Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea Magistrale in Matematica

On the numerical solution of linear matrix equations stemming from finite elements

Tesi di Laurea Magistrale in Analisi Numerica

Relatrice:
Chiar.ma Prof.ssa
Valeria Simoncini

Presentata da:
Daniele Toni

III Sessione
Anno Accademico 2023-2024

A quelli che sono morti, e a quelli che restano.
(H. Murakami)

Abstract

In this work we study and analyze different methods to numerically solve Lyapunov equations. We state and prove several structural properties for this equation, and then we look for numerical methods that exploit these properties and for which also the numerical solution preserves such properties. For this reason, we focus our study on the so called projection methods, for which we give a bound of convergence of the numerical solution found by these methods, for different spaces of projection. Besides, we apply such methods to a particular case of Lyapunov equation, which is obtained by using an appropriate finite elements discretization of the two dimensional Poisson equation. Moreover, we extend and discuss, in the context of iterative methods for the numerical solution of a matrix equations, a well known property on the norm of the total error due to the finite element discretization and the projection method used to obtain the numerical solution.

Sommario

In questo elaborato studiamo metodi risolutivi per equazioni di Lyapunov. In particolare, per tale equazione introduciamo e proviamo proprietà strutturali, e cerchiamo metodi numerici che riescano ad utilizzare tali proprietà, e che restituiscano soluzioni che mantengano tali proprietà, e siano in questo modo interpretabili. Per tale motivo, concentreremo il nostro studio sui metodi proiettivi, e daremo una stima di convergenza della soluzione numerica ottenuta utilizzando tali metodi per vari spazi di proiezione. Inoltre, applichiamo tali risultati ad una particolare equazione di Lyapunov, ottenuta discretizzando mediante elementi finiti, fissando gli appropriati elementi, chiamata equazione di Poisson bidimensionale. Infine, estendiamo e discutiamo nel contesto dei metodi iterativi per la soluzione numerica delle equazioni matriciali una nota proprietà riguardante la norma dell'errore totale commesso discretizzando il problema e risolvendolo numericamente mediante metodi proiettivi.

Contents

Introduction

In the first chapter of the thesis we introduce various preliminary results and concept that turns out to be useful for the rest of the work.

In the second chapter we develop an analytical study of the Poisson equation; in particular we recover the weak form of the problem, and then we introduce a new test set in which we study this problem.

In the third chapter we introduce a finite-dimensional test set, and we introduce the concept of Galerkin solution. Then, we fix a particular basis of the finite-dimensional test set, and we use this basis to derive a Lyapunov equation related to the problem.

In the forth chapter we analyze different properties related to the solution of the Lyapunov equations. Some of properties turns out to be very useful also in the following of this work, as they suggest that use of certain methods, that catches these properties.

In the fifth chapter we introduce different methods to numerically solve the Lyapunov equation. In particular, we introduce the Bartels-Steward algorithm for the small scale setting, and we focus on projection method for the large scale setting. For the last method we study different spaces of projection, and for each of them we estimate the rate of convergence of the approximate solution.

In the sixth chapter we extend in the matrix setting a property related to the error of the solution the Poisson equation, then we give another point of view of the result and we numerically validate the result.

Chapter 1

Preliminaries

In this chapter we recall some preliminary results and concepts that we use in the following chapters. We start with two basic definitions to introduce Sobolev spaces.

Definition 1.1. Let $\Omega \in \mathbb{R}^n$, let $\alpha \in \mathbb{N}^n$ and let $\phi \in C_0^\infty(\Omega)$. We define the α partial derivative of ϕ as

$$D^\alpha \phi := \frac{\partial^{\alpha_1} \dots \partial^{\alpha_n}}{\partial_{x_1} \dots \partial_{x_n}} \phi.$$

Moreover, we define the weight of α as

$$|\alpha| = \sum_{i=1}^{n} \alpha_i$$

Definition 1.2. Let $\Omega \in \mathbb{R}^n$, and let $\alpha \in \mathbb{N}^n$. A function $v : \Omega \to \mathbb{R}$ is weakly differentiable with α-weak partial derivative $u = D^\alpha v$ if for all $\phi \in C_0^\infty(\Omega)$ it holds that:

$$\int_\Omega v D^\alpha \phi \, \mathrm{d}w = (-1)^{|\alpha|} \int_\Omega u\phi \, \mathrm{d}w$$

Definition 1.3 (Sobolev spaces)**.** For any $p \in \mathbb{Z}^+$, we define the Sobolev spaces as:

$$H^p(\Omega) := \{v \in L^2(\Omega) : D^\alpha v \in L^2(\Omega), |\alpha| \leq p\}$$

These spaces can be equipped with the following norms:

$$\|v\|^2_{H^p(\Omega)} := \sum_{|\alpha| \leq p} \|D^\alpha u\|^2_{L^2(\Omega)}, \qquad \|D^\alpha u\|^2_{L^2(\Omega)} = \int_\Omega |D^\alpha u|^2 \, \mathrm{d}w$$

Indeed, the L^2-norm is just a semi-norm, as for $u = 0$ almost everywhere with respect to the Lebesgue measure it holds that

$$\|u\|_{L^2} = 0.$$

As a consequence, we can't control the behaviour of an L^2 function in a zero Lebesgue measure set, such as the boundary. For this reason, we define the compactly supported Sobolev spaces H_0^p as the closure of C_0^∞ with respect to the Sobolev norms

$$H_0^p(\Omega) := \overline{C_0^\infty(\Omega)}^{\|\cdot\|_{H^p(\Omega)}}.$$

Besides, we define H^{-p} as the topological dual of H_0^p,

$$H^{-p}(\Omega) := (H_0^p(\Omega))'.$$

As a consequence, it is a space of distributions. Moreover, it is crucial to observe that the spaces H^{-p} inherit the topology of H^p. We recall from [1] a crucial property of the Sobolev spaces. This property regards the structure of the spaces in terms of their geometry.

Proposition 1.4. *The normed spaces $H^p(\Omega)$, $H_0^p(\Omega)$ are Hilbert spaces.*

In the work we heavily use the Divergence Theorem, which is the following classical result.

Definition 1.5 (Divergence)**.** Let $\Omega \subseteq \mathbb{R}^n$, let $F : \Omega \to \mathbb{R}^n$, $F \in C^1(\overline{\Omega})$. We define the divergence of F as:

$$\operatorname{div}(F) := \nabla \cdot F = \sum_{i=1^n} \frac{\partial F_i}{\partial x_i}$$

Theorem 1.6 (Divergence Theorem)**.** *Let $\Omega \subseteq \mathbb{R}^n$, let $F : \Omega \to \mathbb{R}^n$, $F \in C^1(\overline{\Omega})$. It holds that*

$$\int_\Omega \operatorname{div}(F)(\omega)\ \mathrm{d}\omega = \int_{\partial\Omega^+} F(x) \cdot \nu\ \mathrm{d}H_1(x)$$

Where $\partial\Omega^+$ is the positively oriented boundary, while ν is the outward pointing unit normal vector to $\partial\Omega^+$,

We introduce the definitions of convolution ad mollifiers.

Definition 1.7. (Convolution) Let $f, g \in L^1(\mathbb{R}^n)$. We define the convolution of f and g as the function

$$(f * g)(x) = \int_{\mathbb{R}^n} f(y)g(x-y)\,\mathrm{d}y = \int_{\mathbb{R}^n} f(x-y)g(y)\ \mathrm{d}y$$

Definition 1.8. (Mollifier) We define a mollifier as a function $\phi_\epsilon : \mathbb{R}^n \to \mathbb{R}$ which satisfies the following properties:

i. $\phi_\epsilon \in C^\infty(\mathbb{R}^n)$,

ii. $supp(\phi_\epsilon) \subseteq B_\epsilon(0)$,

iii. $(\phi_\epsilon) \geq 0$,

iv. $\displaystyle\int_{\mathbb{R}^n} \phi_\epsilon(x)\, dx = 1.$

In the following we give a classical example of a mollifier.

Example 1.1. The following function $\phi_\epsilon : \mathbb{R}^n \to \mathbb{R}$ is a mollifier.

$$\tilde{\phi}_\epsilon(x) = \begin{cases} e^{\frac{1}{\|\frac{x}{\epsilon}\|^2 - 1}} & \text{if } \|x\| < 1 \\ 0 & \text{if } \|x\| \geq 1 \end{cases}, \qquad \phi_\epsilon = c_\epsilon \tilde{\phi}_\epsilon.$$

Where the constant c_ϵ is such that the condition *iv.* is satisfied.

We recall a classical result from [10], called Friedrichs' Lemma.

Lemma 1.9 (Friedrichs)**.** *Let $\phi_\epsilon \in C_0^\infty(\mathbb{R}^n)$ be a mollifier, and define the functions $u_\epsilon := u * \phi_\epsilon$, $f_\epsilon := f * \phi_\epsilon$. If $u \in L^2_{loc}(\mathbb{R}^n)$, we have that, for all $K \subseteq \mathbb{R}^n$ compact set*

$$-\Delta u_\epsilon - f_\epsilon \xrightarrow[\epsilon \to 0]{L^2(K)} 0$$

In the following we use the Kronecker product in different contexts.

Definition 1.10 (Kronecker product)**.** Let $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$, $\mathbf{B} \in \mathbb{R}^{m_1 \times m_2}$. We define the operator $\otimes : \mathbb{R}^{n_1 \times n_2} \times \mathbb{R}^{m_1 \times m_1} \to \mathbb{R}^{n_1 m_1 \times n_2 m_2}$ as follows:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{1,1}\mathbf{B} & \cdots & a_{1,n_2}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{n_1,1}\mathbf{B} & \cdots & a_{n_1,n_2}\mathbf{B} \end{bmatrix} \in \mathbb{R}^{n_1 m_1 \times n_2 m_2}.$$

In the following, we state some useful properties of the Kronecker product.

Proposition 1.11. *Let all the algebraic operations stated in the following properties be well posed. The followings hold.*

i. $\mathbf{A} \otimes (\mathbf{B} + \mathbf{C}) = \mathbf{A} \otimes \mathbf{B} + \mathbf{A} \otimes \mathbf{C}$,

ii. $(\mathbf{A} + \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes \mathbf{C} + \mathbf{B} \otimes \mathbf{C}$,

iii. $(k\mathbf{A}) \otimes \mathbf{B} = k(\mathbf{A} \otimes \mathbf{B}) = \mathbf{A} \otimes (k\mathbf{B})$,

iv. $(\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C})$,

v. $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD})$,

vi. $(\mathbf{A} \otimes \mathbf{B})^T = (\mathbf{A}^T \otimes \mathbf{B}^T)$,

vii. $(\mathbf{A} \otimes \mathbf{B})^{-1} = (\mathbf{A}^{-1} \otimes \mathbf{B}^{-1})$.

Definition 1.12. Let $\mathbf{X} = [x_1, ..., x_m] \in \mathbb{R}^{n\times m}$. We define the operator vec : $\mathbb{R}^{n\times m} \to \mathbb{R}^{nm}$ as follows:

$$\text{vec}(\mathbf{X}) = \text{vec}[x_1, ..., x_m] = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}.$$

By using these operators, we can transform the matrix-matrix operations $\mathbf{AX}+\mathbf{XA}$, for $\mathbf{A}, \mathbf{X} \in \mathbb{R}^{n\times n}$ and $\mathbf{AXB}+\mathbf{BXA}$ for $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{m\times m}$ and $\mathbf{X} \in \mathbb{R}^{n\times m}$ into matrix-vector operations, as follows:

$$\text{vec}(\mathbf{AX} + \mathbf{XA}) = (\mathbf{I} \otimes \mathbf{A} + \mathbf{A}^T \otimes \mathbf{I})\text{vec}(\mathbf{X}) \tag{1.1}$$

$$\text{vec}(\mathbf{AXB} + \mathbf{BXA}) = (\mathbf{B}^T \otimes \mathbf{A} + \mathbf{A}^T \otimes \mathbf{B})\text{vec}(\mathbf{X}) \tag{1.2}$$

In the following theorem, we characterize the spectrum of the matrix $(\mathbf{I} \otimes \mathbf{A} + \mathbf{A}^T \otimes \mathbf{I})$.

Theorem 1.13. *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$, and let $\{(v_i, \lambda_i)\}_{i=1,...,n}$ be a basis of eigenpairs of $\mathbf{A}$. It holds that $\{(\lambda_i + \lambda_j, v_i \otimes v_j)\}_{i,j=1,...,n}$ is a basis of eigenpairs of $\mathbf{A} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{A}$.*

Proof. By applying the vector $v_i \otimes v_j$ to the matrix $\mathbf{A} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{A}$, we obtain

$$\begin{aligned} &(\mathbf{A} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{A})(v_i \otimes v_j) \\ =&(\mathbf{A} \otimes \mathbf{I})(v_i \otimes v_j) + (\mathbf{I} \otimes \mathbf{A})(v_i \otimes v_j) \\ =&(\mathbf{A}v_i \otimes \mathbf{I}v_j) + (\mathbf{I}v_i \otimes \mathbf{A}v_j) \\ =&(\lambda_i + \lambda_j)(v_i \otimes v_j) \end{aligned}$$

□

The next theorem is a crucial result for our work.

Theorem 1.14. *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a matrix such that if λ is an eigenvalue of $\mathbf{A}$, then $-\lambda$ is not an eigenvalue of $\mathbf{A}$. For each $\mathbf{C} \in \mathbb{R}^{n\times n}$, there exists a unique matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$ satisfying the following equation*

$$\mathbf{AX} + \mathbf{XA}^T = \mathbf{C} \tag{1.3}$$

Proof. By taking the operator vec to equation (1.3), we obtain the equivalent linear system

$$(\mathbf{I} \otimes \mathbf{A} + \mathbf{A} \otimes \mathbf{I})\text{vec}(\mathbf{X}) = \text{vec}(\mathbf{C}) \tag{1.4}$$

As a consequence of Theorem, 1.13, if the matrix $\mathbf{A}$ satisfies the eigenvalue hypothesis, then the matrix $(\mathbf{I} \otimes \mathbf{A} + \mathbf{A} \otimes \mathbf{I})$ is invertible. Hence, for each $\mathbf{C} \in \mathbb{R}^{n\times n}$ there exists a unique solution $\text{vec}(\mathbf{X})$ of equation (1.4). By taking the unique square matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$ obtained by inverting the operator vec in dimensions $n \times n$ we obtain the result. □

A particular case when the eigenvalue hypothesis holds is when the matrix $\mathbf{A}$ is symmetric positive definite.

In the thesis, we also use the concept of spectral interval of a matrix $\mathbf{A}$, whose eigevalues are real. It is defined as follows.

Definition 1.15 (Spectral interval)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a matrix with real eigenvalues. We define the spectral interval of $\mathbf{A}$ as the interval $[\lambda_{min}(\mathbf{A}), \lambda_{max}(\mathbf{A})]$.

Definition 1.15 can be generalized as follows.

Definition 1.16 (Field of values)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times n}$. We define the field of values of $\mathbf{A}$ as the set of all Rayleigh quotients, that is:

$$W(\mathbf{A}) := \{x^*\mathbf{A}x, x \in \mathbb{C}^n, \|x\| = 1\}$$

This generalization leads us to generalize several statements of the thesis when the involved matrices do not have real eigenvalues, by working with the set $W(\mathbf{A})$ instead of the set $[\lambda_{min}(\mathbf{A}), \lambda_{max}(\mathbf{A})]$.

In the following we use the first Gerschgorin Theorem.

Theorem 1.17 (First Gerschgorin Theorem)**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$. The eigenvalues of $\mathbf{A}$ are contained in the set*

$$\bigcup_{i=1}^{n} K_i, \quad K_i = \{z \in \mathbb{C}, |z - a_{i,i}| \leq \sum_{j=1,\ j\neq i}^{n} |a_{i,j}|\}$$

A fundamental concept we use in the thesis is the so called Krylov space.

Definition 1.18 (Krylov space)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times n}$, and $c \in \mathbb{R}^n$. We define the k-th Krylov space of $\mathbf{A}$ and c as

$$\mathcal{K}_k(\mathbf{A}, c) := span\{c, \mathbf{A}c, \ldots, \mathbf{A}^{k-1}c\}.$$

Definition 1.19 (Block Krylov space)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times n}$, and let $\mathbf{C} \in \mathbb{R}^{n\times p}$ be a tall matrix. We define the k-th Block Krylov space of $\mathbf{A}$ and $\mathbf{C}$ as

$$\mathcal{K}_k^{\square}(\mathbf{A}, \mathbf{C}) := blkspan\{\mathbf{C}, \mathbf{AC}, \ldots, \mathbf{A}^{k-1}\mathbf{C}\}.$$

In the work we need the so called modified Gram-Schmidt algorithm. Given a matrix $\mathbf{U}_k \in \mathbb{R}^{n\times n}$, $\mathbf{U}_k = [u_1, \ldots, u_k]$, such that the columns of $\mathbf{U}_k$ are linearly independents, the modified Gram-Schmidt algorithm generates a matrix with orthonormal columns $\mathbf{V}_k$ such that $Range(\mathbf{V}_k) = Range(\mathbf{U}_k)$, as follows.

As first column we take $v_1 = \frac{u_1}{\|u_1\|}$. Then we take

$$\hat{v}_2 = u_2 - v_1 v_1^T u_2,$$

and we take $v_2 = \frac{\hat{v}_2}{\|\hat{v}_2\|}$. Indeed, we have by construction that v_2 is normalized, and that

$$v_2^T v_1 = u_2^T v_1 - u_2^T v_1 v_1^T v_1 = 0.$$

By induction in j, that is, by supposing that $\mathbf{V}_{j-1}$ is orthonormal, we take the j-th column of V_k as

$$\hat{v}_j = u_j - \mathbf{V}_{j-1}\mathbf{V}_{j-1}^T u_j,$$

and we take

$$v_j = \frac{\hat{v}_j}{\|\hat{v}_j\|}.$$

Hence we have

$$v_j^T \mathbf{V}_{j-1} = u_j^T \mathbf{V}_{j-1} - u_j^T \mathbf{V}_{j-1}\mathbf{V}_{j-1}^T \mathbf{V}_{j-1} = 0.$$

In the implementation, to improve the stability of the algorithm, the orthogonalization step is done vector by vector, as follows. For $i = 1, \ldots, j-1$, we set $h_{i,j-1} = v_i^T u_j$; Then, we update the vector $\hat{v}_j = \hat{v}_j - h_{i,j-1} v_i$.

Algorithm 1: Modified Gram-Schmidt

Data: $\mathbf{U}_k \in \mathbb{R}^{n\times k}$
Result: $\mathbf{v}_k \in \mathbb{R}^{n\times k}$ with orthonormal columns, such that $Range(\mathbf{V}_k) = Range(\mathbf{U}_k)$
1 Set $v_1 = \frac{u_1}{\|u_1\|}$; $\mathbf{V}_0 = \emptyset$;
For $j = 2, \ldots, k$
2 Set $\mathbf{V}_{j-1} = [\mathbf{V}_{j-2}, v_{j-1}]$;
For $i = 1, \ldots, j-1$,
3 Set $h_{i,j-1} = v_i^T u_j$;
4 Update the vector $\hat{v}_j = \hat{v}_j - h_{i,j-1} v_i$
End For
5 Set $v_j = \frac{\hat{v}_j}{\|\hat{v}_j\|}$.
End For
6 Set $\mathbf{V}_k = [\mathbf{V}_{k-1}, v_k]$.

This algorithm can be also extended in block case.

In some results, we use the Frobenius norm of a matrix. It is defined as follows.

Definition 1.20 (Frobenius norm)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times m}$. We define the Frobenius norm of $\mathbf{A}$ as the quantity

$$\|\mathbf{A}\|_F = \sqrt{(\sum_{i,j} a_{ij}^2)}$$

In the thesis, we use the existence of a Riemann mapping for certain sets, which is stated in the following theorem.

Theorem 1.21 (Riemann Mapping)**.** *Let $E \neq \{p \in \mathbb{C}\}$ be a compact set, whose complement $K = \overline{\mathbb{C}} - E$ is simply connected in $\overline{\mathbb{C}}$. Let $D := \{w \in \overline{\mathbb{C}} : |w| \leq 1\}$ be the unit disk. There exists a unique conformal bijection, called Riemann map of E, $\phi : \overline{\mathbb{C}} - E \to \overline{\mathbb{C}} - D$, with $\phi(\infty) = \infty$ and $\phi'(\infty) > 0$.*

Thanks to Theorem 1.21, we are able to define Faber polynomials. We also introduce Chebychev polynomials, which coincide with Faber polynomials in the real case.

Definition 1.22 (Chebychev polynomials of the first kind)**.** We define the Chebychev polynomials of the first kind as $\mathcal{T}_k$ such that $\mathcal{T}_k(t) = \cos(k\cos^{-1}(t))$, for $|t| \leq 1$, and $\mathcal{T}_k(t) = \cosh(k\cosh^{-1}(t))$, for $|t| \geq 1$.

It can be proven that $\mathcal{T}_k$ are real polynomials of degree k.

Definition 1.23 (Faber polynomials)**.** Let $E \neq \{p\}$ be a compact set, whose complement $K = \overline{\mathbb{C}} - E$ is simply connected in $\overline{\mathbb{C}}$, and let ϕ be the Riemann map of E. The k-th Faber polynomial $\Phi_k(z)$ is defined as the polynomial part of the Laurent series of $\phi(z)^k$ at infinity.

These polynomials are important in error theory, for several reasons. In first place, we recall the following result, stated in [9], which provides a decomposition of an analytic function through Faber's polynomials.

Theorem 1.24. *Let K be the closure of a simply connected domain G, bounded by a rectifiable Jordan curve Γ. Given an analytic function f in G, it is uniquely represented by Faber polynomials, that is:*

$$f(z) = \sum_{j=0}^{\infty} a_j \Phi_j(z), \qquad a_j = \frac{1}{2\pi i} \int_{\Gamma} \frac{f(\xi)\phi'(\xi)}{\phi^{j+1}(\xi)} \, \mathrm{d}\xi$$

Moreover, Faber polynomials satisfy the following minimization property, as proved in [31].

Theorem 1.25. *Let $\{p\} \neq E \subset \overline{\mathbb{C}}$ be a compact set. The k-th degree Faber polynomial Φ_k is the unique monic polynomial of degree k which satisfies*

$$\min_{P_k \in \mathcal{P}_k} \|P_k\|_{\infty}$$

where $\mathcal{P}_k$ is the set of degree k monic polynomials in E.

As a consequence, Faber polynomials are an extension of Chebychev polynomials. Indeed, by recalling [22], the lasts satisfies the same property for $E := [-1, 1]$; and since the minimizator is unique, the two polynomials coincide in this case. Moreover, it can be shown that given a real interval $E = [a, b]$, the minimizator is the properly scaled and shifted Chebychev polynomial. We highlight that this property of the Faber polynomials allows us to extend the study of convergences of various numerical methods that involves polynomials, in less constrictive hypotheses on the matrix A.

We also use a generalization of the Faber polynomials, which is the following.

Definition 1.26. Let $\phi : \overline{\mathbb{C}} - E \to \overline{\mathbb{C}} - D$ be the Riemann map of E. We call Takenaka-Malmquist system of rational functions for the cyclic sequence of parameters σ_l the following system:

$$\Phi_l(w) = \frac{\sqrt{1 - |\overline{\phi(\sigma_l)})|^{-2}}}{1 - \overline{\overline{\phi(\sigma_l)}^{-1}} w} B_l(w), \quad l \in \mathbb{N}$$

where the function B_l is a finite length Blanschke product:

$$B_l(w) := \prod_{j=0}^{l-1} \left(\frac{\overline{\phi(\sigma_j)}^{-1} - w}{1 - \overline{\overline{\phi(\sigma_j)}^{-1}} w} \cdot \frac{|\overline{\phi(\sigma_j)}^{-1}|}{\overline{\phi(\sigma_j)}^{-1}} \right), \quad w \in \mathbb{C}.$$

In the following proposition we state some useful properties of the Takenaka-Malmquist system of rational functions, and of the function B_l.

Proposition 1.27. *The system of rational functions introduced in Definition 1.26 satisfy:*

i. $B_l(w)^{-1} = \overline{B_l(\overline{w}^{-1})}$

ii. $|B_l(w)| < 1, \quad |w| < 1,$

iii. $\int_{|w|=1} \Phi_n(w)\Phi_m(w)|dw| = 2\pi\delta_{n,m}, \quad n, m \in \mathbb{N},$

iv. $\max_{n\in\mathbb{N}} \max_{|w|=1} |\Phi_n(w)| \leq \infty.$

We also introduce the Faber-Dzhrbashyan rational function, as follows.

