max_{2 \le h \le P+1} \max_{|z-\gamma_h| = \rho|\gamma_h|} \|(zI - A_h)^{-1}\|.$$

[2]The choice of individuating circular clusters is for the sake of simplicity. Anyway, other more complex sets (such as ellipses) could be chosen, resulting in less simple but more effective relations (see for example [12]).

The model in Proposition 2.5 has qualitative nature. Relation (2.12) shows how at least d_{out} steps have to pass in order to observe a first residual reduction, because GMRES has to process the outliers. Also the number P of clusters influences the trend of the residual norm, as a new reduction is expected to occur each P steps. The residual decrement is ruled by the *asymptotic convergence factor* $\sigma^{P-1}\rho$, and it is faster when the clusters are small and close together, and by the *asymptotic error constant* C, that reflects the non-normality of A and gathers information about the distance between the outliers and the clusters, and the number and radius of the clusters, too.

Chapter 3

A new convergence model

The aim of this chapter is to provide new descriptive bounds for the GMRES residual norm convergence.

The main issue with the bound (2.7) is that, in case of an ill-conditioned matrix, the term $\kappa(X)$ totally inficiates the bound, because of its high magnitude. Sometimes, however, the ill-conditioning of X is due to just a part of it. For instance, it may be caused by a few almost linearly dependent eigenvectors, which is the case we focus on in this thesis. Identifying and isolating the responsible of ill-conditioning may lead to a more accurate description of GMRES convergence. For example, this idea has already been developed in [4], to describe GMRES behavior when the matrix A is close to singular due to a relatively small group of eigenvalues located in a neighborhood of zero. In that article, the derived bound takes into account a sort of splitting of A into its far-from-singular and almost-singular parts.

In our context, however, the eigenvalues do not need to be clustered, nor the few ill-conditioned ones must necessarily be close to zero. The idea is to separate them on the base of the distance between the relative eigenspaces, i.e. to separate the far-from-dependent and almost-dependent invariant subspaces of A. Indeed, eigenvalues alone are not sufficient to describe GMRES behavior, for they do not carry all the necessary information, thus we shifted our focus on strategies that involve eigenspaces, too.

We will begin our discussion introducing the setting and some necessary notions. Initially, we will focus on the simpler situation in which just two eigenspaces are almost parallel. We will present a first convergence model, followed by the analysis of the more specific case in which the system matrix is comparable to a Jordan block matrix. Finally, we will illustrate the derivation of a new convergence model, that reveals to be very descriptive, also generalizing the argument in the case in which two groups of eigenspaces are one the perturbation of the other.

3.1 Setting and notations

Let us consider the unitary matrix

$$\widehat{X} := [x_1, d, x_3, \dots, x_n] \in \mathbb{C}^{n \times n}$$

and the matrix X, defined as

$$X := \widehat{X} \begin{bmatrix} \begin{pmatrix} 1 & 1 \\ & \varepsilon \end{pmatrix} & \\ & I_{n-2} \end{bmatrix} = [x_1, x_1 + \varepsilon d, x_3, \ldots, x_n], \tag{3.1}$$

for $\varepsilon > 0$. Note that X is no longer unitary. Moreover, not only the second column is almost parallel to the first one (for $\varepsilon \ll 1$), but also $\|x_1 + \varepsilon d\| \neq 1$. Let us write $Y^H := X^{-1}$ as the block matrix

$$X^{-1} = \begin{bmatrix} Y_1^H \\ Y_2^H \end{bmatrix}, \qquad Y_1 \in \mathbb{C}^{n\times 2},\, Y_2 \in \mathbb{C}^{n\times(n-2)};$$

for coherence of notation, let $y_1^H, \ldots, y_n^H$ be the rows of X^{-1}.
Let then $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ be distinct values (for brevity, we will often use the notation $a := \lambda_1$ and $b := \lambda_2$), and define $A := X\Lambda X^{-1}$. Using (3.1) we can write

$$A = X\Lambda X^{-1} = \widehat{X} L \widehat{X}^H,$$

where $\Lambda :=\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and L is the block diagonal matrix

$$L := \begin{bmatrix} L_1 & \\ & L_2 \end{bmatrix},$$

with

$$L_1 = \begin{pmatrix} 1 & 1 \\ & \varepsilon \end{pmatrix} \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} \begin{pmatrix} 1 & 1 \\ & \varepsilon \end{pmatrix}^{-1} = \begin{pmatrix} a & \frac{b-a}{\varepsilon} \\ 0 & b \end{pmatrix}, \qquad L_2 := \mathrm{diag}(\lambda_3, \ldots, \lambda_n). \tag{3.2}$$

Remark 3.1. Defining A this way, there are just two ill-conditioned eigenvalues, namely a and b, whose condition numbers are $\mathcal{O}(1/\varepsilon)$, as we will now briefly illustrate. Keep in mind that the condition number of an eigenvalue is defined as

$$\kappa(\lambda) = \frac{1}{|y^H x|}, \tag{3.3}$$

where x and y are respectively the right and the left eigenvectors of unitary norm associated with λ. In our case a and b have as right and left (non-unitary) eigenvectors x_1, y_1, and $x_2 = x_1 + \varepsilon d$, y_2, respectively. Since $Y = X^{-1}$, it holds that $Y_1 = \widehat{X}_1 E^{-H}$, with

$$\widehat{X}_1 := [x_1, d], \qquad E := \begin{pmatrix} 1 & 1 \\ 0 & \varepsilon \end{pmatrix}, \quad \text{and} \quad E^{-H} = \begin{pmatrix} 1 & 0 \\ -1/\varepsilon & 1/\varepsilon \end{pmatrix}.$$

But then

$$y_1 = x_1 - \frac{1}{\varepsilon} d \quad \text{and } y_2 = \frac{1}{\varepsilon} d.$$

Thus

$$\kappa(a)^2 = \left(\frac{\|x_1\|\|y_1\|}{|x_1^H y_1|}\right)^2 = \frac{1+\frac{1}{\varepsilon^2}}{\left(x_1^H x_1 - \frac{1}{\varepsilon} d^H x_1\right)^2} = 1+\frac{1}{\varepsilon^2}$$

$$\kappa(b)^2 = \left(\frac{\|x_2\|\|y_2\|}{|x_2^H y_2|}\right)^2 = \frac{(1+\varepsilon^2)\frac{1}{\varepsilon^2}}{\left(\frac{1}{\varepsilon} d^H x_1 + d^H d\right)^2} = 1+\frac{1}{\varepsilon^2}$$

$$\Rightarrow \kappa(a) = \kappa(b) = \sqrt{1+\frac{1}{\varepsilon^2}}.$$

Therefore $\kappa(a)$ and $\kappa(b)$ blow up as ε decreases, i.e. as the eigenspaces of a and b tend to merge.

In the following, we will very often need to consider matrix valued polynomials. If $A = \widehat{X} L \widehat{X}^H$, for any polynomial p it holds

$$p(A) = \widehat{X} \begin{bmatrix} p(L_1) & \\ & p(L_2) \end{bmatrix} \widehat{X}^H,$$

where

$$p(L_1) = \begin{pmatrix} 1 & 1 \\ & \varepsilon \end{pmatrix} p\left(\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}\right) \begin{pmatrix} 1 & 1 \\ & \varepsilon \end{pmatrix}^{-1} = \begin{pmatrix} p(a) & \frac{p(b)-p(a)}{\varepsilon} \\ 0 & p(b) \end{pmatrix}, \tag{3.4}$$

and $p(L_2) = \mathrm{diag}(p(\lambda_3), \ldots, p(\lambda_n))$. Therefore, going back to the GMRES residual,

$$\|r_m\| = \min_{p \in \mathbb{P}_m^*} \|p(A) r_0\| = \min_{p \in \mathbb{P}_m^*} \|\widehat{X} p(L) \widehat{X}^H r_0\| = \min_{p \in \mathbb{P}_m^*} \|p(L)(\widehat{X}^H r_0)\|.$$

Note also that, since $\widehat{X}$ is unitary, $\|\widehat{X}^H r_0\| = \|r_0\|$. These last two observations explain that, in terms of convergence properties, applying GMRES to the system $Ax = r_0$ is equivalent to applying it to the transformed system $L(\widehat{X}^H x) = (\widehat{X}^H r_0)$. Therefore, without loss of generality, we can directly consider

$$A = \mathrm{blkdiag}(L_1, L_2) \tag{3.5}$$

and rename $\widehat{X}^H r_0$ as r_0.

One more thing we need to introduce is the notion of distance between vector subspaces. We will very concisely give the definition and some basic properties.

3.1.1 Projectors and distance between vector subspaces

Definition 3.1. A *projector* is a linear map $P : \mathbb{C}^n \to \mathbb{C}^n$ such that

$$P^2 = P.$$

It holds:

- Every projector P gives a decomposition of $\mathbb{C}^n$ as

$$\begin{aligned} &\mathbb{C}^n = \mathrm{Im}P \oplus \mathrm{Ker}P, \\ &\forall x \in \mathbb{C}^n, \qquad x = Px + (I - P)x. \end{aligned}$$

- Conversely, given any pair of vector subspaces M and N such that $\mathbb{C}^n = M \oplus N^{\perp}$, there exists a projector P with

$$M = \mathrm{Im}P \quad \text{and} \quad N^{\perp} = \mathrm{Ker}P.$$

- If $N = M$, the associated projector π_M is an orthogonal projector of $\mathbb{C}^n$ over M. Moreover, if the columns of X form an orthonormal basis of M, then π_M has a matrix representation given by

$$\pi_M = XX^H.$$

Definition 3.2. Given the subspaces M, N of $\mathbb{C}^n$, their *distance* is defined as

$$\omega(M, N) := \|\pi_M - \pi_N\|.$$

Note that, once an orthonormal basis of M and N is known, their distance corresponds to a matrix norm.

Definition 3.3. The *distance* between a vector x and a subspace N is given by

$$\mathrm{dist}(x, N) := \|x - \pi_N x\| = \|(I - \pi_N)x\|.$$

Thus, the following result holds

Proposition 3.1.

$$\omega(M, N) = \max \left\{ \max_{\substack{x \in M \\ x^H x = 1}} \mathrm{dist}(x, N), \max_{\substack{y \in N \\ y^H y = 1}} \mathrm{dist}(y, M) \right\}.$$

Moreover, it can be proved that spaces of different dimensions have distance larger than 1; and also that if $\dim(M) = \dim(N)$ and $M \perp N$, then $\omega(M, N) = 1$.

We conclude this introduction with some theoretical results about how perturbation over a Jordan block transfers into the relative eigenvalue and generalized eigenspace.

3.1.2 Perturbations over a Jordan block

In this subsection only A will denote a general square matrix possessing an eigenvalue λ of algebraic multiplicity μ, geometric multiplicity g and index ℓ. Call

$$M = \mathrm{Ker}(A - \lambda I)^{\ell} \quad \text{and} \quad E = \mathrm{Ker}(A - \lambda I)$$

the associated invariant subspace and eigenspace. Consider then a perturbation of A: $A' = A + H$, with $\|H\| = \varepsilon$. Here we will make use of the "prime" to denote the entities related to the perturbed matrix A'.
In [3] a detailed description of the behavior of eigenvalues and eigenspaces in case of perturbation can be found. Here we limit ourselves to just state the main results and show they are effectively satisfied by the matrices L_1 and $L_{1,J}$. First of all we recall that, as may be expected, perturbing A produces a scattering of the μ coincident eigenvalues λ into μ eigenvalues λ'_i, $i = 1, \ldots, \mu$ (not necessarily all distinct) of A'. Naturally, along with this scattering, the generation of new eigenspaces occurs. It is reasonable to think that the new eigenvalues and eigenspaces will not be far from the original ones. Next proposition and theorems describe what the situation is like for what concerns invariant subspaces, eigenvectors and eigenvalues, in this order.

Proposition 3.2. [3, Corollary 4.3.2] *Let M and M' be the invariant subspaces corresponding to λ and $\{\lambda'_i\}_{i=1}^{\mu}$ respectively. Then the distance between these subspaces is of the order of the perturbation (for $\varepsilon \to 0$),*

$$\omega(M, M') = \mathcal{O}(\varepsilon).$$

Remark 3.2. With the expression *invariant subspace corresponding to a set of eigenvalues* we indicate the vector subspace that is the sum of all the generalized eigenspaces associated with those eigenvalues. This is the case for M'. Clearly, when the set of eigenvalues is a singleton, the sum reduces to a unique generalized eigenspace, like M.

Next theorem focuses on eigenvectors (see [3, Theorem 4.3.7], with $j = k = 1$):

Theorem 3.3. *In the previous notations, fix one of the eigenvalues λ'_i of A', and let $E' = Ker(A - \lambda'_i I)$ be the corresponding eigenspace.*
Then for any eigenvector $x' \in E'$ it holds

$$dist(x', E) = \mathcal{O}(\varepsilon^{1/\ell}).$$

In addition, some interesting relations about the position of the scattered eigenvalues are satisfied:

Theorem 3.4. *In the previous notations, it holds:*

$$(i) \ \max_i |\lambda - \lambda'_i| = \mathcal{O}(\varepsilon^{1/\ell}),$$

$$(ii) \ \left| \lambda - \frac{1}{\mu} \sum_{i=1}^{\mu} \lambda'_i \right| = \mathcal{O}(\varepsilon).$$

Gathering the information given by these three results, we can conclude that the distance between the eigenvectors of A and A' is of the order of the distance between the eigenvalues λ and λ'_i, while the distance between the invariant subspaces M and M' is of the same order of the distance between λ and the arithmetic mean of the scattered eigenvalues.

