A numerical algorithm for calculating eigenvalues of preconditioned Toeplitz matrices

Author: Julián Andrés Montenegro Parra

Supervisor: Prof. Manuel Bogoya, Ph.D

Undergraduate thesis submitted for the degree of:

Mathematician

Universidad del Valle

Department of mathematics

Cali, Colombia

A mis padres,
Geovanna Parra y Julián Montenegro,
todo es por y para ustedes.

ACKNOWLEDGMENTS

El presente trabajo de grado es el resultado de múltiples esfuerzos, apoyo y contribuciones de distintas personas a las que quiero aprovechar para agradecer, pues de una manera u otra, este documento representa su aporte.

Quiero iniciar agradeciendo a Dios, por haberme permitido llegar a este punto de la carrera. Ha sido un proceso de continuo aprendizaje y de dificultades, en especial la elaboración de este documento que sin la supervisión del profesor Johan Manuel Bogoya, no hubiera sido posible. Gracias por su paciencia, su dedicación y sus valiosas retroalimentaciones que reflejan su excelente labor como educador.

A mi compañera de vida, María Alejandra Zapata, por su amor y su compañía, que siempre me traen paz y son fuente constante de fuerza y motivación.

A todos los profesores con los que tuve el privilegio de cursar clases, gracias por sus enseñanzas y por compartir su valioso conocimiento. Quiero mencionar especialmente a algunos profesores que, a lo largo de la carrera, me brindaron consejos y estuvieron pendientes de mi progreso: la profesora Luz Victoria De la Pava y el profesor Gonzalo García.

A mis amigos, compañeros de carrera, que hicieron que cada clase fuera más amena y llevadera. En particular, quiero agradecer a Yuhad Olarte, quien me apoyó constantemente en la escritura de este documento.

De manera muy especial, quiero agradecer a Carmencita por su generosidad y por cuidar siempre de mi, permitiéndome concentrarme en mis estudios.

Por último, pero no menos importante, a mis padres, que son mi motor y mi mayor apoyo. Gracias por creer en mí incondicionalmente y por todos sus esfuerzos que me han hecho la persona que soy.

ABSTRACT

This thesis addresses the problem of finding eigenvalues for preconditioned Toeplitz matrices based on a conjecture made by Professor Manuel Bogoya et al. In many cases, working with an ill-conditioned matrix can be computationally expensive. This is measured by what is known as the condition number of the matrix. The effect of preconditioning is to attempt to reduce this condition number so that the matrix is less sensitive to perturbations, that is, to small changes. Various types of numerical algorithms have been implemented to find the spectrum of Toeplitz matrices, and simple-loop theory allows for the derivation of asymptotic expansions for this type of matrix $T_n(f)$ generated by a symbol f. However, this theory is not available in the context of preconditioned Toeplitz matrices.

Key words. Asymptotic symbol, asymptotic expansion, Banach space, Hilbert space, L^p space, bounded operator, Fredholm operator, spectrum.

En este trabajo se aborda el problema de encontrar valores propios para matrices de Toeplitz precondicionadas basados en una conjetura realizada por el profesor Manuel Bogoya et al. En muchas ocasiones, trabajar con una matriz mal condicionada puede resultar muy costoso computacionalmente. Esto se mide con lo que conocemos como el número de condición de la matriz. El efecto de precondicionar es intentar disminuir este número de condición para que la matriz sea menos sensible a perturbaciones, esto es, a pequeños cambios. Distintos tipos de algoritmos numéricos han sido implementados para encontrar el espectro de matrices de Toeplitz y la teoría simple-loop permite derivar expansiones asintóticas para este tipo de matrices $T_n(f)$ generadas por un símbolo f. Sin embargo, esta teoría no está disponible en el contexto de matrices de Toeplitz precondicionadas.

Palabras clave. Símbolo asintótico, expansión asintótica, espacio de Banach, espacio de Hilbert, espacio L^p, operador acotado, operador de Fredholm, espectro.

CONTENTS

Introduction **5**

1 Asymptotic analysis **7**

2 Functional analysis **22**

2.0.1 Metric Spaces 22
2.0.2 Banach spaces 26
2.0.3 Hilbert spaces 29
2.0.4 Linear operators 31

3 Toeplitz matrices **36**

4 Eigenvalue's approximations **48**

5 Numerical implementation **53**

INTRODUCTION

A Toeplitz matrix, named after Otto Toeplitz, is a matrix in which each descending diagonal from left to right is constant, for example, a Toeplitz $n \times n$ matrix has the following form

$$\begin{pmatrix} a_0 & a_{-1} & a_{-2} & \cdots & a_{-n+1} \\ a_1 & a_0 & a_{-1} & \ddots & \vdots \\ a_2 & a_1 & \ddots & \ddots & a_{-2} \\ \vdots & \ddots & \ddots & \ddots & a_{-1} \\ a_{n-1} & \cdots & a_2 & a_1 & a_0 \end{pmatrix}, \tag{1}$$

but it does not have to be a square matrix. For a function $a \in L^1(\mathbb{T})$, where $\mathbb{T}$ is the complex unit circle, its kth Fourier coefficient is given by

$$a_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} a(\mathrm{e}^{\mathrm{i}\sigma})\, \mathrm{e}^{-\mathrm{i}k\sigma}\, \mathrm{d}\sigma,$$

with these coefficients we can form a sequence of Toeplitz matrices $\{T_n(a)\}_{n=1}^{\infty}$ which is generated by the function a (commonly referred to as the symbol of the sequence) and has the form (1). The study of spectral characteristics such as determinants, eigenvalues, singular values, pseudospectra of $T_n(a)$, etc., has been a field of fruitful research for more than a century, starting with Szegő seminal paper [9] in 1915. Toeplitz matrices have numerous applications in areas such as numerical analysis, engineering, stochastic processes, time series analysis, signal processing, quantum and statistical mechanics, and image processing. A very popular application is the numerical solution of differential and integral equations with the approximation technique of Finite Differences, Finite Elements, Finite Volumes, and Isogeometric Analysis because when employing regular gridding in those numerical solutions, the resulting matrices have (almost) pure Toeplitz structure. The available eigensolvers nowadays compute the eigenvalues precisely and reliably for moderately large matrices, i.e, orders in the first thousands, provided they have a “reasonable" condition number.

But several real-world applications involve really large and severely ill-conditioned matrices, i.e, matrices of size $\approx 10^6$ for modeling or the cube root of $\approx 10^{23}$ in applied physics, and in these situations even the most powerful computers in the world fail to correctly find the eigenvalues. It is therefore desirable to have formulas for a direct and precise calculation of the individual eigenvalues of the sequence of Toeplitz matrices regardless of the size.

In the present project, our goal is to understand the known spectral theory for Toeplitz matrices and to numerically find approximations for the eigenvalues of some sequences of Toeplitz matrices with a symbol under specific assumptions.

1

ASYMPTOTIC ANALYSIS

In the present chapter we will get an idea of some important and useful concepts, namely, asymptotic approximations and O notation. For further theory the reader may go to reference [8].

Definition 1.0.1 (Asymptotic symbols)**.**

(i) If $f(x)/\phi(x) \to 1$ as $x \to \infty$, we write $f \sim \phi$ and we say that f is asymptotic to ϕ or ϕ is an asymptotic approximation to f.

(ii) If $f(x)/\phi(x) \to 0$ as $x \to \infty$, we write $f = o(\phi)$, meaning that f has lesser order than ϕ.

(iii) If $|f(x)/\phi(x)|$ is bounded as $x \to \infty$, we write $f = O(\phi)$. The least upper bound is called the **implied constant** of the O term.

The O symbol is sometimes used in an interval $[a, \infty)$, i.e, the corresponding quotient is bounded from a and on. This notation can also be understood when x approaches a finite value, namely $x \to a$, and we refer to a as the distinguished point of the asymptotic (or order) relation. We will assume $x \to \infty$ unless otherwise specified. Some examples with this notation are the following.

Example 1.0.2. We can check that

$$\frac{(x+1)^2}{x^2} = \big(1 + \frac{1}{x}\big)^2 \to 0,$$

as $x \to \infty$, which means we have the asymptotic relation $(x+1)^2 \sim x^2$. We can visualize this in the graph 1.1.

Example 1.0.3. Now, we know x/x^2 is equal to $1/x$ whenever $x \neq 0$ and also we know its limit is 0 when x goes to infinity. Therefore, $\frac{1}{x^2} = o(\frac{1}{x})$. See graph 1.2.

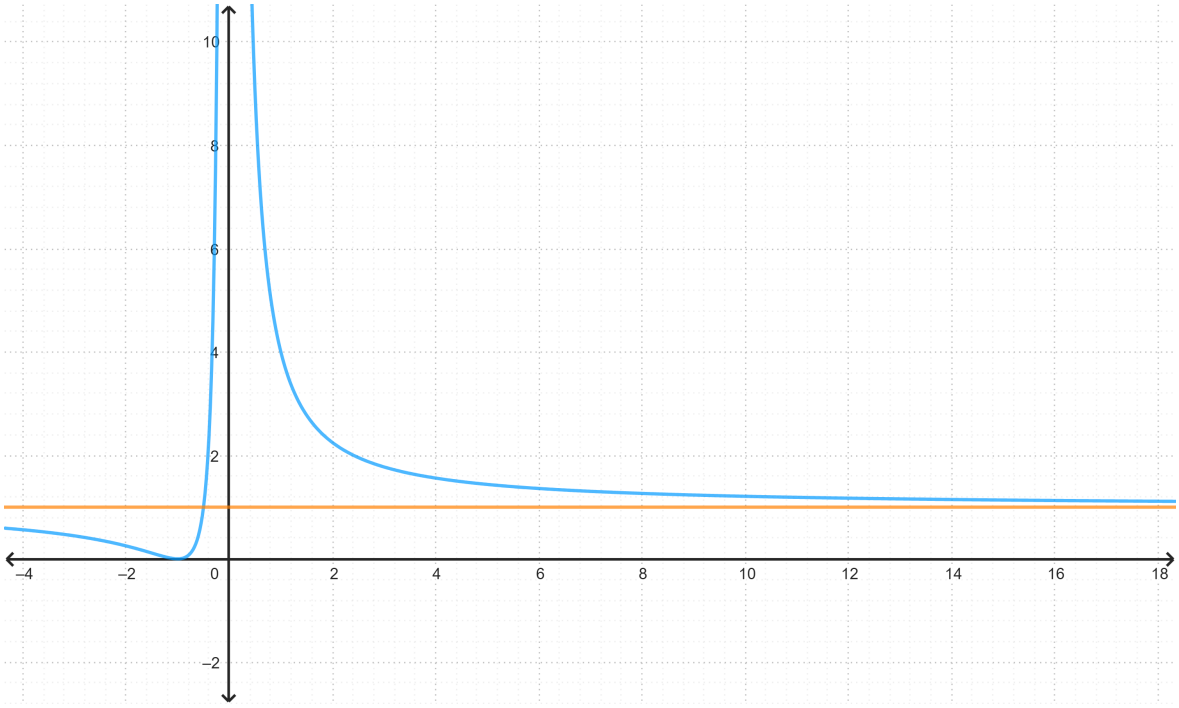


Figure 1.1: Graph of the quotient of the functions $(x+1)^2$ and x^2 in blue. The orange line corresponds to $y = 1$.

Example 1.0.4. From the definition of hiperbolic sine,

$$\sinh(x) = \frac{e^x - e^{-x}}{2},$$

we can show that

$$\frac{\sinh(x)}{e^x} \to \frac{1}{2},$$

as $x \to \infty$. Hence, we have the asymptotic relation $\sinh(x) = O(e^x)$ because the quotient of these functions is bounded, graphically we can see it as in figure 1.3.

Example 1.0.5. In this example, instead of taking $x \to \infty$ we will take $x \to c$ and then, we have the asymptotic relation $\frac{x^2-c^2}{x^2} \sim \frac{2(x-c)}{c}$. Moreover, $\frac{2(x-c)}{c} = O(x-c) = o(1)$. The first relation described ($\sim$) can be seen graphically, with $c = 5$ in figure 1.4.