Definition 1.28. Let $r > 1$, and let ψ be the inverse of the Riemann map of E. Let $\Gamma_r := \{\psi(z) : |z| = r\}$, and let $G_r := \{\psi(z) : |z| < r\}$ be the interior of the curve Γ_r. We define the k-th Faber-Dzhrbashyan rational function as:

$$M_k(z) := \frac{1}{2\pi i} \int_{\Gamma_r} \frac{\Phi_k(\phi(\xi))}{\xi - z} \, \mathrm{d}\xi \quad z \notin \overline{G_r}$$

The next result makes this family of functions useful for us. A proof of the statement can be found in [18].

Proposition 1.29. *Let ϕ be the Riemann map of E, let Φ_j be the $j-$th Takenaka-Malmquist rational function, and let M_j be the j-th Faber-Dzhrbashyan rational function. The following holds.*

$$\frac{\psi'(w)}{\psi(w) - u} = \frac{1}{w} \sum_{j=0}^{\infty} \overline{\Phi_j\Big(\frac{1}{\overline{w}}\Big)} M_j(u), \quad |w| > 1$$

Here we introduce the concepts of plane condenser, and of the Riemann modulus of a condenser. We also state some fundamental properties of these objects. All the proofs can be found in [12].

Definition 1.30. Let F_1 and F_2 be disjoint closed subsets of $\overline{\mathbb{C}}$, each of which has a connected complement. We call (F_1, F_2) a plane condenser.

Definition 1.31. Let (F_1, F_2) be a plane condenser, and set $G := \overline{\mathbb{C}} - (F_1 \cup F_2)$. The unique generalized solution h of the Dirichlet problem $-\Delta h = 0$ in the region G with boundary data equal to 0 on ∂F_1 and 1 on ∂F_2 is called Harmonic measure of the set ∂F_2 relative to the region G.

The following Lemma establishes the regularity of the Harmonic measure of the set ∂F_2 relative to the region G, called H, which is needed in the definition of Riemann modulus.

Lemma 1.32. *Let (F_1, F_2) be a plane condenser, and set $G := \overline{\mathbb{C}} - (F_1 \cup F_2)$. Let $h(z)$ be the harmonic measure of the set ∂F_2 relative to the region G at the point $z \in G$. The function h is harmonic in the region G and at regular (relative to G) points of the sets ∂F_1 and ∂F_2, and it is bounded by 0 and 1.*

Definition 1.33. Let (F_1, F_2) be a condenser, and let γ be a curve that divides $\mathbb{C}$ into two open sets, one of which contains the set F_1, and the other of which contains the set F_2. We define the Riemann modulus of the condenser (F_1, F_2) as:

$$H(F_1, F_2) := c^{-1}(F_1, F_2), \qquad c(F_1, F_2) := \frac{1}{2\pi} \int_\gamma \frac{\partial h}{\partial n} \, \mathrm{d}s.$$

Here the quantity c is called capacity.

The modulus of a condenser (F_1, F_2) is strictly related to the so called third Zolotarev problem, which consists of finding a rational function $R_k(x) := \frac{p_k(x)}{q_k(x)}$ that is as small as possible on the set F_1 while being at least one in absolute value on another set F_2. More precisely, by defining the set $\mathcal{R}_{k,k} := \{r_k(x) = \frac{p_k(x)}{q_k(x)}, p_k, q_k \in \mathcal{P}_k\}$, the solution of the third Zolotarev problem attains the following infimum:

$$R_k(F_1, F_2) := \inf_{r_k \in \mathcal{R}_{k,k}} \frac{\sup_{z \in F_1} |r_k(z)|}{\inf_{z \in F_2} |r_k(z)|}$$

In the following theorem we recall the relation stated in [12] between the Riemann modulus and the solution of the third Zolotarev problem, for a fixed condenser (F_1, F_2).

Theorem 1.34. *Let (F_1, F_2) be a condenser, let $H(F_1, F_2)$ be the Riemann modulus of the condenser; let $R_k(F_1, F_2)$ be the solution of the third Zolotarev problem. It holds that*

$$H(F_1, F_2) = \lim_{k \to \infty} \frac{\log R_k(F_1, F_2)}{k}$$

Moreover, we use the principal and the complementary elliptic integrals of first kind of modulus g. They are defined as follows.

Definition 1.35. We call the principal elliptic integral of the first kind of modulus g the quantity

$$K(g) := \int_0^1 \frac{1}{\sqrt{1-t^2}\sqrt{1-g^2t^2}}\,\mathrm{d}t.$$

We call the complementary elliptic integral of the first kind of modulus g the quantity

$$K'(g) := \int_0^1 \frac{1}{\sqrt{1-t^2}\sqrt{1-(1-g^2)t^2}}\,\mathrm{d}t.$$

Chapter 2

Poisson problem in weak form

2.1 Poisson problem

Given a function $f \in C^0([0,1]^2)$, we want to find $u \in C^2([0,1]^2)$ such that

$$\begin{cases} -\Delta u = f & \text{in } (0,1)^2, \\ u = 0 & \text{in } \partial_D[0,1]^2, \\ \frac{\partial u}{\partial \nu} = 0 & \text{in } \partial_N[0,1]^2. \end{cases}$$

Where $\partial_D[0,1]^2$ is the part of the boundary related to the Dirichlet condition; and it's supposed to be a connected set, while $\partial_N[0,1]^2 = \partial[0,1]^2 \setminus \partial_D[0,1]^2$ is the part of the boundary related to the Neumann condition, and ν is the outward pointing unit normal vector to ∂_N. For the sake of simplicity, we will treat the case of $\partial_D[0,1]^2 = \{0\} \times [0,1) \cup [0,1) \times \{0\}$. A classical approach to tackle the problem is to relax it by taking an appropriate set of functions H, called test set, and by looking for a solution u which satisfies

$$\int_{[0,1]^2} -\Delta u(x,y)\Phi(x,y)\ \mathrm{d}x\ \mathrm{d}y = \int_{[0,1]^2} f(x,y)\Phi(x,y)\ \mathrm{d}x\ \mathrm{d}y, \quad \text{for all } \Phi \in H \tag{2.1}$$

The name test set comes from the fact that we test the function u to be solution with respect to every function in the set H. In our case, we take the set $H = H_0^1([0,1]^2)$. We recall from Definition 1.3 the definition of such a space.

$$H^1(\Omega) := \{v \in L^2(\Omega) : D^\alpha v \in L^2(\Omega), |\alpha| \leq 1\}$$

The space $H^1(\Omega)$ is equipped with the following norm.

$$\|v\|^2_{H^1(\Omega)} := \sum_{|\alpha|\leq 1} \|D^\alpha u\|^2_{L^2(\Omega)}, \qquad \|D^\alpha u\|^2_{L^2(\Omega)} = \int_\Omega |D^\alpha u|^2 dw$$

From this definition we derive:

$$H_0^1(\Omega) := \overline{C_0^\infty(\Omega)}^{\|\cdot\|_{H^1(\Omega)}}, \quad H^{-1}(\Omega) := (H_0^1(\Omega))'$$

We also emphasize a crucial property of the space H, which gives us an equivalent norm for the space H.

Lemma 2.1 (Poincarè Inequality)**.** *Given u in $H_0^1([0,1]^2)$, there exists a constant $C > 0$ which depends only on the domain, such that*

$$\|u\|^2_{L^2([0,1]^2)} \leq C\|\nabla u\|^2_{L^2([0,1]^2)}.$$

Proof. Since $u \in H_0^1([0,1]^2)$, we have that $u(x,0) = 0$. Thus we have:

$$\begin{aligned}
\|u\|^2_{L^2([0,1]^2)} &= \int_0^1 \int_0^1 |u(x,y) - u(x,0)|^2 \, \mathrm{d}x \, \mathrm{d}y \\
&= \int_0^1 \int_0^1 \Big| \int_0^y \frac{\partial u}{\partial y}(x,\eta) \, \mathrm{d}\eta \Big|^2 \, \mathrm{d}x \, \mathrm{d}y \\
&\leq \int_0^1 \int_0^1 \int_0^y \Big| \frac{\partial u}{\partial y}(x,\eta) \Big|^2 \, \mathrm{d}\eta \, \mathrm{d}x \, \mathrm{d}y \\
&\leq \int_0^1 \int_0^1 \int_0^1 |\nabla u(x,\eta)|^2_{\mathbb{R}^2} \, \mathrm{d}\eta \, \mathrm{d}x \, \mathrm{d}y \\
(Holder) &\leq \int_0^1 \int_0^1 \left(\sqrt{\int_0^1 1^2 \, \mathrm{d}\eta} \sqrt{\int_0^1 |\nabla u(x,\eta)|^2_{\mathbb{R}^2} \, \mathrm{d}\eta} \right)^2 \mathrm{d}x \, \mathrm{d}y \\
&= \int_0^1 \int_0^1 \int_0^1 1 \cdot |\nabla u(x,\eta)|^2_{\mathbb{R}^2} \, \mathrm{d}\eta \, \mathrm{d}x \, \mathrm{d}y \\
&= 1 \cdot \int_0^1 \int_0^1 |\nabla u(x,\eta)|^2_{\mathbb{R}^2} \, \mathrm{d}x \, \mathrm{d}\eta
\end{aligned}$$

□

This lemma can be generalized for a Lipschitz domain Ω. However, our proof strongly depends on the fact that the domain is a square. As a corollary of the Poincarè inequality, we can find an equivalent homogeneous norm for $H_0^1([0,1]^2)$. This fact is crucial in our discussion, since it gives us the coercivity, which is defined later, of the form a. Indeed we have the following result.

Corollary 2.2. *Given u in $H_0^1([0,1]^2)$, there exist two positive constants c_1 and C_2, which depend only on the domain, such that*

$$c_1\|\nabla u\|^2_{L^2([0,1]^2)} \leq \|u\|^2_{H_0^1([0,1]^2)} \leq C_2\|\nabla u\|^2_{L^2([0,1]^2)}.$$

Proof. By the Poincarè inequality we have:

$$\begin{aligned}\|u\|^2_{H^1_0([0,1]^2)} \geq \|\nabla u\|^2_{L^2([0,1]^2)} &= \frac{1}{2}\|\nabla u\|^2_{L^2([0,1]^2)} + \frac{1}{2}\|\nabla u\|^2_{L^2([0,1]^2)} \\ &\geq \frac{1}{2C}\|u\|^2_{L^2([0,1]^2)} + \frac{1}{2}\|\nabla u\|^2_{L^2([0,1]^2)} \\ &\geq \min\left\{\frac{1}{2C}, \frac{1}{2}\right\}\|u\|^2_{H^1_0([0,1]^2)}\end{aligned}$$

By taking $c_1 := 1$ and $C_2 := \frac{1}{\min\{\frac{1}{2C}, \frac{1}{2}\}}$, we obtain the result. □

A further relaxation of the problem regards the regularity of the data function f and of the solution u we are searching for. In particular, by taking $H = H^1_0([0,1]^2)$ as a test space, we notice that the integral problem (2.1) is well posed also for a data function $f \in H^{-1}([0,1]^2)$. In the next section, we will discuss problem (2.1), by searching for a solution $u \in H^1_0([0,1]^2)$; in the less restrictive hypotheses for the data function f to be in $H^{-1}([0,1]^2)$. For this setting, we will provide some classical results about existence and uniqueness of a weak solution of the problem.

2.2 Weak form of the Poisson problem

Given $f \in H^{-1}([0,1]^2)$, we want to find $u \in H^1_0([0,1]^2)$, such that, for all $\Phi \in H^1_0([0,1]^2)$, it holds that:

$$\int_{[0,1]^2} -\Delta u(x,y)\Phi(x,y)\ \mathrm{d}x\ \mathrm{d}y = \int_{[0,1]^2} f(x,y)\Phi(x,y)\ \mathrm{d}x\ \mathrm{d}y. \tag{2.2}$$

This is called weaker form, since it has weaker hypotheses in terms of regularity either of the data f and of the searched solution u.

In order to tackle (2.2), we have to use two important results. The first one is the Divergence Theorem, and it will provide us with an equivalent variational form of the problem (2.2), by assuming the existence of solution $u \in H^2_0([0,1]^2)$; while the second one is an extension of Lax-Milgram variational Lemma, due to Stampacchia. This is a fundamental tool for proving existence and uniqueness of the solution of (2.2).

By using the Divergence Theorem and by direct computation, we obtain:

$$\begin{aligned}&\int_{[0,1]^2} \nabla\Phi(x,y)\cdot\nabla u(x,y) + \Delta u(x,y)\Phi(x,y)\ \mathrm{d}x\ \mathrm{d}y \\ =&\int_{[0,1]^2} \mathrm{div}(\Phi(x,y)\nabla u(x,y))\ \mathrm{d}x\ \mathrm{d}y \\ =&\int_{\partial^+[0,1]^2} \Phi(x,y)\nabla u(x,y)\ \mathrm{d}H_1(x,y) = 0\end{aligned}$$

As a consequence, an equivalent form of (2.2) is the following:

$$\int_{[0,1]^2} -\nabla u(x,y)\cdot\nabla\Phi(x,y)\ \mathrm{d}x\ \mathrm{d}y = \int_{[0,1]^2} f(x,y)\Phi(x,y)\ \mathrm{d}x\ \mathrm{d}y \tag{2.3}$$

Since we want to obtain a variational problem, we have to define a linear and a bilinear form related to the problem. In particular, we define the following quantities.

Notation 2.3. Given u, v in $H_0^1([0,1]^2)$ and f in $H^{-1}([0,1]^2)$, we define:

$$a(u,v) = \langle\nabla u, \nabla v\rangle_{L^2([0,1]^2)}, \quad \ell(v) = \langle f, v\rangle_{L^2([0,1]^2)}.$$

By substituting these definitions, the problem (2.3) can be restated as: find $u \in H_0^1([0,1]^2)$, such that, for all $\Phi \in H_0^1([0,1]^2)$, it holds that:

$$a(u,\Phi) = \ell(\Phi) \tag{2.4}$$

In this variational context, continuity and coercivity are fundamental properties of the linear and bilinear forms, which can be defined as follows.

Definition 2.4. Given an Hilbert space H and a bilinear form $\mathfrak{a} : H \times H \to \mathbb{R}$, we say that:

i. $\mathfrak{a}$ is continuous if there exists $c > 0$ so that $|\mathfrak{a}(u,v)| \leq c\|u\|_H\|v\|_H$, for all $u, v \in H$

ii. $\mathfrak{a}$ is coercive if there exists $C > 0$ so that $|\mathfrak{a}(u,u)| \geq C\|u\|_H^2$, for all $u \in H$

Thanks to these properties, we will be able to prove the existence and the uniqueness of the weak solution of the problem (2.3).

Proposition 2.5. *The form $a : H_0^1([0,1]^2) \times H_0^1([0,1]^2) \to \mathbb{R}$ is bilinear, coercive and continuous, while the form $\ell : H_0^1([0,1]^2) \to \mathbb{R}$ is linear and continuous.*

Proof. The proofs of the bilinearity of a and the linearity of ℓ are direct computations, indeed:

$$\begin{aligned}
&\langle\nabla\alpha(u_1+u_2), \nabla v\rangle_{L^2([0,1]^2)} = \alpha\langle\nabla u_1, \nabla v\rangle_{L^2([0,1]^2)} + \alpha\langle\nabla u_2, \nabla v\rangle_{L^2([0,1]^2)},\\
&\langle\nabla u, \nabla\beta(v_1+v_2)\rangle_{L^2([0,1]^2)} = \beta\langle\nabla u, \nabla v_1\rangle_{L^2([0,1]^2)} + \beta\langle\nabla u, \nabla v_2\rangle_{L^2([0,1]^2)},\\
&\langle f, \gamma w_1 + w_2\rangle_{L^2([0,1]^2)} = \gamma\langle f, w_1\rangle_{L^2([0,1]^2)} + \langle f, w_2\rangle_{L^2([0,1]^2)}.
\end{aligned}$$

The coercivity of a follows directly from the right inequality of Corollary 2.2, while the continuity of a and ℓ follow from the left one. □

Corollary 2.6. *The form $a : H_0^1([0,1]^2) \times H_0^1([0,1]^2) \to \mathbb{R}$ induces a norm, called energy norm of a, defined as:*

$$\|u\|_a := \sqrt{a(u,u)}, \quad u \in H_0^1([0,1]^2). \tag{2.5}$$

We now recall a general result, introduced in [28], that will be crucial in several points of our discussion. It is an extension of the well known Lax-Milgram Lemma, and it provides us existence and uniqueness of the weak solution in a certain space.

Theorem 2.7 (Stampacchia)**.** *Let H be a Hilbert space, let $\mathfrak{a} : H \times H \to \mathbb{R}$ be a bilinear, symmetric, continuous and coercive form, and let $K \subseteq H$ be a convex and closed set. For all $\mathfrak{l} \in H'$ there exists an unique $u \in K$ such that:*

$$\mathfrak{a}(u, v) = \mathfrak{l}(v), \text{ for all } v \in K.$$

By using this result, as a corollary, we have a unique weak solution $u \in H_0^1([0,1]^2)$ of the problem (2.4).

Corollary 2.8. *Let f be in $H^{-1}([0,1]^2)$, and let the quantities*

$$a : H_0^1([0,1]^2) \times H_0^1([0,1]^2) \to \mathbb{R},\ a(u,v) = \langle \nabla u, \nabla v \rangle_{L^2([0,1]^2)},$$

$$\ell : H_0^1([0,1]^2) \to \mathbb{R},\ \ell(v) = \langle f, v \rangle_{L^2([0,1]^2)}.$$

There exists a unique $u \in H_0^1([0,1]^2)$ so that

$$a(u, v) = \ell(v), \text{ for all } v \in H_0^1([0,1]^2)$$

Proof. This is a direct consequence of Theorem 2.7 , by setting the quantities $K = H = H_0^1([0,1]^2)$, $\mathfrak{a} = a$, and $\mathfrak{l} = \ell$. □

We can notice that for this special case also Lax-Milgram Theorem would have been sufficient to prove this Corollary. In the future section we will see how to exploit the Theorem 2.7, by taking K as a proper subspace of H.

2.3 A new setting for the variational problem

Starting with this section, we want to use the following linear space on which to find the solution:

$$S := H_0^1([0,1]) \otimes H_0^1([0,1]), \qquad (\phi \otimes \psi)(x,y) = \phi(x)\psi(y), \quad \phi \otimes \psi \in S$$

We also equip S with the following norm:

$$\|\phi \otimes \psi\|_{L^2([0,1]^2)} = \|\phi\|_{L^2([0,1])} \|\psi\|_{L^2([0,1])}, \text{ for } \phi \otimes \psi \in S.$$

This setting has both analytical and computational advantages: indeed, this space is able to catch the important information of the separate direction, and see which one is more central, in dependence of the initial data. Besides, the problem (2.4) can be written in a simpler way by substituting the set S, as we are solving a

problem which is similar to the one-dimensional case. Hence we can write, for $u = u_1 \otimes u_2$, and $v = v_1 \otimes v_2$:

$$\begin{aligned}&\langle \nabla(u_1 \otimes u_2), \nabla(v_1 \otimes v_2)\rangle_{L^2([0,1]^2)}\\ =&\langle Du_1, Dv_1\rangle_{L^2([0,1])}\langle u_2, v_2\rangle_{L^2([0,1])} + \langle u_1, v_1\rangle_{L^2([0,1])}\langle Du_2, Dv_2\rangle_{L^2([0,1]^2)}\end{aligned}$$

For using this space S we have to verify that it is a closed, convex subset of $H_0^1([0,1]^2)$. We will do it in the following lemma.

Lemma 2.9. *S is a closed and convex subset of $H_0^1([0,1]^2)$.*

Proof. We first show that $S \subseteq H_0^1([0,1]^2)$. Let $\hat{\Phi}(x,y) = \phi(x)\psi(y) \in S$, and recall that

$$\|\phi \otimes \psi\|_{L^2([0,1]^2)} = \|\phi\|_{L^2([0,1])}\|\psi\|_{L^2([0,1])}.$$

By computations, we have that:

$$\begin{aligned}\|\phi \otimes \psi\|_{H_0^1([0,1]^2)} =&\Big(\sum_{\alpha\le 1}\|D^\alpha(\phi\otimes\psi)\|_{L^2([0,1]^2)}\Big)\\ =&\|\phi\otimes\psi\|_{L^2([0,1]^2)} + \|D(\phi)\otimes\psi\|_{L^2([0,1]^2)} + \|\phi\otimes D(\psi)\|_{L^2([0,1]^2)}\\ \le&\|\phi\otimes\psi\|_{L^2([0,1]^2)} + \|D(\phi)\otimes\psi\|_{L^2([0,1]^2)} + \|\phi\otimes D(\psi)\|_{L^2([0,1]^2)}+\\ &+\|D(\phi)\otimes D(\psi)\|_{L^2([0,1]^2)}\\ =&\Big(\sum_{\alpha\le 1}\|D^\alpha\phi\|_{L^2([0,1])}\Big)\Big(\sum_{\alpha\le 1}\|D^\alpha\psi\|_{L^2([0,1])}\Big)\\ =&\|\phi\|_{H_0^1([0,1])}\|\psi\|_{H_0^1([0,1])} < \infty.\end{aligned}$$

Hence, $S \subseteq H_0^1([0,1]^2)$. The convexity derives from the fact that S is a linear space, while the closure is a consequence of the structure of the space as a tensor product of two $H_0^1([0,1])$. □

As a consequence of Lemma 2.9 and of Theorem 2.7, we can state the following theorem.

Theorem 2.10. *Let $f \in H^{-1}([0,1]^2)$, and let*

$$\begin{aligned}&a : H_0^1([0,1]^2)\times H_0^1([0,1]^2) \to \mathbb{R},\ a(u,v) = \langle\nabla u, \nabla v\rangle_{L^2([0,1]^2)},\\ &\ell : H_0^1([0,1]^2) \to \mathbb{R},\ \ell(v) = \langle f, v\rangle_{L^2([0,1]^2)}.\end{aligned}$$

There exists a unique $u^s \in S$ such that

$$a(u^s, v) = \ell(v),\ \textit{for all } v \in S \tag{2.6}$$

To justify the variational approach (2.6), an interesting question is how "distant" the weak solution $u^s \in S$ is from the classical solution $u \in C^2([0,1]^2)$ of problem (2.1). A possible answer to this question is formalized next.

Proposition 2.11. *Let the data function $f \in C^0([0,1]^2)$, and let $u \in C^2([0,1]^2)$. It holds that $a(u, \hat{\Phi}) = l(\hat{\Phi})$ for all $\hat{\Phi} \in S$ if and only if $-\Delta u = f$.*

Proof. Let $\Phi(x, y) = \phi(x)\psi(y) \in S$. By Fubini's Theorem, we have

$$\begin{aligned} 0 &= \int_{[0,1]^2} (-\Delta u(x,y) - f(x,y))\Phi(x,y)\ \mathrm{d}x\ \mathrm{d}y \\ &= \int_{[0,1]} \underbrace{\Big(\int_{[0,1]} (-\Delta u(x,y) - f(x,y))\phi(x)\ dx \Big)}_{G(y)} \psi(y)\ \mathrm{d}y \end{aligned}$$

By applying the Fundamental Lemma in the Calculus of Variations to the integral over the variable y, we obtain

$$G(y) = 0.$$

Explicitly writing the function G, and by applying the Fundamental Lemma in the Calculus of Variations, we obtain

$$0 = \int_{[0,1]} (-\Delta u(x,y) - f(x,y))\phi(x)\ \mathrm{d}x \Leftrightarrow -\Delta u(x,y) - f(x,y) = 0$$

□

Indeed, if we suppose to have a regular data f, and a regular solution u, the problems (2.1) and (2.6) are equivalent.

Chapter 3

On the discrete problem obtained by Finite Elements

3.1 Introduction to finite elements

We consider a finite-dimensional subset of S, as approximation space for the solution u^s. For this purpose we define the following spaces.