3.2 A first model

In the introduction to this chapter we explained our decision to separate the eigenvalues based on their conditioning, but also and especially looking at the distance between their eigenspaces. Since all eigenvalues of A are distinct, all eigenspaces are mono-dimensional, thus computing the distances between them only involves their (normalized) generating vectors. The two eigenspaces with eigenvalues a and b have distance of the order of ε, while all the other distances are of order 1. This is easy to see, as $\|x_1 - x_2\| = \varepsilon\|d\| = \varepsilon$, while the distance between any other couple of eigenspaces is 1 due to their reciprocal orthogonality. In order to split the eigenvalues, we opted for building a polynomial written as a product of other polynomials.
Keeping in mind the form of A, let us write the m-th residual norm as

$$\begin{aligned}\|r_m\|^2 &= \min_{p\in\mathbb{P}^*_m} \left\| \begin{bmatrix} p(L_1) & \\ & p(L_2) \end{bmatrix} \begin{bmatrix} r_0^{(1)} \\ r_0^{(2)} \end{bmatrix} \right\|^2 \\ &= \min_{p\in\mathbb{P}^*_m} \left(\|p(L_1)r_0^{(1)}\|^2 + \|p(L_2)r_0^{(2)}\|^2 \right). \end{aligned} \tag{3.6}$$

Here $r_0^{(1)}$and $r_0^{(2)}$ denote respectively the vectors composed by the first two entries of r_0 and the remaining $n-2$ ones.
As a first try, we chose a specific polynomial $p \in \mathbb{P}_m$ to bound the minimum in (3.6), namely the product

$$p(z) = \varphi(z)\psi_{m-2}(z),$$

where $\varphi(z) = (1 - \frac{z}{a})(1 - \frac{z}{b})$ is the minimal polynomial of L_1, whilst ψ_{m-2} is the residual polynomial at step $m-2$ of GMRES applied to the well-conditioned part of the system, that is to the $(n-2)\times(n-2)$ system $L_2x^{(2)} = r_0^{(2)}$. Let s_{m-2} be the corresponding residual at step $m-2$. The idea of writing $p(z)$ as a product of two polynomials has been used in the past for similar purposes, see e.g. [2, 4]. The aim of this choice is to limit the influence of the ill-conditioned part of the matrix. From (3.6) we get:

$$\begin{aligned}\|r_m\|^2 \leq& \|\psi_{m-2}(L_1)\underbrace{\varphi(L_1)r_0^{(1)}}_{=0}\|^2 + \|\varphi(L_2)\psi_{m-2}r_0^{(2)}\|^2 \\ \leq& \|\varphi(L_2)\|^2\|\psi_{m-2}(L_2)r_0^{(2)}\|^2 = \|\varphi(L_2)\|^2\|s_{m-2}\|^2\end{aligned}$$

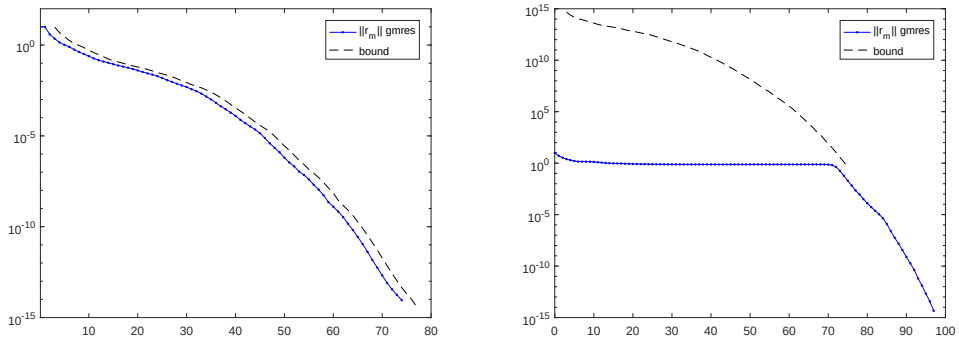

Figure 3.1: GMRES *residual norms and bound (3.7). Data:* $n = 100$, $\varepsilon = 10^{-6}$, $\lambda_3, \ldots, \lambda_n = 98, \ldots, 1$*; left:* $a = 100$, $b = 99$*; right:* $a = 10^{-5}$, $b = 2 \cdot 10^{-5}$.

and thus

$$\|r_m\| \leq \|\varphi(L_2)\| \|s_{m-2}\|. \tag{3.7}$$

In practice the bound (3.7) appears to be informative when the convergence is not too slow in the beginning. However, it presents some limits in case of an initial stagnation, when the setback of having a coefficient of high magnitude springs again.

Figure 3.1 shows two examples. We computed the bound by constructing and evaluating the polynomial φ and by multiplying the results with the residuals obtained running GMRES on the system $L_2 x^{(2)} = r_0^{(2)}$. While in the plot on the left the bound is coherent with the residual history, the situation on the right is totally different. A low initial rate of convergence may be caused, for instance, by a cluster of eigenvalues around the origin, and in our example a and b are set to be both of order 10^{-6}. When a and b are close to zero if compared to the rest of the spectrum, $\|\varphi(L_2)\|$ can be very large, as on the real axis φ represents a parabola which assumes the value 1 on 0 and whose roots are assigned to be a and b. Having observed this, we conclude that bound (3.7) has very little chance to be interesting whenever the eigenvalues on which φ is built are small with respect to the rest of the spectrum. Still, the bound qualitatively describes the trend of the convergence curve after the stagnation phase has been surpassed, i.e. after the information about the ill-conditioned part of the matrix has been processed.

Thus far we have talked about bounds that, in some sense, consider the influence of the spectrum of A and of the relative eigenspaces separately, or even do not take into account either of them at the same time. We have seen how these kinds of relations present some flaws in describing the convergence curve. Therefore we can conclude that the eigenvalues themselves are not sufficient to obtain good, informative relations. Many authors have already pointed this out in several works (see e.g. [10, 11]). To further highlight this issue, we give a theorem, whose main point is that any non-increasing convergence curve is possible for GMRES for a matrix having any prescribed set of eigenvalues. It was originally presented in [1] and [8], where a constructive proof is reported, too.

Theorem 3.5. [10, Theorem 5.7.7] *For any n positive numbers $f_0 \geq f_1 \geq \ldots \geq f_{n-1} > 0$ and any n non-zero complex numbers $\lambda_1, \ldots, \lambda_n$, not necessarily distinct, there exists a matrix $A \in \mathbb{C}^{n\times n}$, with eigenvalues $\lambda_1, \ldots, \lambda_n$, and a vector $b \in \mathbb{C}^n$, with $\|b\| = f_0$, so that* GMRES *applied to $Ax = b$ with x_0 has the residual norms $\|r_m\| = f_m$, $m = 0, 1, \ldots, n-1$.*

It is therefore clear that we should proceed with the purpose of not prematurely separate the information carried by eigenvalues and relative eigenspaces. For example, an approach that follows this idea is to consider the field of values of the matrix (see (2.2)). However, likewise many other proposed bounds, the relations the field of values provides are good for some instances, but rough and completely useless for others, as they do not follow the effective asymptotic convergence. A detailed description of the behavior of different global and local bounds, along with several enlightening examples, can be found in the technical report [5].

3.3 Similarities with the Jordan case

In this section we explore the case in which the ill-conditioning of the system matrix makes it close to a matrix possessing a Jordan block with the corresponding ill-conditioned invariant subspace.
Let $A =$blkdiag(L_1, L_2), with L_1 and L_2 defined in (3.2). In particular,

$$L_1 = \begin{pmatrix} a & \frac{b-a}{\varepsilon} \\ & b \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & \varepsilon \end{pmatrix} \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & \varepsilon \end{pmatrix}^{-1} =: V \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} V^{-1}. \tag{3.8}$$

When a and b are close together, and also when ε is small, the 2×2 matrix L_1 can be read as the perturbation of the following matrix:

$$\begin{aligned} L_{1,J} &:= \begin{pmatrix} a & \frac{b-a}{\varepsilon} \\ \frac{-\varepsilon(b-a)}{4} & b \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ \frac{\varepsilon}{2} & \frac{\varepsilon}{b-a}\left(1 + \frac{b-a}{2}\right) \end{pmatrix} \begin{pmatrix} \frac{a+b}{2} & 1 \\ 0 & \frac{a+b}{2} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ \frac{\varepsilon}{2} & \frac{\varepsilon}{b-a}\left(1 + \frac{b-a}{2}\right) \end{pmatrix}^{-1} \\ &=: V_J \begin{pmatrix} \frac{a+b}{2} & 1 \\ 0 & \frac{a+b}{2} \end{pmatrix} V_J^{-1}, \end{aligned} \tag{3.9}$$

having the single eigenvalue $\frac{a+b}{2}$ of index 2 (see [3, Example 4.2.2]).
The difference between L_1 and $L_{1,J}$ is given by

$$\|L_1 - L_{1,J}\| = \frac{\varepsilon|b-a|}{4},$$

therefore, L_1 is close to the non-diagonalizable matrix $L_{1,J}$ whenever $|b-a|$ or ε are small, that is, whenever the eigenvalues are close to each other, or whenever V and V_J are close to each other. This motivated us to investigate and compare the residual results of GMRES when applied to blkdiag$(L_1, L_2)x = r_0$ and blkdiag$(L_{1,J}, L_2)x = r_0$. For analogy with A, pose $A_J :=$blkdiag$(L_{1,J}, L_2)$.

In the rest of this section we will first discuss the theoretical results presented in Subsection 3.1.2. Later we will continue investigating how GMRES behaves over matrices that are perturbations of Jordan blocks, focusing on the particular case of L_1 and $L_{1,J}$, and will make a few comparisons.

3.3.1 An example: L_1 as a perturbation of the Jordan block $L_{1,J}$

Keeping in mind the results in Proposition 3.2, Theorem 3.3 and Theorem 3.4, we discuss the relations between the invariant subspaces, eigenvectors and eigenvalues of L_1 and $L_{1,J}$. We can write L_1 as

$$L_1 = L_{1,J} + \begin{pmatrix} 0 & 0 \\ \delta & 0 \end{pmatrix}, \qquad \text{with } \delta = \varepsilon \frac{b-a}{4},$$

then the perturbation norm is $|\delta|$. A direct computation of the distances in Proposition 3.2, Theorem 3.3 and Theorem 3.4 is possible. In particular:

- Denoting with v_1, v_2 and $v_{J,1}$, $v_{J,2}$ the columns of the matrices V and V_J, we have: $\omega(\text{span}\{v_1, v_2\},\text{span}\{v_{J,1}, v_{J,2}\}) = \omega(\mathbb{R}^2, \mathbb{R}^2) = 0$, so Proposition 3.2 trivially holds.

- Theorem 3.3 states that $\text{dist}(v_1, v_{J,1}) = \mathcal{O}(|\delta|^{\frac{1}{2}})$ and $\text{dist}(v_2/\|v_2\|, v_{J,1}) = \mathcal{O}(|\delta|^{\frac{1}{2}})$. On the other hand, calling E_J =span$(x_{J,1})$ and proceeding with explicit computations, we have

$$\begin{aligned} \text{dist}(v_1, E_J) = \|(I - \pi_{E_J})v_1\| &= \left\| \left(I - \frac{v_{J,1} v_{J,1}^H}{\|v_{J,1}\|^2} \right) v_1 \right\| \\ &= \left\| \begin{pmatrix} \frac{\varepsilon^2}{4+\varepsilon^2} & \frac{2\varepsilon}{4+\varepsilon^2} \\ \frac{2\varepsilon}{4+\varepsilon^2} & \frac{4}{4+\varepsilon^2} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\| = \frac{\varepsilon}{4+\varepsilon^2} \|(\varepsilon, 2)^T\| \\ &= \frac{\varepsilon}{\sqrt{4+\varepsilon^2}} = \mathcal{O}(\varepsilon) \end{aligned}$$

  and, similarly, $\text{dist}(v_2, E_J) = \mathcal{O}(\varepsilon)$.

- For what concerns eigenvalues, instead, the theory says that $|a - \frac{a+b}{2}| = \mathcal{O}(|\delta|^{\frac{1}{2}})$ and $|b - \frac{a+b}{2}| = \mathcal{O}(|\delta|^{\frac{1}{2}})$, while we obtain

$$\left| a - \frac{a+b}{2} \right| = \left| b - \frac{a+b}{2} \right| = \frac{|b-a|}{2} = \mathcal{O}(|b-a|).$$

What observed in the last two items might seem not completely coherent with the statements of Proposition 3.2 and Theorems 3.3 and 3.4, as it accounts only for either ε or $|b-a|$, while the perturbation δ involves both of them simultaneously. To clarify, note that asymptotically speaking, ε and $|b-a|$ cannot be totally unrelated, as neither L_1 nor $L_{1,J}$ exist if only one of those two quantities tends to zero (to be precise, $L_{1,J}$ would still exist for $|b-a| \to 0$, but it would loose its Jordan form). In particular, both the ratio $(b-a)/\varepsilon$ and its reciprocal must be asymptotically constant. Therefore, $|b-a| = \mathcal{O}(\varepsilon)$ for $\varepsilon \to 0$, and $\varepsilon = \mathcal{O}(|b-a|)$ for $|b-a| \to 0$. This means that $\delta = \mathcal{O}(\varepsilon^2)$, or equivalently, $\delta = \mathcal{O}(|b-a|^2)$, depending on which quantity one focuses on. Thus, in this example, $\mathcal{O}(|\delta|^{\frac{1}{2}}) = \mathcal{O}(\varepsilon) = \mathcal{O}(|b-a|)$, so what we found agrees with the theoretical results.

3.3.2 Comparison: GMRES on $Ax = r_0$ and $A_J x = r_0$

To study the convergence behavior of GMRES we first need to understand the relation between $p(L_1)$ and $p(L_{1,J})$. Recall we already computed $p(L_1)$ from (3.4),

$$p(L_1) = \begin{pmatrix} p(a) & \frac{p(b)-p(a)}{\varepsilon} \\ 0 & p(b) \end{pmatrix}. \tag{3.10}$$

To write $p(L_{1,J})$, we remind that the evaluation of a polynomial over a Jordan block also involves derivatives, so that

$$\begin{aligned} p(L_{1,J}) =& V_J p\left(\begin{pmatrix} \frac{a+b}{2} & 1 \\ 0 & \frac{a+b}{2} \end{pmatrix}\right) V_J^{-1} \\ =& V_J \left(p\left(\frac{a+b}{2}\right)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + p'\left(\frac{a+b}{2}\right)\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\right) V_J^{-1} \\ =& p\left(\frac{a+b}{2}\right)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + p'\left(\frac{a+b}{2}\right) V_J \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} V_J^{-1} \\ =& p\left(\frac{a+b}{2}\right)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + p'\left(\frac{a+b}{2}\right)(b-a)\begin{pmatrix} -\frac{1}{2} & \frac{1}{\varepsilon} \\ -\frac{\varepsilon}{4} & \frac{1}{2} \end{pmatrix} \\ =& p(\xi)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + p'(\xi)\frac{b-a}{2}\begin{pmatrix} -1 & \frac{2}{\varepsilon} \\ -\frac{\varepsilon}{2} & 1 \end{pmatrix}, \end{aligned} \tag{3.11}$$

where $\xi := (a+b)/2$.

For the moment, let us focus on the case in which the distance between the two ill-conditioned eigenvalues is small, i.e. $|b-a| \ll 1$. To proceed, we need some Taylor expansions around ξ:

$$p(a) = p\left(\xi - \frac{b-a}{2}\right) = p(\xi) - p'(\xi)\frac{b-a}{2} + \mathcal{O}(|b-a|^2) \tag{3.12}$$

$$p(b) = p\left(\xi + \frac{b-a}{2}\right) = p(\xi) + p'(\xi)\frac{b-a}{2} + \mathcal{O}(|b-a|^2) \tag{3.13}$$

so that

$$p(b) - p(a) = p'(\xi)(b-a) + \mathcal{O}(|b-a|^2). \tag{3.14}$$

We have the following proposition:

Proposition 3.6. *Consider L_1 and $L_{1,J}$ as defined in (3.8) and (3.9). For fixed $\varepsilon > 0$ and for any polynomial p it holds*

$$p(L_1) = p(L_{1,J}) + p'(\xi)\begin{pmatrix} 0 & 0 \\ \varepsilon\frac{b-a}{4} & 0 \end{pmatrix} + \mathcal{O}(|b-a|^2)\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

and

$$\|p(L_1) - p(L_{1,J})\| = |p'(\xi)|\frac{|b-a|}{4}\varepsilon + \mathcal{O}(|b-a|^2), \tag{3.15}$$

for $|b-a| \to 0$.