The involved variable in an asymptotic relation does not have to be continuous, for example,

$$\sin\left(n\pi + \frac{1}{n}\right) = -\sin\left(\frac{1}{n}\right) = O\left(\frac{1}{n}\right),$$

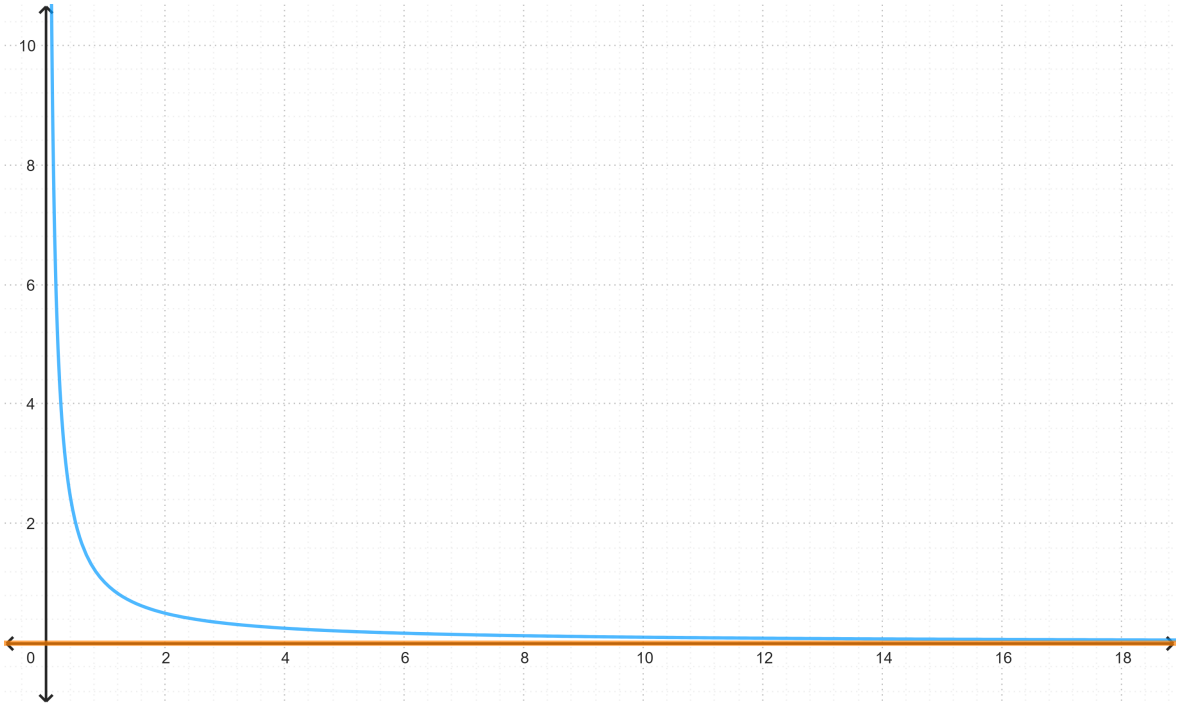


Figure 1.2: Graph of the quotient of the functions $\frac{1}{x^2}$ and $\frac{1}{x}$ in blue. The orange line corresponds to $y = 0$,i.e, the x-axis.

and also, is not always the case that the variable is going to infinity, i.e,

$$\frac{x^2 - c^2}{x^2} \sim \frac{2(x-c)}{c}, \quad \text{as} \quad x \to c.$$

When the variable is complex, one has to specify the path going to infinity, which is done with the concept of sectors. Let S be an infinite sector with $\alpha \leq \text{Arg}(z) \leq \beta$ and let $S(R)$ the intersection between S and $|z| > R$, then the asymptotic relations can be defined within $S(R)$ in the same way already done in the real case. We now give a useful theorem having properties for doing calculations with asymptotic symbols.

Proposition 1.0.6 (Properties.). *Let f, g be functions, then,*

(i) $O(f)O(g) = O(fg)$.

(ii) $O(f)o(g) = o(fg)$.

(iii) $o(f) = O(f)$.

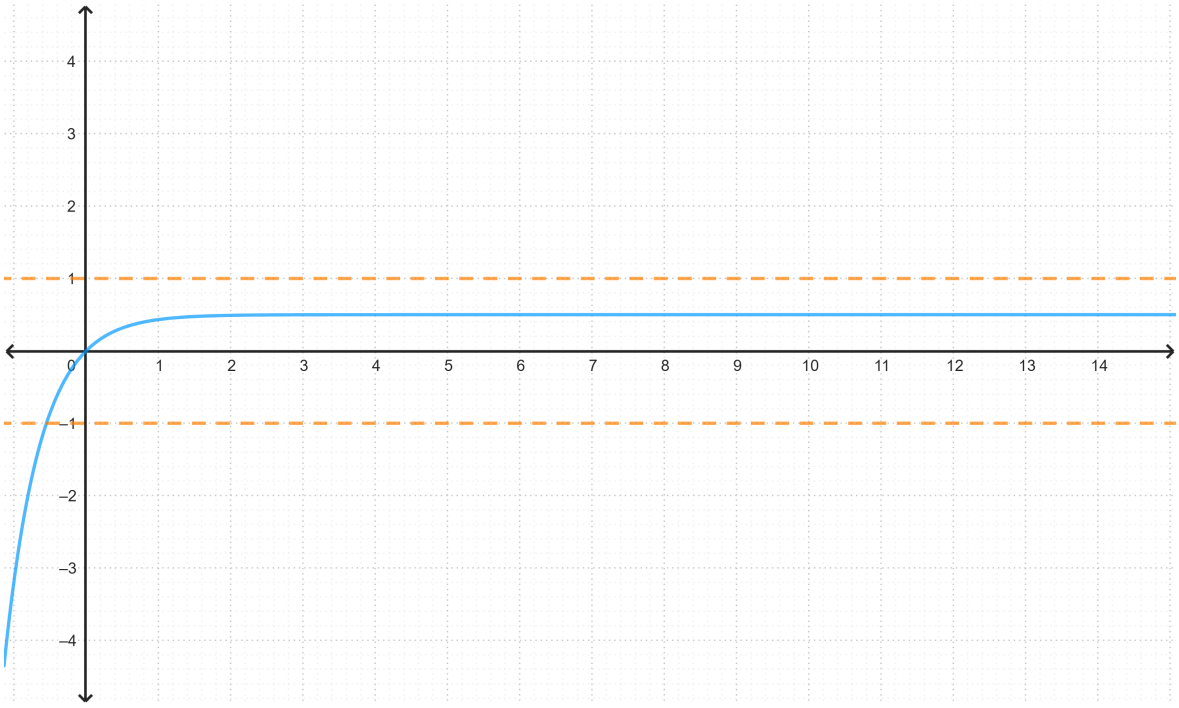


Figure 1.3: Graph of the quotient of the functions $\sinh(x)$ and e^x in blue. The orange dotted lines correspond to $y = 1$ and $y = -1$.

(iv) $O(f) + O(g) = O(|f| + |g|)$.

(v) $O(O(f)) = O(f)$.

(vi) $o(o(f)) = o(f)$

(vii) *If* $f(x) \geq |g(x)|$ *for every sufficiently large* x *then* $O(g) = O(f)$ *and* $o(g) = o(f)$.

(viii) *If* $f(x) \geq |g(x)|$ *for every sufficiently large* x *then* $O(f) + O(g) = O(f)$.

(ix) *If* $f \geq 0$ *then* $f \leq o(g)$ *implies* $f = o(g)$.

(x) $f \sim g$ *if and only if* $f = (1 + o(1))g$, *given that infinity is not a limit point of zeros of* g.

Proof. The proof of all these properties follow from the definitions of the asymptotic symbols. □

The following theorem is useful for some calculations since it concerns to power series.

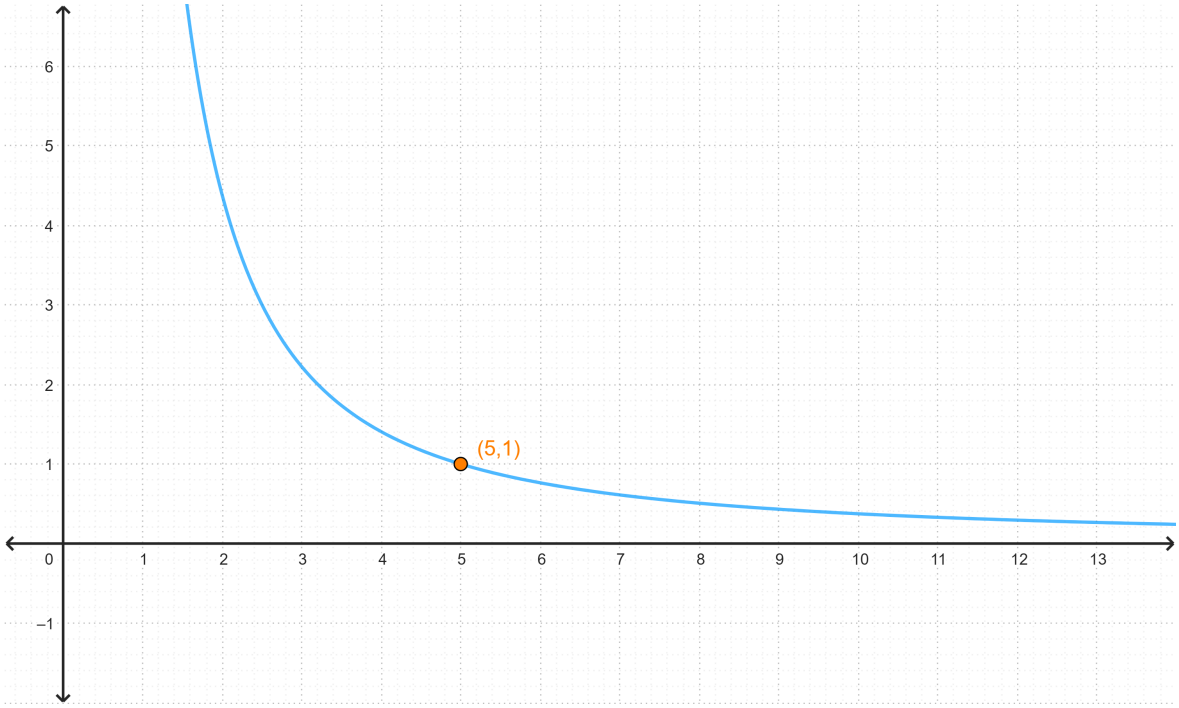


Figure 1.4: Graph of the quotient of the functions $\frac{x^2-25}{x^2}$ and $\frac{2(x-5)}{5}$ in blue.

Theorem 1.0.7. *Let $\sum_{s=0}^{\infty} a_s z^s$ be a convergent series for $|z| < r$ then for every $n \in \mathbb{N}$, $\sum_{s=n}^{\infty} a_s z^s = O(z^n)$ in any disk $|z| < \rho$ such that $\rho < r$.*

Proof. Let $x \in (\rho, r)$, then $a_s x^s \to 0$ as $s \to \infty$ (because of the convergence of the series), hence there exists a constant A such that

$$|a_s| x^s < A$$

for every $s \in \mathbb{N}$. Then we have,

$$\Big| \sum_{s=n}^{\infty} a_s z^s \Big| \leq \sum_{s=n}^{\infty} A \frac{|z|^s}{x^s} = A \frac{x^{1-n}|z|^n}{x - |z|} \leq A \frac{x^{1-n}}{x-\rho} |z|^n$$

which proves $\sum_{s=n}^{\infty} a_s z^s = O(z^n)$. □

Having defined these symbols we are interested in knowing if they are compatible with differentiation and integration. Asymptotic and order relations may be integrated subject to some restrictions. For example, if $f(x) \sim x^z$ where $z \in \mathbb{C}$, then

$$\int_a^{\infty} f(t) \mathrm{d}t \sim \int_a^{\infty} t^z \mathrm{d}t$$

Sometimes, because of the complexity of equations it is of interest to study the asymptotic solutions rather than trying to find the exact root, this proves to be useful since in many cases it is possible to give accurate approximations. Consider for example the equation

$$x + \tanh x = u$$

since $|\tanh x| < 1$, it follows that when $x \to \infty$ the left-hand side is dominated by x. If we rewrite the equation as

$$x = u - \tanh x$$

we can think that $x(u) \sim u$ when $x \to \infty$ which also means

$$x \sim u, \quad \text{as} \quad u \to \infty.$$

Now, using the fact

$$\begin{aligned} \tanh x &= 1 - 2\,\mathrm{e}^{-2x} + 2\,\mathrm{e}^{-4x} - 2\,\mathrm{e}^{-6x} + \cdots \\ \tanh x &= 1 + o(1) \end{aligned}$$

we proceed with the following steps.

Step 1.

$$x = u - 1 + o(1).$$

Step 2.

$$x = u - 1 + O(\mathrm{e}^{-2x}) = u - 1 + O(\mathrm{e}^{-2u}).$$

Step 3.

$$\begin{aligned} x &= u - 1 + 2\,\mathrm{e}^{-2x} + O(\mathrm{e}^{-4x}) \\ &= u - 1 + 2\exp(-2u + 2 + O(\mathrm{e}^{-2u})) + O(\mathrm{e}^{-4u}) \\ &= u - 1 + 2\exp(-2u + 2)\exp(O(\mathrm{e}^{-2u})) + O(\mathrm{e}^{-4u}) \\ &= u - 1 + 2\,\mathrm{e}^{-2u+2}(1 + O(\mathrm{e}^{-2u}) + O(\mathrm{e}^{-4u}) \\ &= u - 1 + 2\,\mathrm{e}^{-2u+2} + O(\mathrm{e}^{-4u}). \end{aligned}$$

Continuing this process we find in each step approximations with errors of steadily diminishing asymptotic order. We will now do the same for the equation $x \tan x = 1$, first we will

change it into a form which is easier for doing calculations,

$$\begin{aligned} x\tan(x) &= 1 \\ \tan(x) &= \frac{1}{x} \\ x &= n\pi + \arctan\Big(\frac{1}{x}\Big). \end{aligned}$$

Now, since $\arctan(1/x)$ is bounded, we have the asymptotic relation $x \sim n\pi$ as $n \to \infty$. Using Taylor's expansion for the tangent $\arctan(\frac{1}{x}) = \frac{1}{x} - \frac{1}{3x^3} + \frac{1}{5x^5} - \frac{1}{7x^7} + \cdots$ we proceed as follows.