Definition 3.1. Let $\{\phi_j(x)\}_{j=1,\dots,n} \subseteq H_0^1([0,1])$, and $\{\psi_k(y)\}_{k=1,\dots,n} \subseteq H_0^1([0,1])$ be two sets of linearly independent functions, and let $h = \frac{1}{n+1}$. We define the finite-dimensional spaces:

$$S_h^x = span\{\phi_1(x), \dots, \phi_n(x)\}, \qquad S_h^y = span\{\psi_1(y), \dots, \psi_n(y)\}$$

Moreover, we define the "tensorized" space

$$S_h = S_h^x \otimes S_h^y.$$

By applying these definitions, a function u_h belonging to S_h can be written as:

$$u_h(x,y) = \sum_{j=1}^{n}\sum_{k=1}^{n} \alpha_{j,k}\phi_j(x)\psi_k(y), \quad u_h \in S_h.$$

As described in Section 1.3, we aim to apply Stampacchia's Theorem by taking $K = S_h$. Besides, in the next lemma we will prove that S_h satisfies the hypotheses of Theorem 2.7, namely that S_h is a closed, convex subset of $H = H_0^1([0,1]^2)$. In particular, we show that S_h is a closed and convex subset of S, which is a subset of H.

Lemma 3.2. *S_h is a closed, convex subset of S.*

Proof. Let $\{\phi_j(x)\}_{j=1,\dots,n} \subset S_h^x$, and $\{\psi_k(y)\}_{k=1,\dots,n} \subset S_h^y$ be two bases, and let $\{\Phi_m(x,y)\}_m \subseteq S_h$ be a sequence of functions that has limit in S. We can write

in a unique each function $\Phi_m \in S_h$ in the sequence as a linear combination of the product of the two bases, and

$$\Phi_m(x,y) = \sum_{j=1}^{n}\sum_{k=1}^{n}(\alpha_{j,k})_m \phi_j(x)\psi_k(y) \xrightarrow[m\to\infty]{} \hat{\Phi}(x,y) \in S.$$

From the limit uniqueness, we have that

$$\hat{\Phi}(x,y) = \sum_{j=1}^{n}\sum_{k=1}^{n}\hat{\alpha}_{j,k}\phi_j(x)\psi_k(y), \text{ for } (\alpha_{j,k})_m \xrightarrow[m\to\infty]{} \hat{\alpha}_{j,k}.$$

Thus, the existence of the limit for the function $\hat{\Phi}$ is equivalent to the existence of the limit of the coefficients $\alpha_{j,k}$. Since $\mathbb{R}^{n\times n}$ is complete, $\hat{\Phi} \in S_h$. The convexity follows from the fact that S_h is a linear subset of S. □

As a consequence, we can state the well-posedness of the following finite-dimensional problem.

Theorem 3.3. *Let $f \in H^{-1}([0,1]^2)$, and let*

$$a : H_0^1([0,1]^2) \times H_0^1([0,1]^2) \to \mathbb{R},\ a(u,v) = \langle \nabla u, \nabla v\rangle_{L^2([0,1]^2)},$$

$$\ell : H_0^1([0,1]^2) \to \mathbb{R},\ \ell(v) = \langle f, v\rangle_{L^2([0,1]^2)}.$$

There exists a unique $u_h \in S_h$ so that

$$a(u_h, v_h) = \ell(v_h),\ \textit{for all } v_h \in S_h \tag{3.1}$$

3.2 Analytical properties of the Galerkin solution

In this section, we first give a different point of view of the finite-dimensional problem, and then we state an analytical result about the closeness of u^s and u_h.

Remark. By using the expression (2.4), and by subtracting it in (3.1), the finite-dimensional problem can be stated as: find $u_h \in S_h$ so that

$$a(u^s - u_h, v) = 0 \text{ for all } v_h \in S_h \tag{3.2}$$

Definition 3.4 (Galerkin solution). The function $u_h \in S_h$ that satisfies condition (3.2), called Galerkin condition will be called Galerkin solution of the problem.

Condition (3.2) gives us a new interpretation of the solution u_h: indeed, we want to find a function u_h such that the error $u^s - u_h$ is orthogonal, under the operator a, with respect to S_h, or in an equivalent way, we want the residual $r(v) := a(u_h, v) - l(v)$ to be zero when $v = v_h \in S_h$. In other words, the solution of the problem (3.1) u_h is the projection, under the operator a, of the solution u^s over S_h.

A direct consequence of (3.2) is stated in the following proposition.

Proposition 3.5 (Céa)**.** *The function $u_h \in S_h$ obtained by the imposing the Galerkin condition (3.2) satisfies*

$$a(u^s - u_h, u^s - u_h) = \min_{v_h \in S_h} a(u^s - v_h, u^s - v_h)$$

Proof. Let $v_h \in S_h$. Since a is bilinear, it holds that

$$\begin{aligned} a(u^s - v_h, u^s - v_h) &= a(u^s - u_h + u_h - v_h, u^s - u_h + u_h - v_h) \\ &= a(u^s - u_h, u^s - u_h) - 2a(u^s - u_h, v_h - u_h) + a(v_h - u_h, v_h - u_h) \end{aligned}$$

Since $v_h - u_h \in S_h$, we have that $a(u^s - u_h, v_h - u_h) = 0$, while by definition of a we have that $a(v_h - u_h, v_h - u_h) \geq 0$. Thus we have that, for all $v_h \in S_h$,

$$a(u^s - u_h, u^s - u_h) \leq a(u^s - v_h, u^s - v_h).$$

By taking the infimum we have the result. □

Thus, the Galerkin solution u_h is the function that minimizes the energy norm of the error $u^s - u_h$ in the space S_h.

We can state the following result on the convergence of the finite-dimensional solution.

Proposition 3.6. *Let a sequence of nested spaces $\{S_{h_j}\}_j$ satisfying $\overline{\bigcup_{j\geq 1} S_{h_j}} = S$ be given. Consider the respective sequence of functions $u_{h_j} \in S_{h_j}$ satisfying (3.2). We have that*

$$\lim_{j\to\infty} \|u^s - u_{h_j}\|_{H_0^1([0,1]^2)} = 0$$

Proof. Since $\bigcup_{j\geq 1} S_{h_j}$ is dense in S , we have that

$$\lim_{j\to\infty} \min_{v_{h_j} \in S_h} \left\|u^s - v_{h_j}\right\|^2_{H_0^1([0,1]^2)} = 0$$

From Corollary 2.2, the bilinear form a induces an equivalent norm of H_0^1, and thus we have that

$$\lim_{j\to\infty} \min_{v_{h_j} \in S_{h_j}} \left\|u^s - v_{h_j}\right\|^2_a = 0$$

From Proposition 3.5 have that u_{h_j} attains such minimum. Hence we conclude. □

In other words, since the residual has to be zero with respect to the subspace S_h, we want to take a sequence of nested spaces $\{S_{h_j}\}_j$ which saturates the entire space S. For the respective sequence of solutions $\{u_{h_j}\}_j$ to (3.2) we have, at the limit, a residual which has to be identically zero.

In the following result we state a more precise formulation of convergence of the method, which is a consequence of Friedrichs' Lemma.

Theorem 3.7. *Let $\{S_{h_j}\}_j$ be a sequence of nested spaces which satisfies $\overline{\bigcup_{j\geq 1} S_{h_j}} = S$ and let the respective sequence of functions $u_{h_j} \in S_{h_j}$ satisfying (3.2). Let $u^s = \lim_{j\to\infty} u_{h_j}$, and f be in $L^2([0,1]^2)$. By using the notations of Lemma 1.9, it holds that:*

$$u_\epsilon \xrightarrow[\epsilon\to 0]{L^2(K)} u^s, \qquad -\Delta u_\epsilon \xrightarrow[\epsilon\to 0]{L^2(K)} f$$

This theorem gives us a more precise result of convergence of the solution found by Galerkin, as we have the convergence of the graph (u_ϵ, f_ϵ) in the L^2 norm.

3.3 Numerical Problem

The purpose of this section is to find a more manageable form of the finite-dimensional problem, and after, to fix a space S_h . In particular, we will derive a matrix equation which is obtained from the problem (3.1), by taking two general bases of S_h^x and S_h^y, and by computing the bilinear form a and the linear form l for each element of the bases. Then, we fix two particular bases of S_h^x and S_h^y, and we explicitly write the matrix equation with respect to these bases. A deeper discussion about the importance of the choice of the bases of S_h^x and S_h^y can be found in [2, chapter 10].

Given $\{\phi_j(x)\}_{1\leq j\leq n} \subseteq S_h^x$, and $\{\psi_k(y)\}_{1\leq k\leq n} \subseteq S_h^y$ two bases, the problem (3.1) is equivalent of finding $u_h \in S_h$ such that

$$a(u_h, \phi_j \otimes \psi_k) = l(\phi_j \otimes \psi_k), \qquad 1 \leq j,k \leq n. \tag{3.3}$$

To explicitly write down the left-hand side, it is useful to define the so called stiffness and mass matrices, while for the right-hand side we define the data matrix. They will be denoted as $\mathbf{K}$, $\mathbf{M}$ and $\mathbf{F}$ respectively.

Definition 3.8. Given $\{\phi_j(x)\}_{1\leq j\leq n} \subseteq S_h^x$, and $\{\psi_k(y)\}_{1\leq k\leq n} \subseteq S_h^y$ two bases, we define:

$$(\mathbf{K}_x)_{i,j} = \langle \phi_i', \phi_j' \rangle_{L^2([0,1])}, \quad (\mathbf{K}_y)_{k,l} = \langle \psi_k', \psi_l' \rangle_{L^2([0,1])}, \quad \mathbf{K}_x, \mathbf{K}_y \in \mathbb{R}^{n\times n}$$
$$(\mathbf{M}_x)_{i,j} = \langle \phi_i, \phi_j \rangle_{L^2([0,1])}, \quad (\mathbf{M}_y)_{k,l} = \langle \psi_k, \psi_l \rangle_{L^2([0,1])}, \quad \mathbf{M}_x, \mathbf{M}_y \in \mathbb{R}^{n\times n}$$
$$(\mathbf{F})_{j,k} = \langle f, \phi_j \otimes \psi_k \rangle_{L^2([0,1]^2)}, \quad \mathbf{F} \in \mathbb{R}^{n\times n}$$

Notice that the matrices $\mathbf{M}$ and $\mathbf{K}$ are obtained from an inner product in one dimension, whereas $\mathbf{F}$ is obtained from an inner product in two dimensions. By these definitions, the following theorem can be stated.

Theorem 3.9. *The function $u_h \in S_h$, $u_h(x,y) = \sum_{j=1}^n \sum_{k=1}^n \alpha_{j,k}\phi_j(x)\psi_k(y)$ is the solution of problem (3.3) if and only if $\mathbf{X}$ is the solution of the following matrix equation.*

$$\mathbf{K}_x\mathbf{X}\mathbf{M}_y + \mathbf{M}_x\mathbf{X}\mathbf{K}_y = \mathbf{F} \tag{3.4}$$

where $\mathbf{K}_x, \mathbf{K}_y$, $\mathbf{M}_x, \mathbf{M}_y$ *and* $\mathbf{F}$ *are the stiffness, mass and data matrices, while* $(\mathbf{X})_{j,k} = \alpha_{j,k}$.

Proof. By using the explicit expression of ∇u_h, the solution of problem (3.3) is the function u_h that satisfies:

$$\sum_{i=1}^{n}\sum_{l=1}^{n} \alpha_{i,l}\langle \phi_i' \otimes \psi_l, \phi_j' \otimes \psi_k\rangle_{L^2([0,1]^2)} + \alpha_{i,l}\langle \phi_i \otimes \psi_l', \phi_j \otimes \psi_k'\rangle_{L^2([0,1]^2)} = \langle f, \phi_j \otimes \psi_k\rangle_{L^2([0,1]^2)}$$

By applying Fubini-Tonelli's Theorem, we obtain the following expressions:

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{l=1}^{n} \alpha_{i,l}\langle \phi_i' \otimes \psi_l, \phi_j' \otimes \psi_k\rangle_{L^2([0,1]^2)} \\ =&\sum_{i=1}^{n}\sum_{l=1}^{n} \alpha_{i,l}\langle \phi_i', \phi_j'\rangle_{L^2([0,1])}\langle \psi_k, \psi_l\rangle_{L^2([0,1])} = (\mathbf{K}_x\mathbf{X}\mathbf{M}_y)_{j,k},\end{aligned}$$

and

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{l=1}^{n} \alpha_{i,l}\langle \phi_i \otimes \psi_l', \phi_j \otimes \psi_k'\rangle_{L^2([0,1]^2)} \\ =&\sum_{i=1}^{n}\sum_{l=1}^{n} \alpha_{i,l}\langle \phi_i, \phi_j\rangle_{L^2([0,1])}\langle \psi_k', \psi_l'\rangle_{L^2([0,1])} = (\mathbf{M}_x\mathbf{X}\mathbf{K}_y)_{j,k}\end{aligned}$$

By summing the two expressions we obtain:

$$\mathbf{K}_x\mathbf{X}\mathbf{M}_y + \mathbf{M}_x\mathbf{X}\mathbf{K}_y = \mathbf{F}$$

□

3.4 Choice of the basis of S_h

In this section, we describe and justify our choice for the tensorized basis of the finite-dimensional set.

Definition 3.10 (Linear finite elements)**.** Set $h = \frac{1}{n+1}$, $x_i = ih$, for $i = 0, \dots, n$, $x_{n+1} = x_n$, and set $y_i = ih$, for $i = 0, \dots, n$, $y_{n+1} = y_n$. We define the linear spaces:

$$S_h^x = span\{\phi_1(x), \dots, \phi_n(x)\}, \quad \phi_j(x) = \begin{cases} \frac{x - x_{j-1}}{h}, & x \in [x_{j-1}, x_j) \\ \frac{x_{j+1} - x}{h}, & x \in [x_j, x_{j+1}] \end{cases}$$

and

$$S_h^y = span\{\psi_1(y), \dots, \psi_n(y)\}, \quad \psi_k(y) = \begin{cases} \frac{y-y_{k-1}}{h}, & y \in [y_{k-1}, y_k) \\ \frac{y_{k+1}-y}{h}, & y \in [y_k, y_{k+1}] \end{cases}$$

There are various advantages of choosing these bases: in first place, fixed ϕ_j, we have that only ϕ_{j-1} and ϕ_{j+1} are not orthogonal, under the operator a, to ϕ_j, and thus the matrices $\mathbf{M}$ and $\mathbf{K}$ are sparse, as they are tridiagonal. Moreover, since the functions ϕ_j are piecewise linear, their first derivatives are easy to compute, and so for the bilinear form $a(\phi_j, \phi_k)$, as suggested in [4]. In the following proposition, which can be found in [4], we fully describe the structure of the two matrices.

Proposition 3.11. *Let the functions ϕ_j and ψ_k be as in Definition 3.10. Then*

$$\langle \phi_j', \phi_{j+1}' \rangle_{L^2([0,1])} = \langle \psi_j', \psi_{j+1}' \rangle_{L^2([0,1])} = -\tfrac{1}{h}, \quad j = 1, \dots, n-1$$

$$\langle \phi_j', \phi_j' \rangle_{L^2([0,1])} = \langle \psi_j', \psi_j' \rangle_{L^2([0,1])} = \tfrac{2}{h}, \quad j = 1, \dots, n-1,$$

$$\langle \phi_n', \phi_n' \rangle_{L^2([0,1])} = \langle \psi_n', \psi_n' \rangle_{L^2([0,1])} = \tfrac{1}{h},$$

$$\langle \phi_j, \phi_{j+1} \rangle_{L^2([0,1])} = \langle \psi_j, \psi_{j+1} \rangle_{L^2([0,1])} = \tfrac{h}{6}, \quad j = 1, \dots, n-1,$$

$$\langle \phi_j, \phi_j \rangle_{L^2([0,1])} = \langle \psi_j, \psi_j \rangle_{L^2([0,1])} = \tfrac{2h}{3}, \quad j = 1, \dots, n-1,$$

$$\langle \phi_n, \phi_n \rangle_{L^2([0,1])} = \langle \psi_n, \psi_n \rangle_{L^2([0,1])} = \tfrac{h}{3}.$$

In an equivalent way, we have that the matrices $\mathbf{K} := \mathbf{K}_x = \mathbf{K}_y$ and $\mathbf{M} := \mathbf{M}_x = \mathbf{M}_y$ have the following forms:

$$\mathbf{K} = \frac{1}{h} \begin{bmatrix} 2 & -1 & & & \\ -1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 2 & -1 \\ & & & -1 & 1 \end{bmatrix}, \qquad \mathbf{M} = \frac{h}{6} \begin{bmatrix} 4 & 1 & & & \\ 1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 4 & 1 \\ & & & 1 & 2 \end{bmatrix}$$

Proof. The proof of this result is done by the explicit computation of the integrals. □

In the following, we derive some properties of the matrices $\mathbf{K}$ and $\mathbf{M}$, that highlights the dependence of their condition number on the discretization.

Proposition 3.12. *Let the matrices $\mathbf{M}$ and $\mathbf{K}$ defined as in Proposition 3.11. It holds that $\mathbf{M}$ and $\mathbf{K}$ are positive definite, and $\kappa(\mathbf{M}) = \mathcal{O}(1)$, while $\kappa(\mathbf{K}) = \mathcal{O}(\frac{1}{h^2})$.*

Proof. We start by showing that $\kappa(\mathbf{M}) = \mathcal{O}(1)$. By definition of the condition number, we have that $\kappa(\mathbf{M}) = \|\mathbf{M}\| \, \|\mathbf{M}^{-1}\| \geq \|\mathbf{I}\| \geq 1$. By using the First Gerschgoring Theorem, we have that $\lambda_{max}(\mathbf{M}) \leq h$, and that $\lambda_{min}(\mathbf{M}) \geq \frac{h}{6}$. Thus

we have that $\kappa(\mathbf{M}) \leq 6$. To show that $\kappa(\mathbf{K}) = \mathcal{O}(\frac{1}{h^2})$, we recall from [32] the following result.

$$\text{spec}(\mathbf{K}) = \frac{1}{h}\left\{2 - 2\cos\frac{2(k-\frac{1}{2})\pi}{2n+1}, \quad k = 1, \ldots, n\right\}.$$

Hence we have that

$$\frac{2}{h} \leq \lambda_{max}(\mathbf{K}) \leq \frac{4}{h}.$$

By using the Taylor expansion for the cosine function, with Lagrange remainder, we also have that

$$\cos\frac{\pi}{n+1} = 1 - \frac{\pi^2}{2(n+1)^2} + \cos\eta\frac{\pi^4}{24(n+1)^4} \quad \eta \in \left(0, \frac{\pi}{n+1}\right),$$

And thus we have that

$$\frac{1}{h}\left(\frac{\pi^2}{(n+1)^2} - \frac{\pi^4}{12(n+1)^4}\right) \leq \lambda_{min}(\mathbf{K}) \leq \frac{1}{h}\left(\frac{\pi^2}{(n+1)^2} + \frac{\pi^4}{12(n+1)^4}\right).$$

Hence, we have that

$$(n+1)^2\left(\frac{96}{48\pi^2 + \pi^4}\right) \leq \kappa(\mathbf{K}) \leq (n+1)^2\left(\frac{192}{48\pi^2 - \pi^4}\right)$$

Since $h = \frac{1}{n}$, we conclude. □

Proposition 3.12 shows that as the grid is refined, the condition number of $\mathbf{K}$ is proportional to $\frac{1}{h^2}$, and so goes to infinity when h tends to zero. A full discussion of this topic can be found in [11], and [27].

Thanks to the positive definiteness of $\mathbf{M}$ and $\mathbf{K}$ stated in Proposition 3.12, the following operator induces a norm, which is the finite-dimensional counterpart of the energy norm of a, as defined in Corollary 2.5.

Corollary 3.13. *The operator $\mathcal{A} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$, $\mathcal{A}\mathbf{X} = \mathbf{KXM} + \mathbf{MXK}$ induces a norm defined as:*

$$\|\mathbf{X}\|_{\mathcal{A}} := \sqrt{trace(\mathbf{X}^T\mathbf{KXM} + \mathbf{X}^T\mathbf{MXK})}, \quad \mathbf{X} \in \mathbb{R}^{n\times n}.$$

To reduce the generalized Lyapunov equation (3.4) into a standard Lyapunov equation, we use a standard technique, recalled in the following lemma.

Lemma 3.14. *Let the matrices $\mathbf{M}$ and $\mathbf{K}$ be as defined in Proposition 3.11, and let $\mathbf{L}$ such that $\mathbf{M} = \mathbf{L}\mathbf{L}^T$. Set $\tilde{\mathbf{K}} = \mathbf{L}^{-1}\mathbf{K}\mathbf{L}^{-T}$, $\tilde{\mathbf{X}} = \mathbf{L}^T\mathbf{X}\mathbf{L}$, and $\tilde{\mathbf{F}} = \mathbf{L}^{-1}\mathbf{F}\mathbf{L}^{-T}$. The matrix equation (3.4) is equivalent to the equation:*

$$\tilde{\mathbf{K}}\tilde{\mathbf{X}} + \tilde{\mathbf{X}}\tilde{\mathbf{K}} = \tilde{\mathbf{F}} \tag{3.5}$$

Proof. The proof is done by properly dividing by $\mathbf{L}$ and $\mathbf{L}^T$, as follows.

$$\begin{aligned}\tilde{\mathbf{F}} &= \mathbf{L}^{-1}(\mathbf{KXM}+\mathbf{MXK})\mathbf{L}^{-T}\\ &= \mathbf{L}^{-1}(\mathbf{KL}^{-T}\mathbf{L}^T\mathbf{XLL}^{-1}\mathbf{M}+\mathbf{ML}^{-T}\mathbf{L}^T\mathbf{XLL}^{-1}\mathbf{K})\mathbf{L}^{-T}\\ &= \mathbf{L}^{-1}(\mathbf{KL}^{-T}\mathbf{L}^T\mathbf{XLL}^T+\mathbf{LL}^T\mathbf{XLL}^{-1}\mathbf{K})\mathbf{L}^{-T}\\ &= \tilde{\mathbf{K}}\tilde{\mathbf{X}}+\tilde{\mathbf{X}}\tilde{\mathbf{K}}\end{aligned}$$

□

In the following, the matrix $\mathbf{L}$ will be taken as the Cholesky factor of $\mathbf{M}$.

Chapter 4

Properties of the Lyapunov Equation

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix, and let $\mathbf{C}_1, \mathbf{C}_2 \in \mathbb{R}^{n\times p}$ be tall matrices. The aim of this chapter is to analyze some properties related to the solution $\mathbf{X}$ of the Lyapunov equation

$$\mathbf{AX} + \mathbf{XA} = \mathbf{C}_1\mathbf{C}_2^T \tag{4.1}$$

We recall that, in our setting, the existence and the uniqueness of such a solution $\mathbf{X}$ are stated in Theorem 1.14.

Some of these properties give us a different perspective of the solution of a Lyapunov equation, and we use them to prove theorems, while others are structural, and they help us for computations, for choosing proper methods to solve equation (3.5). In first place, we will see different ways to express the solution of equation (4.1), and as a corollary we see the connection between the solution of a Lyapunov equation and the energy of a Cauchy problem.

Some of the following properties are stated in a general setting, as they only requires the existence and the uniqueness of a solution $\mathbf{X}$ of (4.1), and they do not exploit the structure of the matrix $\mathbf{A}$.

4.1 Integral forms of the solution of the Lyapunov Equation

The following results give us a characterization of the solution of a Lyapunov equation, by using an integral form.

Proposition 4.1. *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$, let $\mathbf{C}_1, \mathbf{C}_2 \in \mathbb{R}^{n\times p}$ be tall matrices. By defining the functions $x_1(t) := e^{-t\mathbf{A}}\mathbf{C}_1$, and $x_2(t) := e^{-t\mathbf{A}}\mathbf{C}_2$, it holds that:*

$$\mathbf{X} = \int_0^\infty x_1(t)x_2^T(t)dt.$$

Proof. To prove the result, it is sufficient to show that such $\mathbf{X}$ is a solution of the equation, and then we conclude by uniqueness of the solution. By substitution, we have that

$$\begin{aligned}\mathbf{AX}+\mathbf{XA} &= \int_0^\infty \mathbf{A}e^{-t\mathbf{A}}\mathbf{C}_1\mathbf{C}_2^T e^{-t\mathbf{A}} + e^{-t\mathbf{A}}\mathbf{C}_1\mathbf{C}_2^T e^{-t\mathbf{A}}\mathbf{A}dt \\ &= \int_0^\infty -\Big(\frac{d}{dt}(e^{-t\mathbf{A}}\mathbf{C}_1)\Big)\mathbf{C}_2^T e^{-t\mathbf{A}} - e^{-t\mathbf{A}}\mathbf{C}_1\Big(\frac{d}{dt}(\mathbf{C}_2^T e^{-t\mathbf{A}})\Big)dt \\ &= \int_0^\infty -\frac{d}{dt}(e^{-t\mathbf{A}}\mathbf{C}_1\mathbf{C}_2^T e^{-t\mathbf{A}})dt \\ &= -e^{-t\mathbf{A}}\mathbf{C}_1\mathbf{C}_2^T e^{-t\mathbf{A}}|_{t=0}^{t=\infty} = \mathbf{C}_1\mathbf{C}_2^T\end{aligned}$$

□

The result holds also for nonsymmetric $\mathbf{A}$, for the equation $\mathbf{AX}+\mathbf{XA}^T = \mathbf{C}_1\mathbf{C}_2^T$. For the second characterization, we have to use the following lemma.