Proof. Using (3.10), (3.11) and then (3.12), (3.13) and (3.14):

$$\begin{aligned} p(L_1) - p(L_{1,J}) &= \begin{pmatrix} p(a) & \frac{p(b)-p(a)}{\varepsilon} \\ 0 & p(b) \end{pmatrix} - p(\xi) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - p'(\xi) \frac{b-a}{2} \begin{pmatrix} -1 & \frac{2}{\varepsilon} \\ -\frac{\varepsilon}{2} & 1 \end{pmatrix} \\ &= p'(\xi) \begin{pmatrix} -\frac{b-a}{2} & \frac{b-a}{\varepsilon} \\ 0 & \frac{b-a}{2} \end{pmatrix} - p'(\xi) \frac{b-a}{2} \begin{pmatrix} -1 & \frac{2}{\varepsilon} \\ -\frac{\varepsilon}{2} & 1 \end{pmatrix} + \mathcal{O}(|b-a|^2) \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \\ &= p'(\xi) \begin{pmatrix} 0 & 0 \\ \varepsilon \frac{b-a}{4} & 0 \end{pmatrix} + \mathcal{O}(|b-a|^2) \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \end{aligned}$$

and the result follows. □

Proposition 3.6 shows that for any polynomial p such that $|p'(\xi)|$ is not excessively large, the value of $p(L_1)$ is close to that of $p(L_{1,J})$ when the distance between a and b (eigenvalues of L_1) is small. As a consequence, when the eigenvalues a and b are close to each other, the GMRES residual history of $Ax = r_0$ closely follows the one of $A_J x = r_0$ (clearly, for $|b-a| \gg 0$, the behavior may differ).

Proposition 3.7. *Denote with r_m and $r_{m,J}$ the m-th residuals of* GMRES *applied to $Ax = r_0$ and $A_J x = r_0$, respectively, and let $\phi_m(z)$ and $\phi_{m,J}(z)$ be the corresponding residual polynomials. Then*

$$\left| \|r_m\| - \|r_{m,J}\| \right| \leq \max \left\{ \left| \phi'_m \left(\frac{a+b}{2} \right) \right|, \left| \phi'_{m,J} \left(\frac{a+b}{2} \right) \right| \right\} \frac{|b-a|}{4} \varepsilon \|r_0^{(1)}\| + \mathcal{O}(|b-a|^2), \tag{3.16}$$

for $|b-a| \to 0$.

Proof.

$$\begin{aligned} \|r_m\| &= \min_{p \in \mathbb{P}_m} \|p(A) r_0\| \leq \|\phi_{m,J}(A) r_0\| \\ &\leq \|\phi_{m,J}(A_J) r_0\| + \|\phi_{m,J}(A) r_0 - \phi_{m,J}(A_J) r_0\| \\ &= \|r_{m,J}\| + \|(\phi_{m,J}(A) - \phi_{m,J}(A_J)) r_0\| \\ &= \|r_{m,J}\| + \left\| \begin{bmatrix} \phi_{m,J}(L_1) - \phi_{m,J}(L_{1,J}) & 0 \\ 0 & \phi_{m,J}(L_2) - \phi_{m,J}(L_2) \end{bmatrix} r_0 \right\| \\ &= \|r_{m,J}\| + \left\| \begin{bmatrix} \phi_{m,J}(L_1) - \phi_{m,J}(L_{1,J}) & 0 \\ 0 & 0 \end{bmatrix} r_0 \right\| \\ &= \|r_{m,J}\| + \left\| \left(\phi_{m,J}(L_1) - \phi_{m,J}(L_{1,J}) \right) r_0^{(1)} \right\|. \end{aligned}$$

Thus, using (3.15), we obtain:

$$\|r_m\| \leq \|r_{m,J}\| + \left| \phi'_{m,J} \left(\frac{a+b}{2} \right) \right| \frac{|b-a|}{4} \varepsilon \|r_0^{(1)}\| + \mathcal{O}(|b-a|^2),$$

for $|b-a| \to 0$.
In the same way, we can write:

$$\|r_{m,J}\| \leq \|r_m\| + \left| \phi'_m \left(\frac{a+b}{2} \right) \right| \frac{|b-a|}{4} \varepsilon \|r_0^{(1)}\| + \mathcal{O}(|b-a|^2).$$

Combining the last two inequalities we obtain the result. □

Note the presence of the residual polynomials' first derivative in Proposition 3.7, which in some sense is reminiscent of the relation in Theorem 2.4.

Relation (3.7) is illustrated in Figure 4.1, Figure 4.2 and Figure 4.5 in Chapter 4, where GMRES convergence curves relative to the systems $Ax = r_0$ and $A_J x = r_0$ are compared, with very small distance between a and b.

3.3.3 Further comparisons and observations

With the same approach we adopted for A_J, one may wonder wether there are any similarities when GMRES is applied to the two diagonal systems $A_D x = r_0$ and $A_{D_\xi} x = r_0$, where $A_D :=\text{diag}(a, b, \lambda_3, \ldots, \lambda_n)$ and $A_{D_\xi} := \text{diag}(\xi, \xi, \lambda_3, \ldots, \lambda_n)$, with $\xi = \frac{a+b}{2}$. Indeed, for a fixed ε and for $|b-a| \ll 1$, the 2-norm distances

$$\|A - A_D\| = \frac{|b-a|}{\varepsilon} \qquad \text{and} \qquad \|A - A_{D_\xi}\| = \frac{|b-a|}{2}\sqrt{1+\frac{2}{\varepsilon}}$$

are very small. For coherence with the previous notation, let us set $L_{1,D} :=\text{diag}(a, b)$ and $L_{1,D_\xi} :=\text{diag}(\xi, \xi) = \xi I_2$. Moreover, we denote with $\phi_{m,D}$ and ϕ_{m,D_ξ} the m-th GMRES polynomial relative to A_D and A_{D_ξ}, respectively. For the residuals, we use the notations $r_{m,D}$ and r_{m,D_ξ}.

Since $L_1 = V L_{1,D} V^{-1}$, with V defined as in (3.8), studying what happens when $L_{1,D}$ replaces L_1 substantially means to study the diagonal problem without ill-conditioning. On the other hand, $L_{1,D_\xi} = V L_{1,D_\xi} V^{-1}$, so, when $|b-a|$ is small, L_1 can be obtained from L_{1,D_ξ} by slightly perturbing its eigenvalues. The double geometric multiplicity of the eigenvalue ξ allows to always choose an orthonormal basis for the eigenspace, therefore the diagonal problem corresponding to L_{1,D_xi} is not affected by ill-conditioning, either.

Let us now briefly analyze the GMRES behavior over the aforementioned systems. GMRES convergence is driven by polynomials, hence having a small distance between the system matrices does not necessarily imply having similar convergence curves. Instead, we need to compare the corresponding matrix valued polynomials. Since the computations are analogous to those made earlier in this chapter, to give a more slender presentation, we avoid to write them down, directly stating the results.

Proposition 3.8. *With the previous notation, it holds:*

$$\left| \|r_m\| - \|r_{m,D}\| \right| \leq \max\left\{ |\phi'_m(\xi)|, \left|\phi'_{m,D}(\xi)\right| \right\} \frac{|b-a|}{\varepsilon} \|r_0^{(1)}\| + \mathcal{O}(|b-a|^2),$$

$$\left| \|r_m\| - \|r_{m,D_\xi}\| \right| \leq \max\left\{ |\phi'_m(\xi)|, \left|\phi'_{m,D_\xi}(\xi)\right| \right\} \frac{|b-a|}{2}\sqrt{1+\frac{2}{\varepsilon}} \|r_0^{(1)}\| + \mathcal{O}(|b-a|^2),$$

for $|b-a| \to 0$.

For completeness, we also insert the comparison between the Jordan block matrix A_J and the corresponding diagonal one A_{D_ξ}, which holds for any value of $|b-a|$.

Proposition 3.9. *With the previous notation, it holds*

$$\left| \|r_{m,J}\| - \|r_{m,D_\xi}\| \right| \leq \max\left\{ \left|\phi'_{m,J}(\xi)\right|, \left|\phi'_{m,D_\xi}(\xi)\right| \right\} \frac{|b-a|}{2} \left(\frac{1}{4} + \frac{1}{\varepsilon^2} \right) \|r_0^{(1)}\| + \mathcal{O}(|b-a|^2).$$

Figure 4.1 and Figure 4.3 in Chapter 4 show the GMRES convergence curves of the systems $Ax = r_0$, $A_D x = r_0$, $A_{D_\xi} x = r_0$ and $A_J x = r_0$ in the same plots, so to allow a qualitative comparison. Further comments to the plots are presented in Remark 4.3.

Returning now on the comparison with the Jordan case, we would like to point out that, while the theory regarding non-diagonalizable matrices involves not only the polynomial p, but also its first derivative p' (see (2.11)), no derivatives appear when the diagonalizable matrix A is considered. However, if indeed A is a perturbation of A_J, totally forgetting about p' may be misleading. Hence, we expect that the presence of a derivative, or at least a difference ratio, would provide a more descriptive bound. The model problem we discuss in Section 3.4 follows up on this argument, by including a related constraint.

Additional insight towards the inclusion of a constraint is given by some simple bounds for $\|p(L_1)\|$ and $\|p(L_{1,J})\|$, that we present in the following lemmas.

Lemma 3.10. *It holds*

$$\|p(L_1)\| \leq \sqrt{|p(a)|^2 + |p(b)|^2 + \left| \frac{p(b)-p(a)}{\varepsilon} \right|^2}. \tag{3.17}$$

Proof. Thanks to the small dimension of L_1 we can bound $\|p(L_1)\|$ by explicitly computing: $\|p(L_1)\| = \sqrt{\lambda_{\max}(p(L_1)^H p(L_1))}$. The characteristic polynomial of

$$p(L_1)^H p(L_1) = \begin{bmatrix} |p(a)|^2 & \overline{p(a)} \frac{p(b)-p(a)}{\varepsilon} \\ p(a) \overline{\frac{p(b)-p(a)}{\varepsilon}} & |p(b)|^2 + \left| \frac{p(b)-p(a)}{\varepsilon} \right|^2 \end{bmatrix}$$

is

$$q(\lambda) = \lambda^2 - \lambda \left(|p(a)|^2 + |p(b)|^2 + \left| \frac{p(b)-p(a)}{\varepsilon} \right|^2 \right) + \left| p(a)p(b) \right|^2.$$

Since $p(L_1)^H p(L_1)$ is Hermitian and positive definite, its roots are both real and

positive,

$$\begin{aligned}\lambda_{max} =&\frac{1}{2}\left(|p(a)|^2+|p(b)|^2+\left|\frac{p(b)-p(a)}{\varepsilon}\right|^2+\right.\\ &\left.+\sqrt{\left(|p(a)|^2+|p(b)|^2+\left|\frac{p(b)-p(a)}{\varepsilon}\right|^2\right)^2-4\left|p(a)p(b)\right|^2}\right)\\ \leq&\frac{1}{2}\left(|p(a)|^2+|p(b)|^2+\left|\frac{p(b)-p(a)}{\varepsilon}\right|^2+\sqrt{\left(|p(a)|^2+|p(b)|^2+\left|\frac{p(b)-p(a)}{\varepsilon}\right|^2\right)^2}\right)\\ =&|p(a)|^2+|p(b)|^2+\left|\frac{p(b)-p(a)}{\varepsilon}\right|^2.\end{aligned}$$

□

Furthermore, for $|b-a| \ll 1$, using the expansions (3.12), (3.13) and (3.14) to rewrite $p(L_1)$ we also have:

Lemma 3.11. *For $|b-a| \to 0$ it holds*

$$\|p(L_1)\| \leq |p(\xi)| + |p'(\xi)|\frac{|b-a|}{2}\sqrt{\left(1+\frac{2}{\varepsilon^2}\right)+\sqrt{1+\frac{4}{\varepsilon^4}}} + \mathcal{O}(|b-a|^2). \quad (3.18)$$

Proof. Substituting (3.12), (3.13) and (3.14) in (3.10), we can write

$$p(L_1) = p(\xi)I_2 + p'(\xi)\frac{b-a}{2}\begin{pmatrix}-1 & 2/\varepsilon\\ & 1\end{pmatrix} + \mathcal{O}(|b-a|^2).$$

Therefore

$$\|p(L_1)\| \leq |p(\xi)| + |p'(\xi)|\frac{|b-a|}{2}\left\|\begin{pmatrix}-1 & 2/\varepsilon\\ & 1\end{pmatrix}\right\| + \mathcal{O}(|b-a|^2).$$

To complete the proof, observe that the coefficient

$$\sqrt{\left(1+\frac{2}{\varepsilon^2}\right)+\sqrt{1+\frac{4}{\varepsilon^4}}}$$

is computed as the square root of the maximum eigenvalue of

$$\begin{pmatrix}-1 & 2/\varepsilon\\ & 1\end{pmatrix}^H\begin{pmatrix}-1 & 2/\varepsilon\\ & 1\end{pmatrix} = \begin{pmatrix}1 & -2/\varepsilon\\ -2/\varepsilon & 1+4/\varepsilon^2\end{pmatrix}.$$

□

Considering now $p(L_{1,J})$, we have one last lemma

Lemma 3.12. *It holds*

$$\|p(L_{1,J})\| \leq |p(\xi)| + |p'(\xi)|\frac{|b-a|}{2}\left(\frac{2}{\varepsilon}+\frac{\varepsilon}{2}\right). \tag{3.19}$$

Proof. To calculate the bound for $\|p(L_{1,J})\|$ we can rely on its decomposition into two addends,

$$p(L_{1,J}) = p(\xi)I_2 + p'(\xi)\frac{b-a}{2}\begin{pmatrix}-1 & 2/\varepsilon\\ -\varepsilon/2 & 1\end{pmatrix}.$$

Thus

$$\|p(L_{1,J})\| \leq |p(\xi)| + |p'(\xi)|\frac{|b-a|}{2}\left\|\begin{pmatrix}-1 & 2/\varepsilon\\ -\varepsilon/2 & 1\end{pmatrix}\right\|.$$

Like we did before, we can evaluate the 2-norm of the matrix in the right hand side, let it be G, as $\|G\| = \lambda_{\max}(G^TG)$. We have

$$G^TG = \begin{pmatrix}1+\varepsilon/4 & -(2/\varepsilon+\varepsilon/2)\\ -(2/\varepsilon+\varepsilon/2) & 1+\varepsilon/4\end{pmatrix}$$

with characteristic polynomial

$$p(\lambda) = \lambda\left(\lambda - \frac{(\varepsilon^2+4)^2}{4\varepsilon^2}\right).$$

Hence

$$\lambda_{max} = \left(\frac{\varepsilon^2+4}{2\varepsilon}\right)^2 = \left(\frac{2}{\varepsilon}+\frac{\varepsilon}{2}\right)^2,$$

from which the result follows. □

These two bounds have similar form, as they both show a dependence (of the same nature) on the values assumed by p over the eigenvalues, and by $p'(\xi)$ and the difference ratio $\frac{p(b)-p(a)}{b-a}$. This is particularly evident by rewriting (3.17) as

$$\|p(L_1)\| \leq \sqrt{|p(a)|^2 + |p(b)|^2 + \left|\frac{p(b)-p(a)}{b-a}\right|^2\left(\frac{|b-a|}{\varepsilon}\right)^2}.$$

Note that, when ε goes to zero, the coefficient of the first derivative term in (3.18) behaves similarly to the corresponding coefficient in (3.19).