Step 1.

$$x = n\pi + O(x^{-1}) = n\pi + O(n^{-1}).$$

Step 2.

$$\begin{aligned} x &= n\pi + \frac{1}{n\pi + O(n^{-1})} + O(n^{-3}) \\ &= n\pi + \frac{1}{n\pi}\frac{1}{1 + O(n^{-2})} + O(n^{-3}) \\ &= n\pi + \frac{1}{n\pi}(1 + O(n^{-2})) + O(n^{-3}) \\ &= n\pi + \frac{1}{n\pi} + O(n^{-3}). \end{aligned}$$

Step 3.

$$\begin{aligned} x &= n\pi + \frac{1}{x} - \frac{1}{3x^3} + O(x^{-5}) \\ &= n\pi + \frac{1}{n\pi + \frac{1}{n\pi} + O(x^{-3})} - \frac{1}{3}\frac{1}{(n\pi + \frac{1}{n\pi} + O(n^{-3}))^3} + O(x^{-5}) \\ &= n\pi + \frac{1}{n\pi}\frac{1}{1 + \frac{1}{(n\pi)^2} + O(n^{-4})} - \frac{1}{3(n\pi)^3}\frac{1}{\big(1 + \frac{1}{(n\pi)^2 + O(n^{-4})}\big)^3} + O(n^{-5}) \\ &= n\pi + \frac{1}{n\pi}(1 - \frac{1}{(n\pi)^2} + O\Big(n^{-4} + \Big(\frac{1}{(n\pi)^2} + O(n^{-4})\Big)^2\Big) - \frac{1}{3(n\pi)^3}(1 + O(n^{-2})) + O(n^{-5}) \\ &= n\pi + \frac{1}{n\pi} - \frac{1}{(n\pi)^3} + O(n^{-5}) - \frac{1}{3(n\pi)^3} + O(n^{-5}) \\ &= n\pi + \frac{1}{n\pi} - \frac{4}{3(n\pi)^3} + O(n^{-5}). \end{aligned}$$

Before continuing with more examples, we present a theorem which will be of help in some cases for doing asymptotic calculations.

Theorem 1.0.8. *Let f be a continuous and strictly increasing function in an interval (a, ∞) and $f(\psi) \sim \psi$ when $\psi \to \infty$. Let ψ_u be the root $f(\psi_u) = u > f(a)$. Then,*

$$\psi(u) \sim u, \quad \textit{when} \quad u \to \infty.$$

Proof. Since f is strictly increasing then $u_1 > u_2$ if and only if $\psi_{u_1} > \psi_{u_2}$, and so we have $u \to \infty$ if and only if $\psi \to \infty$. By hypothesis, $\psi = f(\psi)(1 + o(1))$. Replacing by ψ_u yields

$$\psi_u = f(\psi_u)(1 + o(1)) = u(1 + o(1))$$

Note that the first $o(1)$ is associated with $\psi \to \infty$ while the second one is associated with $u \to \infty$, the change in the limit from ψ to u can be done due to the continuity of f. □

Let us illustrate this theorem for solving $x^2 - \log(x) = u$. Here the error term $\log(x)$ is not bounded, if we make $\psi = x^2$ then we end up with the equation

$$\psi - \frac{1}{2}\log(\psi) = u.$$

Now taking $f(\psi) = \psi - \frac{1}{2}\log(\psi)$ we have the necessary hypothesis for applying the Theorem 1.0.8 since f is strictly increasing when $\psi > \frac{1}{2}$ and then $\psi \sim u$ when $u \to \infty$ so $x = u^{\frac{1}{2}}(1 + o(1))$, and going back to the original equation yields,

$$\begin{aligned} x^2 = u + \log(x) &= u + \log(u^{1/2}) + \log(1 + o(1)) \\ &= u + \log(u^{1/2}) + o(1) \\ &= u + \frac{1}{2}\log(u) + o(1). \end{aligned}$$

Taking square root we get,

$$\begin{aligned} x &= (u + \frac{1}{2}\log(u) + o(1))^{1/2} \\ &= u^{1/2}\Big(1 + \frac{1}{2}\frac{\log(u)}{u} + o\Big(\frac{1}{u}\Big)\Big)^{1/2} \\ &= u^{1/2} + 1 + \frac{1}{4}\frac{\log(u)}{u} + o\Big(\frac{1}{u}\Big) + o\Big(\frac{1}{u^2}\Big) \\ &= u^{1/2} + 1 + \frac{1}{4}\frac{\log(u)}{u} + o\Big(\frac{1}{u}\Big). \end{aligned}$$

In the following examples we continue to illustrate the technique of "steps" for obtaining a desired order in some equations.

Example 1.0.9. Consider the equation $x\tan(x) = u$ with $0 < x < \frac{\pi}{2}$ and the following transformation

$$\begin{aligned}\cot(x) &= \frac{x}{u}\\ x &= \cot^{-1}\Big(\frac{x}{u}\Big) = \frac{\pi}{2} - \tan^{-1}\Big(\frac{x}{u}\Big)\\ x &= \frac{\pi}{2} - \frac{x}{u} + \frac{1}{3}\frac{x^3}{u^3} - \frac{1}{5}\frac{x^5}{u^5} + \frac{1}{7}\frac{x^7}{u^7} - \frac{1}{9}\frac{x^9}{u^9} + O\Big(\frac{x^{11}}{u^{11}}\Big).\end{aligned}$$

Then, we obtain a fourth order approximation with the following steps.

Step 1.

When $u \to \infty$, $x \to \frac{\pi}{2}$ implies $O(\frac{x}{u}) = O(\frac{1}{u})$.

Step 2.

$$x = \frac{\pi}{2} + O\Big(\frac{1}{u}\Big).$$

Step 3.

$$\begin{aligned}x &= \frac{\pi}{2} - \frac{x}{u} + O\Big(\frac{1}{u^3}\Big)\\ &= \frac{\pi}{2} - \frac{1}{u}\Big(\frac{\pi}{2} - \frac{x}{u} + O\Big(\frac{1}{u^3}\Big)\Big)\\ &= \frac{\pi}{2}\Big(1 - \frac{1}{u} + \frac{1}{u^2}\Big) + O\Big(\frac{1}{u^4}\Big).\end{aligned}$$

Step 4.

$$\begin{aligned}x &= \frac{\pi}{2} - \frac{x}{u} + \frac{x^3}{3u^3} + O\Big(\frac{1}{u^5}\Big)\\ &= \frac{\pi}{2} - \frac{\pi}{2u}\Big(1 - \frac{1}{u} + \frac{1}{u^2} + O\Big(\frac{1}{u^4}\Big)\Big) + \frac{\pi^3}{24u^3}\Big(1 - \frac{1}{u} + \frac{1}{u^2} + O\Big(\frac{1}{u^4}\Big)\Big)^3 + O\Big(\frac{1}{u^5}\Big)\\ &= \frac{\pi}{2}(1 - u^{-1} + u^{-2}) - \frac{\pi}{2}u^{-3} + \frac{\pi^3}{24u^3}\Big(1 + O\Big(\frac{1}{u}\Big)\Big) + O\Big(\frac{1}{u^5}\Big)\\ &= \frac{\pi}{2}(1 - u^{-1} + u^{-2}) - \Big(\frac{\pi}{2} - \frac{\pi^3}{24}\Big)u^{-3} + O(u^{-4}).\end{aligned}$$

Example 1.0.10. For solving the equation $\tan(x) = x$ numerically we will use the fact that for $x > 0$,

$$\arctan(x) = \frac{\pi}{2} - \tan^{-1}\Big(\frac{1}{x}\Big)$$

and so $x = n\pi + \tan^{-1}(x) = \pi(n + 1/2) - \arctan(\frac{1}{x})$. Using the Taylor expansion of $\tan^{-1}$ we arrive at,

$$x = \pi\big(n + \frac{1}{2}\big) - \frac{1}{x} + \frac{1}{3x^3} - \frac{1}{5x^5} + \frac{1}{7x^7} + O(x^{-9}).$$

Let $u = \pi(n + 1/2)$ and consider $f(x) = x + \tan^{-1}(\frac{1}{x})$, we are looking for roots for $f(x) = u$. Since f is increasing, continuous and $f(x) \sim x$ then by Theorem 1.0.8 it follows that $x(u) \sim u$ when $u \to \infty$. Now we work on the approximation steps until we get to some desired order of the error.

Step 1.

$$\begin{aligned} x &= u - \frac{1}{u(1+o(1))} + O(u^{-3}) \\ &= u - \frac{1}{u}(1+o(1)) + O(u^{-3}) \\ &= u - u^{-1} + O(u^{-1}). \end{aligned}$$

Step 2.

$$\begin{aligned} x &= u - \frac{1}{x} + \frac{1}{3x^3} + O(u^{-5}) \\ &= u - \frac{1}{u - \frac{1}{u} + O(u^{-1})} + \frac{1}{3(u - \frac{1}{u} + O(u^{-1}))^3} + O(u^{-5}) \\ &= u - \frac{1}{u}\frac{1}{1 + u^{-2} + O(u^{-2})} + \frac{1}{3u^3}\frac{1}{(1 - u^{-2} + O(u^{-2}))^3} + O(u^{-5}) \\ &= u - \frac{1}{u}(1 - u^{-2} + O(u^{-2})) + \frac{1}{3u^3}(1 + O(u^{-2})) + O(u^{-5}) \\ &= u - u^{-1} - u^{-3} + O(u^{-3}) + \frac{1}{3}u^{-3} + O(u^{-5}) \\ &= u - u^{-1} - \frac{2}{3}u^{-3} + O(u^{-3}). \end{aligned}$$

Step 3.

$$\begin{aligned} x &= u - \frac{1}{x} + \frac{1}{3x^3} - \frac{1}{5x^5} + O(u^{-7}) \\ &= u - (u - u^{-1} - \frac{2}{3}u^{-3} + O(u^{-3}))^{-1} + \frac{1}{3}(u - u^{-1} - \frac{2}{3}u^{-3} + O(u^{-3}))^{-3} + O(u^{-5}) \\ &= u - u^{-1}(1 - u^{-2} - \frac{2}{3}u^{-4} + O(u^{-4}))^{-1} + \frac{1}{3}u^{-3}(1 - u^{-2} - \frac{2}{3}u^{-4} + O(u^{-4}))^{-3} + O(u^{-5}) \\ &= u - u^{-1}(1 + u^{-2} + O(u^{-4})) + \frac{1}{3}u^{-3}(1 + O(u^{-2})) + O(u^{-5}) \\ &= u - u^{-1} - u^{-3} + O(u^{-5}) + \frac{1}{3}u^{-3} + O(u^{-5}) \\ &= u - u^{-1} - \frac{2}{3}u^{-3} + O(u^{-5}). \end{aligned}$$

Example 1.0.11. Consider the following equations

$$M(x)\cos(\phi(x)) = \cos(x) + o(1) \qquad M(x)\sin(\phi(x)) = \sin(x) + o(1)$$

with $M > 0$ and ϕ a continuous function. Then we prove $M(x) = 1 + o(1)$ and $\phi(x) = x + 2n\pi + o(1)$.

In fact, if we square both equations and then add them up we get,

$$\begin{aligned} M^2(x)\left(\cos^2(\phi(x)) + \sin^2(\phi(x)\right) &= (\cos(x) + o(1))^2 + (\sin(x) + o(1))^2 \\ &= \cos^2(x) + \sin^2(x) + (\cos(x) + \sin(x))o(1) + o(1) \\ &= 1 + o(1), \end{aligned}$$

and then $M(x) = (1 + o(1))^{1/2} = 1 + o(1)$. Replacing this in any of the initial equations, we obtain,

$$(1 + o(1))\cos(\phi(x)) = \cos(x) + o(1) \longrightarrow \cos(\phi(x)) = \cos(x) + o(1),$$

$$\phi(x) = \arccos(\cos(x) + o(1)).$$

Using Taylor's formula $f(x + h) = f(x) + f'(x)h \cdots = f(x) + O(h)$ with $f(x) = \cos^{-1}(x)$ and $h = o(1)$, we arrive at

$$\phi(x) = \cos^{-1}(\cos(x)) + O(o(1)) = x + 2n\pi + o(1).$$

Example 1.0.12. Find approximations for the roots of the equation $x\,\mathrm{e}^{1/x} = \mathrm{e}^u$.

By doing algebraic manipulations we find the equation

$$\frac{1}{x} + \log(x) = u.$$

Taking $\psi = 1/x$ and $f(\psi) = \psi - \log(x)$ we have the hypothesis for Theorem 1.0.8 since f is increasing, continuous and $f(\psi) \sim \psi$. If we consider $f(\psi) = u$ then $\psi(u) \sim u$ when $u \to \infty$, hence,

$$\frac{1}{x} = u(1 + o(1)) \longrightarrow x = \frac{1}{u}(1 + o(1))$$

and going back to the original equation, we obtain,

$$\begin{aligned} u &= \frac{1}{x} + \log(x) \\ u &= \frac{1 + x\log(x)}{x} \\ x &= \frac{1}{u}(1 + x\log(x)). \end{aligned}$$

Step 1.

$$\begin{aligned} x &= \frac{1}{u}\Big[1+\frac{1}{u}(1+o(1))\log\Big(\frac{1}{u}(1+o(1)\Big)\Big] \\ &= \frac{1}{u}+\frac{1}{u^2}[(1+o(1))(-\log(u)+o(1))] \\ &= \frac{1}{u}+\frac{1}{u^2}[-\log(u)+o(1)\log(u)+o(1)] \\ &= \frac{1}{u}-\frac{\log(u)}{u^2}+o(1)O\Big(\frac{\log(u)}{u^2}\Big). \end{aligned}$$

Step 2. We break this step in three items, starting in (i) with our main equation and then we calculate an approximation for $x\log(x)$ in (ii) with the one found in **Step 1**, we do the same in (iii) for the $\log$ part needed for (ii).

(i) $x=\frac{1}{u}(1+x\log(x))$.