Lemma 4.2. *Set $\mathbf{A} \in \mathbb{R}^{n\times n}$, and let $z \in \overline{\mathbb{C}}$ such that $Re(z) < \mu$, for all $\mu \in [\lambda_{min}(\mathbf{A}), \lambda_{max}(\mathbf{A})]$. It holds that:*

$$(-z\mathbf{I}+\mathbf{A})^{-1} = \int_0^\infty e^{-(-z\mathbf{I}+\mathbf{A})t}dt$$

Proposition 4.3. *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$, let $\mathbf{C}_1, \mathbf{C}_2 \in \mathbb{R}^{n\times p}$ be tall matrices. Let Γ be a closed contour in $\overline{\mathbb{C}} - ([\lambda_{min}(\mathbf{A}), \lambda_{max}(\mathbf{A})] \cup [-\lambda_{max}(\mathbf{A}), -\lambda_{min}(\mathbf{A})])$ which is homeotopic to the imaginary axis $i\mathbb{R}$ being passed from $-i\infty$ to $-i\infty$. It holds that:*

$$X = \frac{1}{2\pi i}\int_\Gamma (z\mathbf{I}-\mathbf{A})^{-1}\mathbf{C}_1\mathbf{C}_2^T(z\mathbf{I}-\mathbf{A})^{-*}dz \tag{4.2}$$

Proof. We notice that, since Γ is homeotopic to the imaginary axis $i\mathbb{R}$ being passed from $-i\infty$ to $-i\infty$, and since the matrix $\mathbf{A}$ is symmetric, equation (4.2) is equivalent to the following equation.

$$\mathbf{X} = \frac{1}{2\pi i}\int_\Gamma (z\mathbf{I}-\mathbf{A})^{-1}\mathbf{C}_1\mathbf{C}_2^T(-z\mathbf{I}-\mathbf{A})^{-1}dz \tag{4.3}$$

We thus prove equation (4.3). It is sufficient to show that $\mathbf{X}$ is equivalent to the solution obtained in Proposition 4.1, that is $\mathbf{X} = \int_0^\infty e^{-t\mathbf{A}}\mathbf{C}_1\mathbf{C}_2^T e^{-t\mathbf{A}}dt$. To prove the result, we treat the two exponentials in different ways. For the one to the left, we use Lemma 4.2, while for the other we use Cauchy's Integral Formula.
Indeed, we have that:

$$e^{-t\mathbf{A}} = \frac{1}{2\pi i}\int_\Gamma e^{zt}(z\mathbf{I}+\mathbf{A})^{-1}dz$$

Hence,

$$\begin{aligned}
\mathbf{X} &= \int_0^\infty e^{-t\mathbf{A}}\mathbf{C}_1\mathbf{C}_2^T e^{-t\mathbf{A}}dt \\
&= \frac{1}{2\pi i}\int_0^\infty e^{-t\mathbf{A}}\mathbf{C}_1\mathbf{C}_2^T\Big[\int_\Gamma e^{zt}(z\mathbf{I}+\mathbf{A})^{-1}dz\Big]dt \\
(\mathbf{I} \text{ commutes}) &= \frac{1}{2\pi i}\int_\Gamma \Big[\int_0^\infty e^{-(-z\mathbf{I}+\mathbf{A})t}dt\Big]\mathbf{C}_1\mathbf{C}_2^T(z\mathbf{I}+\mathbf{A})^{-1}dz \\
&= \frac{1}{2\pi i}\int_\Gamma (-z\mathbf{I}+\mathbf{A})^{-1}\mathbf{C}_1\mathbf{C}_2^T(z\mathbf{I}+\mathbf{A})^{-1}dz
\end{aligned}$$

□

As discussed in [19], for a sufficiently large R, we can take the curve $\Gamma := \gamma_1 \sqcup \gamma_2$, where γ_1 is the line in the complex plane $r(s) = is$, $s \in [-R, R]$, and γ_2 is the semicircle $c(t) = Re^{it}$, $t \in [-\frac{\pi}{2}, \frac{\pi}{2}]$.

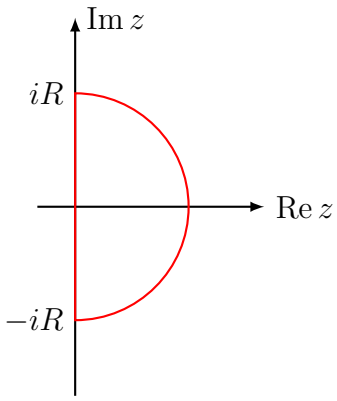

Figure 4.1: Curve $\Gamma := \gamma_1 \sqcup \gamma_2$

We have

$$\mathbf{X} = \lim_{R\to\infty}\int_{\gamma_1\sqcup\gamma_2}(z\mathbf{I}-\mathbf{A})^{-1}\mathbf{C}_1\mathbf{C}_2^T(z\mathbf{I}-\mathbf{A})^{-*}dz$$

We split the integral in the two curves γ_1 and γ_2. Then, by direct computations, it holds that

$$\begin{aligned}
&\lim_{R\to\infty}\int_{\gamma_2}(z\mathbf{I}-\mathbf{A})^{-1}\mathbf{C}_1\mathbf{C}_2^T(z\mathbf{I}-\mathbf{A})^{-*}dz \\
&= \lim_{R\to\infty}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}(Re^{it}\mathbf{I}-\mathbf{A})^{-1}\mathbf{C}_1\mathbf{C}_2^T(Re^{it}\mathbf{I}-\mathbf{A})^{-*}iRe^{it}dt = 0
\end{aligned}$$

Hence we also obtain

$$\mathbf{X} = \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} (z\mathbf{I} - \mathbf{A})^{-1} \mathbf{C}_1 \mathbf{C}_2^T (z\mathbf{I} - \mathbf{A})^{-*} dz. \tag{4.4}$$

As a consequence of these results, when $\mathbf{C} := \mathbf{C}_1 = \mathbf{C}_2$, the solution of (4.1) is strictly related to the total output energy of the following Cauchy problem, as shown in [15].

Theorem 4.4. *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix, let $\mathbf{C} \in \mathbb{R}^{n\times p}$ be a tall matrix. Let the Cauchy problem:*

$$\frac{dx}{dt}(t) = -\mathbf{A}x(t), \qquad x(0) = \mathbf{C}$$

The solution X of the Lyapunov equation $\mathbf{AX} + \mathbf{XA} = \mathbf{CC}^T$ satisfies

$$E^2 = trace(\mathbf{X})$$

Where $E = \left(\int_0^\infty \|x(t)\|^2 \, dt \right)^{\frac{1}{2}}$ represents the total output energy of the Cauchy system.

Proof. First, we notice that the solution of the Cauchy problem is $x(t) = e^{-t\mathbf{A}}\mathbf{C}$, which is defined in $t \in [0, \infty)$, since $\mathbf{A}$ is symmetric positive definite. To prove the result we use Proposition 4.1. Indeed,

$$\begin{aligned} E^2 &= \int_0^\infty \|x(t)\|^2 \, dt \\ &= \int_0^\infty trace(x(t)^T x(t)) dt \\ &= \int_0^\infty trace(x(t)x(t)^T) dt \\ &= trace \int_0^\infty (x(t)x(t)^T) dt = trace(\mathbf{X}) \end{aligned}$$

□

4.2 Low-rank property of the solution of the Lyapunov equation

The last class of properties we investigate regards the eigenvalues and rank decay of the solution of the Lyapunov equation (4.1), under the assumption of $\mathbf{C} := \mathbf{C}_1 = \mathbf{C}_2$. In particular, we see two results. The first one is stated in [23].

Theorem 4.5. *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix, let $\mathbf{C} \in \mathbb{R}^{n\times p}$ be a tall matrix, and let $\mathbf{X}$ be the solution of the Lyapunov equation*

$$\mathbf{AX} + \mathbf{XA} = \mathbf{CC}^T.$$

There exists a matrix $\mathbf{X}_k$ *of rank at most* kp *such that*

$$\|\mathbf{X} - \mathbf{X}_k\|_F \leq \frac{8\,\|\mathbf{C}\|_F}{\lambda_{min}(\mathbf{A})} \exp\left(\frac{-k\pi^2}{\log\left(8\kappa(\mathbf{A})\right)}\right).$$

Where $\kappa(\mathbf{A})$ *is the condition number of the matrix* $\mathbf{A}$*.*

Under the hypothesis of a well-conditioned matrix $\mathbf{A}$, the bound indicates an exponential decay of the distance between $\mathbf{A}$ and a rank kp matrix. Otherwise, when the matrix $\mathbf{A}$ is ill-conditioned, the bound deteriorates. The second result, introduced in [16], fixes this problem.

Theorem 4.6. *Let* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *be a symmetric positive definite matrix, and take* $\delta > 0$ *such that spec*$(\mathbf{A}) \subseteq (\delta, \infty]$*, let* $\mathbf{C} \in \mathbb{R}^{n\times p}$ *be a tall matrix, and let* $\mathbf{X}$ *be the solution of the Lyapunov equation:*

$$\mathbf{AX} + \mathbf{XA} = \mathbf{CC}^T$$

For every $k \in \mathbb{N}$ *there exists a matrix* $\mathbf{X}_k$ *of rank at most* $(2k+1)p$ *such that*

$$\|\mathbf{X} - \mathbf{X}_k\|_F \leq \frac{C_{St}\,\|\mathbf{C}\|_F^2}{2\delta} \exp\big(-\sqrt{k}\pi\big).$$

Where C_{St} *is independent of* k*.*

Notice that, as a counterpart, the last bound is less efficient in case of well-conditioned problems. In the following figures we show the comparison of the bounds for different values of $\kappa(\mathbf{A})$, by ignoring the constants. Moreover, as stated in [16], in our setting the quantity C_{St} can be numerically estimated as $C_{St} \approx 2.75$.

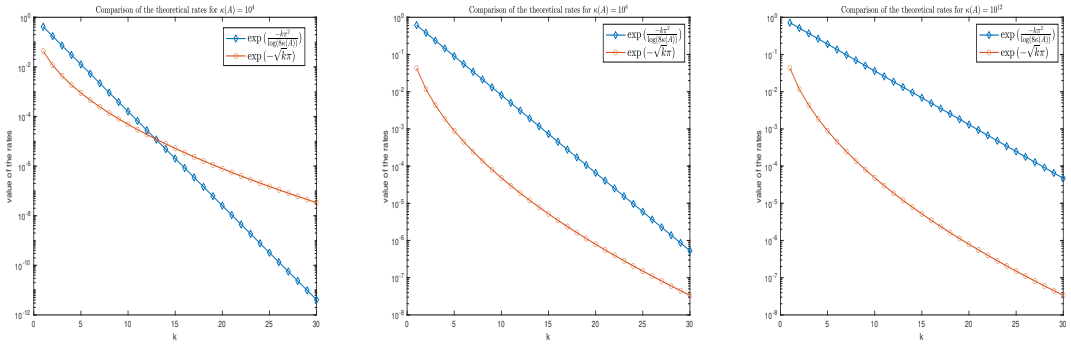

Figure 4.2: Comparison between the theoretical bounds of Theorem 4.5 and Theorem 4.6, for $k = 1, \ldots, 30$, and for $\kappa(\mathbf{A}) = 10^4, 10^8, 10^{12}$.

As a consequence of these properties, we are allowed to search for a low rank approximation of the solution $\mathbf{X}$.

Chapter 5

Numerical solution of the Lyapunov equation

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix, and let $\mathbf{C} \in \mathbb{R}^{n\times n}$. We are interested in analyzing numerical procedures for solving the matrix equation

$$\mathbf{AX} + \mathbf{XA} = \mathbf{C} \tag{5.1}$$

This equation corresponds to (3.5) for $\mathbf{A} = \tilde{\mathbf{K}}$, $\mathbf{C} = \tilde{\mathbf{F}}$ and $\mathbf{X} = \tilde{\mathbf{X}}$. We split the analysis in two different cases, which are the small scale problem and the large scale problem.

5.1 The small scale problem

In the small case setting, a robust and efficient method for solving equation (5.1) is the Bartels–Stewart algorithm, introduced in [3], which computes the solution $\mathbf{X}$ by the following procedure. Firstly, we compute the spectral decomposition of the matrices $\mathbf{A}$, that is

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$$

where $\mathbf{\Lambda}$ is diagonal and $\mathbf{U}$ is orthonormal, that is $\mathbf{U}^T = \mathbf{U}^{-1}$. Then, we set $\mathbf{G} := \mathbf{U}^T\mathbf{C}\mathbf{U}$ and $\mathbf{Y} = \mathbf{U}^T\mathbf{X}\mathbf{U}$, and we multiply equation (5.1) by $\mathbf{U}$ from the right and by $\mathbf{U}^T$ from the left. Hence, we obtain

$$\begin{aligned} \mathbf{G} &= \mathbf{U}^T(\mathbf{AX} + \mathbf{XA})\mathbf{U} && (5.2)\\ &= \mathbf{\Lambda}(\mathbf{U}^T\mathbf{X}\mathbf{U}) + (\mathbf{U}^T\mathbf{X}\mathbf{U})\mathbf{\Lambda} && (5.3)\\ &= \mathbf{\Lambda}\mathbf{Y} + \mathbf{Y}\mathbf{\Lambda} && (5.4) \end{aligned}$$

We are left to solve equation (5.4). This is done by exploting the diagonal

structure of $\mathbf{\Lambda}$. Indeed, we have

$$y_{i,j} = \frac{g_{i,j}}{\lambda_i + \lambda_j} \tag{5.5}$$

In the end, we recover the solution $\mathbf{X} = \mathbf{U}\mathbf{Y}\mathbf{U}^T$. The full procedure is summarized in Algorithm 2.

For the sake of completeness, we recall that in case the matrix $\mathbf{A}$ is not symmetric, instead of working with the spectral decomposition of $\mathbf{A}$, we work with the Schur decomposition of $\mathbf{A}$, which is $\mathbf{A} = \mathbf{U}\mathbf{R}\mathbf{U}^*$, where $\mathbf{R}$ is upper triangular and $\mathbf{U}$ is unitary, that is $\mathbf{U}^* = \mathbf{U}^{-1}$.

Algorithm 2: Bartels-Stewart Algorithm for a symmetric positive definite $\mathbf{A}$

Data: $\mathbf{A}, \mathbf{C} \in \mathbb{R}^{n\times n}$
Result: $\mathbf{X} \in \mathbb{R}^{n\times n}$ solution of $\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A} = \mathbf{C}$
1 Compute the spectral decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$.
2 Solve $y_{i,j} = \frac{g_{i,j}}{\lambda_i+\lambda_j}$.
3 Compute $\mathbf{X} = \mathbf{U}\mathbf{Y}\mathbf{U}^T$.

5.2 Kronecker formulation for the large scale problem

For the large scale setting, the Schur decomposition may require a prohibitive amount of computer memory. As a consequence, we have to look for other numerical methods to solve equation (5.1).

The naive approach to numerically solve equation (5.1) in the large scale setting is by means of the Kronecker product and the vec operator, as defined in Definition 1.10, and in Definition 1.12. This procedure yields the following linear system

$$(\mathbf{I} \otimes \mathbf{A} + \mathbf{A}^T \otimes \mathbf{I})\mathrm{vec}(\mathbf{X}) = \mathrm{vec}(\mathbf{C}). \tag{5.6}$$

However, this approach has several disadvantages. In the first place, the dimensions of the linear system are the square of the dimension of the problem; as a consequence, also if the dimensions of the matrices involved in equation (5.1) are moderate, the dimensions of the linear system (5.6) can be huge. Moreover, by using this transformation, if we do not determine the solution $\mathbf{X}$ with high accuracy, the spectral and low rank properties of the solution will be lost. Instead, it may be advantageous to solve the original matrix equation, instead of the equivalent transformed linear system. We recall that a common way to solve the matrix equation is by using the so called Alternating-Direction Implicit (ADI) method, introduced in [29], although we do not treat it in this work.

5.3 Projection methods for the large scale problem

Under the further assumption for the matrix $\mathbf{C}$ to be low rank, that is, there exists $p \ll n$, and there exist two matrices $\mathbf{C}_1, \mathbf{C}_2 \in \mathbb{R}^{n\times p}$ such that $\mathbf{C} = \mathbf{C}_1\mathbf{C}_2^T$, a possibility to numerically solve equation (5.1) is by using projection methods. The main idea of this class of methods is to search a low rank approximate solution $\mathbf{X}_k$, of the solution $\mathbf{X}$, whose existence is guaranteed by Theorem 4.5, and Theorem 4.6.

In particular, we first consider two sequences of nested linear spaces, namely $\{\mathcal{V}_k\}_{1\le k\le n}$ and $\{\mathcal{W}_k\}_{1\le k\le n}$, each of them has dimension $m := m(k,p)$, and such that $\mathcal{V}_k \subseteq \mathcal{V}_{k+1}$ and $\mathcal{W}_k \subseteq \mathcal{W}_{k+1}$, and such that there exists $\bar{k}$ such that $\mathcal{V}_{\bar{k}}$ and $\mathcal{W}_{\bar{k}}$ have dimension n. Such spaces exists for a sufficiently general matrices $\mathbf{A}$, $\mathbf{C}_1$ and $\mathbf{C}_2$. Then, by starting from $k = 1$, we build two matrices with orthonormal columns $\mathbf{V}_k$ and $\mathbf{W}_k$, whose columns form a basis of $\mathcal{V}_k$, and $\mathcal{W}_k$ respectively. We remark that since we consider a sequence of nested spaces, we want to avoid to rebuild the entire bases in each iteration; instead, we want to obtain the new bases by updating the bases of the previous iteration. We then want to find a matrix $\mathbf{X}_k$ in the form $\mathbf{X}_k = \mathbf{V}_k\mathbf{Y}\mathbf{W}_k^T$, with the matrix $\mathbf{Y} \in \mathbb{R}^{m\times m}$ such that the residual of the problem $\mathbf{R}_k := R(X_k) = \mathbf{A}\mathbf{X}_k + \mathbf{X}_k\mathbf{A} - \mathbf{C}_1\mathbf{C}_2^T$ is orthogonal to the space $\mathcal{W}_k \otimes \mathcal{V}_k$. In other words, we determine a matrix $\mathbf{Y} \in \mathbb{R}^{m\times m}$ such that the matrix $\mathbf{X}_k = \mathbf{V}_k\mathbf{Y}\mathbf{W}_k^T$ satisfies a Galerkin condition for equation (5.1), with respect to the space $\mathcal{W}_k \otimes \mathcal{V}_k$. In particular, since the columns of the matrices $\mathbf{V}_k$ and $\mathbf{W}_k$ for a basis of $\mathcal{V}_k$, and $\mathcal{W}_k$, an equivalent condition for the Galerkin condition with respect to the space $\mathcal{W}_k \otimes \mathcal{V}_k$ is to ask that the vectorized residual $\mathrm{vec}(\mathbf{R}_k)$ is orthogonal to the matrix $\mathbf{W}_k \otimes \mathbf{V}_k$. This leads us to the following projected problem

$$(\mathbf{W}_k \otimes \mathbf{V}_k)^T \mathrm{vec}(\mathbf{R}_k) = 0 \tag{5.7}$$

By taking the unique square matrix $\mathbf{R}_k$ obtained by inverting the operator vec in dimensions $n \times n$, we also obtain

$$\mathbf{V}_k^T(\mathbf{A}\mathbf{X}_k + \mathbf{X}_k\mathbf{A} - \mathbf{C}_1\mathbf{C}_2^T)\mathbf{W}_k = 0 \tag{5.8}$$

By rearranging the terms, by explictly writing $\mathbf{X}_k$ and since the matrices $\mathbf{V}_k$ and $\mathbf{W}_k$ have orthonormal columns, equation (5.8) becomes

$$(\mathbf{V}_k^T\mathbf{A}\mathbf{V}_k)\mathbf{Y} + \mathbf{Y}(\mathbf{W}_k^T\mathbf{A}\mathbf{W}_k) = (\mathbf{V}_k^T\mathbf{C}_1)(\mathbf{C}_2^T\mathbf{W}_k) \tag{5.9}$$

Since $\mathbf{Y}$ is the solution of a Lyapunov equation, and since the matrices $\mathbf{V}_k$ and $\mathbf{V}_k$ have orthonormal columns, the following representation holds.

Lemma 5.1. *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$, let $\mathbf{C}_1, \mathbf{C}_2 \in \mathbb{R}^{n\times p}$ be tall matrices. By defining the functions $x_{k,1}(t) := \mathbf{V}_k e^{-t\mathbf{H}_k}\mathbf{V}_k^T\mathbf{C}_1$, $x_{k,2}(t) := \mathbf{W}_k e^{-t\mathbf{T}_k}\mathbf{W}_k^T\mathbf{C}_2$, it holds that:*

$$\mathbf{X}_k = \int_0^\infty x_{k,1}(t)x_{k,2}^T(t)dt.$$

Proof. Let $\mathbf{Y}$ be the solution of equation (5.9). From Proposition 4.1, it holds that

$$\mathbf{Y} = \int_0^\infty e^{-t\mathbf{H}_k}\mathbf{V}_k^T\mathbf{C}_1\mathbf{C}_2^T\mathbf{W}_k e^{-t\mathbf{T}_k}dt.$$

Hence we have

$$\mathbf{X}_k = \mathbf{V}_k\mathbf{Y}\mathbf{W}_K^T = \mathbf{V}_k\left(\int_0^\infty e^{-t\mathbf{H}_k}\mathbf{V}_k^T\mathbf{C}_1\mathbf{C}_2^T\mathbf{W}_k e^{-t\mathbf{T}_k}dt\right)\mathbf{W}_k^T = \int_0^\infty x_{k,1}(t)x_{k,2}^T(t)dt.$$

Where the last equality holds since the matrices $\mathbf{V}_k$ and $\mathbf{W}_k$ do not depend on t. □

Moreover, if we set the matrices $\mathbf{H}_k := \mathbf{V}_k^T\mathbf{A}\mathbf{V}_k$ and $\mathbf{T}_k := \mathbf{W}_k^T\mathbf{A}\mathbf{W}_k$, it holds that the set $[\lambda_{min}(\mathbf{H}_k), \lambda_{max}(\mathbf{H}_k)] \cup [-\lambda_{max}(\mathbf{T}_k), -\lambda_{min}(\mathbf{T}_k)]$ is contained in the set $[\lambda_{min}(\mathbf{A}), \lambda_{max}(\mathbf{A})] \cup [-\lambda_{max}(\mathbf{A}), -\lambda_{min}(\mathbf{A})]$, and so we obtain the analogous representation of Proposition 4.3 for the projected solution $\mathbf{X}_k$, for all curves Γ which satisfy the hypotheses of Proposition 4.3. In the following lemma we state the result.