3.4 A new convergence model

As noted in the first section of this chapter, it is not restrictive to consider the linear system $Ax = r_0$, where A is the block diagonal matrix $L = \text{blkdiag}(L_1, L_2)$. The GMRES residual norm is thus given by:

$$\|r_m\| = \min_{p\in\mathbb{P}_m^*}\|p(L)r_0\|.$$

Here we study what happens if we shrink the polynomial set used in the GMRES residual problem, by adding the constraint

$$p(b) = \frac{p(b) - p(a)}{\varepsilon}. \tag{3.20}$$

Recalling the bound for $\|p(L_1)\|$ in (3.17), this constraint ensures that the ratio $(p(b) - p(a))/\varepsilon$ will not blow up, hence it limits the influence of ε.

Even if a and b are far from each other, what matters for GMRES seems to be the similarity of the values the minimizing polynomial assumes on them. This motivates the choice of the constraint (3.20). Indeed, once reformulated as

$$p(a) = (1 - \varepsilon)p(b) \tag{3.21}$$

or also, when $p(b) \neq 0$, as

$$\frac{p(b) - p(a)}{p(b)} = \varepsilon,$$

the constraint can be interpreted as a request for $p(a)$ and $p(b)$ to be as close as ε, in a relative sense.

Another remark is that, by asking $(p(b) - p(a))/\varepsilon$ to assume a precise value, we are implicitly imposing conditions regarding the quotient $(p(b) - p(a))/(b - a)$, because

$$\frac{p(b) - p(a)}{\varepsilon} = \frac{p(b) - p(a)}{b - a}\frac{b - a}{\varepsilon}.$$

Observe that, when a and b are close to each other, using (3.13) and (3.14) shows that condition (3.20) differs by $\mathcal{O}(|b - a|^2)$ from the request

$$\frac{2p(\xi)}{2 - \varepsilon} = p'(\xi)\frac{b - a}{\varepsilon},$$

that is, we are implicitly imposing conditions on the first derivative of the minimal polynomial.

Finally, we can state the following bound for the GMRES residual norm.

Theorem 3.13. *Let $\Lambda = diag(a, b, \lambda_3, \dots, \lambda_n)$ be the diagonal matrix containing the eigenvalues of A. Then at every step m the* GMRES *residual norm satisfies*

$$\|r_m\| \leq \sqrt{3} \min_{\substack{p \in \mathbb{P}_m^* \\ p(b) = \frac{p(b) - p(a)}{\varepsilon}}} \|p(\Lambda) r_0\|. \tag{3.22}$$

Proof.

$$\begin{aligned} \|r_m\|^2 &= \min_{p \in \mathbb{P}_m^*} \|p(A) r_0\|^2 = \min_{p \in \mathbb{P}_m^*} \left(\|p(L_1) r_0^{(1)}\|^2 + \|p(L_2) r_0^{(2)}\|^2 \right) \\ &\leq \min_{\substack{p \in \mathbb{P}_m^* \\ p(b) = \frac{p(b) - p(a)}{\varepsilon}}} \left(\|p(L_1) r_0^{(1)}\|^2 + \|p(L_2) r_0^{(2)}\|^2 \right) \end{aligned}$$

With the constraints (3.20), the expression in (3.10) gives

$$\begin{aligned}\|p(L_1)r_0^{(1)}\|^2 &= \left\|\begin{pmatrix} p(a) & p(b) \\ & p(b)\end{pmatrix} r_0^{(1)}\right\|^2 = \left\|\begin{pmatrix} 1 & 1 \\ & 1\end{pmatrix}\begin{pmatrix} p(a) & \\ & p(b)\end{pmatrix} r_0^{(1)}\right\|^2 \\ &\leq 3\left\|\begin{pmatrix} p(a) & \\ & p(b)\end{pmatrix} r_0^{(1)}\right\|^2.\end{aligned} \tag{3.23}$$

Therefore

$$\begin{aligned}\|r_m\|^2 &\leq \min_{\substack{p\in\mathbb{P}_m^* \\ p(b)=\frac{p(b)-p(a)}{\varepsilon}}} 3\left(\left\|\begin{pmatrix} p(a) & \\ & p(b)\end{pmatrix} r_0^{(1)}\right\|^2 + \|p(L_2)r_0^{(2)}\|^2\right) \\ &= \min_{\substack{p\in\mathbb{P}_m^* \\ p(b)=\frac{p(b)-p(a)}{\varepsilon}}} 3\|p(\Lambda)r_0\|^2,\end{aligned}$$

from which the result follows. □

Remarks:

1. The constant 3 is an upper bound for $\frac{3+\sqrt{5}}{2} \approx 2.6$, the squared norm of the 2×2 Jordan matrix
$$\begin{pmatrix} 1 & 1 \\ & 1\end{pmatrix}.$$

2. The bound in (3.22) is surprisingly clean, involving nothing but the eigenvalues of A and the initial residual. We would like to point out that the alternative constraint
$$\frac{p(b)-p(a)}{\varepsilon} = p(a)$$
could have been considered. However, adopting it does not yield an equally simple bound, as (3.23) would become:
$$\|p(L_1)r_0^{(1)}\|^2 = \left\|\begin{pmatrix} p(a) & p(a) \\ & p(b)\end{pmatrix} r_0^{(1)}\right\|^2 = \left\|\begin{pmatrix} p(a) & \\ & p(b)\end{pmatrix}\begin{pmatrix} 1 & 1 \\ & 1\end{pmatrix} r_0^{(1)}\right\|^2.$$
This time the norm of the Jordan matrix cannot be bounded without separating the norms of $\mathrm{diag}(p(a), p(b))$ and $r_0^{(1)}$ as well.

3. The bound does not depend on how close a and b are. In other words, it can be descriptive irrespective of the distribution of the eigenvalues. So indeed, relation (3.22) is an alternative to eigenvalue-clustering-based bounds, that may not be descriptive as they take into account only information regarding eigenvalues.

4. The bound in Theorem 3.13 can be further estimated as
$$\|r_m\| \leq \sqrt{3} \min_{\substack{p\in\mathbb{P}_m^* \\ p(b)=\frac{p(b)-p(a)}{\varepsilon}}} \max_{i=1,\ldots,n} |p(\lambda_i)|\|r_0\|,$$
which leads to a constrained min-max polynomial problem.

Numerical experiments in Chapter 4 illustrate the descriptive value of this new bound in several cases.

Remark 3.3. Using the relation (3.21), we can rewrite (3.22) as

$$\|r_m\| \leq \sqrt{3} \min_{\substack{p\in\mathbb{P}^*_m \\ p(a)=(1-\varepsilon)p(b)}} \|p(\widehat{\Lambda})\widehat{r}_0\|$$

where

$$\widehat{\Lambda} = \mathrm{diag}(b, b, \lambda_3, \ldots, \lambda_n) \quad \text{and} \quad \widehat{r}_0 = ((1-\varepsilon)r_{0,1}, r_{0,2}, \ldots, r_{0,n})^T.$$

Indeed, thanks to (3.21) we have

$$\begin{aligned}
\|p(\Lambda)r_0\|^2 &= \sum_{i=1}^{n} p(\lambda_i)^2 r_{0,i}^2 = p(a)^2 r_{0,1}^2 + p(b)^2 r_{0,2}^2 + \sum_{i=3}^{n} p(\lambda_i)^2 r_{0,i}^2 \\
&= p(b)^2(1-\varepsilon)^2 r_{0,1}^2 + p(b)^2 r_{0,2}^2 + \sum_{i=3}^{n} p(\lambda_i)^2 r_{0,i}^2 \\
&= \left\| \begin{pmatrix} p(b) & & & & \\ & p(b) & & & \\ & & p(\lambda_3) & & \\ & & & \ddots & \\ & & & & p(\lambda_n) \end{pmatrix} \begin{pmatrix} (1-\varepsilon)r_{0,1} \\ r_{0,2} \\ r_{0,3} \\ \vdots \\ r_{0,n} \end{pmatrix} \right\|^2 .
\end{aligned}$$

3.5 Generalization to more than two ill-conditioned eigenspaces

Let us now consider the case in which the ill-conditioning of the eigenvector matrix is due to more than two eigenvectors.
We proceed by generalizing the previous model problem. To maintain the same notation, let $\widehat{X}$ be a square unitary matrix of the form:

$$\widehat{X} = [X_1, D, X_3]$$

where this time, X_1, D and X_3 are tall rectangular matrices. Let

$$X_2 := X_1 + DE,$$

for some square and non-singular matrix E, $E = \varepsilon E_0$, with the hypothesis that both $\|E_0\|$ and $\|E_0^{-1}\|$, and thus $\kappa(E_0)$ as well, are moderate. In this way the columns of X_1 and X_2 span closely related subspaces. Define

$$X := [X_1, X_2, X_3] = \widehat{X} \begin{bmatrix} I & I & \\ & E & \\ & & I \end{bmatrix},$$

and for our choice of X_2, the matrix X is non-singular. We again denote with Λ the diagonal matrix containing the eigenvalues of A, similar to A through X:

$$A = X\Lambda X^{-1}$$

and we use the block partitioning

$$\Lambda = \begin{bmatrix} \Lambda_1 & & \\ & \Lambda_2 & \\ & & \Lambda_3 \end{bmatrix}$$

conforming to X_1, X_2 and X_3, respectively. We have

$$A = X\Lambda X^{-1} = \widehat{X} \begin{bmatrix} I & I & \\ & E & \\ & & I \end{bmatrix} \begin{bmatrix} \Lambda_1 & & \\ & \Lambda_2 & \\ & & \Lambda_3 \end{bmatrix} \begin{bmatrix} I & -E^{-1} & \\ & E^{-1} & \\ & & I \end{bmatrix} \widehat{X}^H.$$

To generalize (3.20) we take the new matrix constraint

$$p(\Lambda_2) = \frac{p(\Lambda_2) - p(\Lambda_1)}{\varepsilon}. \tag{3.24}$$

The following theorem extends Theorem 3.13 to the case of a more general eigenspace ill-conditioning situation. As already explained in Section 3.1, without loss of generality we can suppose to directly have

$$A = \begin{bmatrix} I & I & \\ & E & \\ & & I \end{bmatrix} \begin{bmatrix} \Lambda_1 & & \\ & \Lambda_2 & \\ & & \Lambda_3 \end{bmatrix} \begin{bmatrix} I & -E^{-1} & \\ & E^{-1} & \\ & & I \end{bmatrix} = \begin{bmatrix} \Lambda_1 & (\Lambda_2 - \Lambda_1)E^{-1} & \\ & E\Lambda_2 E^{-1} & \\ & & \Lambda_3 \end{bmatrix} \tag{3.25}$$

and rename $\widehat{X}^H r_0$ as r_0.

Theorem 3.14. *Consider X, Λ, A as defined before. Let $\mathcal{J} := blkdiag(I, E_0, I)$. Then at every step m the* GMRES *residual norm satisfies*

$$\|r_m\| \leq \|\mathcal{J}\| \min_{\substack{p \in \mathbb{P}_m^* \\ p(\Lambda_2) = \frac{p(\Lambda_2) - p(\Lambda_1)}{\varepsilon}}} \left\| p(\Lambda)\mathcal{J}^{-1} r_0 \right\|, \tag{3.26}$$

where $\|\mathcal{J}\| \leq \sqrt{2 + \|E_0\|}$ is moderate, since $\|E_0\|$ is moderate by construction.

Proof. Following what was done in the proof of Theorem 3.13, we have

$$\begin{aligned} \|r_m\|^2 &= \min_{p \in \mathbb{P}_m^*} \|p(A)r_0\|^2 = \min_{p \in \mathbb{P}_m^*} \left\| \begin{bmatrix} p(\Lambda_1) & (p(\Lambda_2) - p(\Lambda_1))E^{-1} & \\ & Ep(\Lambda_2)E^{-1} & \\ & & p(\Lambda_3) \end{bmatrix} r_0 \right\|^2 \\ &= \min_{p \in \mathbb{P}_m^*} \left\| \begin{bmatrix} p(\Lambda_1) & \frac{p(\Lambda_2) - p(\Lambda_1)}{\varepsilon} E_0^{-1} & \\ & E_0 p(\Lambda_2) E_0^{-1} & \\ & & p(\Lambda_3) \end{bmatrix} r_0 \right\|^2 \\ &\leq \min_{\substack{p \in \mathbb{P}_m^* \\ p(\Lambda_2) = \frac{p(\Lambda_2) - p(\Lambda_1)}{\varepsilon}}} \left\| \begin{bmatrix} p(\Lambda_1) & \frac{p(\Lambda_2) - p(\Lambda_1)}{\varepsilon} E_0^{-1} & \\ & E_0 p(\Lambda_2) E_0^{-1} & \\ & & p(\Lambda_3) \end{bmatrix} r_0 \right\|^2. \end{aligned}$$

Now let us focus on the upper triangular block. If we ask (3.24) to hold, we have

$$\begin{bmatrix} p(\Lambda_1) & \frac{p(\Lambda_2)-p(\Lambda_1)}{\varepsilon} E_0^{-1} \\ & E_0 p(\Lambda_2) E_0^{-1} \end{bmatrix} = \begin{bmatrix} p(\Lambda_1) & p(\Lambda_2) E_0^{-1} \\ & E_0 p(\Lambda_2) E_0^{-1} \end{bmatrix} = \begin{bmatrix} I & I \\ & E_0 \end{bmatrix} \begin{bmatrix} p(\Lambda_1) & \\ & p(\Lambda_2) \end{bmatrix} \begin{bmatrix} I & \\ & E_0^{-1} \end{bmatrix}. \tag{3.27}$$

Therefore

$$\|r_m\|^2 \leq \min_{\substack{p \in \mathbb{P}_m^* \\ p(\Lambda_2) = \frac{p(\Lambda_2)-p(\Lambda_1)}{\varepsilon}}} \left\| \begin{bmatrix} I & I & \\ & E_0 & \\ & & I \end{bmatrix} \begin{bmatrix} p(\Lambda_1) & & \\ & p(\Lambda_2) & \\ & & p(\Lambda_3) \end{bmatrix} \begin{bmatrix} I & & \\ & E_0^{-1} & \\ & & I \end{bmatrix} r_0 \right\|^2$$

and we finally find

$$\|r_m\| \leq \|\mathcal{J}\| \min_{\substack{p \in \mathbb{P}_m^* \\ p(\Lambda_2) = \frac{p(\Lambda_2)-p(\Lambda_1)}{\varepsilon}}} \left\| p(\Lambda) \begin{bmatrix} I & & \\ & E_0^{-1} & \\ & & I \end{bmatrix} r_0 \right\|.$$

□

Remark 3.4. Again, the bound of Theorem 3.14 can be further estimated as

$$\|r_m\| \leq \kappa(\mathcal{J}) \min_{\substack{p \in \mathbb{P}_m^* \\ p(\Lambda_2) = \frac{p(\Lambda_2)-p(\Lambda_1)}{\varepsilon}}} \max_{i=1,\dots,n} |p(\lambda_i)| \|r_0\|. \tag{3.28}$$

We next show that the bound simplifies by assuming the matrix E_0 is unitary:

Corollary 3.15. *Consider X, Λ, A as previously defined, and let the matrix E_0 be unitary. Then, at every step m, the* GMRES *residual norm satisfies:*

$$\|r_m\| \leq \sqrt{3} \min_{\substack{p \in \mathbb{P}_m^* \\ p(\Lambda_2) = \frac{p(\Lambda_2)-p(\Lambda_1)}{\varepsilon}}} \left\| p(\Lambda) \mathcal{J}^H r_0 \right\|.$$

Remark 3.5. Observe that this time, since E_0 is unitary, the matrix $\mathcal{J}^H$ is unitary as well, thus the norm of r_0 is preserved when the matrix-vector multiplication is performed. In particular, (3.28) holds with $\kappa(\mathcal{J}) = 1$.