(ii) $x\log(x)=\Big(\frac{1}{u}-\frac{\log(u)}{u^2}+O\big(\frac{\log(u)}{u^2}\big)\Big)\log\Big(\frac{1}{u}-\frac{\log(u)}{u^2}+O\big(\frac{\log(u)}{u^2}\big)\Big)$.

(iii)

$$\begin{aligned} \log\Big(\frac{1}{u}-\frac{\log(u)}{u^2}+O\Big(\frac{\log(u)}{u^2}\Big)\Big) &= -\log(u)+\log\Big(1-\frac{\log(u)}{u^3}+O\Big(\frac{\log(u)}{u^3}\Big)\Big) \\ &= -\log(u)-\frac{\log(u)}{u^3}+O\Big(\frac{\log(u)}{u^3}\Big). \end{aligned}$$

Replacing (iii) in (ii),

$$\begin{aligned} x\log(x) &= \Big(\frac{1}{u}-\frac{\log(u)}{u^2}+O\big(\frac{\log(u)}{u^2}\big)\Big)\Big(-\log(u)-\frac{\log(u)}{u^3}+O\big(\frac{\log(u)}{u^3}\big)\Big) \\ &= -\frac{\log(u)}{u}+\frac{(\log(u))^2}{u^2}+\frac{1}{u}O\Big(\frac{\log(u)}{u^2}\Big) \\ &= -\frac{\log(u)}{u}+\frac{(\log(u))^2}{u^2}+O\Big(\frac{\log(u)}{u^2}\Big). \end{aligned}$$

Finally, replacing the last equality in (i), we obtain

$$\begin{aligned} x=\frac{1}{u}(1+x\log(x)) &= \frac{1}{u}\Big(1-\frac{\log(u)}{u}+\frac{(\log(u))^2}{u^2}+O\Big(\frac{\log(u)}{u^2}\Big)\Big) \\ &= \frac{1}{u}-\frac{\log(u)}{u^2}+\frac{(\log(u))^2}{u^3}+O\Big(\frac{\log(u)}{u^3}\Big). \end{aligned}$$

This is one approach for solving this problem but we can also try without natural logarithm with the equation $x=\mathrm{e}^u\,\mathrm{e}^{-1/x}$. Since $\mathrm{e}^{-1/x}=1+o(1)$ then $x=\mathrm{e}^u(1+o(1))$. Using the Taylor's formula for the exponential,

$$\mathrm{e}^{-1/x}=1-\frac{1}{x}+\frac{1}{2x^2}-\frac{1}{3!x^3}+\cdots$$

we can then proceed as follows.

Step 1.

$$\begin{aligned} x &= \mathrm{e}^u(1 - \frac{1}{x} + O(x^{-2})) \\ &= \mathrm{e}^u \left(1 - \frac{1}{\mathrm{e}^u(1 + o(1))} + O(\mathrm{e}^{-2u})\right) \\ &= \mathrm{e}^u \left(1 - \frac{1 + o(1)}{\mathrm{e}^u} + O(\mathrm{e}^{-2u})\right) \\ &= \mathrm{e}^u -1 + o(1) + O(\mathrm{e}^{-u}) \\ &= \mathrm{e}^u -1 + O(\mathrm{e}^{-u}). \end{aligned}$$

Step 2.

$$\begin{aligned} x &= \mathrm{e}^u \left(1 - \frac{1}{x} + \frac{1}{2x^2} + O(x^{-3})\right) \\ &= \mathrm{e}^u \left(1 - \frac{1}{\mathrm{e}^u(1 - \mathrm{e}^{-u} + O(\mathrm{e}^{-2u}))} + \frac{1}{2\,\mathrm{e}^{2u}(1 - \mathrm{e}^{-u} + O(\mathrm{e}^{-2u}))^2} + O(\mathrm{e}^{-3u})\right) \\ &= \mathrm{e}^u -1 + \frac{1}{2}\,\mathrm{e}^{-u} + O(\mathrm{e}^{-2u}). \end{aligned}$$

Finally, we end this section of asymptotic analysis with a topic known as **Asymptotic expansions** due to Poincaré (1886). If we have the following function,

$$F(x) = \int_0^\infty e^{-xt} \cos(t) \mathrm{d}t$$

with $x > 0$, then we can approximate the integral using Taylor's series for cosine,

$$\begin{aligned} F(x) &= \int_0^\infty e^{-xt} \Big(1 - \frac{t^2}{2!} + \frac{t^4}{4!} - \frac{t^6}{6!} + \cdots\Big) \mathrm{d}t \\ &= \frac{1}{x} - \frac{1}{x^3} + \frac{1}{x^5} - \frac{1}{x^7} \cdots \\ &= \sum_{n=0}^{\infty} (-1)^n \frac{1}{x^{2n+1}}, \end{aligned}$$

this series converges when $x > 1$, and then $F(x) = \frac{x}{x^2+1}$. If we try doing the same with the

function,

$$\begin{aligned} G(x) &= \int_0^\infty \frac{e^{-xt}}{1+t}\mathrm{d}t \\ &= \int_0^\infty e^{-xt}(1-t+t^2-t^3+\cdots)\mathrm{d}t \\ &= \frac{1}{x}-\frac{1}{x^2}+\frac{2!}{x^3}-\frac{3!}{x^4}+\cdots, \end{aligned}$$

this series diverges for infinite values of x. In the former example, this process was valid because we could change the integral symbol with the sum symbol of the series since it is uniformly continuous in any bounded interval of t. In the latter case, the expansion of $(1+t)^{-1}$ diverges when $t \geq 1$, then in the interval of integration we can not change the integral with the sum. Nevertheless, what we did give good approximations of the function. If we call $\epsilon_n(x) = G(x) - g_n(x)$ where,

$$g_n(x) = \frac{1}{x}-\frac{1}{x^2}+\frac{2!}{x^3}-\frac{3!}{x^4}+\cdots+(-1)^{n-1}\frac{(n-1)!}{x^n},$$

it can be proved that

$$|\epsilon_n(x)| < \frac{n!}{x^{n+1}}$$

and so, the error is less than the first few omitted terms of the series. These ideas show the power of an asymptotic expansion.

Definition 1.0.13. Let f be a function of complex or real variable. If for any $n \in \mathbb{N}$,

$$f(z) = a_0 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_{n-1} z^{-(n-1)} + R_n(z)$$

with $R_n(z) = O(z^{-n})$ when $z \to \infty$ in an unbounded region, then we say that $\sum a_s z^{-s}$ is an asymptotic expansion of f and we write,

$$f(z) \sim a_0 + a_1 z^{-1} + a_2 z^{-2} + \cdots .$$

Theorem 1.0.14. *A neccesary and sufficient condition for a function f to have asymptotic expansion is*

$$z^n R_n(z) \to a_n,$$

when $z \to \infty$ uniformly with respect to $\mathrm{Arg}(z)$.

Proof. Write

$$f(z) = a_0 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_{n-1} z^{-(n-1)} + R_n(z).$$

If $z^n R_n(z) \to a_n$ then $z^n R_n(z) = a_n(1 + o(1))$, equivalently, $R_n(z) = \frac{a_n}{z^n}(1 + o(1))$ implying $R_n(z) = O(z^{-n})$. Now, if $R_n(z) = O(z^{-n})$ for every $n \in N$ then,

$$\begin{aligned} z^n R_n(z) &= z^n\big(\frac{a_n}{z^n} + R_{n+1}(z)\big) \\ &= a_n + z^n R_{n+1}(z) \\ &= a_n + O(z^n)O(z^{-(n+1)}) \\ &= a_n + O(z^{-1}) \\ &= a_n + o(1), \end{aligned}$$

proving what we wanted. □

2

FUNCTIONAL ANALYSIS

In the present chapter we give the basic concepts of functional analysis. We follow [5] and [6] as main references.

2.0.1 Metric Spaces

When studying basic calculus in $\mathbb{R}$ the central concept is the limit because it allows to define differentiation and integration, two important topics in analysis. This operation of taking the limit is associated with an idea of closeness which is measured with a distance function. We want then, to abstract the concept of distance for more general settings.

Definition 2.0.1. A **metric** or **distance** in a set X is a function

$$d\colon X \times X \to \mathbb{R}_0^+$$

satisfying:

(i) $d(x, y) = 0$ iff $x = y$.

(ii) **Symmetric:** $d(x, y) = d(y, x)$.

(iii) **Triangular inequality:** $d(x, z) \leq d(x, y) + d(y, z)$.

As one can easily check, these properties are satisfied by the usual distance in $\mathbb{R}^n$.

Definition 2.0.2. A **metric space** is a set X together with a metric d. We write (X, d) or simply X if there is no possible confusion with the metric taken.

The following are some examples of metric spaces.

Example 2.0.3. Let X be any non-empty set and consider the function

$$d(x,y) = \begin{cases} 0, & x = y \\ 1, & x \neq y \end{cases}.$$

Then,

(i) by definition $d(x,y)$ only when $x = y$.

(ii) by definition this functions is symmetrical.

(iii) let $x, y, z \in X$. If $x = z$ then $d(x,z) = 0$ and so the triangle inequality is satisfied since $d(x,y) + d(y,z) \geq 0$. If $x \neq z$ then either $x \neq y$ or $y \neq z$, in both cases $d(x,z) = 1 \leq d(x,y) + d(y,z)$.

Example 2.0.4. Consider the set ℓ^∞ of all bounded infinite real sequences with arbitrary elements

$$x = (x_1, x_2, \ldots), \quad y = (y_1, y_2, \ldots),$$

and let

$$d(x,y) = \sup_k |x_k - y_k|.$$

The proof that this is indeed a metric follows from the fact that the absolute value satisfies the properties of a metric, along with the properties of the supremum, which together ensure the triangle inequality.

Example 2.0.5. Let ℓ^p $(1 \leq p < \infty)$ be the set of infinite sequences $(x_k)_{k=1}^\infty$ in $\mathbb{R}$ (or $\mathbb{C}$) such that

$$\sum_{k=1}^{\infty} |x_k|^p < \infty,$$

and

$$d(x,y) = \Big(\sum_{k=1}^{\infty} |x_k - y_k|^p \Big)^{\frac{1}{p}}.$$

The proof of the first and second properties for d to be a metric are the same as in Example 9. Before all this, we have to proof that the definition of d makes sense because it's possible that the sum in d doesn't even converge.

From Minkowski's inequality it follows

$$\big(\sum_{k=1}^{n} |x_k - z_k|^p \big)^{\frac{1}{p}} \leq \big(\sum_{k=1}^{n} |x_k|^p \big)^{\frac{1}{p}} + \big(\sum_{k=1}^{n} |y_k|^p \big)^{\frac{1}{p}},$$

and since $x, y \in \ell_p$

$$\big(\sum_{k=1}^{\infty} |x_k|^p\big)^{\frac{1}{p}} < \infty, \qquad \big(\sum_{k=1}^{\infty} |y_k|^p\big)^{\frac{1}{p}} < \infty,$$

implying

$$\big(\sum_{k=1}^{\infty} |x_k - z_k|^p\big)^{\frac{1}{p}} < \infty.$$

This also proved the triangle inequality since it's the same as for finite index but when $n \to \infty$ (which is possible because of what was just proven).

The concept of sequence is also of great interest since it is related with the topology of the space we are working with, we can use it to characterize closed sets, accumulation points and continuity. The notion of convergence of a sequence is again closely related to the metric of the space.

Definition 2.0.6. A sequence (x_n) in a metric space (X, d) converges to x if for every $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that

$$n > N \longrightarrow d(x_n, x) < \varepsilon$$

We write this as

$$\lim_{n\to\infty} x_n = x.$$

Some desired properties of convergent sequences in $\mathbb{R}$ such as the uniqueness and boundedness of the limit, also hold in more general settings such as this one. We recall that a set $M \subseteq X$ is said to be bounded if

$$\sup_{x,y\in M} d(x, y) < \infty,$$

which is equivalent to saying that M is contained in a ball.

Proposition 2.0.7. *Every convergent sequence in a metric space is bounded and has a unique limit.*

The following proposition will be particularly useful for proving the completeness of metric spaces.

Proposition 2.0.8. *If $y_n \to y$ and for every $n \in \mathbb{N}$, $d(x_n, y_n) < k$ then $d(x_n, y) \leq k$*

Proof. Let $\varepsilon > 0$, and using the triangle inequality we have

$$d(x_n, y) \leq d(x_n, y_n) + d(y_n, y) < k + \varepsilon$$

therefore, $d(x_n, y) \leq k$ □

Definition 2.0.9. A sequence (x_n) is called a **Cauchy sequence** if for every $\varepsilon > 0$ there is $N \in \mathbb{N}$ such that

$$n, m > N \longrightarrow d(x_n, x_m) < \varepsilon.$$

A metric space is said to be **complete** if every Cauchy sequence is convergent.

We will later show some examples illustrating the fact that not every metric space is complete, i.e, not every Cauchy sequence is convergent. However, the converse is true, as stated in the following theorem.

Theorem 2.0.10. *Every convergent sequence in a metric space is a Cauchy sequence.*

Before exploring some examples of complete (and incomplete) metric spaces we give some important theorems which exhibit the relationship of sequences with the topology of the space.