Lemma 5.2. *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$, let $\mathbf{C}_1, \mathbf{C}_2 \in \mathbb{R}^{n\times p}$. Let Γ be a closed contour in $\overline{\mathbb{C}} \setminus ([\lambda_{min}(\mathbf{A}), \lambda_{max}(\mathbf{A})] \cup [-\lambda_{max}(\mathbf{A}), -\lambda_{min}(\mathbf{A})])$ which is homeotopic to the imaginary axis $i\mathbb{R}$ being passed from $-i\infty$ to $-i\infty$. It holds that:*

$$\mathbf{X}_k = \frac{1}{2\pi i}\int_\Gamma \mathbf{V}_k(z\mathbf{I} - \mathbf{H}_k)^{-1}\mathbf{V}_k^T\mathbf{C}_1\mathbf{C}_2^T\mathbf{W}_k(z\mathbf{I} - \mathbf{T}_k)^{-*}\mathbf{W}_k^T dz \tag{5.10}$$

Proof. Let $\mathbf{Y}$ be the solution of equation (5.9). From Proposition 4.3, it holds that

$$\mathbf{Y} = \frac{1}{2\pi i}\int_\Gamma (z\mathbf{I} - \mathbf{H}_k)^{-1}\mathbf{V}_k^T\mathbf{C}_1\mathbf{C}_2^T\mathbf{W}_k(z\mathbf{I} - \mathbf{T}_k)^{-*}dz.$$

Hence we have

$$\mathbf{X}_k = \mathbf{V}_k\mathbf{Y}\mathbf{W}_K^T = \frac{1}{2\pi i}\int_\Gamma \mathbf{V}_k(z\mathbf{I} - \mathbf{H}_k)^{-1}\mathbf{V}_k^T\mathbf{C}_1\mathbf{C}_2^T\mathbf{W}_k(z\mathbf{I} - \mathbf{T}_k)^{-*}\mathbf{W}_K^T dz$$

Where the last equality holds since the matrices $\mathbf{V}_k$ and $\mathbf{W}_k$ do not depend on t. It is sufficient to apply Proposition 4.3 to equation (5.9). □

We also notice that the matrices involved in equation (5.9) have dimensions $m \times m$, instead of $n \times n$. As a consequence, we are left to solve a smaller problem, for instance by Algorithm 2. The last step is to iterate the process in k, until the norm of the error $\mathbf{E}_k := \mathbf{X} - \mathbf{X}_k$ is less than a fixed tolerance. The effectiveness of the method thus depends on the number of iterations, which is associated with the approximation space dimension, and the computational cost of each iteration. The complete procedure is outlined in Algorithm 3.

Algorithm 3: Galerkin Projection Methods

Data: $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{C}_1, \mathbf{C}_2 \in \mathbb{R}^{n\times p}$
Result: $\mathbf{X}_k \in \mathbb{R}^{n\times n}$ low rank approximation of the solution $\mathbf{X}$ of $\mathbf{AX} + \mathbf{XA} = \mathbf{C}_1\mathbf{C}_2^T$
1 Choose two nested sequences of linear spaces $\{\mathcal{V}_k\}_k$, and $\{\mathcal{W}_k\}_k$,
For $k = 1, \ldots,$ convergence,
2 Compute two matrices $\mathbf{V}_k$ and $\mathbf{W}_k$, whose columns are orthonormal and form a basis of $\mathcal{V}_k$, and $\mathcal{W}_k$,
3 Solve the equation $(\mathbf{V}_k^T\mathbf{A}\mathbf{V}_k)\mathbf{Y} + \mathbf{Y}(\mathbf{W}_k^T\mathbf{A}\mathbf{W}_k) = (\mathbf{V}_k^T\mathbf{C}_1)(\mathbf{C}_2^T\mathbf{W}_k)$,
4 Estimate $\|\mathbf{X} - \mathbf{X}_k\|$
End For

We highlight two key points of Algorithm 3. The first one is item 1, where we have to choose the spaces in which we project equation (5.1). We propose three different choices for the spaces. In particular, we investigate the Krylov space, introduced in Definition 1.19, and two generalizations of the Krylov space. The second point is item 2, where the expansion of the Krylov space at each iteration is performed by a generalization of the Gram-Schmidt process, called in this context the Arnoldi algorithm, which we describe in the next section.

5.4 Arnoldi algorithm

Given a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $c \in \mathbb{R}^n$, in this section we treat the modified Gram-Schmidt algorithm in the case where the basis comes from a Krylov space, which we recall

$$\mathcal{K}_k(\mathbf{A}, c) := span\{c, \mathbf{A}c, \ldots, \mathbf{A}^{k-1}c\}.$$

In this case, the procedure is called Arnoldi algorithm, and it creates a matrix with orthonormal columns $\mathbf{V}_k = [v_1, \ldots, v_k]$, whose columns form a basis of $\mathcal{K}_k(\mathbf{A}, c)$, as follows.

As first column we take $v_1 = \frac{c}{\|c\|}$. We compute $p_2 := \mathbf{A}v_1$, and then we orthogonalize p_2 with respect to v_1, as

$$\hat{v}_2 = w_2 - v_1 h_{1,1}, \qquad h_{1,1} := v_1^T p_2.$$

We then orthonormalize $\hat{v}_2$, by setting $v_2 = \frac{\hat{v}_2}{h_{2,1}}$, where $h_{2,1} := \|\hat{v}_2\|$. By repeating the iteration in j, we obtain the following recurrence. First, we compute $p_j := \mathbf{A}v_{j-1}$ and then we orthogonalize p_j with respect to $\mathbf{V}_{j-1} := [v_1, \ldots, v_{j-1}]$, as follows

$$\hat{v}_j = p_j - \mathbf{V}_{j-1}\begin{bmatrix} h_{1,j-1} \\ \vdots \\ h_{j-1,j-1} \end{bmatrix}, \qquad h_{i,j-1} := v_i^T p_j, \quad i = 1, \ldots, j-1. \tag{5.11}$$

We then orthonormalize $\hat{v}_j$, by setting $v_j = \frac{\hat{v}_j}{h_{j,j-1}}$ where $h_{j,j-1} := \|\hat{v}_j\|$. To improve the stability of the algorithm, the orthogonalization step is done vector by vector,

in a for loop, as explained in the modified Gram-Schmidt algorithm. By a simple computation, it is possible to verify that the computational cost of the Arnoldi algorithm is $\mathcal{O}(nk^2)$.

Notation 5.3. We denote the matrix of the orthogonalization coefficients

$$\underline{\mathbf{H}}_k := (h_{i,j})_{\substack{i=1,\dots,k+1,\\ j=1,\dots,k}} \quad \underline{\mathbf{H}}_k = \begin{bmatrix} \mathbf{H}_k \\ h_{k+1,k} e_k^T \end{bmatrix}.$$

In the following we analyze the structure of the matrix $\mathbf{H}_k$ determined by the Arnoldi iteration (5.11), and we state some important properties of the algorithm.

Remark. The matrix $\mathbf{H}_k$ is upper Hessenberg.

Theorem 5.4 (Arnoldi relation)**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$, and $c \in \mathbb{R}^n$. The matrices $\mathbf{V}_k$ and $\mathbf{H}_k$ determined by the Arnoldi iteration (5.11) satisfy:*

$$\mathbf{A}\mathbf{V}_k = \mathbf{V}_k \mathbf{H}_k + v_{k+1} h_{k+1,k} e_k^T$$

Proof. From the Arnoldi iteration (5.11), we have that

$$\mathbf{A}v_j = v_{j+1} h_{j+1,j} + \mathbf{V}_j \begin{bmatrix} h_{1,j} \\ \vdots \\ h_{j-1,j} \end{bmatrix}.$$

By taking $j = 1, \dots, k$, we obtain the result. □

As a corollary of this formula, we can state the following consequences.

Remark. Since the columns of $\mathbf{V}_n$ are n independent vectors in $\mathbb{R}^n$ and v_{n+1} is orthogonal to $\mathbf{V}_n$, we have that $v_{n+1} = 0$. As a consequence, it holds that:

$$\mathbf{A}\mathbf{V}_n = \mathbf{V}_n \mathbf{H}_n.$$

Remark. Since $\mathbf{V}_k$ has orthonormal columns, and since v_{k+1} is orthogonal to $\mathbf{V}_k$, it holds that:

$$\mathbf{V}_k^T \mathbf{A}\mathbf{V}_k = \mathbf{H}_k.$$

In the following we comment the symmetric case, studied by Lanczos in [20].

Lemma 5.5. *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a symmetric matrix. The matrix $\mathbf{H}_k \in \mathbb{R}^{k\times k}$ determined by the Arnoldi iteration is tridiagonal.*

Proof. We recall that $\mathbf{V}_k^T \mathbf{A}\mathbf{V}_k = \mathbf{H}_k$. As a consequence, $\mathbf{H}_k$ is symmetric. The only possibility for $\mathbf{H}_k$ to be symmetric and upper Hessenberg is to be tridiagonal. □

As an important consequence, it is possible so simplify the inner loop of the algorithm. Indeed, it holds that $h_{i,j-1} = 0$ for $i = 1, \dots, j-2$, and thus the

Algorithm 4: Arnoldi Iteration

Data: $\mathbf{A} \in \mathbb{R}^{n\times n}$, $c \in \mathbb{R}^n$
Result: $\mathbf{V}_k \in \mathbb{R}^{n\times k}$ whose columns are orthonormal and form a basis for $\mathcal{K}_k(\mathbf{A}, c)$
Set $v_1 = \frac{c}{\|c\|}$; $\mathbf{V}_0 = \emptyset$;
For $j = 2, \dots, k$
  Set $\mathbf{V}_{j-1} = [\mathbf{V}_{j-2}, v_{j-1}]$;
  Compute $p_j = \mathbf{A}v_{j-1}$;
  Obtain $h_{i,j-1}$ and v_j from p_j with the modified Gram-Schmidt algorithm.
End For
Set $\mathbf{V}_k = [\mathbf{V}_{k-1}, v_k]$.

orthogonalization procedure involves just the last two terms. We set $p_j = \mathbf{A}v_{j-1}$, and we orthogonalize p_j as

$$\hat{v}_j = p_j - h_{j-1,j-1}v_{j-1} - h_{j-2,j-1}v_{j-2}.$$

Moreover, this simplification makes the computational cost of the algorithm lower, as it is $\mathcal{O}(nk)$.

Given $\mathbf{C} \in \mathbb{R}^{n\times p}$, the previous algorithm can be generalized to the case of Block Krylov space $\mathcal{K}_k(\mathbf{A}, \mathbf{C})$, by setting $\mathbf{P}_j = \mathbf{A}\mathbf{V}_{j-1} \in \mathbb{R}^{n\times p}$ and the block $\mathbf{H}_{i,j} = \mathbf{V}_i^T\mathbf{P}_j$. Differently from the vector case, the tall matrix $\mathbf{V}_j$ is obtained by the tall matrix $\mathbf{P}_j$ by using the reduced QR factorization, and by taking the block $\mathbf{H}_{j+1,j}$ as the reduced triangular matrix resulting from the reduced QR factorization. In this setting, the matrix $\mathbf{H}_k$ is no longer Hessenberg, but block tridiagonals, with $p \times p$ blocks. We also remark that a discussion about the loss of orthogonality of the method can be found in [21], where some modifications are implemented in special cases.

5.5 Convergence Analysis of the Galerkin method with Krylov space

In this section we analyze the convergence of Algorithm 3 when the approximation space is the polynomial Krylov subspace; which we recall, for a given matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $c \in \mathbb{R}^n$.

$$\mathcal{K}_k(\mathbf{A}, c) = span\{c, \mathbf{A}c, \dots, \mathbf{A}^{k-1}c\}.$$

In our notation, we consider $\mathbf{C} := c_1c_2^T$, and we set the spaces of projection $\mathcal{V}_k := \mathcal{K}_k(\mathbf{A}, c_1)$, and $\mathcal{W}_k := \mathcal{K}_k(\mathbf{A}, c_2)$; moreover we set $\mathbf{H}_k := \mathbf{V}_k^T\mathbf{A}\mathbf{V}_k$, $\mathbf{T}_k := \mathbf{W}_k^T\mathbf{A}\mathbf{W}_k$ and $\mathbf{X}_k := \mathbf{V}_k\mathbf{Y}\mathbf{W}_k^T$, where $\mathbf{V}_k$ and $\mathbf{W}_k$ are matrices with orthonormal columns, whose columns form a basis for $\mathcal{V}_k$ and $\mathcal{W}_k$ and are obtained by the Arnoldi algorithm. For this choice, we study the convergence rate of the approximate solution, which can be found in [26], and we provide a numerical example in

our setting. In order to tackle the proof of the convergence statement, we have to introduce some preliminary results. By using Proposition 4.1, we write the solution of the Lyapunov equation (4.1) as follows:

$$\mathbf{X} = \int_0^\infty e^{-t\mathbf{A}} c_1 c_2^T e^{-t\mathbf{A}} dt.$$

We also recall Lemma 5.1, which gives us the following representation for the approximated solution $\mathbf{X}_k$.

$$X_k = \int_0^\infty x_{k,1}(t) x_{k,2}^T(t) dt.$$

As a consequence of these representations for $\mathbf{X}$ and $\mathbf{X}_k$, if we consider c_1 and c_2 having unit norm, we have that:

$$\begin{aligned}
\|\mathbf{X} - \mathbf{X}_k\| &\leq \int_0^\infty \left\| x_1 x_2^T - x_{k,1} x_{k,2}^T \right\| dt \\
&= \int_0^\infty \left\| x_1 (x_2 - x_{k,2})^T - (x_{k,1} - x_1) x_{k,2}^T \right\| dt \\
&\leq \int_0^\infty \|x_1\| \|x_2 - x_{k,2}\| + \|x_{k,1} - x_1\| \|x_{k,2}\| dt \\
&\leq \int_0^\infty e^{-t\lambda_{min}(\mathbf{A})} \|x_2 - x_{k,2}\| + e^{-t\lambda_{min}(\mathbf{A})} \|x_{k,1} - x_1\| dt
\end{aligned}$$

Where the final bound derives from the definitions of x_1 and $x_{k,2}$, and from the fact that $[\lambda_{min}(\mathbf{T}_k), \lambda_{max}(\mathbf{T}_k)] \subseteq [\lambda_{min}(\mathbf{A}), \lambda_{max}(\mathbf{A})]$. By defining the matrices $\tilde{\mathbf{A}} := \mathbf{A} + \lambda_{min}(\mathbf{A})$, $\tilde{\mathbf{H}}_k := \mathbf{V}_k^T \tilde{\mathbf{A}} \mathbf{V}_k$, $\tilde{\mathbf{T}}_k := \mathbf{W}_k^T \tilde{\mathbf{A}} \mathbf{W}_k$ and the quantities $\tilde{x}_1 := e^{-t\tilde{\mathbf{A}}} c_1$, $\tilde{x}_2 := e^{-t\tilde{\mathbf{A}}} c_2$, $\tilde{x}_{k,1} := \mathbf{V}_k e^{-t\tilde{\mathbf{H}}_k} \mathbf{V}_k^T c_1$, $\tilde{x}_{k,2} := \mathbf{W}_k e^{-t\tilde{\mathbf{T}}_k} \mathbf{W}_k^T c_2$, i.e, as the quantities $x_1, x_2, x_{k,1}, x_{k,2}$ related to the matrix $\tilde{\mathbf{A}}$, we are able to write the previous error as the error in the approximation of the exponential of the shifted matrix with the Krylov subspace solution, as follows:

$$e^{-t\lambda_{min}(\mathbf{A})} \|x_{k,1} - x_1\| = \|\tilde{x}_{k,1} - \tilde{x}_1\|, \quad e^{-t\lambda_{min}(\mathbf{A})} \|x_2 - x_{k,2}\| = \|\tilde{x}_2 - \tilde{x}_{k,2}\|.$$

In the end, we obtain:

$$\|\mathbf{X} - \mathbf{X}_k\| \leq \int_0^\infty (\|\tilde{x}_{k,1} - \tilde{x}_1\| + \|\tilde{x}_2 - \tilde{x}_{k,2}\|) dt. \tag{5.12}$$

For deriving a final bound on the norm of the error, we recall Faber polynomials, introduced in Definition 1.23, and their properties, stated in Theorem 1.24 and in Theorem 1.25. Moreover, we need to state the following technical lemma.

Lemma 5.6. *Let Φ_k be the Faber polynomial of degree k, and let f be an analytic function, with $f(z) = \sum_{n=0}^{\infty} a_n \Phi_n(z)$ as a series expansion. If the expansion of $f(\mathbf{A})$, $f(\mathbf{H}_k)$ and $f(\mathbf{T}_k)$ are well defined, we have that:*

$$\left\| f(\mathbf{A})c_1 - \mathbf{V}_k f(\mathbf{H}_k)\mathbf{V}_k^T c_1 \right\| \leq \sum_{j=k}^{\infty} |a_j|(\|\Phi_j(\mathbf{A})\| + \|\Phi_j(\mathbf{H}_k)\|)$$

and

$$\left\| f(\mathbf{A})c_2 - \mathbf{W}_k f(\mathbf{T}_k)\mathbf{W}_k^T c_2 \right\| \leq \sum_{j=k}^{\infty} |a_j|(\|\Phi_j(\mathbf{A})\| + \|\Phi_j(\mathbf{T}_k)\|)$$

Proof. To prove the result it is sufficient to show one of the two inequalities. We have that:

$$f(\mathbf{A})c_1 - \mathbf{V}_k f(\mathbf{H}_k)\mathbf{V}_k^T c_1 = \sum_{j=0}^{\infty} a_j(\Phi_j(\mathbf{A})c_1 - \mathbf{V}_k \Phi_j(\mathbf{H}_k)\mathbf{V}_k^T c_1)$$

By splitting the sum, we have that

$$\begin{aligned}
&\sum_{j=0}^{\infty} a_j(\Phi_j(\mathbf{A})c_1 - \mathbf{V}_k \Phi_j(\mathbf{H}_k)\mathbf{V}_k^T c_1) \\
=&\sum_{j=0}^{k-1} a_j(\Phi_j(\mathbf{A})c_1 - \mathbf{V}_k \Phi_j(\mathbf{H}_k)\mathbf{V}_k^T c_1) + \sum_{j=k}^{\infty} a_j(\Phi_j(\mathbf{A})c_1 - \mathbf{V}_k \Phi_j(\mathbf{H}_k)\mathbf{V}_k^T c_1)
\end{aligned}$$

We are left to prove that $\sum_{j=0}^{k-1} a_j(\Phi_j(\mathbf{A})c_1 - \mathbf{V}_k \Phi_j(\mathbf{H}_k)\mathbf{V}_k^T c_1) = 0$. This fact follows from Arnoldi relation and from the structure of $\mathbf{H}_k$. Indeed, for $j = 0, \ldots, k-1$, we have that $\mathbf{A}^j c_1 = \mathbf{V}_k \mathbf{H}_1^j \mathbf{V}_k^T c_1$; as a consequence, $\Phi_j(\mathbf{A})c_1 = \mathbf{V}_k \Phi(\mathbf{H}_k)\mathbf{V}_k^T c_1$, for $j = 0, \ldots, k-1$. □

In the lemma above, the role of Faber polynomials is to guarantee the well posedness of a polynomial expansion for an analytic f, but the same proof can be done for other polynomials, if we know that the expansion of the function f via these polynomials is convergent. For our case, we highlight that Faber polynomials coincide with Chebychev ones.

We are now able to state the main theorem about the rate of convergence of the method for the Block Krylov subspace as projection space.

Theorem 5.7. *Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a symmetric positive definite matrix, let $\tilde{\kappa} := \frac{\lambda_{max}(\mathbf{A}) + \lambda_{min}(\mathbf{A})}{2\lambda_{min}(\mathbf{A})}$ and let $c_1, c_2 \in \mathbb{R}^n$. It holds that*

$$\|\mathbf{X} - \mathbf{X}_k\| \leq 2\frac{\sqrt{\tilde{\kappa}} + 1}{\lambda_{min}(\mathbf{A})\sqrt{\tilde{\kappa}}} \left(\frac{\sqrt{\tilde{\kappa}} - 1}{\sqrt{\tilde{\kappa}} + 1} \right)^k$$

Proof. By using (5.12), we are left to estimate the quantities $\int_0^\infty \|\tilde{x}_{k_1} - \tilde{x}_1\|\, dt$, and $\int_0^\infty \|\tilde{x}_2 - \tilde{x}_{k_2}\|\, dt$. We estimate the first one, the second can be done in the same way. By using Theorem 1.24, it holds that both x_1, $x_{k,1}$ can be written as Faber expansions, which here coincides with the Chebyshev expansion. Thanks to [8, Formula (4.2)], and by defining the quantities $\hat{\mathbf{A}} := \frac{\lambda_{max}(\mathbf{A})+\lambda_{min}(\mathbf{A})}{\lambda_{max}(\mathbf{A})-\lambda_{min}(\mathbf{A})}\mathbf{I} - \frac{2}{\lambda_{max}(\mathbf{A})-\lambda_{min}(\mathbf{A})}\mathbf{A}$, and $\hat{\mathbf{H}}_k := \frac{\lambda_{max}(\mathbf{A})+\lambda_{min}(\mathbf{A})}{\lambda_{max}(\mathbf{A})-\lambda_{min}(\mathbf{A})}\mathbf{I} - \frac{2}{\lambda_{max}(\mathbf{A})-\lambda_{min}(\mathbf{A})}\mathbf{H}_k$, we obtain:

$$\tilde{x}_1 = 2e^{-t\frac{\lambda_{max}(\mathbf{A})+3\lambda_{min}(\mathbf{A})}{2}} \sum_{j=0}^{\infty}{}' I_j\Big(t\frac{\lambda_{max}(\mathbf{A}) - \lambda_{min}(\mathbf{A})}{2}\Big)\mathcal{T}_j(\hat{\mathbf{A}})c_1,$$

$$\tilde{x}_{k_1} = \mathbf{V}_k(2e^{-t\frac{\lambda_{max}(\mathbf{A})+3\lambda_{min}(\mathbf{A})}{2}} \sum_{j=0}^{\infty}{}' I_j\Big(t\frac{\lambda_{max}(\mathbf{A}) - \lambda_{min}(\mathbf{A})}{2}\Big)\mathcal{T}_j(\hat{\mathbf{H}}_k)\mathbf{V}_k^T c_1),$$

where I_j is the Bessel function of an imaginary argument, and the prime over the sum indicates that the first term is divided by two. By the definitions of $\hat{\mathbf{A}}$ and $\hat{\mathbf{H}}_1$, it holds that $\left\|\mathcal{T}_j(\hat{\mathbf{A}})\right\|, \left\|\mathcal{T}_j(\hat{\mathbf{H}}_k)\right\| \leq 1$. By using Lemma 5.6, we have that

$$\|\tilde{x}_{k_1} - \tilde{x}_1\| \leq 4e^{-t\frac{\lambda_{max}(\mathbf{A})+3\lambda_{min}(\mathbf{A})}{2}} \sum_{j=k}^{\infty} I_j\Big(t\frac{\lambda_{max}(\mathbf{A}) - \lambda_{min}(\mathbf{A})}{2}\Big)$$

By taking the integral, and by setting the quantities $q := \frac{\tilde{\kappa}+1}{\tilde{\kappa}-1}$, $\rho := q + \sqrt{q^2-1}$, thanks to [14, Formula (6.611.4)], we obtain:

$$\begin{aligned}
&\int_0^\infty \|\tilde{x}_{k_1} - \tilde{x}_1\|\, dt \\
\leq &\int_0^\infty 4e^{-t\frac{\lambda_{max}(\mathbf{A})+3\lambda_{min}(\mathbf{A})}{2}} \sum_{j=k}^{\infty} I_j\Big(t\frac{\lambda_{max}(\mathbf{A}) - \lambda_{min}(\mathbf{A})}{2}\Big)dt \\
(6.611.4) = &\frac{4}{\sqrt{(\frac{3\lambda_{min}(\mathbf{A})+\lambda_{max}(\mathbf{A}))}{2})^2 - (\frac{\lambda_{max}(\mathbf{A})-\lambda_{min}(\mathbf{A})}{2})^2}} \sum_{j=k}^{\infty} \frac{1}{\rho^j} \\
= &\frac{2}{\lambda_{min}(\mathbf{A})\sqrt{\tilde{\kappa}}} \sum_{j=0}^{\infty} \frac{1}{\rho^{j+k}} = \frac{2}{\lambda_{min}(\mathbf{A})\sqrt{\tilde{\kappa}}} \frac{\rho}{\rho-1}\frac{1}{\rho^k} \\
= &\frac{\sqrt{\tilde{\kappa}}+1}{\lambda_{min}(\mathbf{A})\sqrt{\tilde{\kappa}}}\Big(\frac{\sqrt{\tilde{\kappa}}-1}{\sqrt{\tilde{\kappa}}+1}\Big)^k
\end{aligned}$$

Where we used the facts that:

$$\frac{\lambda_{max}(\mathbf{A}) + 3\lambda_{min}(\mathbf{A})}{2\lambda_{min}(\mathbf{A})} = \tilde{\kappa} + 1, \qquad \frac{\lambda_{max}(\mathbf{A}) - \lambda_{min}(\mathbf{A})}{2\lambda_{min}(\mathbf{A})} = \tilde{\kappa} - 1.$$

and

$$\rho = \frac{\tilde{\kappa}+1}{\tilde{\kappa}-1} + \sqrt{\frac{4\tilde{\kappa}}{(\tilde{\kappa}-1)^2}} = \frac{\tilde{\kappa}+1+2\sqrt{\tilde{\kappa}}}{\tilde{\kappa}-1} = \frac{(\sqrt{\tilde{\kappa}}+1)^2}{(\sqrt{\tilde{\kappa}}-1)(\sqrt{\tilde{\kappa}}+1)} = \frac{\sqrt{\tilde{\kappa}}+1}{\sqrt{\tilde{\kappa}}-1}$$

By repeating the same computations on the second term, we obtain the result. □

We remark that the bound involves the condition number of the shifted matrix $\tilde{\mathbf{A}} := \mathbf{A} + \lambda_{min}(\mathbf{A})$, instead of the the condition number of the matrix $\mathbf{A}$. Besides, if we call κ the condition number of $\mathbf{A}$, we have that the condition number of $\tilde{\mathbf{A}}$ satisfies $\frac{1}{2}\kappa \leq \tilde{\kappa} \leq \kappa$. As a consequence, the bound has a faster convergence to zero than the same bound relative to $\mathbf{A}$. As a counterpart, the rate of convergence strongly depends on the condition number $\tilde{\kappa}$, and we might expect a slow convergence if the problem is not well-conditioned.