Next, we consider the case of a diagonal matrix E_0. Of course, this is a very special case, for requiring E_0 to be diagonal implies having an X_2 with a quite peculiar structure: each column of X_2 is almost parallel to the corresponding column of X_1, and it is not almost a linear combination of all columns of X_1.

Theorem 3.16. *Consider X, Λ, A as previously defined, and let the matrix E_0 be diagonal. Then, at every step m, the* GMRES *residual norm satisfies:*

$$\|r_m\| \leq \sqrt{2 + \|E_0^{-1}\|} \min_{\substack{p \in \mathbb{P}_m^* \\ p(\Lambda_2) = \frac{p(\Lambda_2)-p(\Lambda_1)}{\varepsilon}}} \|p(\Lambda) r_0\|.$$

Proof. The passages mirror those in the proof of Theorem 3.14. If E_0 is diagonal, since $E_0 p(\Lambda_2) E_0^{-1} = E_0 E_0^{-1} p(\Lambda_2)$, then the matrix in 3.27 can be rewritten as

$$\begin{bmatrix} p(\Lambda_1) & p(\Lambda_2)E_0^{-1} \\ & p(\Lambda_2) \end{bmatrix} = \begin{bmatrix} I & E_0^{-1} \\ & I \end{bmatrix} \begin{bmatrix} p(\Lambda_1) & \\ & p(\Lambda_2) \end{bmatrix}.$$

Therefore it holds

$$\|r_m\|^2 \leq \min_{\substack{p \in \mathbb{P}_m^* \\ p(\Lambda_2) = \frac{p(\Lambda_2) - p(\Lambda_1)}{\varepsilon}}} \left\| \begin{bmatrix} I & E_0^{-1} & \\ & I & \\ & & I \end{bmatrix} p(\Lambda) r_0 \right\|^2,$$

from which the result follows. □

Chapter 4

Numerical evidence: constrained minimization problem

In this chapter we report on our computational experience aimed at illustrating the sharpness of the new convergence bound obtained in Chapter 3.

For $m \in \{1, \ldots, n\}$ consider the minimization problems that appear in Theorem 3.13 and Theorem 3.14:

$$\min_{\substack{p \in \mathbb{P}_m^* \\ p(b)=\frac{p(b)-p(a)}{\varepsilon}}} \|p(\Lambda) r_0\| \qquad \text{and} \qquad \min_{\substack{p \in \mathbb{P}_m^* \\ p(\Lambda_2)=\frac{p(\Lambda_2)-p(\Lambda_1)}{\varepsilon}}} \left\|p(\Lambda)\mathcal{J}^{-1} r_0\right\| \tag{4.1}$$

(remember that $\mathcal{J}$:=blkdiag(I, E_0, I), with moderate $\|E_0\|$ and $\|E_0^{-1}\|$). For later reference let n_1 be the number of columns of E_0. Since the former problem is a particular case of the latter, here we will focus on the more general case, from which the derivation of the procedure for the other case is immediate.

Note that in our model no hypothesis of reality over the eigenvalues is assumed, therefore it is natural to provide numerical evidence involving complex spectra. To be able to compare (4.1) with the actual GMRES residual norm as the iteration proceeds, we need to numerically solve the constrained minimization problem in (4.1). The procedure used in our experiments is discussed in Section 4.1.

4.1 Solving the constrained minimization problem

In order to have the means to compute all quantities and closely follow the evolution of the polynomial, we limit ourselves to small dimension problems, namely $n = 50$ (in Examples 5.1 and 5.2 $n = 49$ for construction reasons). Among the representations we took in account, the most stable way to evaluate p was given by the following form:

$$p(z, c) = \prod_{j=1}^{m} \left(1 - \frac{z}{c_j}\right),$$

where the dependence on the (complex) vector of the polynomial roots $c = (c_1, \dots, c_m)^T$ is highlighted. Note that writing p in this form also allows to implicitly impose co-monicity. From now on, we will denote with p_{con} the solution of the constrained minimization problem in (4.1).

Remark 4.1. In section 1.4, through Theorem 1.9, we introduced the harmonic Ritz values as the roots of the GMRES residual polynomial. A small modification of the GMRES algorithm allows to compute such roots on the fly, by solving the generalized eigenvalue problem (1.14). The relatively modest computational efforts to obtain them (for small dimension problems), along with the parallelism with the roots c_j of the sought polynomial p_{con}, makes the harmonic Ritz values a good initial guess candidate when solving unconstrained system (4.2) through iterative methods.

We can now proceed with the derivation of the numerical procedure to solve the constrained problem in (4.1). Setting $\widetilde{r}_0 := \mathcal{J}^{-1} r_0$ and identifying $\mathbb{C}$ with $\mathbb{R}^2$, the function $F : \mathbb{R}^{2m} \to \mathbb{R}$ to be minimized is given by

$$\begin{aligned} F(c) :=& \|p(\Lambda, c)\widetilde{r}_0\|^2 \\ =& \sum_{h=1}^{n} \left(\overline{p(\lambda_h, c)\widetilde{r}_0^{(h)}}\right) \left(p(\lambda_h, c)\widetilde{r}_0^{(h)}\right), \end{aligned}$$

while the constraint function $\phi : R^{2m} \to R^{n_1}$, $\phi(c) = (\phi_1(c), \dots, \phi_{n_1}(c))^T$ has components

$$\begin{aligned} \phi_k(c) :=& \left|(1-\varepsilon)p(\lambda_{n_1+k}, c) - p(\lambda_k, c)\right|^2 \\ =& \left((1-\varepsilon)\overline{p(\lambda_{n_1+k}, c)} - \overline{p(\lambda_k, c)}\right) \left((1-\varepsilon)p(\lambda_{n_1+k}, c) - p(\lambda_k, c)\right). \end{aligned}$$

Both F and ϕ are $\mathbb{R}$-differentiable with respect to the real and imaginary parts of the roots $c_j =: x_j + iy_j$, so we can address the minimization problem making use of Lagrange multipliers, which give rise to the non-linear system

$$\begin{cases} \nabla F(c) = \sum_{k=1}^{n_1} \mu_k \nabla \phi_k(c) \\ \phi(c) = 0 \end{cases} \tag{4.2}$$

in the unknowns x_j, y_j and μ_k, for $j = 1, ..., m$ and $k = 1, ..., n_1$.

4.2 Non-linear equations algorithms

When addressing a non-linear system $F(x) = 0$, where $F : \Omega \to \mathbb{R}$ is a continuously differentiable function defined over a convex open set $\Omega \subset \mathbb{R}^n$, it is natural to think of Newton-like methods. We decided to use the Matlab built-in function `fsolve`, that implements several variants of such algorithms.

As it is well known, the k-th Newton iteration basically consists in:

Solve $J_F(x_k)d_k = -F(x_k)$

Update $x_{k+1} = x_k + d_k$

where $J_F(x_k)$ is the Jacobian matrix of F at the point x_k. However, some drawbacks may arise when adopting the method in this raw form. The local convergence is probably the one that most inficiates the method's usefulness. To overcome this disadvantage, Newton algorithm is combined with *trust-region techniques*, that grant global convergence.

In particular, the results we report in this thesis were obtained through the *Trust-Region Dogleg* and the *Levenberg-Marquardt* algorithms (depending on which behaved best in terms of convergence and stability). Keeping in mind that the study of these methods falls beyond the scope of this thesis, for completeness we would like to briefly recall their principal features. First of all, let us introduce the main ideas behind the trust-region approach. After this we will shortly describe the two cited algorithms.

In this context, trust-region techniques are used to solve the non-linear minimization problem

$$\min_x f(x),$$

where $f(x) := \frac{1}{2}\|F(x)\|^2$. Indeed, every solution x^* of $F(x) = 0$ is a minimizer for $f(x)$. For simplicity from now on we assume $\Omega = \mathbb{R}^n$.

The idea behind the method can be summarized as follows. For every iteration:

- Approximate f with a quadratic model $m_k(d)$, which must reasonably reflect the behavior of f in a neighborhood N of x_k. Such N is the so-called *trust region*. In this case the quadratic model is given by

$$\begin{aligned} m_k(d) =& \frac{1}{2}\|F(x_k) + J_F(x_k)d\|^2 = f(x_k) + \nabla f(x_k)^T d + \frac{1}{2} d^T J_F(x_k)^T J_F(x_k) d \\ =& \frac{1}{2} F(x_k)^T F(x_k) + J_F(x_k)^T F(x_k) d + \frac{1}{2} d^T J_F(x_k)^T J_F(x_k) d. \end{aligned}$$

- Compute a trial step d_k which minimizes m_k over N (defined as a spherical set centered in x_k, with radius $\delta > 0$):

$$m_k(d_k) = \min_{d \in N} m_k(d) = \min_{\substack{d \in \mathbb{R}^n \\ \|d\| \leq \delta}} m_k(d) \qquad \text{(Trust-region subproblem).} \quad (4.3)$$

  Since m_k is a quadratic model, if N is opportunely chosen then $f(x_k + d_k) < f(x_k)$.
  We note that sometimes, due to a poor scaling of the problem, the trust-region is built as an ellipsoidal set, making use of a diagonal scaling matrix D: $N = \{d \in \mathbb{R}^n \text{ s.t. } \|Dd\| \leq \delta\}$.

- The new iterate $x_{k+1} := x_k + d_k$ is accepted when a sufficient reduction of f occurs. More precisely, d_k is considered a good step when the coefficient

$$\rho_f(d_k) = \frac{f(x_k) - f(x_k + d_k)}{m_k(0) - m_k(d_k)}$$

is larger than a certain fixed value in $(0, \frac{1}{4})$. In this case, $\rho_f(d_k)$ also influences the choice of the new trust-region radius.
If the achieved reduction is not sufficiently important, the current point remains x_k and the trust region N is shrunk; then the trial step computation is repeated.

The following theorem characterizes the solution of the trust-region subproblem (4.3).

Theorem 4.1. [13, Lemma 10.3] *Given a continuously differentiable function $F : \mathbb{R}^n \to \mathbb{R}$, a point $x \in \mathbb{R}^n$ and $\delta > 0$, the vector d is a solution of the trust-region subproblem*

$$\min_{\substack{d\in\mathbb{R}^n \\ \|d\|\leq\delta}} \|F(x) + J_F(x)d\|^2$$

if and only if there exists a scalar $\lambda \leq 0$ such that

$$\begin{cases} \left(J_F(x)^T J_F(x) + \lambda I\right) d = -J_F(x)^T F(x), \\ \lambda(\delta - \|d\|) = 0. \end{cases} \tag{4.4}$$

Remark 4.2. The trust-region step d falls between the Gauss-Newton step d_{GN} and the Cauchy step d_C. Recall that $d_{GN} := -J_F(x)^{-1}F(x)$, while d_C is instead defined as the minimum of m_k over the trust-region, along the steepest descent direction:

$$d_C := \begin{cases} -\overline{\tau}\nabla F(x) & \text{if } \overline{\tau}\|\nabla F(x)\| \leq \delta \\ -\delta\dfrac{\nabla F(x)}{\|\nabla F(x)\|} & \text{otherwise} \end{cases}, \qquad \overline{\tau} = \frac{\|\nabla F(x)\|^2}{\|J_F(x)\nabla F(x)\|^2}.$$

The statement can be deduced from (4.4): when $\lambda = 0$, then $d = -J_F(x)^{-1}F(x)$, while if $\lambda \neq 0$, then $\|d\| = \delta$ and $\lim_{\lambda\to+\infty} \|d\| = 0$. Moreover for a sufficiently large λ, we have $d \approx -\frac{1}{\lambda}\nabla F(x)$.

The two Matlab algorithms previously introduced differ in the way the trust-region subproblem (4.3) is coped with.

Trust-Region-Dogleg algorithm

As the name suggests, the Trust-Region-Dogleg algorithm adopts a dogleg strategy to solve (4.3). It consists in building the step d as the convex combination of d_C and d_{GN}:

$$d = d_C + \lambda(d_{GN} - d_C)$$

where λ is the largest value in $[0, 1]$ such that $\|d\| \leq \delta$. If $J_F(x_k)$ is (nearly) singular, d is just the Cauchy direction. In this way the dogleg algorithm is more robust then the Gauss-Newton method with a line search. In addition, it is efficient since it requires only one linear solve per iteration (for the computation of the Gauss-Newton step), that may be performed using an LU factorization of $J_F(x_k)$.

Levenberg-Marquardt algorithm

The Levenberg-Marquardt algorithm takes as step d the solution of (4.4), exploiting the fact that it represents the normal equations for the unconstrained linear least-squares problem

$$\min_d \frac{1}{2} \left\| \begin{bmatrix} J_F \\ \sqrt{\lambda} I \end{bmatrix} d + \begin{bmatrix} F \\ 0 \end{bmatrix} \right\|^2 .$$

This problem may be solved through a QR factorization of the coefficient matrix which involves both Householder reflections and Givens rotation for computational efficiency reasons.

4.3 Numerical experiments

In this section we present a few examples through which the results of Theorem 3.13 and Theorem 3.14 are tested. The selection was made with the purpose of covering a relatively wide range of settings while remaining concise.

The proposed tables and plots were obtained applying GMRES to the system $Ax = r_0$, with A as in (3.5) or (3.25), to get the convergence curve, and making use of the Matlab function `fsolve`, with the algorithms described in Section 4.2, to solve the non-linear system (4.2). The tolerance on the residual norm for GMRES was set to 10^{-8}, and so were set the tolerances involved in the `fsolve` function. In order to have comparable results for the different choices of eigenvalues, ε was set to 10^{-4} in all cases.