Example 2.0.11. Let (x_n) be a Cauchy sequence in ℓ^∞, let $x_k^{(n)}$ denote the k-th component of the n-th element of the sequence (x_n). Recall that in this space the points are bounded real sequences (let's say $|a_k^{(n)}| < M_n$ for $k \geq 1$), and the distance function were defined as

$$d(x, y) = \sup_k |x_k - y_k|.$$

Then, since (x_n) is Cauchy, for every $\varepsilon > 0$ there exists N such that for $n, m > N$

$$d(x_m, x_n) = \sup_k |x_k^{(m)} - x_k^{(n)}| < \varepsilon,$$

so that for every $k \geq 1$

$$|x_k^{(m)} - x_k^{(n)}| < \varepsilon, \tag{2.1}$$

then the sequence

$$\left(x_k^{(1)}, x_k^{(2)}, x_k^{(3)}, \ldots, x_k^{(n)}, \ldots\right)$$

is a real Cauchy sequence. Let y_k be its limit and $y = (y_k)$, we now prove that y is in ℓ^p. This follows from

$$|y_k| \leq |y_k - x_k^{(n)}| + |x_k^{(n)}| \leq \varepsilon + M_n$$

hence, $y = (y_k)$ is a bounded sequence then $y \in \ell^\infty$, moreover letting $n \to \infty$ in equation (2.1) yields

$$|x_k^{(m)} - y_k| < \varepsilon,$$

this shows $x_m \to y$ proving that ℓ^∞ is complete.

Example 2.0.12. Let (x_n) be a Cauchy sequence in the space ℓ^p. Then for every $\varepsilon > 0$ there is N such that for $m, n > N$ we have

$$d(x_m, x_n) = \big(\sum_{k=1}^{\infty} |x_k^{(m)} - x_k^{(n)}|^p\big)^{\frac{1}{p}} < \varepsilon. \tag{2.2}$$

Then for a fixed k the sequence (x_k^m) is a real Cauchy sequence, say it converges to y_k and let $y = (y_k)$. We prove $y \in \ell^p$. From triangle inequality we know

$$\sum_{k=1}^{r} |y_k|^p < \sum_{k=1}^{r} |y_k - x_k^{(m)}|^p + \sum_{k=1}^{r} |x_j^{(m)}|^p. \tag{2.3}$$

Letting $n \to \infty$ in inequality (2.2) , we get

$$\sum_{k=1}^{\infty} |x_k^{(m)} - y_k|^p < \varepsilon^p < \infty,$$

(which also proved $x_m \to y$) and since $x_m \in \ell^p$ then

$$\sum_{k=1}^{\infty} |x_k^{(m)}| < \infty,$$

therefore, from inequality (2.3) we conclude

$$\sum_{k=1}^{\infty} |y_k|^p < \infty,$$

and so $y \in \ell^p$. Since (x_m) was an arbitrary Cauchy sequence in ℓ^p, this proves its completeness.

2.0.2 Banach spaces

Now, we are interested in studying the relation between vector spaces and metric spaces, with this goal in mind we define the concept of **norm** which tries to abstract the properties of the absolute value in $\mathbb{R}$.

Definition 2.0.13. Let X be a vector space, a function

$$\|\cdot\| : X \to \mathbb{R}_0^+,$$

is called a **norm** if it satisfies:

(i) $\|x\| = 0$ if and only if $x = 0$,

(ii) $\|\lambda x\| = |\lambda| \, \|x\|$,

(iii) $\|x + y\| \leq \|x\| + \|y\|$,

for any scalar λ and $x, y \in X$. A vector space X together with a norm $\|\cdot\|$ is called a **normed space**.

The norm naturally induces a metric in the space given by

$$d(x, y) = \|x - y\| .$$

Definition 2.0.14. A normed space X is said to be a **Banach space** if it is complete with respect to the metric induced by the norm.

Theorem 2.0.15. *The norm is a continuous function.*

Many of the examples given on the previous section can be viewed also as normed spaces or, moreover, Banach spaces.

Example 2.0.16. $\mathbb{R}^n$ with its usual norm.

Example 2.0.17. ℓ^p with the p-norm given by

$$\|x\|_p = \big(\sum_{i=1}^{\infty} |x_i|^p\big)^{\frac{1}{p}}.$$

Example 2.0.18. ℓ^∞ with the norm given by

$$\|x\|_\infty = \sup_i |x_i|.$$

Note that, in the previous section we proved that the examples above were complete with the induced metric. It follows that they are all Banach spaces. An interesting fact about the p-norms is that if we think of $\mathbb{R}^2$ with these norms then we will have different shapes for the unit circle as shown in figure 2.1.

Now we give some theorems which shows the connection between linear algebra and the norm.

Lemma 2.0.19. Let $x_1, x_2, \ldots, x_n$ be linearly independent in a normed space X. Then, there exists $c > 0$ such that for any linear combination the following holds

$$\|\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n\| \geq c(|\alpha_1| + |\alpha_2| + \cdots + |\alpha_n|).$$

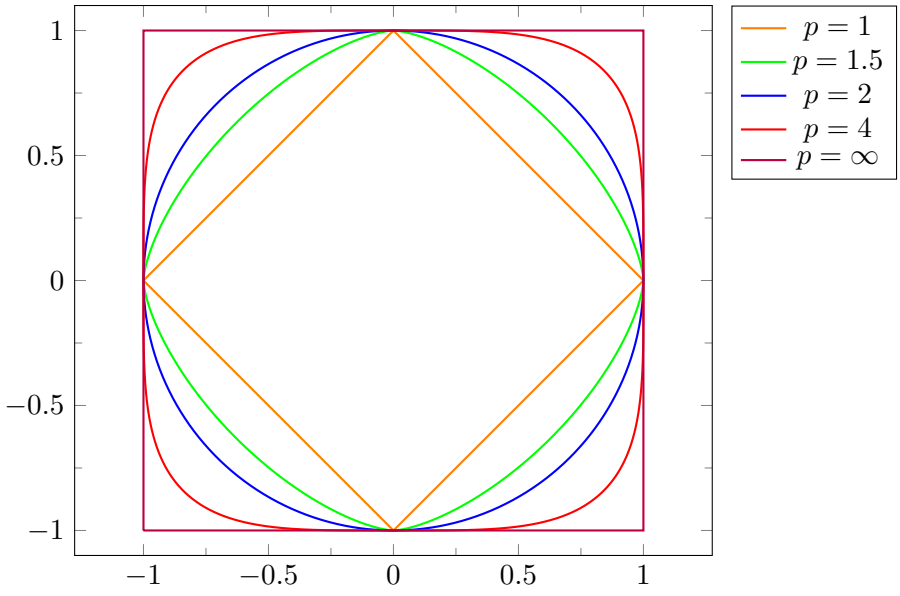


Figure 2.1: Unit balls in different p-norms. As $p \to \infty$, the shape transitions from a diamond ($p = 1$) to a square ($p = \infty$).

Proof. For a proof the reader may consult [6, Lemma 2.4-1, page 72]. □

Theorem 2.0.20. *Every finite dimensional normed space is complete.*

Proof. Let X be a finite dimensional normed space and let $\{e_1, \ldots, e_n\}$ be a basis for it. Consider (x_m) a Cauchy sequence in X. Since $x_n \in X$, then,

$$x_m = a_1^{(m)} e_1 + a_2^{(m)} e_2 + \cdots + a_n^{(m)} e_n.$$

Now, for every $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that if $m, r > N$ we have $\|x_m - x_r\| < \varepsilon$. Using the lemma above,

$$\varepsilon > \|x_m - x_r\| = \Big|\Big| \sum_{i=1}^{n} (a_i^{(m)} - a_i^{(r)}) e_i \Big|\Big| \geq c\Big(\sum_{i=1}^{n} |a_i^{(m)} - a_i^{(r)}| \Big),$$

hence,

$$|a_i^{(m)} - a_i^{(r)}| \leq \frac{\varepsilon}{c}.$$

This means the sequence $(a_i^{(m)})$ is a Cauchy sequence for every $1 \leq i \leq n$. Therefore, it converges say to a_i. If we let $x = a_1 e_1 + a_2 e_2 + \cdots + a_n e_n$ then we can see that $x_m \to x$, showing X is complete. □

2.0.3 Hilbert spaces

We already defined what a metric is by doing a generalization of the concept of distance in $\mathbb{R}$, likewise, we abstracted the concept of length of a vector by defining the norm in a vector space. Another familiar operation is the dot product in R^n, given $x = (x_i), y = (y_i) \in \mathbb{R}^n$ we have

$$x \cdot y = \sum_{i=1}^{n} x_i y_i,$$

which can be used to write the norm

$$\|x\| = \sqrt{x \cdot x},$$

and also to give a necessary and sufficient condition for vectors to be orthogonal, namely, $x \cdot y = 0$.

Definition 2.0.21. Let X be a complex(or real) vector space. A function

$$\langle \cdot\, , \cdot \rangle \colon X \times X \to \mathbb{C},$$

is called an **inner product** if it satisfies

(i) $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$,

(ii) $\langle \alpha x, y\rangle = \alpha\langle x, y\rangle$,

(iii) $\langle x, y\rangle = \overline{\langle y, x\rangle}$,

(iv) $\langle x, x\rangle = 0 \longleftrightarrow x = 0$.

It follows, from these properties that

$$\langle x, \alpha y\rangle = \overline{\langle \alpha y, x\rangle} = \overline{\alpha\langle y, x\rangle} = \overline{\alpha}\overline{\langle y, x\rangle} = \overline{\alpha}\langle x, y\rangle.$$

A vector space with an inner product is called **inner product space**.

The inner product naturally induces a norm,

$$\|x\| = \sqrt{\langle x, x\rangle}.$$

We will prove it is indeed a norm. For this, we first prove the *Schwarz's inequality*

$$|\langle x, y\rangle| \leq \|x\| \, \|y\| \, .$$

If $y = 0$, then it holds since $\langle x, 0\rangle = 0$ and moreover it is an equality. If $y \neq 0$, then, for every $\alpha \in \mathbb{C}$

$$\begin{aligned} 0 \leq \|x - \alpha y\|^2 &= \langle x - \alpha y, x - \alpha y\rangle \\ &= \langle x, x - \alpha y\rangle - \langle \alpha y, x - \alpha y\rangle \\ &= \langle x, x\rangle - \overline{\alpha}\langle x, y\rangle - \langle \alpha y, x\rangle + \langle \alpha y, \alpha y\rangle \\ &= \langle x, x\rangle - \overline{\alpha}\langle x, y\rangle - \alpha\langle y, x\rangle + \alpha\overline{\alpha}\langle y, y\rangle \\ &= \langle x, x\rangle - \overline{\alpha}\langle x, y\rangle - \alpha[\langle y, x\rangle - \overline{\alpha}\langle y, y\rangle]. \end{aligned}$$

Letting $\alpha = \langle y, x\rangle / \langle y, y\rangle$ we obtain,

$$0 \leq \|x\|^2 - \frac{|\langle x, y\rangle|^2}{\|y\|^2},$$

multiplying both sides by $\|y\|^2$ yields the inequality we wanted.

Remark 2.0.22. Note that equality happens when $y = 0$ or $\|x - \alpha y\|^2 = 0$ ($x = \alpha y$), hence when $\{x, y\}$ is a linearly dependent set.

Proposition 2.0.23. *If X is an inner product space then*

$$\|x\| = \sqrt{\langle x, x\rangle},$$

is a norm.

Proof. We need to prove the three properties of being a norm, first, $\|x\| = 0$ if and only if $\langle x, x\rangle = 0$ which is equivalent to $x = 0$. Second, we have

$$\|\alpha x\| = \sqrt{\langle \alpha x, \alpha x\rangle} = \sqrt{\alpha\overline{\alpha}\langle x, x\rangle} = \sqrt{|\alpha|^2 \|x\|^2} = |\alpha| \|x\|,$$

and lastly for the triangle inequality,

$$\begin{aligned} \|x + y\|^2 &= \langle x + y, x + y\rangle \\ &= \langle x, x\rangle + \langle x, y\rangle + \langle y, x\rangle + \langle y, y\rangle \\ &\leq \|x\|^2 + 2\|x\| \|y\| + \|y\|^2 \\ &= (\|x\| + \|y\|)^2. \end{aligned}$$

Therefore, $||x||$ defines a norm. □

An inner product space is called **Hilbert space** if the normed space associated is a Banach space, i.e, if the space is complete with the metric induced by the inner product norm.

Proposition 2.0.24. *If X is an inner product space then, the Parallelogram's equality holds*

$$\|x+y\|^2 + \|x-y\|^2 = 2(\|x\|^2 + \|y\|^2).$$

This proposition gives us a criterion for deciding whether a norm comes from an inner product or not.

2.0.4 Linear operators

From linear algebra we know that the homomorphisms between vector spaces are called linear transformations which are functions that preserves the operations. In this context, a function between metric spaces will be called **operator** and then a linear transformation between normed spaces is called **linear operator**.

Definition 2.0.25. Let X, Y be normed spaces and $\mathrm{T}\colon X \to Y$. The operator T is said to be bounded, if there exists $c \in \mathbb{R}$ such that

$$\|\mathrm{T}x\| \leq c\,\|x\|\,,$$

upon division by $\|x\|$ we see that c is at least as big as the supremum of

$$\frac{\|\mathrm{T}x\|}{\|x\|}.$$

Define the norm of the operator as

$$\|\mathrm{T}\| = \sup_{x \in X\setminus\{0\}} \frac{\|\mathrm{T}(x)\|}{\|x\|}.$$

The set of all bounded linear operators $\mathrm{T}\colon X \to Y$ is denoted by $\mathcal{B}(X,Y)$.

Remark 2.0.26. We will use $\mathrm{T}x$ or $\mathrm{T}(x)$ indistinctively for x evaluated in the operator T.