Example 5.1. We consider the matrix $\mathbf{A} = \tilde{\mathbf{K}} \in \mathbb{R}^{n\times n}$, as defined in Lemma 3.14, for $n = 2^l$, and $l = 4, \ldots, 10$, where $2^{10} = 1024$. In Table 5.1, we show the quantity $\frac{\sqrt{\tilde{\kappa}}-1}{\sqrt{\tilde{\kappa}}+1}$ as the dimensions of the problem grows.

n	$\kappa(\mathbf{A})$	$\frac{\sqrt{\tilde{\kappa}}-1}{\sqrt{\tilde{\kappa}}+1}$
2^4	1.6115e+03	9.3131e-01
2^5	6.3773e+03	9.6482e-01
2^6	2.5343e+04	9.8218e-01
2^7	1.0101e+05	9.9103e-01
2^8	4.0328e+05	9.9550e-01
2^9	1.6116e+06	9.9775e-01
2^{10}	6.4433e+06	9.9887e-01

Table 5.1: Rate of convergence from Theorem 5.7 for Example 5.1.

Example 5.2. We consider the matrix $\mathbf{A} = \tilde{\mathbf{K}} \in \mathbb{R}^{n\times n}$, as defined in Lemma 3.14, of dimension $n = 2^6 = 64$, and we take as data the normalized matrix of $\mathbf{C} = \mathbf{L}^{-1}vv^T\mathbf{L}^{-T}$, where v is the vector of all ones. The following figure shows the behaviour of the error norm compared with the rate of convergence $\left(\frac{\sqrt{\tilde{\kappa}}-1}{\sqrt{\tilde{\kappa}}+1}\right)^k$ as k increases.

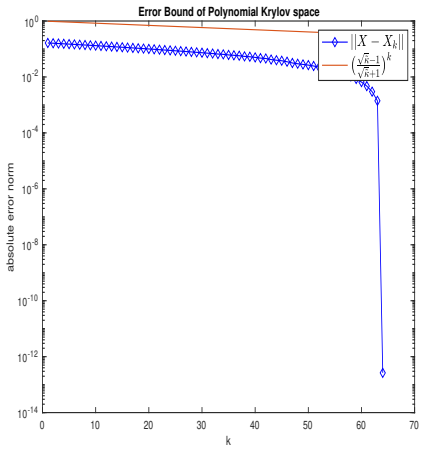

Figure 5.1: Comparison between the true error norm and the theoretical rate of convergence for Example 5.2.

By ignoring the constants, the bound catches well the initial behaviour of the iterates until superlinear convergence takes place, and so it is sharp.

We emphasize that the rank of the data matrix $\mathbf{C} := \mathbf{C}_1\mathbf{C}_2^T$ can be greater than one. For this scenario, we use as approximation space the Block Krylov space, which we recall from Definition 1.19 that for a given $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{C} \in \mathbb{R}^{n\times p}$ we set

$$\mathcal{K}_k^{\square}(\mathbf{A}, \mathbf{C}) = blkspan\{\mathbf{C}, \mathbf{AC}, \ldots, \mathbf{A}^{k-1}\mathbf{C}\}.$$

In particular, for this scenario we set the spaces of projection $\mathcal{V}_k := \mathcal{K}_k^{\square}(\mathbf{A}, \mathbf{C}_1)$, and $\mathcal{W}_k := \mathcal{K}_k^{\square}(\mathbf{A}, \mathbf{C}_2)$. Unfortunately, the convergence analysis for the block case is fare more involved than for the vector case, because it requires the use of matrix polynomials. We thus limit ourselves to working with the case when $\mathbf{C}$ is a rank one matrix, so that we can dwell with vector Krylov subspaces. A convergence analysis for this scenario can be found on [5].

5.6 Arnoldi algorithm for the Extended Krylov space

Given $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $c \in \mathbb{R}^n$, we want to enrich the previously studied Krylov space, and catch more important spectral information of the matrix $\mathbf{A}$. For this reason we introduce the so called Extended Krylov space, which is defined as follows.

Definition 5.8 (Extended Krylov space)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times n}$, and $c \in \mathbb{R}^n$. We define the k-th Extended Krylov space of $\mathbf{A}$ and c as

$$\mathcal{EK}_k(\mathbf{A}, c) = \mathcal{K}_k(\mathbf{A}, c) + \mathcal{K}_k(\mathbf{A}^{-1}, \mathbf{A}^{-1}c). \tag{5.13}$$

Indeed, it is well known that the Krylov space catches well the behaviour of the largest eigenvalues of the matrix $\mathbf{A}$, see e.g. [20]. Hence, we want to use this extension to catch also the behaviour of the smallest eigenvalues of $\mathbf{A}$. In this section we describe the procedure to obtain a matrix with orthonormal columns $\mathbf{V}_k = [v_1, \dots, v_k]$, whose columns form a basis of $\mathcal{EK}_k(\mathbf{A}, c)$.

As first pair of columns we take $p_1 = [c, \mathbf{A}^{-1}c]$. Then we take v_1 as the orthonormal pair of vectors obtained by applying the block Gram-Schmidt procedure to p_1. We then compute $p_2 := [\mathbf{A}v_1^{(1)}, \mathbf{A}^{-1}v_1^{(2)}]$, where $v_1^{(1)}$ is the first column of v_1 and $v_1^{(2)}$ is the second column of v_1; and we orthonormalize p_2 with respect to v_1, by applying the block Gram-Schmidt procedure. By repeating the iteration in j, we obtain the following recurrence. First, we compute $p_j := [\mathbf{A}v_{j-1}^{(1)}, \mathbf{A}^{-1}v_{j-1}^{(2)}]$; then we orthonormalize p_j with respect to $\mathbf{V}_{j-1} := [v_1, \dots, v_{j-1}]$ by applying the block Gram-Schmidt procedure. Also for this space, it can be shown that the matrix $\mathbf{H}_k := \mathbf{V}_k^T\mathbf{A}\mathbf{V}_k$, $\mathbf{H}_k \in \mathbb{R}^{2k\times 2k}$ is upper Hessenberg, see e.g. [25], and thus $\mathbf{H}_k$ is block tridiagonal when the matrix $\mathbf{A}$ is symmetric.

Algorithm 5: Arnoldi Iteration for the Extended Krylov space

Data: $\mathbf{A} \in \mathbb{R}^{n\times n}$, $c \in \mathbb{R}^n$
Result: $\mathbf{V}_k \in \mathbb{R}^{n\times 2k}$ whose columns are orthonormal and form a basis for $\mathcal{EK}_k(\mathbf{A}, c)$
1 Set $p_1 = [c, \mathbf{A}^{-1}c]$, $\mathbf{V}_0 = \emptyset$;
2 Obtain v_1 from p_1 with the modified block Gram-Schmidt algorithm;
For $j = 2, \dots, k$
3 Set $\mathbf{V}_{j-1} = [\mathbf{V}_{j-2}, v_{j-1}]$;
4 Compute $p_j = [\mathbf{A}v_{j-1}^{(1)}, \mathbf{A}^{-1}v_{j-1}^{(2)}]$;
5 Obtain v_j from p_j with the modified block Gram-Schmidt algorithm.
End For
6 Set $\mathbf{V}_k = [\mathbf{V}_{k-1}, v_k]$.

As for Algorithm 4, Algorithm 5 can be generalized in case of tall matrices $\mathbf{C}$, instead of vectors c.

5.7 Convergence Analysis of the Galerkin method with Extended Krylov space

In this section we analyze the convergence of Algorithm 3 when the approximation space is the Extended Krylov subspace; introduced in Definition 5.8. In our notation, we consider $\mathbf{C} := c_1c_2^T$, and we set the spaces of projection $\mathcal{V}_k := \mathcal{EK}_k(\mathbf{A}, c_1)$, and $\mathcal{W}_k := \mathcal{EK}_k(\mathbf{A}, c_2)$; moreover we set $\mathbf{H}_k := \mathbf{V}_k^T\mathbf{A}\mathbf{V}_k$, $\mathbf{T}_k := \mathbf{W}_k^T\mathbf{A}\mathbf{W}_k$ and $\mathbf{X}_k := \mathbf{V}_k\mathbf{Y}\mathbf{W}_k^T$, where $\mathbf{V}_k$ and $\mathbf{W}_k$ are matrices with orthonormal columns, whose columns form a basis for $\mathcal{V}_k$ and $\mathcal{W}_k$ and are obtained by the Arnoldi algorithm. In particular, in this section we derive the convergence rate of the solution that can be found in [17]. In order to derive the final bound,

we have to introduce some preliminary results. By using Proposition 4.3, we write the solution of the Lyapunov equation (4.1) as follows:

$$\mathbf{X} = \frac{1}{2\pi i}\int_{\Gamma}(z\mathbf{I}-\mathbf{A})^{-1}c_1c_2^T(z\mathbf{I}-\mathbf{A})^{-*}dz$$

Or equivalently:

$$\mathbf{X} = \frac{1}{2\pi i}\int_{-i\infty}^{i\infty}(z\mathbf{I}-\mathbf{A})^{-1}c_1c_2^T(z\mathbf{I}-\mathbf{A})^{-*}dz$$

We also recall Lemma 5.2 for the following representation of the approximate solution $\mathbf{X}_k$.

$$\mathbf{X}_k = \frac{1}{2\pi i}\int_{\Gamma}\mathbf{V}_k(z\mathbf{I}-\mathbf{H}_k)^{-1}\mathbf{V}_k^Tc_1c_2^T\mathbf{W}_k(z\mathbf{I}-\mathbf{T}_k)^{-*}\mathbf{W}_k^Tdz.$$

We now want to express the quantity $(z\mathbf{I}-\mathbf{A})^{-1}$ in a way so that we are able to replicate Lemma 5.6 in the case of Extended Krylov. For this reason, we use a generalization of the previously defined Faber polynomials, which are called Takenaka-Malmquist rational functions, introduced in Definition 1.26, and Faber-Dzhrbashyan rational function, introduced in Definition 1.28.

Moreover, we notice that in our case, Definition 1.26 can be reduced to a simpler form. Indeed, we have $\sigma_{2l} = \infty$, and $\sigma_{2l+1} = 0$. By definition, we have that $\phi(\infty) = \infty$, and $\phi(0) < 0$, and thus we find out that:

$$B_{2l}(w) = \left(\frac{-w}{1}\right)^l\left(-\frac{1-\phi(0)w}{\phi(0)-w}\right)^l$$

As a consequence, we obtain the following system of functions:

$$\Phi_{2l}(w) = B(w)^l, \qquad \Phi_{2l+1}(w) = -\frac{\sqrt{1-|\phi(0)|^{-2}}}{1-|\phi(0)|^{-1}w}wB(w)^l, \quad l\in\mathbb{N} \tag{5.14}$$

where the function B is:

$$B(w) := w\left[\frac{1-\phi(0)w}{\phi(0)-w}\right], \quad w\in\mathbb{C}.$$

By applying Proposition 1.29, we are able to express the quantity $(z\mathbf{I}-\mathbf{A})^{-1}$ in a more manageable way.

Lemma 5.9. *Let $\mathbf{A}$, let $z\notin E := [\lambda_{min}(\mathbf{A}), \lambda_{max}(\mathbf{A})]$. Let ϕ be the Riemann map of E, let Φ_j be the $j-$th Takenaka-Malmquist rational function, and let M_j be the j-th Faber-Dzhrbashyan rational function. The followings hold:*

$$(z\mathbf{I}-\mathbf{A})^{-1} = \frac{1}{\phi(z)\psi'(\phi(z))}\sum_{j=0}^{\infty}\overline{\Phi_j(\phi(z)^{-1})}M_j(\mathbf{A}), \quad z\notin E$$

Proof. Set $u := \mathbf{A}$ and $w := \phi(z)$. Since $\phi = \psi^{-1}$, the result follows by direct substitution. □

The same computation can be done for $(z\mathbf{I}-\mathbf{A})^{-*}$, $(z\mathbf{I}-\mathbf{H}_k)^{-1}$ and $(z\mathbf{I}-\mathbf{T}_k)^{-*}$, and as a consequence, by substituting the quantities found in Proposition 5.9 in (4.3), and in (5.10), we obtain the following expressions for the matrices $\mathbf{X}$ and $\mathbf{X}_k$.

$$\mathbf{X} = \frac{1}{2\pi i}\sum_{j=0}^{\infty}\sum_{l=0}^{\infty} a_{j,l} M_j(\mathbf{A}) c_1 c_2^T M_l(\mathbf{A}), \tag{5.15}$$

and

$$\mathbf{X}_k = \frac{1}{2\pi i}\mathbf{V}_k\sum_{j=0}^{\infty}\sum_{l=0}^{\infty} a_{j,l} M_j(\mathbf{H}_k)\mathbf{V}_k^T c_1 c_2^T \mathbf{W}_k M_l(\mathbf{T}_k)\mathbf{W}_k^T, \tag{5.16}$$

where the coefficients $a_{j,l}$ are defined as follows.

$$a_{j,l} := \int_\Gamma \frac{\Phi_j(\phi(z)^{-1})\Phi_l(\phi(-z)^{-1})}{\phi(z)\psi'(\phi(z))\phi(-z)\psi'(\phi(-z))}\,\mathrm{d}z. \tag{5.17}$$

Thanks to the uniform boundedness of M_j on the set $[\lambda_{min}(\mathbf{A}), \lambda_{max}(\mathbf{A})]$ and to [6, Theorem 2], we have that

$$\max_{j\in\mathbb{N}} \max\{\|M_j(\mathbf{A})\|, \|M_j(\mathbf{H}_k)\|, \|M_j(\mathbf{T}_k)\|\} < \infty.$$

By having the quantities $(z\mathbf{I}-\mathbf{A})^{-1}$ and $(z\mathbf{I}-\mathbf{A})^{-*}$ written in this form, we are able to use the property of the exactness of Extended Block Krylov subspace for M_j, when $j \le 2k-1$, proven in [17], and thus we obtain the following bound.

$$\|\mathbf{X}-\mathbf{X}_k\| \lesssim \sum_{j,l\in\mathbb{N},\ \max\{j,l\}\ge 2k} |a_{j,l}| \le \sum_{j=2k}^{\infty}\sum_{l=0}^{\infty}|a_{j,l}| + \sum_{j=0}^{\infty}\sum_{l=2k}^{\infty}|a_{j,l}| \tag{5.18}$$

In the following lemma we estimate the coefficients $a_{j,l}$.

Lemma 5.10. *Set $\mu := \sqrt{\max_{z\in i\mathbb{R}}|B(\phi(z)^{-1})|} < 1$. The coefficients $a_{j,l}$ defined in equation (5.17) satisfy:*

$$a_{j,l} \lesssim \mu^{j+l}$$

Proof. From (4.4), we can take the curve $\Gamma := i\mathbb{R}$. From the definition of the Takenaka-Malmquist functions, and by the properties of the Riemann mapping, we have that

$$a_{j,l} \lesssim \int_{i\mathbb{R}} \frac{|\Phi_j(\phi(z)^{-1})\Phi_l(\phi(-z)^{-1})|}{|\phi(z)\psi'(\phi(z))\phi(-z)\psi'(\phi(-z))|}\,|\mathrm{d}z| \lesssim \int_{i\mathbb{R}} \frac{|B(\phi(z)^{-1})|^{\frac{j+l}{2}}}{\max\{1,|z|^2\}}\,|\mathrm{d}z| \lesssim \mu^{j+l}$$

□

In the following we define the quantity ρ, which turns out to be the rate of convergence of the approximate solution for this space of projection.

Definition 5.11. We define the function $h(w) := B\Big(\phi\big(-\psi(w^{-1})\big)^{-1}\Big)$, and we set

$$\rho := \max_{|w|=1} |h(w)|.$$

In the following lemma we explicitly compute the quantity ρ by exploiting the assumption that the matrix $\mathbf{A}$ is a symmetric positive definite matrix.

Lemma 5.12. *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix and let ρ be the quantity introduced in Definition 5.11. It holds that*

$$\rho = \left(\frac{\sqrt[4]{\kappa}-1}{\sqrt[4]{\kappa}+1}\right)^2$$

Proof. Let the quantities $c := \frac{\lambda_{max}(\mathbf{A})+\lambda_{min}(\mathbf{A})}{2}$, and $r := \frac{\lambda_{max}(\mathbf{A})-\lambda_{min}(\mathbf{A})}{2}$. Since $\mathbf{A}$ is symmetric and positive definite, we have that the compact set given by the spectral interval of $\mathbf{A}$, $E := [\lambda_{min}(\mathbf{A}), \lambda_{max}(\mathbf{A})]$, is a segment in the real axis, and thus we have the explicit expression of the Riemann map of E, ϕ and for the inverse of the Riemann map of E, ψ, which are respectively

$$\phi(z) = \left(\frac{z-c}{r} + \sqrt{\frac{(z-c)^2-r^2}{r^2}}\right)^{-1}, \quad z \in \overline{\mathbb{C}} - E$$

and

$$\psi(w) = c + r\frac{w^2+1}{2w}, \quad w \in \overline{\mathbb{C}} - D.$$

Hence, by defining $l_1 := \phi(-\lambda_{min}(\mathbf{A}))^{-1} \in \mathbb{R}$, and $l_2 := \phi(-\lambda_{max}(\mathbf{A}))^{-1} \in \mathbb{R}$, we have that:

$$\max_{|w|=1} \Big|B\Big(\phi\big(-\psi(w)\big)^{-1}\Big)\Big| = \max_{\lambda\in[\lambda_{min}(\mathbf{A}),\lambda_{max}(\mathbf{A})]} \Big|B\Big(\phi\big(-\lambda\big)^{-1}\Big)\Big| = \max_{u\in[l_1,l_2]} |B(u)|$$

Hence, we have to look for the behaviour of the function B on a real segment. By definition of B, we have that $B(u) \in \mathbb{R}$ for $u \in [l_1, l_2]$. Moreover, we have that $B(u) > 0$ for $u \in [\phi(0)^{-1}, 0]$, and since the Riemann map ϕ is monotonically increasing in $\mathbb{R}$, it holds that $\phi(0)^{-1} < l_1$. Thus, we have that

$$\max_{u\in[l_1,l_2]} |B(u)| = \max_{u\in[l_1,l_2]} B(u).$$

To find the maximum in the segment of the function B, we search for its critical points. By deriving the function B, we have that

$$DB(u) = \frac{\phi(0) - u - 2\phi(0)^2u + 2\phi(0)u^2 + u - \phi(0)u^2}{(\phi(0)-u)^2} = \phi(0)\frac{u^2 - 2\phi(0)u + 1}{(\phi(0)-u)^2}$$

Therefore, the two critical points of B are $u_\pm := \phi(0) \pm \sqrt{\phi(0)^2 - 1}$. By a sign analysis, we find out that u_+ is a local maximum. By explicit computation, we obtain that

$$\phi(0) = -\frac{\sqrt{\lambda_{max}(\mathbf{A})} + \sqrt{\lambda_{min}(\mathbf{A})}}{\sqrt{\lambda_{max}(\mathbf{A})} - \sqrt{\lambda_{min}(\mathbf{A})}} = -\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}.$$

And

$$\begin{aligned} B(u_+) &= \frac{\phi(0) + \sqrt{\phi(0)^2-1} - \phi(0)(2\phi(0)^2 + 2\phi(0)\sqrt{\phi(0)^2-1} - 1)}{-\sqrt{\phi(0)^2-1}} \\ &= \frac{-2\phi(0)(\sqrt{\phi(0)^2-1})^2 + (-2\phi(0)^2+1)\sqrt{\phi(0)^2-1}}{-\sqrt{\phi(0)^2-1}} \\ &= 2\phi(0)^2 - 1 - 2\phi(0)\sqrt{\phi(0)^2-1} = u_+^2. \end{aligned}$$

By writing u_+ in terms of κ, we have that

$$u_+ = -\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1} + \frac{2\sqrt[4]{\kappa}}{\sqrt{\kappa}-1} = -\frac{\sqrt[4]{\kappa}-1}{\sqrt[4]{\kappa}+1}$$

By monotonicity of the Riemann map ϕ, we have that $\frac{-\lambda_{max}(\mathbf{A})-c}{r} < \phi(0) < \frac{-\lambda_{min}(\mathbf{A})-c}{r}$. Hence $u_+ \in [l_1, l_2]$, and so we conclude.

□

We are now able to state the main theorem about the rate of convergence of the approximate solution for this projection space.

Theorem 5.13. *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix and let $c_1, c_2 \in \mathbb{R}^n$. It holds that*

$$\|\mathbf{X} - \mathbf{X}_k\| \lesssim k\left(\frac{\sqrt[4]{\kappa}-1}{\sqrt[4]{\kappa}+1}\right)^{2k}$$

Proof. We prove the result by showing that $\|X - X_k\| \lesssim k\rho^k$, where ρ is the quantity defined in Definition 5.11. Then we apply Lemma 5.12 to conclude. Set $E := [\lambda_{min}(\mathbf{A}), \lambda_{max}(\mathbf{A})]$, let ψ be the inverse of the Riemann map of E. Set $\Gamma = \psi(\partial D)$, and set $z := \psi\left(\frac{1}{w}\right)$. By differentiating z, it holds that $\mathrm{d}z = -D\psi\left(\frac{1}{w}\right)\frac{1}{w^2}\,\mathrm{d}w$. By substituting z in (5.17), we have that

$$\begin{aligned} a_{j,l} &= \int_{|w|=1} \Phi_j(w) \frac{\Phi_l(\phi(-\psi(\frac{1}{w}))^{-1})}{w\phi(\psi(\frac{1}{w}))\psi'(\phi(-\psi(\frac{1}{w})))}\,\mathrm{d}w \\ &= \int_{|w|=1} \Phi_j(w) h(w)^{\left[\frac{l}{2}\right]} g_l(w)\,\mathrm{d}w \\ &= \int_{|w|=1} \Phi_j(w) h(w)^{\left[\frac{l}{2}\right]} g_l(w) i e^{iArg(w)}\,|\mathrm{d}w| \end{aligned}$$

Where:

$$g_l(w) := \begin{cases} \frac{1}{wuD\psi(u)}\big|_{u=\phi\left(-\psi\left(\frac{1}{w}\right)\right)}, & l \text{ is even}, \\ \frac{1}{wD\psi(u)}\frac{\sqrt{1-\phi(0)^{-2}}}{1-\phi(0)u^{-1}}\big|_{u=\phi\left(-\psi\left(\frac{1}{w}\right)\right)}, & l \text{ is odd}. \end{cases}$$

Moreover, we notice that g_l is a smooth function in the unit circumference, and thus it is limited. By substituting ρ, by taking the maximum in the circumference of $|h|$ in the integral, and by Parseval Theorem for an orthogonal decomposition, we have that

$$\sum_{j=0}^{\infty} a_{j,l}^2 \lesssim \rho^l.$$

And for the same reasoning,

$$\sum_{j=0}^{\infty}\sum_{l=2k}^{\infty} a_{j,l}^2 \lesssim \rho^{2k}.$$

By repeating the computations for $E := [-\lambda_{max}(\mathbf{A}), -\lambda_{min}(\mathbf{A})]$, for ψ as the inverse of the Riemann map of E and $\Gamma = \psi(\partial D)$, we obtain that

$$\sum_{j=2k}^{\infty}\sum_{l=0}^{\infty} a_{j,l}^2 \lesssim \rho^{2k}.$$

Thus we have that:

$$\sum_{j,l\in\mathbb{N},\ \max\{j,l\}\geq 2k} a_{j,l}^2 \lesssim \rho^{2k}.$$

To conclude the proof, we set the quantity $j_0 = \left[k\frac{\log\rho}{\log\mu}\right]$, where μ is the quantity found in Lemma 5.10, and we split the set $\{j,l\in\mathbb{N},\ \max\{j,l\}\geq 2k\}$ into the sets $\{j,l\in\mathbb{N},\ \max\{j,l\}\geq 2k, j+l<j_0\}$ and $\{j,l\in\mathbb{N},\ \max\{j,l\}\geq 2k, j+l\geq j_0\}$. For the first set, we have that

$$\sum_{j,l\in\mathbb{N},\ \max\{j,l\}\geq 2k, j+l<j_0} |a_{j,l}| \lesssim j_0\sqrt{\sum_{j,l\in\mathbb{N},\ \max\{j,l\}\geq 2k} |a_{j,l}|^2} \lesssim k\rho^k.$$

While for the second set, by using the geometric series we have that

$$\sum_{j,l\in\mathbb{N},\ \max\{j,l\}\geq 2k, j+l\geq j_0} |a_{j,l}| \leq j_0\mu^{j_0} \lesssim k\rho^k$$

Where we substitute j_0 in the last inequality. By summing the two pieces we find the result:

$$\sum_{j,l\in\mathbb{N},\ \max\{j,l\}\geq 2k} |a_{j,l}| \lesssim k\rho^k.$$

□

Remark. The bound is particularly useful when the condition number κ is big. Indeed we have:

$$\left(\frac{\sqrt[4]{\kappa}-1}{\sqrt[4]{\kappa}+1}\right)^2 = \left(1-\frac{2}{\sqrt[4]{\kappa}+1}\right)^2 = 1-\frac{4}{\sqrt[4]{\kappa}+1}+o\left(\frac{1}{(\sqrt[4]{\kappa}+1)}\right) \approx \left(\frac{\sqrt[4]{\kappa}-3}{\sqrt[4]{\kappa}+1}\right)$$

As a consequence, the convergence of the method mildly depends on the condition number of the matrix $\mathbf{A}$, and this allows us to have good convergence properties also for ill-conditioned problems.