We begin by illustrating the result of Theorem 3.13. The first examples focus on problems with exclusively real eigenvalues, while the subsequent ones embrace the situation in which the spectrum lies in the complex plane. Subsequently, bound of Theorem 3.14 is tested on a few more general cases where ill-conditioning is extended to more than two eigenvalues (in the plots, ill-conditioned eigenvalues are highlighted using red and light blue colors). In all the examples, E_0 is a random unitary matrix.
Data in the tables illustrate and compare GMRES residual norm $\|r_k\|$ with the computed values of bound (3.22) or (3.26), depending on the number of ill-conditioned eigenvalues, sampled every five GMRES iterations. To facilitate the comparison, also the relative difference between the bound and the residual norm is reported. We report just the firt three significative digits of any value, to avoid an excessively heavy presentation. We would like to point out that, although some of these values are negative, they are actually identifiable with zero, as the corresponding absolute differences (from which they are computed) are of the same order of machine precision. In addition, the number of iterations `fsolve` required to terminate is given, too.

Example 1. The spectrum is located on the positive real axis. The aim is to test the quality of bound (3.22) as the position of the two eigenvalues a and b changes. More precisely, we present three variants in which spectra are built considering all the natural numbers between 1 and 50, and subsequently ascribing the ill-conditioning to the pairs $(a, b) = (49, 50)$, $(a, b) = (1, 50)$ and

$(a, b) = (30.5, 30.5001)$; in this last case, a and b substitute the eigenvalues 1 and 2. With this construction a and b are not outliers with respect to the other eigenvalues.

In the first two variants of Example 1 the two ill-conditioned eigenvalues are chosen to be relatively distant from each other. In particular, in the first case a and b are on the right extremum of the spectrum, while in the second case they delimit the remaining eigenvalues. The polynomial that solves the constrained minimization problem of Theorem 3.13 gives a bound that strictly follows the GMRES convergence curve. The small distance between a and b of the third case seems not to affect the quality of the results, either. Tables 4.1, 4.2 and 4.3 show very low relative difference values: the bound sticks to the GMRES residual norm up to the fifteenth decimal digit, with values that are sometimes very close to machine precision.

The next example further explores the situation in which the distance between a and b is of the same order of the perturbation ε, (i.e. when both the eigenvalues and their relative eigenspaces are very close to each other).

Example 2. In addition to the small distance between a and b, in this example we locate them outside the interval containing the other eigenvalues. Apart from a and b, the spectrum is composed by the natural numbers between 3 and 50. Case 1 analyzes the bound on the positive definite, nearly-singular problem given by $a = 10^{-4}$, $b = 2 \cdot 10^{-4}$, while case 2 looks into the indefinite problem characterized by $a = -20$ and $b = -19.9999$.

Like for the third case of Example 1, the proximity of a and b does not inficiate the sharpness of the bound, as the obtained values still reflect the behavior of the convergence curve. The small magnitude of the two eigenvalues in Example 2.1, however, provokes some instability and difficulties in the convergence of `fsolve`, as it is peculiarly noticeable in Figure 4.1 and Table 4.4. In the current and in the later examples, we will mark the possible unreliability of a result (assessed on the base of the `fsolve` warnings) by using asterisks in the tables and different colors in the plots (the black and the blue circles indicate respectively that `fsolve` stopped with a message of missing convergence or too many iterations done).

Since the distance between a and b is very small, both cases of Example 2 are suitable to qualitatively show how similarly GMRES behaves when A is replaced with A_J (see Section 3.3 and relation (3.16)). This effect can be appreciated in Figure 4.1 and Figure 4.2, as well as later in Figure 4.5, where the GMRES residual norms of the system $A_J x = r_0$ (cyan line) show a trend that is very akin to the one related to the original system $Ax = r_0$.

Remark 4.3. As already pointed out in Chapter 3, when the eigenvalues are close together, looking at the small distances (2-norm) between the matrices A, $A_D :=\text{diag}(a, b, \lambda_3, \ldots, \lambda_n)$ and $A_{D_\xi} :=\text{diag}(\xi, \xi, \lambda_3, \ldots, \lambda_n)$, where $\xi = \frac{a+b}{2}$, one may expect a similar behavior of the GMRES residual norms for the systems

k	$iter_{fsolve}$	$\|p_{con}(L)r_0\|$	$\|r_k\|$	$\frac{\|p_{con}(L)r_0\|-\|r_k\|}{\|r_k\|}$
5	1	1.45e-01	1.45e-01	6.50e-15
10	2	4.49e-02	4.49e-02	5.10e-15
15	1	1.08e-02	1.08e-02	3.83e-13
20	1	1.57e-03	1.57e-03	7.88e-15
25	1	1.26e-04	1.26e-04	4.31e-15
30	1	5.12e-06	5.12e-06	2.85e-12
35	1	9.36e-08	9.36e-08	1.92e-12

Table 4.1: *Example 1, case 1. Eigenvalues uniformly distributed in* $[1, 50]$*,* $a = 50$*,* $b = 49$*.*

k	$iter_{fsolve}$	$\|p_{con}(L)r_0\|$	$\|r_k\|$	$\frac{\|p_{con}(L)r_0\|-\|r_k\|}{\|r_k\|}$
5	1	1.91e-01	1.91e-01	4.05e-11
10	3	4.31e-02	4.31e-02	1.05e-10
15	2	1.09e-02	1.09e-02	1.51e-10
20	1	1.57e-03	1.57e-03	1.77e-10
25	1	1.26e-04	1.26e-04	2.00e-10
30	1	5.12e-06	5.12e-06	2.23e-10
35	1	9.36e-08	9.36e-08	2.21e-10

Table 4.2: *Example 1, case 2. Eigenvalues uniformly distributed in* $[1, 50]$*,* $a = 1$*,* $b = 50$*.*

k	$iter_{fsolve}$	$\|p_{con}(L)r_0\|$	$\|r_k\|$	$\frac{\|p_{con}(L)r_0\|-\|r_k\|}{\|r_k\|}$
5	0	6.93e-02	6.93e-02	-3.43e-14
10	0	6.71e-03	6.71e-03	2.06e-14
15	0	5.03e-04	5.03e-04	-6.67e-15
20	0	2.63e-05	2.63e-05	5.90e-14
25	0	8.58e-07	8.58e-07	3.20e-15
30	0	1.52e-08	1.52e-08	-1.89e-13

Table 4.3: *Example 1, case 3.* $a = 30.5$*,* $b = 30.5001$*, other eigenvalues uniformly distributed in* $[3, 50]$*.*

k	$iter_{fsolve}$	$\|p_{con}(L)r_0\|$	$\|r_k\|$	$\frac{\|p_{con}(L)r_0\|-\|r_k\|}{\|r_k\|}$
5	1	1.85e-01	1.85e-01	6.13e-13
10	4	1.43e-01	1.43e-01	3.15e-11
15	8	1.41e-01	1.41e-01	2.93e-09
20	13	1.41e-01	1.41e-01	7.57e-08
25	2	1.41e-01	1.41e-01	2.02e-13
30	100	1.41e-01	1.41e-01	9.57e-06
35	2	8.30e+03 (*)	1.41e-01	5.87e+04 (*)
40	1	7.41e-02 (*)	6.58e-02	1.26e-01 (*)
45	1	1.02e-04	1.02e-04	5.55e-06

Table 4.4: *Example 2, case 1.* $a = 10^{-4}$, $b = 2 \cdot 10^{-4}$, *other eigenvalues uniformly distributed in* $[3, 50]$.

k	$iter_{fsolve}$	$\|p_{con}(L)r_0\|$	$\|r_k\|$	$\frac{\|p_{con}(L)r_0\|-\|r_k\|}{\|r_k\|}$
5	1	1.70e-01	1.70e-01	-4.07e-12
10	1	3.80e-02	3.80e-02	-1.83e-15
15	1	3.41e-03	3.41e-03	3.82e-16
20	1	2.27e-04	2.27e-04	1.31e-15
25	1	1.01e-05	1.01e-05	2.29e-08
30	1	2.61e-07 (*)	2.61e-07	1.24e-05 (*)

Table 4.5: *Example 2, case 2.* $a = -20$, $b = -19.9999$, *other eigenvalues uniformly distributed in* $[3, 50]$.

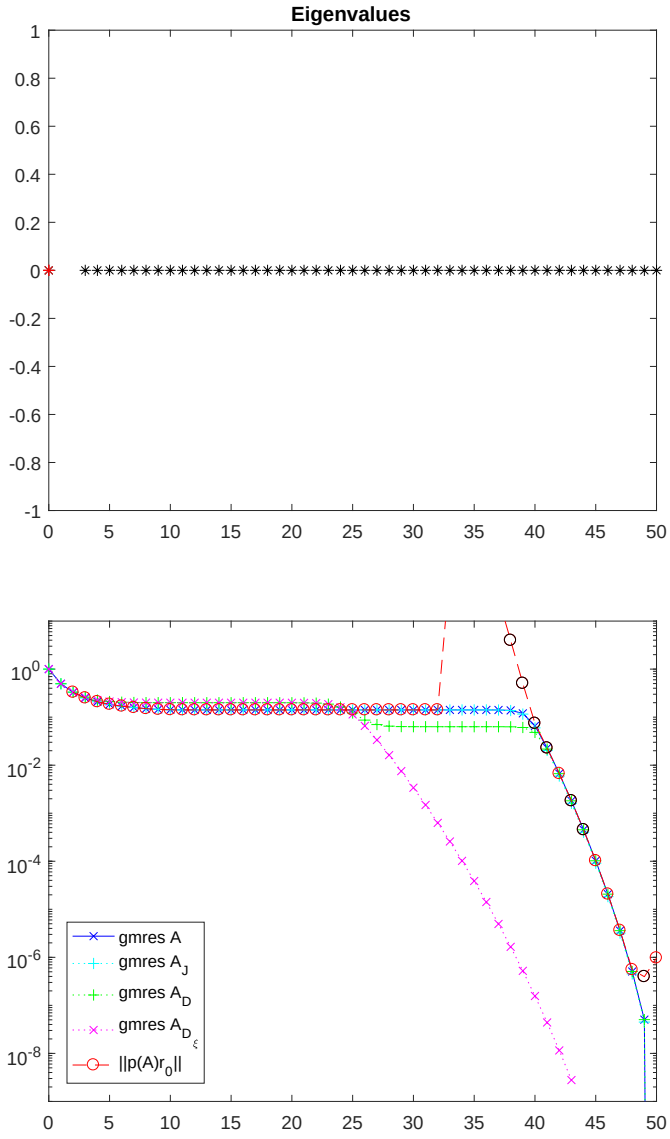

Figure 4.1: *Example 2, case 1. Top: Eigenvalues; Bottom: Convergence curve for* $Ax = r_0$ *(blue) and corresponding bound (red, black circles indicate unreliable results). Comparison is made with the convergence curves for* $A_J x = r_0$ *(cyan),* $A_D x = r_0$ *(green) and* $A_{D_\xi} x = r_0$ *(magenta).*

$Ax = r_0$, $A_D x = r_0$ and $A_{D_\xi} x = r_0$. Figure 4.1 and Figure 4.3 prove wrong this idea, since the blue, green and magenta lines show quite different trends. A possible explanation is given by the relations presented in Proposition 3.8: both the inequalities show the perturbation ε as a denominator, thus the smaller the value it assumes, the wider the distance between the residual norms is allowed to be.

The same holds for the comparison between the convergence curves of the systems $A_J x = r_0$ and $A_{D_\xi} x = r_0$, visible in Figure 4.1 and described in Proposition 3.9.

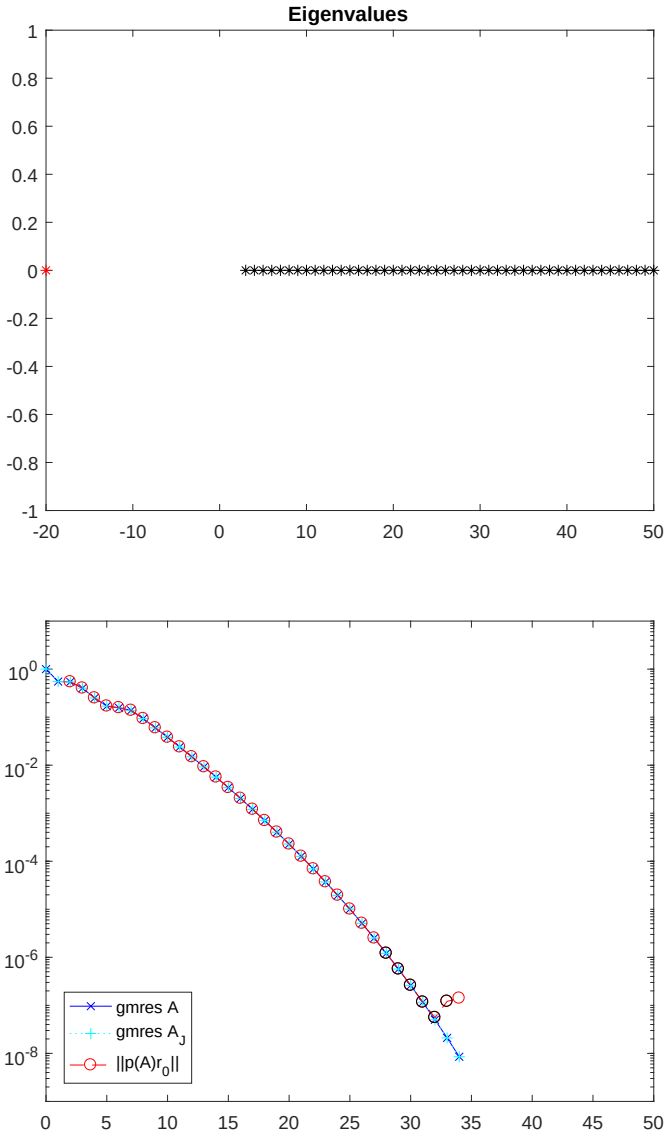

Figure 4.2: *Example 2, case 2. Top: Eigenvalues; Bottom: Convergence curve for* $Ax = r_0$ *(blue) and corresponding bound (red, black circles indicate unreliable results). Comparison is made with the convergence curves for* $A_J x = r_0$ *(cyan).*

Note that, in Figure 4.3, none of the diagonal cases shows some stagnation, in agreement with the fact that they are not affected by ill-conditioning (see Subsection 3.3.3). In Figure 4.1, anyway, the very small magnitude of the eigenvalues a, b and ξ results equally in an initial phase of very slow convergence.