Proposition 2.0.27. *Let T be a bounded linear operator, then*

(i) $\|\mathrm{T}\| = \sup_{\|x\|=1} \|\mathrm{T}(x)\|$.

(ii) The norm $\|\mathrm{T}\|$ is indeed a norm defined on the space of operators.

Proof.

(i) For each $x \in X \setminus \{0\}$, let $x = ky$, where $\|y\| = 1$ and $k \in \mathbb{C}$ then,

$$\frac{\|\mathrm{T}x\|}{\|x\|} = \frac{\|\mathrm{T}(ky)\|}{\|ky\|} = \frac{|k|\,\|\mathrm{T}y\|}{|k|\,\|y\|} = \|\mathrm{T}y\|\,.$$

Proving what we wanted.

(ii) This follows from properties of the supremum. □

Theorem 2.0.28. *If a normed space X is finite dimensional then every linear operator on X is bounded.*

Proof. Let X be a finite dimensional vector space and let $\{e_1, e_2, \dots, e_n\}$ be a basis for X. Suppose $x = a_1e_1 + a_2e_2 + \cdots + a_ne_n$ is any element in X, then,

$$\|\mathrm{T}x\| = \left\|\mathrm{T}\Big(\sum_{i=1}^{n} a_ie_i\Big)\right\| \leq \sum_{i=1}^{n} |a_i| \, \|\mathrm{T}e_i\| \leq \max_{1\leq i\leq n} \|\mathrm{T}e_i\| \sum_{i=1}^{n} |a_i| \leq \max_{1\leq i\leq n} \|\mathrm{T}e_i\| (c\,\|x\|)$$

for some positive constant c according to 2.0.19. Therefore, T is bounded. □

Theorem 2.0.29. *Let $\mathrm{T}\colon X \to Y$ be a linear operator, then T is continuous if and only if T is bounded.*

Proof. First, suppose T is bounded. Let $\varepsilon > 0$, then,

$$\|\mathrm{T}x - \mathrm{T}y\| = \|\mathrm{T}(x-y)\| \leq \|\mathrm{T}\| \, \|x-y\| < \|\mathrm{T}\| \, \delta$$

and so, letting $\delta = \epsilon / \|\mathrm{T}\|$ yields the continuity.

Conversely, consider T to be continuous, i.e, given $\varepsilon > 0$ and $x_0 \in \operatorname{Dom}\mathrm{T}$ there exists $\delta > 0$ such that

$$\|x - x_0\| < \delta \quad \to \quad \|\mathrm{T}x - \mathrm{T}x_0\| < \varepsilon.$$

Let $x - x_0 = \frac{\delta}{\|y\|} y$, for some non-zero $y \in \operatorname{Dom} T$. Then,

$$\|\mathrm{T}x - \mathrm{T}x_0\| = \|\mathrm{T}(x - x_0)\| = \frac{\delta}{\|y\|} \|\mathrm{T}y\| < \varepsilon.$$

Implying T is bounded since $\|\mathrm{T}y\| < \frac{\varepsilon}{\delta} \|y\|$. □

We will now state and prove some important result which are going to be very useful in the next section, but before that we explain a little bit the L^p spaces.

Definition 2.0.30. Let (X, Σ, μ) be a measure space and $1 \leq p < \infty$. For a measurable function $f\colon X \to \mathbb{C}$ the L^p norm is defined as

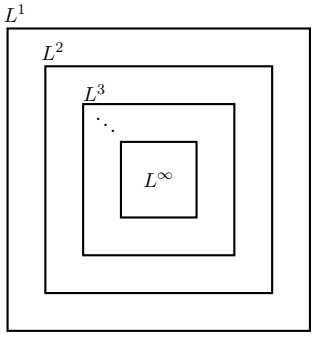
$$||f||_p = \Big(\int_X |f|^p \, \mathrm{d}\mu\Big)^{\frac{1}{p}}.$$

For this function to defined a norm we consider the relation in X given by $f \sim g$ if $f = g$ almost everywhere. The set of all such functions f satisfying $||f||_p < \infty$ is denoted $L^p(X, \mu)$. Separetely, we define the space $L^\infty(X, \mu)$ using the norm

$$||f||_\infty = \operatorname{ess\,sup} |f| = \{a \in \mathbb{R}\colon \mu(|f|^{-1}(a, \infty)) = 0\}.$$

Remark 2.0.31. Sometimes $L^p(X, \mu)$ is simply written as L^p when it is clear from the context in which measure space we are working.

A very interesting property of L^p spaces is that whenever the space X is of finite measure, they are embedded one in another, i.e, if $1 \leq p < q < \infty$ then, $L^q(X, \mu) \subset L^p(X, \mu)$ and so, we have the following pictorial representation



For the proof of this fact, we use Hölder's inequality. Given that a, b are conjugate exponents, that is,

$$\frac{1}{a} + \frac{1}{b} = 1.$$

We have the inequality,

$$\int_X |fg| \, \mathrm{d}\mu \leq \Big(\int_X |f|^a \, \mathrm{d}\mu \Big)^{\frac{1}{a}} \Big(\int_X |g|^b \, \mathrm{d}\mu \Big)^{\frac{1}{b}}.$$

Suppose $h \in L^q(X, \mu)$. Taking $f = |h|^p$, $g = 1$, $a = q/p$ and $b = q/(q-p)$ yields

$$\begin{aligned} \int_X |h|^p \, \mathrm{d}\mu &\leq \Big(\int_X (|h|^p)^{\frac{q}{p}} \, \mathrm{d}\mu \Big)^{\frac{p}{q}} \Big(\int_X 1^{\frac{q}{q-p}} \, \mathrm{d}\mu \Big)^{\frac{q-p}{p}} \\ &= \Big(\int_X |h|^q \Big)^{\frac{p}{q}} \Big(\int_X \mathrm{d}\mu \Big)^{\frac{q-p}{p}} \\ &= (||h||_q)^p (\mu(X))^{\frac{q-p}{p}} < \infty. \end{aligned}$$

Hence, $h \in L^p(X, \mu)$, proving that L^p spaces are nested.

Lemma 2.0.32. Let $M_f \colon L^2 \to L^2$ denote the multiplication operator. Then M_f is bounded if and only if $f \in L^\infty$, moreover

$$||M_f|| = ||f||_\infty = \operatorname{ess\,sup} |f|,$$

i.e, the operator norm of M_f is the essential supremum of the absolute value of f.

Proof. First, we know $f \leq \operatorname{ess\,sup} |f|$ almost everywhere, then, for every $g \in L^2$,

$$||M_f g||^2 = \int f^2 g^2 \mathrm{d}\mu \leq \operatorname{ess\,sup} f^2 \int g^2 \mathrm{d}\mu = \operatorname{ess\,sup} f^2 ||g||^2.$$

Which is equivalent to

$$||M_f g|| \leq \operatorname{ess\,sup} |f| ||g||.$$

Hence,

$$||M_f|| = \sup_g \frac{||M_f g||}{||g||} \leq \sup_g \frac{\operatorname{ess\,sup} |f| ||g||}{||g||} = \operatorname{ess\,sup} |f|,$$

obtaining one side of the inequality

$$||M_f|| \leq \operatorname{ess\,sup} |f|.$$

For the other direction of the inequality, note that since $f(x) \leq \operatorname{ess\,sup} |f|$ almost everywhere, it follows that, for any $\varepsilon > 0$ sufficiently small, the set

$$A = \{x : f(x) \geq \operatorname{ess\,sup} |f| - \varepsilon \geq 0\}$$

has measure greater than 0. Let $B \subseteq A$ such that $\mu(B) < \infty$, then, if I_B denotes the characteristic function on B,

$$||M_f I_B||^2 = \int f^2 I_B^2 \, \mathrm{d}\mu \geq (\operatorname{ess\,sup} |f| - \varepsilon) \int I_B^2 \, \mathrm{d}\mu = (\operatorname{ess\,sup} |f| - \varepsilon) ||I_B||^2.$$

Therefore,

$$\frac{||M_f I_B||}{||I_B||} \geq (\operatorname{ess\,sup} |f| - \varepsilon)^{\frac{1}{2}}.$$

Since this holds for every $\varepsilon > 0$ we will have,

$$\frac{||M_f I_B||}{||I_B||} \geq \operatorname{ess\,sup} |f|$$

and then,

$$||M_f|| \geq \frac{||M_f I_B||}{||I_B||} = \operatorname{ess\,sup} |f|. \qquad \square$$

Lemma 2.0.33. Suppose $\{A_n\}$ is a sequence in $\mathcal{B}(X, Y)$ with the property that

$$\sup_n \|A_n x\| < \infty,$$

for each $x \in X$. Then $\sup_n \|A_n\| < \infty$.

Proof. For a proof the reader may refer to [4, Theorem 14.3, page 83]. $\square$

Theorem 2.0.34 (Uniform boundedness principle). *Let X, Y be Banach spaces and suppose $F \subseteq \mathcal{B}(X,Y)$ with the property that $\sup_{A\in F} \|Ax\| < \infty$. Then, $\sup_{A\in F} \|A\| < \infty$.*

Proof. Assume $\sup_{A\in F} \|A\| = \infty$. Then, for every $n \in \mathbb{N}$ there exists $A_n \in F$ such that $\|A_n\| \geq n$, now, using the hypothesis

$$\sup_{n\in\mathbb{N}} \|A_n x\| \leq \sup_{A\in F} \|Ax\| < \infty.$$

Therefore, by the preceding lemma, $\sup_n \|A_n\| < \infty$ which contradicts $\|A_n\| \geq n$ for every $n \in \mathbb{N}$. □

Theorem 2.0.35 (Banach-Steinhaus). *Let $\{A_n\} \subset \mathcal{B}(X,Y)$ be such that for every $x \in X$, $\{A_n x\}$ is convergent. Then, the operator defined by $Ax = \lim_{n\to\infty} A_n x$ satisfies the following inequality,*

$$||A|| \leq \sup_n ||A_n||.$$

Proof. Let $\{A_n x\}$ be convergent for every $x \in X$. Then this sequence is bounded and thus $\sup_n \|A_n x\| < \infty$. By the uniform boundedness principle, we have $\sup_n \|A_n\| < \infty$. We also have the inequality

$$\|A_n x\| \leq \|A_n\| \|x\| \leq \sup_n \|A_n\| \|x\|.$$

Taking limits yields,

$$\|Ax\| = \lim_{n\to\infty} \|A_n x\| \leq \sup_n \|A_n\| \|x\|.$$

Therefore, $\|A\| \leq \sup_n \|A_n\|$. □

Remark 2.0.36. Another idea for the inequality in Banach-Steinhaus Theorem is the following. If $\{A_n\}$ are bounded operators then

$$\begin{aligned} \|Ax\| &= \lim_{n\to\infty} \|A_n x\| \\ &= \liminf_{n\to\infty} \|A_n x\| \\ &\leq \left(\liminf_{n\to\infty} \|A_n\|\right) \|x\|, \end{aligned}$$

which implies $\|A\| \leq \liminf_{n\to\infty} \|A_n\| \leq \sup_n \|A_n\|$.

3

TOEPLITZ MATRICES

We are interested in the study of complex matrices of the following type,

$$A = \begin{pmatrix} a_0 & a_{-1} & a_{-2} & \dots \\ a_1 & a_0 & a_{-1} & \dots \\ a_2 & a_1 & a_0 & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}$$

the first question which naturally arises in this context is, when does this matrix induce a linear operator in a specific sequence space ℓ^p ($1 \leq p \leq \infty$)?, in particular, we want to answer this question in ℓ^2 since this will give us the chance to use some tools of L^2 space (which is the completion of ℓ^2). Having this in mind, let $x = (x_i) \in \ell^2$ and let $b = (b_i)$ be any row of A. Then, since Ax is done by perfoming a row of A dot product the vector x, we have the following,

$$\begin{aligned} b \cdot x &= \sum_{i=1}^{\infty} b_i x_i \\ &\leq \sum_{i=1}^{\infty} |b_i x_i| \\ &\leq \Big(\sum_{i=1}^{\infty} |b_i|^2\Big)^{1/2} \Big(\sum_{i=1}^{\infty} |x_i|^2\Big)^{1/2}. \end{aligned}$$

Hence, a sufficient condition for each component of Ax to be well defined (finite) is $b \in \ell^2$. Moreover, we want to know when $Ax \in \ell^2$, since this would mean $A\colon \ell^2 \to \ell^2$ is a linear operator. It is not difficult to check that if $y = Ax$ then $y_n = \sum_{i=0}^{\infty} a_{-i+(n-1)} x_i$. Now, we turn to the question of when

$$\sum_{i=1}^{\infty} |y_i|^2 < \infty.$$

In other words, when is $y = Ax$ in ℓ^2. For this, consider the following,

$$\begin{aligned}\sum_{n=1}^{\infty} |y_n|^2 &= \sum_{n=1}^{\infty} \Big| \sum_{i=0}^{\infty} a_{-i+(n-1)} x_i \Big|^2 \\ &\leq \sum_{n} \big(\sum_{i} |a_{-i+(n-1)} x_i| \big)^2 \\ &\leq \sum_{n} \Big[\big(\sum_{i} |a_{-i+(n-1)}|^2 \big) \big(\sum_{i} |x_i|^2 \big) \Big] \\ &\leq \big(\sum_{i} |x_i|^2 \big) \big(\sum_{n} \sum_{i} |a_{-i+(n-1)}|^2 \big),\end{aligned}$$

therefore, a condition for $y = Ax$ to be in ℓ^2 is that

$$\sum_{n=1}^{\infty} \sum_{i=0}^{\infty} |a_{-i+(n-1)}|^2 < \infty.$$

We already gave a partial answer to our first concern. Next, we want to know in what cases A is a **bounded** linear operator. Why bounded? Because it is a desirable property since for linear operators **boundedness** and **continuity** are the same. An answer to this question is given by a classical result due to Otto Toeplitz.