Example 5.3. We consider the matrix $\mathbf{A} = \tilde{\mathbf{K}} \in \mathbb{R}^{n\times n}$ as defined in Lemma 3.14, for $n = 2^l$, and $l = 4, \ldots, 10$, where $2^{10} = 1024$. In the next table, we show the quantity $\left(\frac{\sqrt[4]{\kappa}-1}{\sqrt[4]{\kappa}+1}\right)^2$ as the dimensions of the problem grow.

n	$\kappa(\mathbf{A})$	$\left(\frac{\sqrt[4]{\kappa}-1}{\sqrt[4]{\kappa}+1}\right)^2$
2^4	1.6115e+03	5.2906e-01
2^5	6.3773e+03	6.3795e-01
2^6	2.5343e+04	7.2783e-01
2^7	1.0101e+05	7.9883e-01
2^8	4.0328e+05	8.5316e-01
2^9	1.6116e+06	8.9378e-01
2^{10}	6.4433e+06	9.2367e-01

Table 5.2: Rate of convergence from Theorem 5.13 for Example 5.3.

Example 5.4. We consider the matrix $\mathbf{A} = \tilde{\mathbf{K}} \in \mathbb{R}^{n\times n}$ as defined in Lemma 3.14 of dimension $n = 2^6 = 64$, and we take as data the normalized matrix of $\mathbf{C} = \mathbf{L}^{-1}vv^T\mathbf{L}^{-T}$, where v is the vector of all ones. The following figure shows the behaviour of the error norm compared with the rate of convergence $\left(\frac{\sqrt[4]{\kappa}-1}{\sqrt[4]{\kappa}+1}\right)^{2k}$ as k increases.

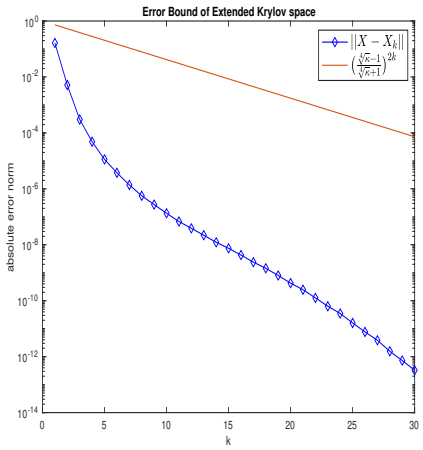

Figure 5.2: Comparison between the true error norm and the theoretical rate of convergence for Example 5.4.

The method performs better than expected by the rate. Indeed, we see that after few iterations the convergence has a faster decay than the rate. This phenomenon is called superlinear convergence to the exact solution.

Also for this space, it is useful for our application to introduce the block version of this space, when the data matrix $\mathbf{C} := \mathbf{C}_1\mathbf{C}_2^T$ has rank greater than one.

Definition 5.14 (Extended Block Krylov space)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times n}$, and let $\mathbf{C} \in \mathbb{R}^{n\times p}$ be a tall matrix. We define the Extended Block Krylov space as:

$$\mathcal{EK}_k^{\square}(\mathbf{A}, \mathbf{C}) = \mathcal{K}_k^{\square}(\mathbf{A}, \mathbf{C}) + \mathcal{K}_k^{\square}(\mathbf{A}^{-1}, \mathbf{A}^{-1}\mathbf{C}).$$

In particular, for this scenario we set the spaces of projection $\mathcal{V}_k := \mathcal{EK}_k^{\square}(\mathbf{A}, \mathbf{C}_1)$, and $\mathcal{W}_k := \mathcal{EK}_k^{\square}(\mathbf{A}, \mathbf{C}_2)$. A convergence analysis for this setting can be found on [5].

5.8 Arnoldi algorithm for the Rational Krylov space

Given $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $c \in \mathbb{R}^n$, we want to improve the previously studied Extended Krylov space. Here we introduce the so called Rational Krylov space, which is defined as follows.

Definition 5.15 (Rational Krylov space)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times n}$, $c \in \mathbb{R}^n$ and let $\sigma \in \mathbb{R}^k$. We define the k-th Rational Krylov space of $\mathbf{A}$, c and σ as

$$\mathcal{RK}_k(\mathbf{A}, c, \sigma) = span\{c, (\mathbf{A} - \sigma_1)^{-1}c, \ldots, \prod_{j=1}^{k-1}(\mathbf{A} - \sigma_j)^{-1}c\}. \tag{5.19}$$

Indeed, the matrices involved to create the Extended Krylov space are only $\mathbf{A}$ and $\mathbf{A}^{-1}$, while here we consider matrices of the form $(\mathbf{A}-\sigma_i)^{-1}$, for some shifts σ_i.

In this section, we describe the procedure to obtain a matrix with orthonormal columns $\mathbf{V}_k = [v_1, \ldots, v_k]$, whose columns form a basis of $\mathcal{RK}_k(\mathbf{A}, c, \sigma)$.

As first column we take $v_1 = \frac{c}{\|c\|}$. We then compute $p_2 := (\mathbf{A}-\sigma_1)^{-1}v_1$, and we orthonormalize w_2 with respect to v_1, by applying the Gram-Schmidt procedure. By repeating the iteration in j, we obtain the following recurrence. First, we compute $p_j := (\mathbf{A}-\sigma_j)^{-1}v_{j-1}$; then we orthonormalize p_j with respect to $\mathbf{V}_{j-1} := [v_1, \ldots, v_{j-1}]$ by applying the Gram-Schmidt procedure.

Algorithm 6: Arnoldi Iteration for the Rational Krylov space

Data: $\mathbf{A} \in \mathbb{R}^{n\times n}$, $c \in \mathbb{R}^n$, $\sigma \in \mathbb{R}^k$
Result: $\mathbf{V}_k \in \mathbb{R}^{n\times k}$ whose columns are orthonormal and form a basis for $\mathcal{RK}_k(\mathbf{A}, c, \sigma)$
1 Set $v_1 = \frac{c}{\|c\|}$, $\mathbf{V}_0 = \emptyset$;
For $j = 2, \ldots, k$
2 Set $\mathbf{V}_{j-1} = [\mathbf{V}_{j-2}, v_{j-1}]$;
3 Compute $p_j = (\mathbf{A}-\sigma_j)^{-1}v_{j-1}$;
4 Obtain v_j from p_j with the modified Gram-Schmidt algorithm.
End For
5 Set $\mathbf{V}_k = [\mathbf{V}_{k-1}, v_k]$.

Also Algorithm 6 can be generalized in case of tall matrices $\mathbf{C} \in \mathbb{R}^{n\times p}$, instead of vectors $c \in \mathbb{R}^n$.

5.9 Convergence Analysis of the Galerkin method with Rational Krylov space

In this section we analyze the convergence of Algorithm 3 when the approximation space is the Rational Krylov subspace; which we recall, for a given matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, for $c \in \mathbb{R}^n$ and $\sigma \in \mathbb{R}^k$.

$$\mathcal{RK}_k(\mathbf{A}, c, \sigma) = span\{c, (\mathbf{A}-\sigma_1)^{-1}c, \ldots, \prod_{j=1}^{k-1}(\mathbf{A}-\sigma_j)^{-1}c\}.$$

In our notation, we consider $\mathbf{C} := c_1 c_2^T$ we set the spaces of projection $\mathcal{V}_k := \mathcal{RK}_k(\mathbf{A}, c_1)$, and $\mathcal{W}_k := \mathcal{RK}_k(\mathbf{A}, c_2)$; moreover we set $\mathbf{H}_k := \mathbf{V}_k^T\mathbf{A}\mathbf{V}_k$, $\mathbf{T}_k := \mathbf{W}_k^T\mathbf{A}\mathbf{W}_k$ and $\mathbf{X}_k := \mathbf{V}_k\mathbf{Y}\mathbf{W}_k^T$, where $\mathbf{V}_k$ and $\mathbf{W}_k$ are matrices with orthonormal columns, whose columns form a basis for $\mathcal{V}_k$ and $\mathcal{W}_k$ and are obtained by the Arnoldi algorithm. We follow [7] for the analysis of the convergence rate of the approximate solution, where an optimal choice of the parameters σ_j, with respect

to the equilibrium measure for the set $E := \phi([-\lambda_{max}(\mathbf{A}), -\lambda_{min}(\mathbf{A})])^{-1}$, is considered. In particular, we ask the parameters to be uniformly distributed in the equilibrium measure.

Definition 5.16. Let $\psi : \mathbb{C} - D \to \mathbb{C} - E$ be the inverse of the Riemann map of a set E. We define the equilibrium measure of E as the push-forward under ψ of the uniform measure $\frac{\mathrm{d}\theta}{2\pi}$ on ∂D.

The techniques used in this section are similar to the ones used for the convergence rate result for the Extended Block Krylov space; indeed, the role of the parameters σ_j in the latter is to have an explicit and easy to treat form of the Blaschke product B, while the preliminary results are stated for both Takenaka-Malmquist and Faber-Dzhrbashyan rational function defined for general parameters. Moreover, we derive asymptotical error bounds so we at once account for the whole sequence of shifts σ_j.

With the previous definitions, we recall that:

$$\|\mathbf{X} - \mathbf{X}_k\| \lesssim \sum_{j,l\in\mathbb{N},\ \max\{j,l\}\geq k} |a_{j,l}|,$$

where

$$a_{j,l} := \int_\Gamma \frac{\Phi_j(\phi(z)^{-1})\Phi_l(\phi(-z)^{-1})}{\phi(z)\psi'(\phi(z))\phi(-z)\psi'(\phi(-z))}\ \mathrm{d}z.$$

The goal of this frame is to estimate before the quantities $a_{i,j}$ with the norm of the Blaschke product B_k for an admissible choice of parameters, and after estimate the norm B_k for an optimal choice of parameters. For this reason, we introduce the following lemma, which estimates the values of Blaschke products near the set E. A proof of the statement can be found in [7].

Lemma 5.17. *Let $C(E)$ be the set of continuous functions of the compact set E. Uniformly in k and w, it holds that:*

$$|B_k(w)| \leq \|B_k\|_{C(E)}\,[1 + o(1)]^k,\ \textit{as } dist(w, \partial E) \to 0$$

As a consequence of this technical lemma, we give a first result for the convergence of the error.

Proposition 5.18. *Let the parameters σ_l be uniformly separated from the set $[\lambda_{min}(\mathbf{A}), \lambda_{max}(\mathbf{A})]$. The following estimate holds.*

$$\limsup_{k\to\infty} \|\mathbf{X} - \mathbf{X}_k\|^{\frac{1}{k}} \lesssim \limsup_{k\to\infty} \|B_k\|_{C(E)}^{\frac{1}{k}} < 1 \tag{5.20}$$

Proof. The right inequality follows from the uniform separation of the parameters and from property *ii.* of Proposition 1.27. Moreover, by property *ii.* the bound can be obtained for every $B^s := B_1(\cdot; \sigma_s)$, and thus we have that

$$q := \sup_{s\in\mathbb{N}} \|B^s\|_{C(E)} < 1.$$

As a consequence, it holds that

$$\begin{aligned}\|B_l\|_{C(E)} &\leq q^l, \quad l \in \mathbb{N},\\ \|B_{l_1+l_2}\|_{C(E)} &\leq q^{l_1} \|B_{l_2}\|_{C(E)}, \quad l_1, l_2 \in \mathbb{N}.\end{aligned}$$

For the left inequality, we want to use Lemma 5.17. For this purpose, we define $\rho := 1 + \epsilon$, for $\epsilon \in (0, 1)$, and we set $\Gamma := \Gamma_\rho$, as defined in Definition 1.28. We now want to estimate the coefficients $a_{j,l}$, as done in Lemma 5.10:

$$\begin{aligned}|a_{j,l}| &\lesssim \int_{\Gamma_\rho} \frac{|\Phi_j(\phi(z)^{-1})\Phi_l(\phi(-z)^{-1})|}{|\phi(z)\psi'(\phi(z))\phi(-z)\psi'(\phi(-z))|} \, |\mathrm{d}z|\\ &\lesssim \int_{\Gamma_\rho} |B_j(\phi(z)^{-1})B_l(\phi(-z)^{-1})| \, |\mathrm{d}z|\\ &\lesssim \|B_j\|_{C(\phi(\Gamma_\rho)^{-1})} \|B_l\|_{C(\phi(-\Gamma_\rho)^{-1})}\end{aligned}$$

By using the quantity q, we obtain

$$\sum_{j=0}^{\infty}\sum_{l=k}^{\infty} |a_{j,l}| \lesssim \sum_{j=0}^{\infty}\sum_{l=k}^{\infty} q^j q^{l-k} \|B_l\|_{C(\phi(-\Gamma_\rho)^{-1})} \lesssim \|B_l\|_{C(\phi(-\Gamma_\rho)^{-1})}$$

And, by taking $\Gamma := -\Gamma_\rho$, we also obtain

$$\sum_{j=k}^{\infty}\sum_{l=0}^{\infty} |a_{j,l}| \lesssim \sum_{j=k}^{\infty}\sum_{l=0}^{\infty} q^{j-k} q^l \|B_l\|_{C(\phi(-\Gamma_\rho)^{-1})} \lesssim \|B_l\|_{C(\phi(-\Gamma_\rho)^{-1})}$$

As a consequence, we have that

$$\sum_{j,l\in\mathbb{N},\ \max\{j,l\}\geq k} |a_{j,l}| \lesssim \|B_l\|_{C(\phi(-\Gamma_\rho)^{-1})}.$$

By taking $\epsilon \to 0$, we have that $dist(u, \partial E) \to 0$ for $u \in \phi(-\Gamma_\rho)^{-1}$, so that we satisfy the hypotheses of Lemma 5.17. Hence,

$$\|X - X_k\| \lesssim \sum_{j,l\in\mathbb{N},\ \max\{j,l\}\geq k} |a_{j,l}| \lesssim \|B_k\|_{C(\phi(-\Gamma_\rho)^{-1})} \lesssim \|B_k\|_{C(E)} (1 + o(1))^k.$$

By extracting the k-th root, and taking the lim sup, we conclude the proof. □

To conclude the estimate, it remains to study the quantity on the right-hand side, for an optimal choice of parameters. For this reason, we recall the concepts of condenser (F_1, F_2) and of Riemann modulus of a condenser (F_1, F_2) introduced in Definition 1.30 and in Definition 1.33.

A plane condenser consists in a pair of disjoint closed subset of $\mathbb{C}$, F_1 and F_2, each of which has a connected complement. It is denoted as (F_1, F_2).

The Riemann modulus of the condenser (F_1, F_2) is defined as:

$$H(F_1, F_2) := c^{-1}(F_1, F_2), \qquad c(F_1, F_2) := \frac{1}{2\pi} \int_\gamma \frac{\partial h}{\partial n} \, \mathrm{d}s.$$

Here $G := \overline{\mathbb{C}} - (F_1 \cup F_2)$, and h is the Harmonic measure of the set ∂F_2 relative to the region G, introduced in Definition 1.31, while γ is a curve that divides $\mathbb{C}$ into two open sets, one of which contains the set F_1, and the other of which contains the set F_2 .

In our setting, we take $F_1 := E$, and $F_2 := F = \phi([-\lambda_{max}(\mathbf{A}), -\lambda_{min}(\mathbf{A})])$. The reason for such F_2 belongs to the structure of the Blaschke product B and the property $i.$ of Proposition 1.27. Indeed, we have that:

$$\|B_k\|_{C(E)} = \left\|B_k^{-1}\right\|_{C(F)} \tag{5.21}$$

In the following statement we give a first bound of the quantity in the right of the inequality of the Theorem 5.20.

Lemma 5.19. *Let $H(E, F)$ be the Riemann modulus of the condenser (E, F). Let the parameters σ_j be chosen so that the sequence $\overline{\phi(\sigma_j)}^{-1} \in \partial E$ is uniformly distributed with respect to the equilibrium measure of E. Then the following error estimate holds.*

$$\limsup_{k\to\infty} \|\mathbf{X} - \mathbf{X}_k\|^{\frac{1}{k}} \lesssim H(E, F)^{-\frac{1}{2}}$$

Proof. From Theorem 5.20, we are left to show that $\limsup_{k\to\infty} \|B_k\|_{C(E)}^{\frac{1}{k}} \lesssim H(E, F)^{-\frac{1}{2}}$. Since $\overline{\phi(\sigma_j)} \in \partial F$, it is uniformly distributed with respect to the equilibrium measure of F, work [30] gives us

$$\begin{aligned} H(E, F)^{-1} &= \limsup_{k\to\infty} \|B_k\|_{C(E)}^{\frac{1}{k}} \left\|B_k^{-1}\right\|_{C(F)}^{\frac{1}{k}} \\ (5.21) &= \limsup_{k\to\infty} \|B_k\|_{C(E)}^{\frac{1}{k}} \|B_k\|_{C(E)}^{\frac{1}{k}} = \limsup_{k\to\infty} \|B_k\|_{C(E)}^{\frac{2}{k}} \end{aligned}$$

□

We now want to give a second bound of the quantity in the right of the inequality of the Theorem 5.20, by exploiting the assumption that the matrix $\mathbf{A}$ is symmetric and positive definite. Without loss of generality, by taking as matrix $\mathbf{A}$ the matrix $\frac{1}{\lambda_{max}(\mathbf{A})}\mathbf{A}$, we consider $[\lambda_{min}(\mathbf{A}), \lambda_{max}(\mathbf{A})] \subset [0, 1]$. The following result gives us the rate of convergence of the method, for an optimal choice of parameters.

Theorem 5.20. *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix, with $[\lambda_{min}(\mathbf{A}), \lambda_{max}(\mathbf{A})] = [\delta, 1]$. Let the parameters σ_j be chosen so that the sequence $\overline{\phi(\sigma_j)}^{-1} \in \partial E$ is uniformly distributed with respect to the equilibrium measure of E, and let $c_1, c_2 \in \mathbb{R}^n$. It holds that*

$$\limsup_{k\to\infty} \|\mathbf{X} - \mathbf{X}_k\|^{\frac{1}{k}} \lesssim \exp\left(\frac{-\pi K(g)}{2K'(g)}\right)$$

where K and K' are respectively the principal and the complementary elliptic integral of first kind of modulus g, for the following quantity g:

$$g = \sqrt{\frac{(d-a)(c-b)}{(d-b)(c-a)}}$$

where:

$$a = -\frac{3+\delta}{1-\delta} - \sqrt{\left(\frac{3+\delta}{1-\delta}\right)^2 - 1}, \quad d = \frac{1}{a},$$
$$b = -\frac{1+3\delta}{1-\delta} - \sqrt{\left(\frac{1+3\delta}{1-\delta}\right)^2 - 1}, \quad c = \frac{1}{b}.$$

Proof. We recall the explicit expression for the Riemann map for the set $E := [\lambda_{min}(\mathbf{A}), \lambda_{max}(\mathbf{A})]$, which is

$$\phi(z) = \left(\frac{z-c}{r} + \sqrt{\frac{(z-c)^2 - r^2}{r^2}}\right)^{-1}, \quad z \in \overline{\mathbb{C}} - (\delta, 1),$$

where $c := \frac{1+\delta}{2}$, and $r := \frac{1-\delta}{2}$. By direct substitution, we obtain that $F = [a, b]$ and $E = [c, d]$. From Lemma 5.19 and [13], the result follows. □

Example 5.5. We consider the matrix $\mathbf{A} = \tilde{\mathbf{K}} \in \mathbb{R}^{n\times n}$, as defined in Lemma 3.14, for $n = 2^l$, and $l = 4, \ldots, 10$, where $2^{10} = 1024$. In the next table, we show the quantity $\exp\left(\frac{-\pi K(g)}{2K'(g)}\right)$ as the dimensions of the problem grow.

n	$\kappa(\mathbf{A})$	$\exp\left(\frac{-\pi K(g)}{2K'(g)}\right)$
2^4	1.6115e+03	2.4957e-01
2^5	6.3773e+03	2.8721e-01
2^6	2.5343e+04	3.2078e-01
2^7	1.0101e+05	3.5103e-01
2^8	4.0328e+05	3.7852e-01
2^9	1.6116e+06	4.0367e-01
2^{10}	6.4433e+06	4.2682e-01

Table 5.3: Rate of convergence from Theorem 5.20 for Example 5.5.

Example 5.6. We consider the matrix $\mathbf{A} = \tilde{\mathbf{K}} \in \mathbb{R}^{n\times n}$, as defined in Lemma 3.14, of dimension $n = 2^6 = 64$, and we take as right-hand side the normalized matrix of $\mathbf{C} = \mathbf{L}^{-1}vv^T\mathbf{L}^{-T}$, where v is the vector of all ones. The following figure shows the behaviour of the error norm compared with the rate of convergence $\exp\left(\frac{-\pi K(g)}{2K'(g)}\right)^k$ as k increases.

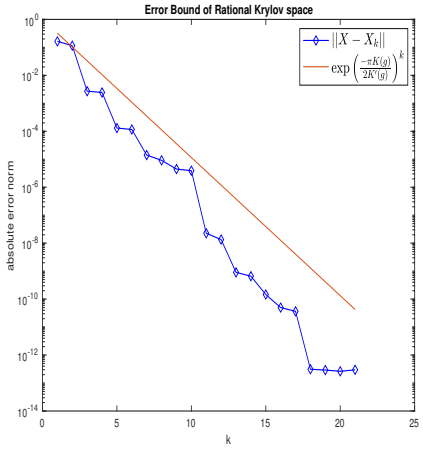

Figure 5.3: Comparison between the true error norm and the theoretical rate of convergence, related to Example 5.6.

The theoretical rate has a good behaviour also when the problem is ill-conditioned. As a consequence, when a preconditioning technique is not possible or a good preconditioner is not available, the Rational Krylov space is a possible substitute; since few more iterations are necessary when the dimensions of the problem grow.

We emphasize that also for this space, it is useful to introduce its block version, when the data matrix $\mathbf{C} := \mathbf{C}_1\mathbf{C}_2^T$ has rank greater than one.

Definition 5.21 (Rational Block Krylov space)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times n}$, and let $\mathbf{C} \in \mathbb{R}^{n\times p}$ be a tall matrix. We define the Rational Krylov space as:

$$\mathcal{RK}_k^{\square}(\mathbf{A}, \mathbf{C}, \sigma) = blkspan\{\mathbf{C}, (\mathbf{A} - \sigma_1)^{-1}\mathbf{C}, \ldots, \prod_{j=1}^{k-1}(\mathbf{A} - \sigma_j)^{-1}\mathbf{C}\}$$

In this scenario, we set the spaces of projection $\mathcal{V}_k := \mathcal{RK}_k^{\square}(\mathbf{A}, \mathbf{C}_1)$, and $\mathcal{W}_k := \mathcal{RK}_k^{\square}(\mathbf{A}, \mathbf{C}_2)$. For this block version, we refer to [5] for a convergence analysis.