As anticipated before, our experiments encompass cases characterized by a complex spectrum, too. For each of the next three proposed spectral environments, the instance in which just two eigenspaces are almost parallel is compared to the more general case where two groups of eigenspaces are one the perturbation of the other.

k	$iter_{fsolve}$	$\|p_{con}(L)r_0\|$	$\|r_k\|$	$\frac{\|p_{con}(L)r_0\|-\|r_k\|}{\|r_k\|}$
5	750	3.46e-01 (*)	3.46e-01	8.41e-04 (*)
10	417	2.22e-01 (*)	2.22e-01	4.56e-05 (*)
15	305	1.60e-02 (*)	1.56e-02	2.68e-02 (*)
20	17	1.62e-03	1.62e-03	6.48e-04
25	1	1.26e-04	1.26e-04	5.36e-05
30	1	5.12e-06	5.12e-06	1.72e-05
35	1	9.36e-08	9.36e-08	4.01e-05

Table 4.6: *Example 3, case 1. Purely imaginary eigenvalues.* $a = i$, $b = 2i$, *other eigenvalues uniformly distributed in* $[3i, 50i]$.

k	$iter_{fsolve}$	$\|\mathcal{J}\|\|p_{con}(L)\mathcal{J}^{-1}r_0\|$	$\|r_k\|$	$\frac{\|\mathcal{J}\|\|p_{con}(L)\mathcal{J}^{-1}r_0\|-\|r_k\|}{\|r_k\|}$
5	731	6.59e-01 (*)	6.84e-01	-3.59e-02 (*)
10	382	6.36e-01 (*)	5.28e-01	2.04e-01 (*)
15	343	4.02e-01 (*)	3.79e-01	6.20e-02 (*)
20	205	6.75e-03 (*)	4.37e-03	5.43e-01 (*)
25	80	5.40e-04	1.71e-04	2.16e+00
30	0	9.98e-06	3.69e-06	1.70e+00
35	0	2.86e-07	4.10e-08	5.97e+00

Table 4.7: *Example 3, case 2. Same (purely imaginary) eigenvalues of case 1. The ill-conditioned eigenvalues are gathered in the sets* $\{i, \dots, 5i\}$ *and* $\{6i, \dots, 10i\}$.

Note that, in these hypotheses, the GMRES residual polynomial may be complex valued, differently from what happened in the previous examples, in which having all real eigenvalues resulted in the reality of the residual polynomial as well.

Example 3. As a counterpart for the previous experimental settings, this example involves purely imaginary eigenvalues, namely the numbers $i, 2i, \dots, 50i$. In the first instance bound (3.22) is checked, with $a = i$ and $b = 2i$. In the second one, instead, ill-conditioning is attributed to the subsets $\{i, \dots, 5i\}$ and $\{6i, \dots, 10i\}$, as represented also in Figure 4.3 and Figure 4.4.

Even if `fsolve` warns about some convergence problems in the first iterations, sharpness of bound (3.22) remains remarkable: the digits in Table 4.6 exhibit a behavior that is coherent with the GMRES residual norm up to the fifth decimal place. On the other hand, the results of Example 3.2 regarding (3.26) are slightly loose if compared to all the previous experiments. Since this phenomenon is also encountered in later tests, it may be attributed to the wider dimension of the ill-conditioned space. In any case, data for bound (3.26) in Table 4.7 are of the same order of the residual norms, which remains a definitively positive result.

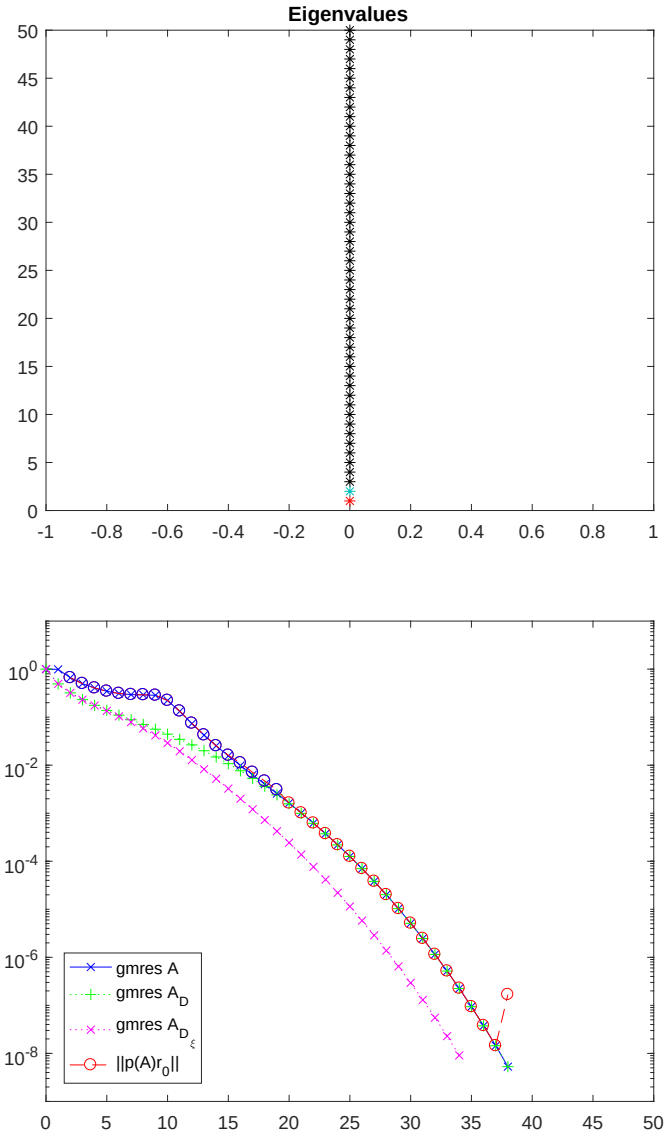

Figure 4.3: *Example 3, case 1. Top: Eigenvalues; Bottom: Convergence curve for* $Ax = r_0$ *(blue) and corresponding bound (red, black circles indicate unreliable results). Comparison is made with the convergence curves for* $A_D x = r_0$ *(green) and* $A_{D_\xi} x = r_0$ *(magenta).*

Example 4. The spectra of this example are composed by random, uniformly distributed eigenvalues located in a square region of the complex plane. Like in Example 3, the inequalities of both Theorem 3.13 and Theorem 3.14 are assessed. This time the latter is tested on a system matrix having two groups of ten eigenvalues each responsible for ill-conditioning.

Scaling the spectrum with the 10^{-4} factor permitted to generate one additional evidence for the comparison with the Jordan case (see Example 2).

What is immediately noticeable when looking at data in Table 4.8 and Table 4.9 is that `fsolve` stopped without performing any iteration. As suggested

k	$iter_{fsolve}$	$\|p_{con}(L)r_0\|$	$\|r_k\|$	$\frac{\|p_{con}(L)r_0\|-\|r_k\|}{\|r_k\|}$
5	0	1.18e-01	1.18e-01	2.39e-12
10	0	1.71e-02	1.71e-02	-8.84e-13
15	0	2.48e-03	2.48e-03	2.61e-12
20	0	2.67e-04	2.67e-04	-8.70e-13
25	0	2.09e-05	2.09e-05	1.57e-12
30	0	7.59e-07	7.59e-07	7.33e-12
35	0	1.95e-08	1.95e-08	1.02e-09

Table 4.8: *Example 4, case 1. Eigenvalues (pseudo-)randomly chosen in* $[0, 10^{-4}] \times [0, 10^{-4}]i \in \mathbb{C}$.

k	$iter_{fsolve}$	$\|\mathcal{J}\|\|p_{con}(L)\mathcal{J}^{-1}r_0\|$	$\|r_k\|$	$\frac{\|\mathcal{J}\|\|p_{con}(L)\mathcal{J}^{-1}r_0\|-\|r_k\|}{\|r_k\|}$
5	0	1.00e-01	9.40e-01	6.29e-02
10	0	9.56e-01	8.84e-01	8.16e-02
15	0	2.66e-02	2.58e-02	3.02e-02
20	0	1.95e-03	1.14e-03	7.09e-01
25	0	5.32e-04	1.18e-04	3.50e+00
30	0	1.21e-05	5.21e-06	1.32e+00
35	0	3.52e-07	1.04e-07	2.38e+00

Table 4.9: *Example 4, case 2. Eigenvalues (pseudo-)randomly chosen in* $[0, 10^{-4}] \times [0, 10^{-4}]i \in \mathbb{C}$, *the two sets of eigenvalues responsible for ill-conditioning have (both) cardinality ten.*

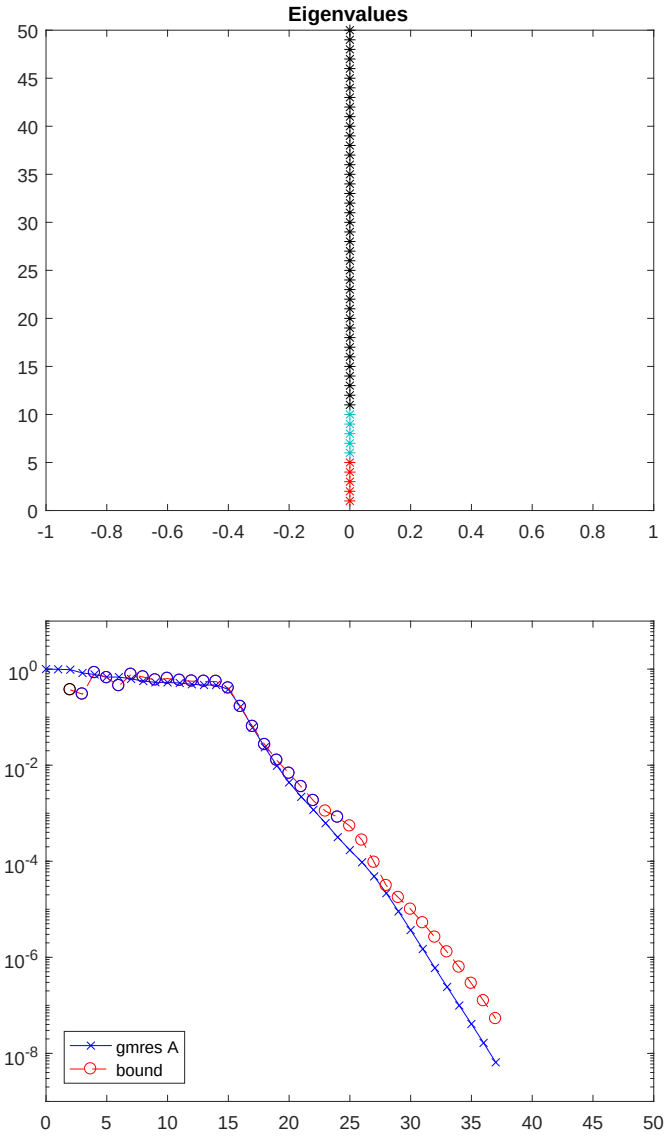

Figure 4.4: *Example 3, case 2. Top: Eigenvalues; Bottom: Convergence curve for $Ax = r_0$ (blue) and corresponding bound (red, black circles indicate unreliable results).*

in Remark 4.1, the initial guess was chosen to be the vector of the harmonic Ritz values. This means that, within the given tolerances, the minimizer polynomial is indeed the GMRES residual polynomial. We will further discuss this important issue later.

For what concerns the quality of the results, it is confirmed that the bounds not only provide a threshold curve for the GMRES convergence history, but also quantitatively describe its trend.

Example 5. The last example we present refers to a PDE. In particular, the considered system matrix A is similar to the one obtained from the discretization

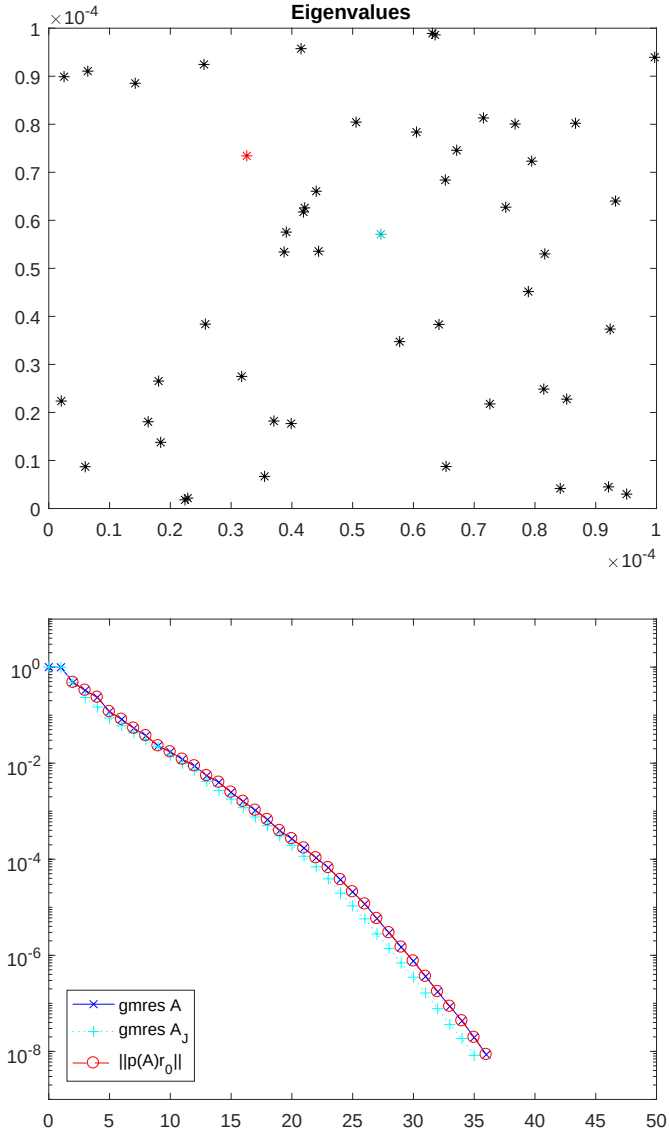

Figure 4.5: *Example 4, case 1. Top: Eigenvalues; Bottom: Convergence curve for $Ax = r_0$ (blue) and corresponding bound (red). Comparison is made with the convergence curves for $A_J x = r_0$ (cyan).*

of the 2D stationary advection-diffusion problem

$$\begin{cases} -u_{xx} - u_{yy} + f(x)u_x + g(y)u_y = 1 & (x,y) \in (0,1)^2 \\ u(x,y) = 0 & \text{for } (x,y) \in \Gamma = \partial(0,1)^2, \end{cases}$$

where $f(x) = -2.5x$ and $g(y) = -50y$. With this choice for the coefficient functions, A is real with all complex conjugate eigenvalues. This results in a real valued GMRES polynomial, despite the complex nature of the spectrum. Ill-conditioning is assigned as follows: when calculating bound (3.22) we have $b = \bar{a}$, while for bound (3.26) the two groups of eigenvalues responsible for ill-

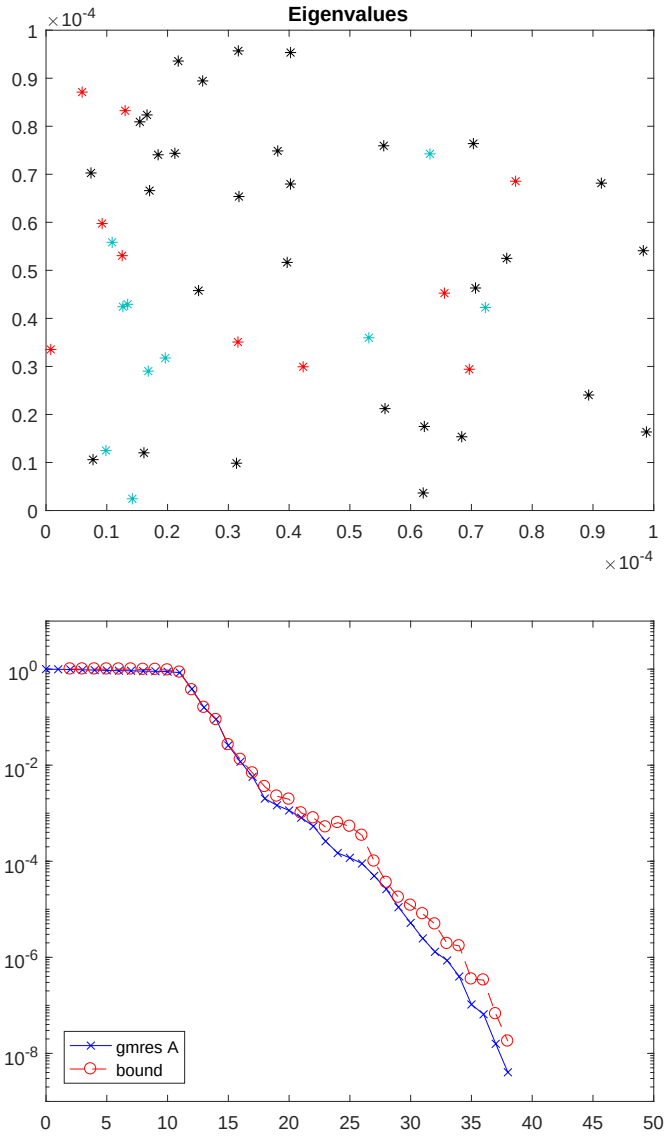

Figure 4.6: *Example 4, case 2. Top: Eigenvalues; Bottom: Convergence curve for* $Ax = r_0$ *(blue) and corresponding bound (red).*

conditioning are composed by five elements, chosen so that complex conjugate values belonged to the same set (see Figure 4.8).