We state and prove now one of the main theorems of this thesis.

Theorem 3.0.1 (Toeplitz 1911)**.** *The matrix A defined at the beginning of this section defines a bounded operator on ℓ^2 if and only if the numbers a_n are the Fourier coefficients of some function $a \in L^\infty(\mathbb{T})$,*

$$a_n = \frac{1}{2\pi} \int_0^{2\pi} a(\mathrm{e}^{\mathrm{i}\,\theta}) \mathrm{e}^{-\mathrm{i}\,n\theta} \mathrm{d}\theta, \qquad n \in \mathbb{Z}.$$

where $\mathbb{T}$ denotes the complex unit circle.

Proof. By Lemma 2.0.32 we know that the multiplication operator is bounded if and only if $a \in \mathrm{L}^\infty$. Consider the set $\{e_n\}_{n\in\mathbb{Z}}$ where

$$e_n(t) = \frac{1}{\sqrt{2\pi}} t^n, \quad t \in \mathbb{T}.$$

We know it is an orthonormal basis for $L^2(\mathbb{T})$. We want to see what the matrix representation

of the multiplication operator is. For this we shall study the behavior of its columns,

$$\begin{aligned} M_a(e_k) &= \sum_{j\in\mathbb{Z}} a_j e_j \\ ae_k &= \sum_{j\in\mathbb{Z}} a_j e_j \\ a_n &= ae_k \cdot e_n \\ a_n &= \int_{\mathbb{T}} a(t)e_k(t)\overline{e_n(t)}\mathrm{d}t \\ a_n &= \frac{1}{2\pi}\int_0^{2\pi} a(e^{\mathrm{i}\,\theta})e^{\mathrm{i}(k-n)\theta}\mathrm{d}\theta. \end{aligned}$$

We have already found a formula for the components of the k-th column, if we repeat this for the $(k+1)$-th column,

$$b_n = \frac{1}{2\pi}\int_0^{2\pi} a(e^{\mathrm{i}\,\theta})e^{\mathrm{i}(k+1-n)\theta}\mathrm{d}\theta = \frac{1}{2\pi}\int_0^{2\pi} a(e^{\mathrm{i}\,\theta})e^{\mathrm{i}(k-(n-1))\theta}\mathrm{d}\theta = a_{n-1},$$

in consequence, the n-th component of a column equals the $(n-1)$-th component of the previous one and so we have the following matrix representation

$$L_a = \begin{pmatrix} \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & a_0 & a_{-1} & a_{-2} & a_{-3} & a_{-4} & \dots \\ \dots & a_1 & a_0 & a_{-1} & a_{-2} & a_{-3} & \dots \\ \dots & a_2 & a_1 & a_0 & a_{-1} & a_{-2} & \dots \\ \dots & a_3 & a_2 & a_1 & a_0 & a_{-1} & \dots \\ \dots & a_4 & a_3 & a_2 & a_1 & a_0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{pmatrix}$$

where the a_n's are the Fourier coefficients and then L_a defines a bounded operator on ℓ^2 if and only if $a \in L^\infty$ and $||L_a|| = ||a||_\infty$. The matrix A we are looking for can be viewed as the compression of L_a to $\ell^2(\mathbb{Z}_+)$, this implies

$$||A|| \leq ||L_a||.$$

For a natural number n, consider the projection $P_n\colon \ell^2 \to \ell^2$ which acts as follows

$$(x_k)_{k\in\mathbb{Z}} \to (\dots, 0, 0, x_{-n}, \dots, x_1, x_2, \dots).$$

The matrix representation of the operator P_n is given by $(0|I)$ where I is the identity matrix starting in the $(-n)$-th column, then

$$P_n L_a P_n = \left(\begin{array}{c|c} 0 & 0 \\ \hline 0 & A \end{array}\right),$$

the divisions being in the $(-n)$-th column and row. Then, we can see that $(P_nL_aP_n)|_{\mathrm{Im}P_n}$ is essentially the same as A. This proves

$$||A|| = ||S_nL_aS_n||,$$

and since $S_n \to I$ strongly, it follows that $S_nL_aS_n \to L_a$ strongly. Therefore, by Banach-Steinhauss Theorem 2.0.35,

$$||L_a|| \leq \sup_n ||P_nL_aP_n|| = ||A||.$$

In conclusion, we proved A is bounded if and only if $a \in L^\infty$ and $||A|| = ||a||_\infty$. □

We denote both the matrix A and the operator it induces in l^2 by T_a, the function a in this context is known as the symbol of the **Toeplitz operator** T_a.

In matrix theory, it is always useful to think in terms of the spectrum of matrices since the eigenvalues and eigenvectors carry important information. There are many applications of these concepts, which vary from pure mathematics (eigenvalues are used to solve ordinary differential equations) to applied mathematics (optimization, numerical analysis, techniques in statistics such as PCA, etc.). From a class in basic linear algebra, we know that working with matrices and linear transformations is the same in the finite-dimensional case, and it is worth noting that when constructing the isomorphism between the vector space of matrices and the vector space of linear transformations, we use the basis of the spaces involved. It is of no surprise then that if we want to generalize this behavior, we would first have to define the concept of a basis for infinite-dimensional vector spaces. For this problem, several answers have been given (Hamel basis, Schauder basis, etc.). Moreover, it is also known that compact linear operators have been found to be similar in spectral properties to finite-dimensional matrices, and thus, having them is a tool for studying the eigenvalues and eigenvectors of linear operators. Another property of great value is self-adjointness since it provides information about the eigenvalues (which are real), and we also have properties like orthogonality, which allows us to decompose spaces into a direct sum of eigenspaces. For this type of operator, there is also a developed theory from which a well-known theorem comes, the Spectral Theorem.

Returning to Toeplitz operators, so far we know when they are bounded. For further investigation, we would want to have information about their spectral decomposition, which, in fact, turns out that neither of the two tools already mentioned are available, and this is due to the two following propositions. First, we recall two definitions, the one of compact operator and self-adjoint operator.

Definition 3.0.2 (Compact operator)**.** An operator $K \in \mathcal{B}(X, Y)$ is compact if for every bounded sequence $\{x_n\}$ in X, the sequence $\{Kx_n\}$ has a convergent subsequence.

Definition 3.0.3 (Self-adjoint operator)**.** An operator K is said to be self-adjoint if $K^* = K$.

Proposition 3.0.4 (Gohberg 1952)**.** *The only compact Toeplitz operator is the zero operator.*

Proof. Let $a \in L^\infty$ and suppose T_a is compact. Let Q_n be the projection

$$(x_0, x_1, \ldots) \to (0, \ldots, 0, x_n, x_{n+1}).$$

As $Q_n \to 0$ strongly, it follows $||Q_n T_a Q_n|| \to 0$ but $Q_n T_a Q_n|_{\mathrm{Im} Q_n}$ has the same matrix as T_a, then $||T_a|| = ||Q_n T_a Q_n|| = 0$. □

Proposition 3.0.5. *T_a is self-adjoint if and only if a is real-valued.*

Proof. T_a is self-adjoint if and only if $a_n = \overline{a_{-n}}$ which happens if and only if $a(t) = \overline{a(t)}$ which is equivalent to a being real-valued. □

We will now look for tools and concepts which can be of help for studying Toeplitz matrices, in particular we delve into some topics of functional analysis such as C^*-algebras and index theory of Fredholm operators.

A vector space V is a set together with an addition operation and a scalar product satisfying some properties. If we also have a multiplication operator within V such that it is associative, compatible with scalars, and distributive over addition, then we say V is an **algebra**. A **Banach algebra** is a Banach space X which is also an algebra satisfying $\|ab\| \leq \|a\| \, \|b\|$. We say that a conjugate linear map $a \to a^*$ is an involution if $a^{**} = a$ and $(ab)^* = b^* a^*$. Finally, a C^***-algebra** is a Banach algebra with an involution satisfying $\|a^* a\| = \|a\|^2$. If there is an element acting as an identity for multiplication, then we have a unital C^*-algebra. Two important examples of C^*-algebras are given by the set $\mathcal{B}(H)$ of bounded linear operators and the collection $\mathcal{K}(H)$ of compact linear operators, where H is a Hilbert space.

Example 3.0.6. Let H be a Hilbert space. Since $\mathrm{Hom}(H)$ is a vector space and $\mathcal{B}(H) \subseteq \mathrm{Hom}(H)$ is closed under addition and scalar multiplication, we know it is a vector (sub)space and if we also consider the norm operator then we have it is a Banach space. Taking multiplication as the composition yields,

$$\|(F_1 \circ F_2)x\| = \|F_1(F_2 x)\| \leq \|F_1\| \, \|F_2 x\| \leq \|F_1\| \, \|F_2\| \, \|x\| \, ,$$

whenever $F_1, F_2 \in \mathcal{B}(H)$. Hence, it is also a Banach algebra. When constructing the Hilbert-adjoint operator by means of the equality,

$$\langle Fx, y \rangle = \langle x, F^* y \rangle$$

there are many properties it satisfies, one of these (which we are interested in) are:

(i) $F^{**} = F$

(ii) $(F_1 \circ F_2)^* = F_2^* \circ F_1^*$

(iii) $\|F^* \circ F\| = \|F\|^2$

saying that the passage to adjoint operator serves as involution in $\mathcal{B}(H)$, making it into a C^*-algebra.

Example 3.0.7. Again, let H be a Hilbert space. We know every compact operator is bounded, then $\mathcal{K}(H) \subseteq \mathcal{B}(H)$. Using the characterization of compactness of an operator as sending every bounded sequence to a sequence (in the codomain) which has a convergent subsequence, we can prove this set is closed under addition, scalar multiplication and composition. Moreover, the properties of the involution do not depend on whether the operator is compact or not, hence, the collection of compact linear operators is a C^*-algebra. Being an algebra is in a way being a ring, then, we can talk about ideals, it turns out, $\mathcal{K}(H)$ is an ideal of $\mathcal{B}(H)$, (this can be prove using again a sequence argument).

Example 3.0.8. The set L^∞ is also a C^*-algebra with passage to complex conjugate as involution.

A very interesting property of unital C^*-algebras is given in the next theorem and allows us to define what is known as the spectrum of an element without specifying in which subalgebra or algebra we are working in.

Theorem 3.0.9. *Let $B \leq A$ where A, B are C^*-algebras sharing the same identity. Then, an element is invertible in A if and only if it is invertible in B*

Proof. For a proof the reader may want to check [7, Theorem 2.1.11, Page 41]. □

Definition 3.0.10. Let $\mathcal{A}$ be a C^*-algebra with identity e. We define the spectrum of an element as

$$\operatorname{sp}_{\mathcal{A}} a = \{\lambda \in \mathbb{C} \colon a - \lambda e \text{ is not invertible in } \mathcal{A}\}.$$

By the previous theorem we can write this set as $\operatorname{sp} a$.

Finally, we close this part of C^*-algebras with the definition of morphism and a theorem giving some properties to have in mind.

Definition 3.0.11. A morphism F of C^*-algebras is a linear operator preserving the product and involution, i.e,

(i) $F(ab) = F(a)F(b)$

(ii) $F(a^*) = F(a)^*$

Theorem 3.0.12. *A morphism $F\colon A \to B$ of unital C^*-algebras satisfies the following*

(i) $\|F(a)\| \leq \|a\|$,

(ii) $\operatorname{Im} F \leq \mathcal{B}$,

(iii) If F is injective then it preserves spectra and norm.

Definition 3.0.13. Let X, Y be Banach spaces. An operator A in $\mathcal{B}(X, Y)$ is called a **Fredholm operator** if the dimensions of $\operatorname{Ker} A$ and $\operatorname{Coker} A$ are finite. Then, the index defined by,

$$\operatorname{Ind} A = \dim \operatorname{Ker} A - \dim \operatorname{Coker} A$$

is a well-defined value.

Remark 3.0.14. We will use the usual notation of $\eta(A)$ and $d(A)$ for denoting the dimension of the kernel of A and dimension of cokernel of A, respectively.

We are going to state some results which will be of help for proving the theorem summarizing some well known properties of the index of a Fredholm operator.

Definition 3.0.15. A closed subspace M of X is said to be **complemented** in X if there exists a closed subspace N such that,

(i) $X = M + N$

(ii) $M \cap N = 0$

In this case we say X is a direct sum of M and N, and we write $X = M \oplus N$

Proposition 3.0.16. *An operator $A \in \mathcal{B}(X, Y)$ has a closed range if and only if there exists $c > 0$ such that*

$$\|Ax\| \geq c\,\|x\|,$$

for every $x \in X$.

Proposition 3.0.17. *An operator $A \in \mathcal{B}(X, Y)$ is a Fredholm operator if and only if it has the representation*

$$A = D + F,$$

where F has finite rank and D is a one-sided invertible Fredholm operator.