Chapter 6

Analysis for the energy norm of the error

In this chapter, we prove and discuss an extension of [24, property (2.7)], which is the following. Let u_h be the function satisfying Galerkin condition (3.2), let $\tilde{x}_k \in \mathbb{R}^{n^2}$ be the approximate solution of the linear system (5.6) formed in the $k-$th iteration of the Conjugate Gradients method, and let $\tilde{\mathbf{X}}_k \in \mathbb{R}^{n\times n}$ be the square matrix obtained from x_k by inverting the operator vec in dimensions $n \times n$, and let $\mathbf{X}_k := \mathbf{L}^{-T}\tilde{\mathbf{X}}_k\mathbf{L}^{-1}$, where $\mathbf{L}$ is the Cholesky factor of the matrix $\mathbf{M}$ of equation (3.4). The following holds.

$$\left\|u^s - u_h^{(k)}\right\|_a^2 = \|u^s - u_h\|_a^2 + \|\mathbf{X} - \mathbf{X}_k\|_{\mathcal{A}}^2 \tag{6.1}$$

The aim of this extension is to fill the gap between the convergence of the finite-dimensional solution u_h to the solution u^s, which is known for every dimension of the domain, and the convergence of the numerical solution u_h^k, which is formed in the $k-$th iteration of a projection method, to the finite-dimensional solution u_h. Typically, this type of convergence is studied for the solution of the linear system obtained by using the Conjugate Gradients methods on the linear sistem associated with the Kronecker formulation of the matrix equation.

Theorem 6.1. *Let the function u_h be the function satisfying Galerkin condition (3.2), and let the matrix $(\mathbf{X})_{j,l} := \alpha_{j,l}$, $\mathbf{X} \in \mathbb{R}^{n\times n}$ be the solution of the matrix equation (3.4). Let $\mathbf{L}$ be the Cholesky factor of the matrix $\mathbf{M}$ of equation (3.4). Let $\tilde{\mathbf{X}}_k$ be the approximated solution formed in the $k-$th iteration of (3.5), by using a projection method, and set $(\mathbf{X}_k)_{j,l} := (\mathbf{L}^{-T}\tilde{\mathbf{X}}_k\mathbf{L}^{-1})_{j,l} = \alpha_{j,l}^{(k)}$. Set also $u_h^{(k)} := \sum_{j=1}^n \sum_{l=1}^n \alpha_{j,l}^{(k)} \varphi_j \otimes \psi_l$,. The following equality holds.*

$$\left\|u^s - u_h^{(k)}\right\|_a^2 = \|u^s - u_h\|_a^2 + \|\mathbf{X} - \mathbf{X}_k\|_{\mathcal{A}}^2 \,. \tag{6.2}$$

Proof. Firstly, we observe that $u_h - u_h^k \in S_h$, since the coefficients $\alpha_{j,k}^{(k)}$ are projected on a linear subspace of $\mathbb{R}^n$. Since u_h is the Galerkin solution, from (3.2) we have that $a(u^s - u_h, u_h - u_h^k) = 0$. Thus:

$$\begin{aligned} a(u^s - u_h^k, u^s - u_h^k) &= a(u^s - u_h + u_h - u_h^k, u^s - u_h + u_h - u_h^k) \\ &= a(u^s - u_h, u^s - u_h) + 2a(u^s - u_h, u_h - u_h^k) + a(u_h - u_h^k, u_h - u_h^k) \\ &= a(u^s - u_h, u^s - u_h) + a(u_h - u_h^k, u_h - u_h^k). \end{aligned}$$

By applying Fubini's Theorem we have that,

$$\begin{aligned} & a(u_h - u_h^k, u_h - u_h^k) \\ =& \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{m=1}^{n}\sum_{l=1}^{n}(\alpha_{i,j} - \alpha_{i,j}^r)(\alpha_{m,l} - \alpha_{m,l}^k)\cdot \\ & \cdot \left(\langle \phi_i', \phi_m' \rangle_{L^2([0,1])} \langle \psi_j, \psi_l \rangle_{L^2([0,1])} + \langle \phi_i, \phi_m \rangle_{L^2([0,1])} \langle \psi_j', \psi_l' \rangle_{L^2([0,1])}\right) \\ =& \|\mathbf{X} - \mathbf{X}_k\|_{\mathcal{A}}^2 \end{aligned}$$

□

We highlight some important implications of Theorem 6.1. In first place, we notice that the terms of the result have different nature. In the left-hand side, we have the total error, while in the right-hand side we have separated the analytical error, which only depends on the chosen space S_h, from the numerical error, which only depends on the method used to solve the matrix equation. Moreover, we observe that the result is not a bound, but an equality. This fact implies that we have the exact control of the energy norm of the total error, in terms of the discretization error and the algebraic ones. As a consequence, it is useless to solve well the matrix equation if we choose a space S_h that cannot approximate well the real solution, and as a counterpart it is useless to take a well-refined space S_h if we are not able to solve well the matrix equation.

Lastly, we derive a minimization property of the numerical solution as a consequence of this result.

Corollary 6.2. *Let the function u_h be the function which satisfy Galerkin condition (3.2), and let the matrix $(\mathbf{X})_{j,l} := \alpha_{j,l}$, $\mathbf{X} \in \mathbb{R}^{n\times n}$ be the solution of the matrix equation (3.4). Let $\mathbf{L}$ be the Cholesky factor of the matrix $\mathbf{M}$ of equation (3.4). Let $\tilde{\mathbf{X}}_k$ be the approximated solution formed in the $k-$th iteration of (3.5), by using a projection method, and set $(\mathbf{X}_k)_{j,l} := (\mathbf{L}^{-T}\tilde{\mathbf{X}}_k\mathbf{L}^{-1})_{j,l} = \alpha_{j,l}^{(k)}$. Set also $u_h^{(k)} := \sum_{j=1}^n \sum_{l=1}^n \alpha_{j,l}^{(k)} \varphi_j \otimes \psi_l$,. The following minimization property holds.*

$$\left\|u^s - u_h^k\right\|_a^2 = \min_{v\in S_h} \|u^s - v\|_a^2 + \min_{\substack{\mathbf{S}\in\mathbb{R}^{m\times m} \\ \mathbf{Z}:=\mathbf{L}^{-T}\mathbf{V}_k\mathbf{S}\mathbf{W}_k\mathbf{L}^{-1}}} \|\mathbf{X} - \mathbf{Z}\|_{\mathcal{A}}^2 .$$

Proof. From Theorem 6.1, we derive the error equality. From Lemma 3.5, we derive the first minimum. The second follows from the same computations, since we are imposing a Galerkin condition over the space $\mathcal{V}_k \otimes \mathcal{W}_k$. □

This result emphasizes the quality and the optimality of both approaches, as we minimize both the discretization and the algebraic error, with respect to the same norm.

6.1 Validation of the result for the Conjugate Gradients method

In the following we illustrate Property 6.1 by taking a general example. In particular, we show the behaviours of the two terms involved in the total error, and we measure the mismatch of the equality, as $\delta_{n,k} := \left\|u - u_h^k\right\|_a^2 - (\|u - u_h\|_a^2 + \|\mathbf{X} - \mathbf{X}_k\|_{\mathcal{A}}^2)$.

Example 6.1. We want to find $u \in C^2([0,1]^2)$ such that:

$$\begin{cases} -\Delta u = -2[(x^2 - 2x) + (y^2 - 2y)] & \text{in } (0,1)^2, \\ u = 0 & \text{in } \partial_D[0,1]^2, \\ \frac{\partial u}{\partial \nu} = 0 & \text{in } \partial_N[0,1]^2. \end{cases}$$

The analytical expression of the solution u is given by $u(x,y) = (x^2-2x)(y^2-2y)$.

In the next tables we show the numerical illustration of property 6.1, and the behaviour of the different components of the errors in terms of the iteration k, and the dimension of the space S_h, n^2. In Table 6.1, we fix the dimension of the space S_h, and we look for the history of the errors, as k increases.

k	$\|u - u_h^k\|_a^2$	$\|u - u_h\|_a^2$	$\|\mathbf{X} - \mathbf{X}_k\|_{\mathcal{A}}^2$	$\delta_{n,k}$
5	1.0511e+00	2.1702e-05	1.0511e+00	6.8012e-13
10	7.7364e-01	2.1702e-05	7.7362e-01	5.0671e-13
15	5.8822e-01	2.1702e-05	5.8820e-01	3.7281e-13
20	4.5610e-01	2.1702e-05	4.5608e-01	-1.1380e-13
⋮	⋮	⋮	⋮	⋮
110	7.5352e-04	2.1702e-05	7.3181e-04	-1.3737e-13
115	3.0781e-04	2.1702e-05	2.8611e-04	-1.9085e-13
120	1.0227e-04	2.1702e-05	8.0570e-05	-1.5303e-13
125	3.9651e-05	2.1702e-05	1.7950e-05	-3.7528e-13
⋮	⋮	⋮	⋮	⋮
495	2.1702e-05	2.1702e-05	1.8113e-25	-2.5979e-13

Table 6.1: Term in (6.1) and mismatch of the equality for Example 6.1.

We remark that for Example 6.1, after $k \approx 125$, the most relevant term of the total error is the discretization one. Hence, we can fix the maximum iteration $k \approx 125$ for the computation of the approximate solution.

In Table 6.2 we show a good matching between the chosen dimension of the space S_h and the relative iteration of the Conjugate Gradients method, k.

n^2	k	$\|u-u_h^k\|_a^2$	$\|u-u_h\|_a^2$	$\|\mathbf{X}-\mathbf{X}_k\|_{\mathcal{A}}^2$	$\delta_{n,k}$
2^4	2^2	4.6549e-02	2.2326e-02	2.4224e-02	3.9552e-16
2^6	2^3	7.9905e-03	5.5620e-03	2.4285e-03	-5.7766e-15
2^8	2^4	1.8384e-03	1.3893e-03	4.4914e-04	-1.1191e-14
2^{10}	2^5	4.4272e-04	3.4725e-04	9.5473e-05	-2.0125e-14
2^{12}	2^6	1.1507e-04	8.6808e-05	2.8266e-05	-4.2560e-14
2^{14}	2^7	3.3890e-05	2.1702e-05	1.2188e-05	1.3039e-13
2^{16}	2^8	1.2791e-05	5.4258e-06	7.3655e-06	4.2277e-14

Table 6.2: Term in (6.1) and mismatch of the equality for Example 6.1, for different values of n and k.

6.2 Validation of the result for projection methods

In the following we illustrate Theorem 6.1 by taking a general example. In particular, we show the behaviours of the two terms involved in the total error, and we measure the mismatch of the equality, as $\delta_{n,k} := \left\|u-u_h^k\right\|_a^2 - (\|u-u_h\|_a^2 + \|\mathbf{X}-\mathbf{X}_k\|_{\mathcal{A}}^2)$. In particular, we solve the following example via projection methods, by testing different spaces of projection.

Example 6.2. We want to find $u \in C^2([0,1]^2)$ such that:

$$\begin{cases} -\Delta u = -[2\pi^2\cos(2\pi x)\sin(\pi y)^2 + 2\pi^2\sin(\pi x)^2\cos(2\pi y)] & \text{in } (0,1)^2, \\ u = 0 & \text{in } \partial_D[0,1]^2, \\ \frac{\partial u}{\partial \nu} = 0 & \text{in } \partial_N[0,1]^2. \end{cases}$$

The analytical expression of the solution u is given by $u(x,y) = \sin(\pi x)^2\sin(\pi y)^2$. Moreover, we numerically see the data matrix for this example has rank 2.

In Table 6.3 we show the history of the errors, for the Block Krylov space, in terms of n^2, the dimension of the space S_h.

n^2	k	$\|u-u_h^k\|_a^2$	$\|u-u_h\|_a^2$	$\|\mathbf{X}-\mathbf{X}_k\|_{\mathcal{A}}^2$	$\delta_{n,k}$
2^6	3	1.8870e-01	1.8867e-01	2.3632e-05	-1.9984e-15
2^8	3	4.7463e-02	4.7462e-02	1.4326e-06	4.6352e-15
2^{10}	3	1.1885e-02	1.1885e-02	5.2213e-07	1.0644e-14
2^{12}	3	2.9725e-03	2.9724e-03	6.7752e-08	9.5909e-14
2^{14}	3	7.4329e-04	7.4328e-04	5.9683e-09	-9.0883e-14

Table 6.3: Term in 6.1 and mismatch of the equality for Example 6.2, for different n and k.

In Table 6.4 we show the history of the errors, for the Extended Block Krylov space, in terms of the dimension of the space S_h, n^2.

n^2	k	$\|u-u_h^r\|_a^2$	$\|u-u_h\|_a^2$	$\|\mathbf{X}-\mathbf{X}_k\|_{\mathcal{A}}^2$	$\delta_{n,k}$
2^6	3	1.8867e-01	1.8867e-01	1.5083e-06	-9.4369e-16
2^8	3	4.7462e-02	4.7462e-02	4.7613e-11	3.2058e-15
2^{10}	3	1.1885e-02	1.1885e-02	5.1414e-10	7.4107e-15
2^{12}	3	2.9724e-03	2.9724e-03	2.4011e-11	1.6229e-13
2^{14}	3	7.4328e-04	7.4328e-04	2.5645e-12	-5.7584e-13

Table 6.4: Term in 6.1 and mismatch of the equality for Example 6.1, for different n and k.

In the next table we show the history of the errors, for the Rational Block Krylov space, in terms of the dimension of the space S_h, n^2.

n^2	k	$\|u-u_h^r\|_a^2$	$\|u-u_h\|_a^2$	$\|\mathbf{X}-\mathbf{X}_k\|_{\mathcal{A}}^2$	$\delta_{n,k}$
2^6	3	1.8868e-01	1.8867e-01	2.0028e-06	-1.9151e-15
2^8	3	4.7462e-02	4.7462e-02	5.1961e-09	-8.0699e-15
2^{10}	3	1.1885e-02	1.1885e-02	1.4452e-08	-5.1764e-15
2^{12}	3	2.9724e-03	2.9724e-03	1.8183e-09	1.7350e-13
2^{14}	3	7.4328e-04	7.4328e-04	1.6668e-10	-6.5870e-13

Table 6.5: Term in 6.1 and mismatch of the equality for Example 6.1, for different n and k.

For each space, with a few number of iterations, we obtain an algebraic error which is under the discretization one. Hence, it is sufficient to solve a matrix equation of small dimension.

Conclusions and perspectives

In this work we studied a finite element discretization of a two dimensional Poisson problem, for a fixed space of elements, and we derived a matrix equation, in the form of a Lyapunov equation, related to the discretization. Then, we introduced several properties of the Lyapunov equation, which help us have a deeper knowledge of the structure of the equation. Then, we discussed different methods to numerically solve the Lyapunov equation, in the small and the large scale scenarios. For the large scale case, we focused on projection methods, which give us a good intepretation of the numerical solution, and we analyzed the rate of convergence of the method by taking different spaces of projection. Besides, we generalized to the matrix case an elegant result stated in [24] about the norm of the total error, which is due to the finite element discretization and the projection method used to obtain a numerical solution. We also explained the importance of this result, and we gave a second perspective of this, in term of two minimization problems. We analyzed only the Poisson problem, while it is possible to extend the study to different equations. Besides, in the work we only treat two dimensional case of the equation, while it could be possible to extend the study for higher dimensional cases.

Ringraziamenti

Visto che i ricordi tendono sempre a svanire, a diventare opachi o a perdersi nel tempo, vorrei dedicare un piccolo spazio di questo lavoro per rendere indelebili le persone a cui ho lasciato parte del mio cuore durante questi cinque anni.

I primi sentiti ringraziamenti vanno alla Professoressa Valeria Simoncini, che mi ha sempre lasciato molte più opportunità di quante fossi in grado di cogliere. La ringrazio per avermi sempre cercato di trasmettere l'importanza di fare ricerca e per essermi sempre venuta incontro, capendo le situazioni e non trascurando il lato umano della vita. Le devo molto.

Vorrei poi ringraziare mia mamma Claudia, mio padre Angelo ed i miei nonni Libera e Pietro. Mi hanno sempre sostenuto da lontano nel mio percorso e nella mia crescita, hanno fatto di tutto per rimanere coinvolti in ciò che facevo, e mi hanno accolto sempre con gioia quando ritornavo a casa. Terrò nel cuore tutte le partite a Scala, i viaggi a Bologna, i messaggi, le torte per colazione ed i sughi di nonna. Un altro ringraziamento va a mia sorella Cinzia ed al suo compagno Andrea, per le tante chiacchierate durante l'ora di sonno dei miei nipotini. Mi hanno dato sempre tanti consigli, e tanti vecchi aneddoti dei loro anni da studenti che mi hanno fatto vedere la città con tante sfumature diverse.

Inoltre vorrei ringraziare il mio coinquilino Marco, che conosco ormai da cinque anni. Per me è diventato un Fratello, e grazie a lui ho ritrovato in Bologna una seconda Casa. Ricorderò tutti i momenti assieme, i giri per la città, le camminate sui colli, le serie tv, i momenti tristi passati assieme, ed il costante sottofondo musicale che permeava la casa. Un ringraziamento va anche a tante persone meravigliose che ho conosciuto grazie a lui, come Alice, che ha salvato una mia cara piantina, Camilla, ed Alessandra.

Devo tanto anche a tutti coloro che hanno dato un volto a Bologna, che l'hanno resa così piena e bella. Scordardomi già ora di tanti di loro, vorrei ringraziare Gaia, Domenico, Marco M., Tommaso, Giovanni, Matteo, Eugenio, Saverio, Sara R., Leonardo, Mattia, Marco P., Giacomo, Federico, Alberto, Alessandro, Giulia, Simone, Sara H., Camilla, Francesca, Luca, Evanne, Louise ed Anthony.

Un grazie anche a tutti i miei vecchi amici di Castelnovo con cui ho mantenuto tanti bei rapporti, Lorenzo, Matteo, Federico, Giulia, Alessia, Francesca, Marco e Luca.

Ringrazio anche tutti i posti che mi hanno fatto sentire in pace con me stesso e in cui sono riusciuto ad esprimere una bella parte di me, come il Black Fire Pub, l'Anarchico ed il Guernelli.

Vorrei inoltre dedicare un grande ringraziamento ad una persona speciale, Luna. In questi quattro anni siamo cresciuti insieme come persone e come coppia, ed insieme facciamo sempre emergere il migliore lato di noi. La ringrazio per tutti i momenti, i viaggi, le esperienze gioiose, e per avere sempre uno sguardo profondo e pieno di entusiasmo sulle cose, in modo da scoprire sempre insieme nuovi lati del mondo; ha reso ogni momento di questi anni unico e stimolante. La ringrazio con tutto il cuore per essermi sempre stata vicina, in ogni momento, anche quando il dolore era davvero tanto. Per me è una presenza fondamentale. Un grazie va anche alla sua famiglia, per avermi sempre accolto con affetto.

Infine vorrei dedicare una memoria a Tommaso, per ciò che è stato.

Bibliography

[1] R. A. Adams and J. J. F. Fournier. *Sobolev spaces*, volume 140. Elsevier/Academic Press, Amsterdam, second edition, 2003.

[2] K. E. Atkinson and W. Han. *Theoretical Numerical Analysis: A Functional Analysis Framework*. Springer, 2001.

[3] R. H. Bartels and G. W. Stewart. Solution of the matrix equation ax + xb = c. *Commun. ACM*, 15(9):820–826, sep 1972.

[4] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*, volume 15 of *Texts in Applied Mathematics*. Springer, 2008.

[5] A. A. Casulli. Block rational Krylov methods for matrix equations and matrix functions. *PhD thesis, Dipartimento di Matematica, Scuola Normale Superiore, Pisa*, 2024.

[6] M. Crouzeix. Numerical range and functional calculus in Hilbert space. *Journal of Functional Analysis*, 244:668–690, 03 2007.

[7] V. Druskin, L. Knizhnerman, and V. Simoncini. Analysis of the rational Krylov subspace and adi methods for solving the Lyapunov equation. *SIAM J. Numerical Analysis*, 49:1875–1898, 01 2011.

[8] V. L. Druskin and L. A. Knizhnerman. Two polynomial methods of calculating functions of symmetric matrices. *USSR Computational Mathematics and Mathematical Physics*, 29(6):112–121, 1989.

[9] G. Faber. Über Tschebyscheffsche polynome. *Journal für die reine und angewandte Mathematik*, 150:79–106, 1920.

[10] K. Friedrichs. Eine invariante Formulierung des Newtonschen Gravititationsgesetzes und des Grenzüberganges vom Einsteinschen zum Newtonschen Gesetz. *Math. Ann.*, 98:566–575, 1927.

[11] T. Gergelits, K.-A. Mardal, B. F. Nielsen, and Z. Strakos. Laplacian preconditioning of elliptic PDEs: Localization of the eigenvalues of the discretized operator. *SIAM Journal on Numerical Analysis*, 57(3):1369–1394, 2019.

[12] A. A. Gonchar. Zolotarev problems connected with rational functions. *Mat. Sb.*, 78(120):623–635, 1969.

[13] A. A. Gonchar. On the speed of rational approximation of some analytic functions. *Mathem. Digest (Matem. Sbornik)*, 34:131–145, 1978.

[14] I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series, and Products*, volume 29. Academic Press, 1980.

[15] L. Grubisic and H. Hakula. High order approximations of the operator Lyapunov equation have low rank. *BIT Numerical Mathematics*, 62, 04 2022.

[16] L. Grubisic and D. Kressner. On the eigenvalue decay of solutions to operator Lyapunov equations. *Systems and Control Letters*, 73, 11 2014.

[17] L. Knizhnerman and V. Simoncini. Convergence analysis of the extended Krylov subspace method for the Lyapunov equation. *Numerische Mathematik*, 118:567–586, 07 2011.

[18] G. Kocharyan. On approximation by rational functions in the complex domain. *Izv. AN Arm. SSR, ser. fiz.-matem*, 11(4), 1958.

[19] P. Lancaster. Explicit solutions of linear matrix equations. *SIAM Review*, 12(4):544–566, 1970.

[20] C. Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of Research of the National Bureau of Standards.*, 45(4):255–282, 1950.

[21] C. C. Paige, M. Rozložník, and Z. Strakos. Modified Gram-Schmidt (MGS), Least Squares, and Backward Stability of MGS-GMRES. *SIAM Journal on Matrix Analysis and Applications*, 28(1):264–284, 2006.

[22] T. J. Rivlin. *Chebyshev polynomials*. Courier Dover Publications, 2020.

[23] J. Sabino. Solution of large-scale Lyapunov equations via the block modified smith method. *PhD thesis, Department of Computational and Applied Mathematics, Rice University, Houston*, 2007.

[24] D. Silvester, H. Elman, and A. Wathen. *Finite Elements and Fast Iterative Solvers: With Applications in Incompressible Fluid Dynamics*. Oxford University Press, 2005.

[25] V. Simoncini. A new iterative method for solving large-scale Lyapunov matrix equations. *SIAM J. Sci. Comput.*, 29:1268–1288, 2007.

[26] V. Simoncini and V. Druskin. Convergence analysis of projection methods for the numerical solution of large Lyapunov equations. *SIAM Journal on Numerical Analysis*, 47(2):828–843, 2009.

[27] V. Simoncini and D. Toni. On some structural properties of generalized Lyapunov eigenproblems and application to operator preconditioning. *Bollettino dell'Unione Matematica Italiana*, 17:625–645, 2024.

[28] G. Stampacchia. Formes bilineaires coercitives sur les ensembles convexes. *C. R. Acad. Sci., Paris*, 258, 1964.

[29] P. D. W. and H. H. Rachford. The numerical solution of parabolic and elliptic differential equations. *Journal of the Society for Industrial and Applied Mathematics.*, 3(1):28–41, 1955.

[30] J. L. Walsh. *Interpolation and Approximation by Rational Functions in the Complex Domain*. American Mathematical Society, 1965.

[31] H. Widom. Extremal polynomials associated with a system of curves in the complex plane. *Advances in Mathematics*, 3(2):127–232, 1969.

[32] W.-C. Yueh. Eigenvalues of several tridiagonal matrices. *Applied Mathematics E-Notes [electronic only]*, 5:66–74, 2005.