The situation reported in Table 4.10, Table 4.11, Figure 4.7 and Figure 4.8 is totally similar to the two previous examples. The computed bound sticks to the convergence curve in the first case, and looses a bit of its sharpness in the second instance, when ill-conditioning is extended, still maintaining a nice behavior.

A common remark for case 2 of Examples 3, 4 and 5 is that convergence of both GMRES and `fsolve` encountered a few more difficulties, presumably due to the larger dimension of the ill-conditioned part of the system matrix. On the one hand, the residual norm history shows a quite extended initial phase of

k	$iter_{fsolve}$	$\|p_{con}(L)r_0\|$	$\|r_k\|$	$\frac{\|p_{con}(L)r_0\|-\|r_k\|}{\|r_k\|}$
5	29	8.54e-02 (*)	8.54e-02	-1.98e-11 (*)
10	409	8.88e-03 (*)	8.88e-03	1.72e-08 (*)
15	32	9.33e-04 (*)	9.33e-04	1.14e-11 (*)
20	1	7.47e-05	7.47e-05	-1.15e-11
25	1	2.88e-06	2.88e-06	1.61e-11
30	1	4.08e-08	4.08e-08	4.15e-09

Table 4.10: *Example 5, case 1. Complex conjugate eigenvalues, $b = \bar{a}$.*

k	$iter_{fsolve}$	$\|\mathcal{J}\|\|p_{con}(L)\mathcal{J}^{-1}r_0\|$	$\|r_k\|$	$\frac{\|\mathcal{J}\|\|p_{con}(L)\mathcal{J}^{-1}r_0\|-\|r_k\|}{\|r_k\|}$
5	641	1.07e-01 (*)	9.50e-01	-8.87e-01 (*)
10	353	1.37e-02 (*)	1.34e-02	2.74e-02 (*)
15	3	1.25e-03	1.13e-03	1.01e-01
20	0	1.50e-04	5.90e-05	1.54e+00
25	0	9.21e-06	1.67e-06	4.52e+00
30	0	9.93e-08	2.56e-08	2.88e+00

Table 4.11: *Example 5, case 2. Same eigenvalues of case 1. The two groups of eigenvalues responsible for ill-conditioning are chosen so that complex conjugate values belonged to the same set.*

stagnation. On the other hand, `fsolve` computations relative to the same first iterations are affected by some instability. Indeed, at a first glance, Figure 4.4, Figure 4.6 and Figure 4.8 seem not to agree with the result in Theorem 3.14, as the bound appears to be located beneath the convergence curve. However, keeping in mind that only the red marks indicate convergence of `fsolve`, the obtained values verify the theoretical predictions and represent a significantly good bound for the GMRES convergence curve.

In conclusion, thanks to the numerical evidence provided in this section, Theorem 3.13 and Theorem 3.14 indeed seem to dispense an accurate model for GMRES convergence behavior in the specific case in which ill-conditioning is attributable to some specific eigenvalues and eigenspaces.

We would like to notice how in most of our experimentations it happened that `fsolve` stopped immediately, without doing any iteration, which means that the GMRES residual polynomial (our initial guess) solved the constrained minimization problem (within the specified tolerances). This has great significance for the convergence model presented in this thesis, for it supports the fact that constraints (3.20) and (3.24) not only provide a bound for the GMRES residual norms, but actually describe the GMRES residual polynomial behavior.

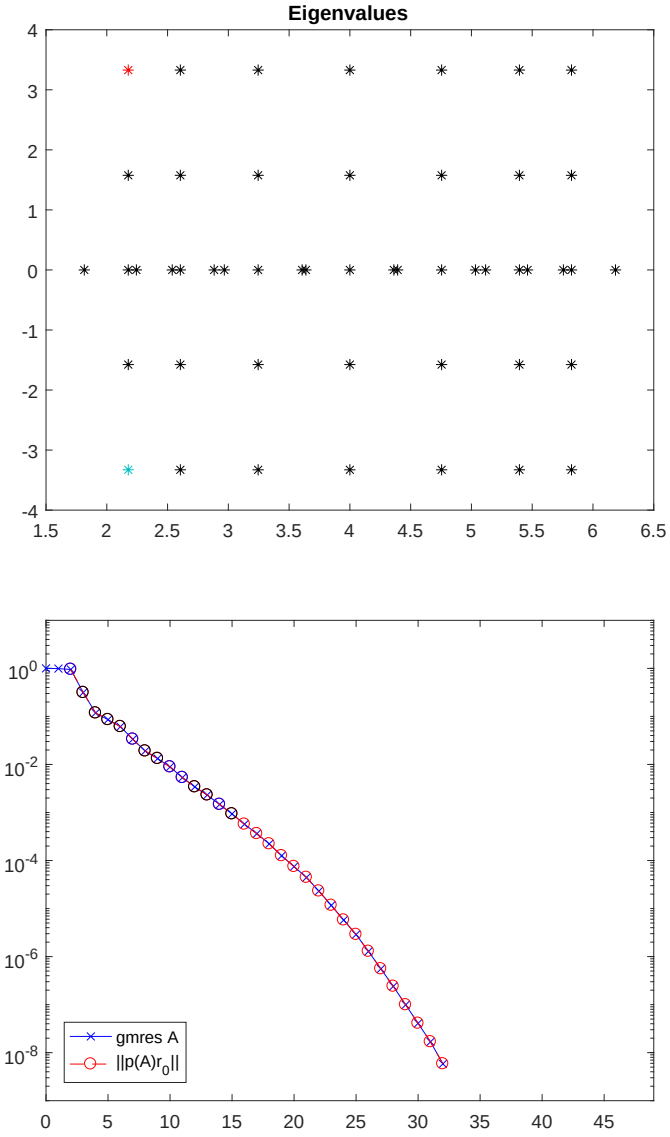

Figure 4.7: *Example 5, case 1. Top: Eigenvalues; Bottom: Convergence curve for* $Ax = r_0$ *(blue) and corresponding bound (red, black circles indicate unreliable results).*

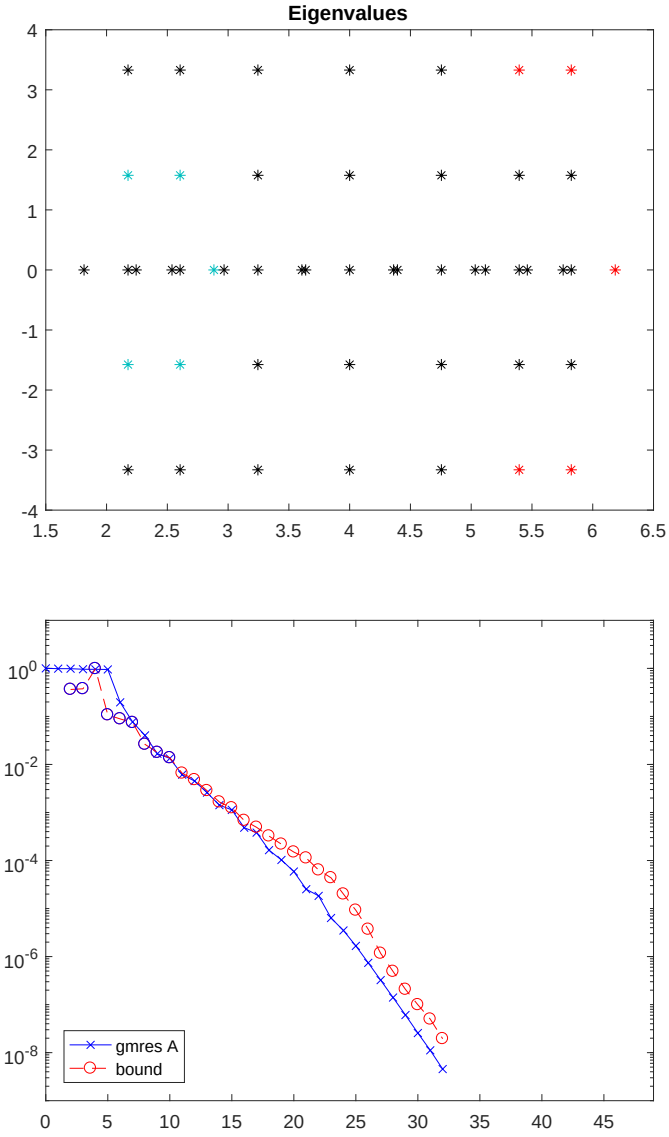

Figure 4.8: *Example 5, case 2. Top: Eigenvalues; Bottom: Convergence curve for* $Ax = r_0$ *(blue) and corresponding bound (red, black circles indicate unreliable results).*

Bibliography

[1] M. Arioli, V. Pták and Z. Strakoš, *Krylov sequences of maxima2l length and convergence of GMRES*, BIT, 38(1998), pp. 636-643.

[2] S. L. Campbell, I. C. F. Ipsen, C. T. Kelley and C. D. Meyer, *GMRES and the Minimal Polynomial*, BIT Numer. Math., 36(1996), pp. 664-675.

[3] F. Chatelin, *Eigenvalues of Matrices*, Wiley, Chichester, England, 2^{nd} ed., 1993.

[4] L. Eldén and V. Simoncini, *Solving Ill-Posed Linear Systems with GMRES and a Singular Preconditioner*, SIAM J. Matrix Anal. Appl., 33(2012), pp. 1369-1394.

[5] M. Embree, *How descriptive are GMRES convergence bounds?*, Technical report, Oxford University Computing Laboratory, Oxford, UK, (anno??).

[6] B. Fisher and R. Freund, *Chebyshev polynomials are not always optimal*, J. Approx. Theory, 65(1991), pp. 261-272.

[7] R. W. Freund, *Quasi-Kernel Polynomials and Convergence Results for Quasi-Minimal Residual Iterations*, RIACS Technical report 92.07, 1992.

[8] A. Greenbaum and Z. Strakoš, *Predicting the behavior of finite precision Lanczos and conjugate gradient computations*, SIAM J. Matrix Anal. Appl., 13(1992), pp. 121-137.

[9] C. Lanczos, *Solution of systems of linear equations by minimized iterations*, J. Reasearch Nat. Bur. Standards, 49(1952), pp. 33-53.

[10] J. Liesen and Z. Strakoš, *Krylov Subspace Methods, Principles and Analysis*, Oxford University Press, Oxford, UK, 1^{st} ed., 2013.

[11] J. Liesen and P. Tichý, *Convergence analysis of Krylov subspace methods*, GAMM-Mitt., 27(2004), pp. 153-173.

[12] T. A. Manteuffel, *Adaptive procedure for estimating parameters for the nonsymmetric Tchebychev iteration*, Numer. Math.,31(1978), pp. 183-208.

[13] J. Nocedal and S.J. Wright, *Numerical Optimization*, Springer-Verlag New York, Inc, 1^{st} ed., 1999.

[14] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS, Boston, 1^{st} ed., 1996.

[15] V. Simoncini, *On the Convergence of Restarted Krylov Subspace Methods*, SIAM J. Matrix Anal. Appl., 22(2000), pp. 430-452.

[16] V. Simoncini and D. B. Szyld, *Recent computational developments in Krylov subspace methods for linear systems*, Numer. Linear Algebra Appl., 14(2007), pp. 1-59.

[17] L. N. Trefethen , *Approximation theory and numerical linear algebra*, in Algorithms for approximation, II (Shrivenham, 1998), Chapman and Hall, London, 1990, pp. 336-360.

Aknowledgments

I would like to sincerely thank Professor Valeria Simoncini, supervisor of this thesis, for the time she dedicated to me during the last months, for her support, availability and precious suggestions, that allowed me to elaborate this thesis.

I am also grateful for Margherita Porcelli's helpful insights regarding non-linear equations.

With love I thank my family, for the encouragment and the economical support I received during all the course of my studies.

Finally, a big "thank you" to all my friends, who donated me wonderful memories and helped me to grow during the last years. Thanks to Giulia and Arianna for being such good listeners. Thanks to Anna and the other mathematicians friends, with whom I shared the passion for Mathematics. Thanks to Jasper, who taught me the importance of smiling. Thanks to my musician friends, for having often reminded me to let my voice be heard.

Ringraziamenti

Vorrei sinceramente ringraziare la Professoressa Valeria Simoncini per tutto il tempo dedicatomi in questi mesi, per il suo supporto, la disponibilità ed i preziosi consigli, che mi hanno permesso di elaborare questa tesi.

Vorrei inoltre esprimere la mia gratitudine a Margherita Porcelli, che gentilmente mi ha fornito nozioni e consigli utili riguardo la risoluzione di sistemi non lineari.

Ringrazio di cuore i miei familiari per avermi accompagnata ed incoraggiata durante tutto il percorso universitario con il loro sostegno, sia morale che economico.

Infine, un grande "grazie" a tutti i miei amici, che mi hanno regalato giornate memorabili e mi hanno aiutato a crescere durante gli anni dell'Università. Grazie a Giulia e Arianna, per il loro ascolto e affetto. Grazie ad Anna e gli altri amici matematici, con cui ho condiviso la passione per la Matematica. A Jasper, per avermi insegnato a sorridere. Ai miei amici musicisti, per avermi spesso ricordato di far valere la mia voce.