Theorem 3.0.18. *A necessary and sufficient condition for an operator $A \in \mathrm{B}(X,Y)$ to have a left inverse is that* $\ker A = \{0\}$ *and* $\operatorname{Im} A$ *is closed and complemented in Y. In this case, the operator AL is a projection onto* $\operatorname{Im} A$ *and where L is a left inverse of A.*

Proof. Suppose A have a left inverse L. Then, $Ax = 0$ implies $x = LAx = 0$. Meaning $\ker A = \{0\}$, moreover, note that if $y \in Y$ then $y = ALy + (y - ALy)$ with $ALy \in \operatorname{Im} A$ and $y - ALy \in \ker AL = \ker L$. If $y = Ax \in \ker AL$ then,

$$0 = ALy = AL(Ax) = Ax = y.$$

Hence $Y = \operatorname{Im} A \oplus \ker AL = \operatorname{Im} A \oplus \ker L$. Now, let $\ker A = \{0\}$ and $Y = \operatorname{Im} A \oplus M$ where both are closed subspaces of Y. Consider $P\colon Y \to \operatorname{Im} A$ be the natural projection and A_1 to be the operator A when restricted the codomain to $\operatorname{Im} A$, this makes A_1 a bijective bounded operator and $A^{-1}P$ a bounded left inverse of A since

$$A^{-1}PAx = A^{-1}Ax = x,$$

proving what we wanted. □

Lemma 3.0.19. Given $A \in \mathcal{B}(X,Y)$ and $B \in \mathcal{B}(Y,X)$, where X, Y and Z are Banach spaces. Suppose that $AB = F$ and $BA = G$ are Fredholm operators. Then $A$and B are Fredholm operators.

Proof. Since $\operatorname{Im} F \subset \operatorname{Im} A$ and $\ker A \subset \ker G$ it follows that $d(A) \leq d(F) < \infty$ and $\nu(A) \leq \nu(G) < \infty$. Thus A is a Fredholm operator. Similaly for B. □

Theorem 3.0.20. *(Properties) Let $A\colon X \to Y$ and $B\colon Y \to Z$ be Fredholm operators, then the following holds*

(i) $\operatorname{Im} A$ *is closed.*

(ii) BA *is a Fredholm operator and* $\operatorname{Ind} BA = \operatorname{Ind} A + \operatorname{Ind} B$.

(iii) The adjoint operator A' of A is also Fredholm and $\operatorname{Ind} A' = -\operatorname{Ind} A$.

Proof. (i) Since the kernel and cokernel are finite dimensional, then there exists subspaces M, N such that

$$X = M \oplus \mathrm{Ker}A, \qquad Y = N \oplus \mathrm{Im}A.$$

Consider the function $B\colon M \times N \to Y$ given by $B(u,v) = Au + v$, with $\|(u,v)\| = \|u\| + \|v\|$. Note that it is inyective since $B(u,v) = 0$ implies $Au = -v$, i.e, $Au, v \in$

$\mathrm{Im}A \cap N = \{0\}$. Then $v = 0$ and $u \in \mathrm{Ker}A \cap M = \{0\}$. Therefore, $B(u, v)$ if and only if $(u, v) = (0, 0)$. On the other hand, note that

$$\|B(u,v)\| \|Au + v\| \leq \|A\| \|u\| + \|v\| \leq \sup\{\|A\|, 1\}(\|u\| + \|v\|).$$

In consequence, B is bounded. Now, since Y is direct sum of $\mathrm{Im}\, A$ and N we see B is also surjective. Hence, B is a bounded invertible operator and

$$\left\|B^{-1}\right\| \|Au\| \geq \left\|B^{-1}(Au)\right\| = \|(u, 0)\| \|u\|.$$

Therefore, $\mathrm{Im}A$ is closed, since $\|Au\| \geq \frac{1}{\|B^{-1}\|} \|u\|$.

(ii) If A is Fredholm then we know $\ker A$ and $\mathrm{coker}\, A$ are finite dimensional. Then,

$$\ker BA = \ker A \oplus M$$

since $\ker A \subseteq \ker BA$ implying the former can be complemented with a closed subspace M. Note that if $y \in \mathrm{Im}\, A \cap \ker B$ then $y = Ax$ and $BAx = 0$, meaning $x \in \ker BA$. Letting $x = n + m$ where $n \in \ker A$ and $m \in M$ yields $y = Ax = An + Am = Am$. This means the restriction of A to M is onto. Moreover, if $m \in M$ and $Am = 0$ then $m \in \ker A \cap M = \{0\}$. Hence A is bijective and

$$\dim M = \dim(\mathrm{Im}\, A \cap \ker B) \leq \eta(B) < \infty.$$

Let $N_1 = \mathrm{Im}\, A \ker B$ and $n_1 = \dim N_1$, then

$$\eta(BA) = \eta(A) + n_1.$$

Now, since $n_1 < \infty$ we have it is complemented in $\ker B$ with a subspace say N_2. Since $d(A)$ is finite then $\mathrm{Im}\, A$ is complemented in Y, and since $\mathrm{Im}\, A \cap N_2 = \{0\}$ then there exists a finite dimensional subspace N_3 such that

$$Y = \mathrm{Im}\, A \oplus N_2 \oplus N_3.$$

Therefore, $d(a) = n_2 + n_3$.

Now, since $\ker B = N_1 \oplus N_2 \subseteq \mathrm{Im}\, A \oplus N_2$ it follows that B restricted to N_3 is one to one (therefore $\dim BN_3 = \dim N_3$) and if $y \in Y$, say $y = Ax + x_2 + x_3$ implies $By = BAx + Bx_2 + Bx_3 = BAx + Bx_3$ and then,

$$\mathrm{Im}\, B = \mathrm{Im}\, BA \oplus BN_3.$$

From this and the fact that $d(B) < \infty$ we obtain,

$$Z = \operatorname{Im} BA \oplus BN_3 \oplus W,$$

for some finite dimensional subspace W and this implies

$$d(BA) = \dim BN_3 + \dim W = \dim N_3 + d(B) = n_3 + d(B) < \infty.$$

Thus we conclude BA is a Fredholm operator and

$$\begin{aligned} \operatorname{ind}(BA) = \eta(BA) + d(BA) &= \eta(A) + n_1 - d(B) - n_3 \\ &= \eta(A) + \eta(B) - n_2 - d(B) - n_3 \\ &= \eta(A) - n_2 - n_3 + \eta(B) - d(B) \\ &= \operatorname{ind}(A) + \operatorname{ind}(B). \end{aligned}$$

(iii) Since A is a Fredholm Operator then $A = D + F$ where D is a one-sided invertible Fredholm operator and F has finite rank. Note that $A' = D' + F'$ so we shall show D' is a Fredholm Operator. If L is a left inverse of D then $LD = I$ and DL is a projection onto $\operatorname{Im} A$, which implies $F = I - LD$ is a projection onto the subspace complementing $\operatorname{Im} A$. Since $d(D) < \infty$ it follows that F is a finite rank operator. Since F is of finite rank then so it is F', by analogous of Proposition 3.0.17 we know $L'D' = I - F'$ is a Fredholm operator. Therefore, D' is Fredholm by Lemma 3.0.19. Similarly if D has right inverse. Therefore, A' is a Fredholm operator. The equality in the index follow from the fact

$$\eta(A') = d(A), \qquad d(A') = \eta(A).$$

□

Lemma 3.0.21. If $A_n = T_n(a) + P_nKP_n + W_nLW_n + C_n$ is a sequence such that $\sup ||A_n|| < \infty$ then

$$A_n \to T(a) + K$$

strongly.

Proof. For a proof the reader may refer to [3, Proposition 2.5, page 19] □

Lemma 3.0.22. Let K and L be compact operators and let $\{C_n\}$ be a sequence of $n \times n$ matrices such that $||C_n|| \to 0$. Then

$$\{P_nKP_n + W_nLW_n + C_n\} \in \mathcal{O},$$

where $\mathcal{O}$ stands for the set of al sequence $\{K_n\}$ such that $\frac{||K_n||_1}{n} \to 0$ and $\sup ||K_n|| < \infty$

Proof. For a proof the reader may refer to [3, Lemma 5.9, page 89] □

Theorem 3.0.23 (Szegő). *Let a be a continuous real-valued symbol. Then,*

$$\frac{1}{n}\sum_{j=1}^{n} f(\lambda_j(T_n(a))) \to \frac{1}{2\pi}\int_0^{2\pi} f(a(e^{i\theta}))\,\mathrm{d}\theta \tag{3.1}$$

for every $f \in C_0(\mathbb{R})$.

Proof. We know $\{T_n^k(a)\} \in S(C)$ for every $k \in \mathbb{N}$, that is,

$$T_n^k(a) = T_n(b) + P_nKP_n + W_nLW_n + C_n.$$

Passage to the strong limit $n \to \infty$ yields, $T^k(a) = T(b) + K$ according to Lemma 3.0.21. Since we know $T^k(a) = T(a^k) + K_1$ with K_1 a compact operator then $T(a^k - b)$ is compact. Therefore, $b = a^k$ and then,

$$T_n^k(a) = T_n(a^k) + P_nKP_n + W_nLW_n + C_n.$$

from Lemma 3.0.22 we have,

$$T_n^k(a) = T_n(a^k) + K_n,$$

with $K_n \in \mathcal{O}$. Now,

$$\frac{|\operatorname{tr} K_n|}{n} \le \frac{\|K_n\|_1}{n} = o(1)$$

whence

$$\frac{\operatorname{tr} T_n^k(a)}{n} = \frac{\operatorname{tr} T_n(a^k)}{n} + o(1).$$

Since the trace of a matrix is equal to the sum of the diagonal elements and also the sum of its eigenvalues,

$$\operatorname{tr} T_n^k(a) = \sum_{j=1}^{n} (\lambda_j(T_n(a)))^k,$$

$$\operatorname{tr} T_n(a^k) = n(a^k)_0 = \frac{n}{2\pi}\int_0^{2\pi} (a(e^{i\theta}))^k\,\mathrm{d}\theta,$$

we obtain (3.1) for $f(x) = x^k$ and so, this proves int he case f polynomial. Now, let $f \in C_0(\mathbb{R})$, $m = \min|a|, M = \max|a|$. Given $\varepsilon > 0$, there is a polynomial p such that $|f(x) - p(x)| < \varepsilon$ for $x \in [m, M]$. Since $\lambda_j(T_n(a)) \in [m, M]$ for al j and n then

$$\frac{1}{n}\sum_{j=1}^{n} |f(\lambda_j(T_n(a))) - p(\lambda_j(T_n(a)))| < \frac{1}{n} n\varepsilon = \varepsilon$$

and we also have,

$$\frac{1}{2\pi}\int_0^{2\pi}\left|f(a(e^{i\theta}))-p(a(e^{i\theta}))\right|\mathrm{d}\theta<\varepsilon.$$

The above together with the fact that we have the desired limit for polynomials yields,

$$\left|\frac{1}{n}\sum_{j=1}^{n}f(\lambda_j(T_n(a)))-\frac{1}{2\pi}\int_0^{2\pi}f(a(e^{i\theta}))\,\mathrm{d}\theta\right|<3\varepsilon,$$

when n is large. □

4

EIGENVALUE'S APPROXIMATIONS

This chapter concerns the implementation of a method found in the article [2] for the approximation of eigenvalues of some types of matrix which are the product of an inverse Toeplitz matrix and another Toeplitz matrix. We write a conjecture formulated by M. Bogoya et al. and then explain the method and each crucial part of the code which was done in Mathematica.

Conjecture 4.0.1. Let g, l be two even functions with $g > 0$ in $(0, \pi)$, and suppose $f = l/g$ is monotone increasing over $(0, \pi)$ and sufficiently smooth. Set $X_n = T_n^{-1}(g)T_n(l)$. Then, for some integer $\alpha \geq 1$, some $\beta \in (0, 1]$, and every $j = 1, \ldots, n$, as $n \to \infty$ the following asymptotic expansion holds:

$$\lambda_j(X_n) = f(s_{j,n}), \quad s_{j,n} = \theta_{j,n} + \sum_{k=1}^{\alpha} \rho_k(\theta_{j,n}) h^{\beta k} + E_{k,n,\alpha}, \tag{4.1}$$

where

- The numbers $s_{j.n}$ are arranged in nondecreasing order.
- $h = 1/(n+1)$ and $\theta_{j,n} = \pi j h$.
- The coefficients ρ_k are continuous functions from $(0, \pi)$ to $\mathbb{R}$ depending only on f.
- $E_{j,n,\alpha} = O(h^{(\alpha+1)\beta})$ is the remainder that satisfies the inequality

  $$|E_{j,n,\alpha}| \leq \kappa h^{(\alpha+1)\beta}$$

  for some constant κ depending only on α, g, l.

This conjecture came as an improvement of some other results obtained by the authors of the paper we are working with [2]. In fact, a specific case of Conjecture 1.1 of this paper

was formally proved by Bogoya, Böttcher, Grudsky, and Maximenko. Following conjecture 4.0.1, we shall use a numerical algorithm to approximate the eigenvalues of any matrix X_n with the given hypothesis. We can break the algorithm into the following steps. First, a precomputing phase in which we will have the eigenvalues of X_n for some small values of n. Second, an extrapolation phase that aims to determine (approximately) the values of $\rho_k(\theta_{j,n})$ having n fixed and $1 \leq j \leq n$ and also with specific α. Third, the interpolation phase in which we estimate the functions ρ_k for any value in the interval $[0, \pi]$. Finally, the fourth step concerns the examination of the errors between the exact eigenvalues and the calculation using the process described in the conjecture.

In the conjecture, we see $\theta_{j,n} = \pi j h = \frac{\pi j}{n+1}$. One of the key details in this method is to realize that for some combinations of j, n we get the same value of $\theta_{j,n}$,

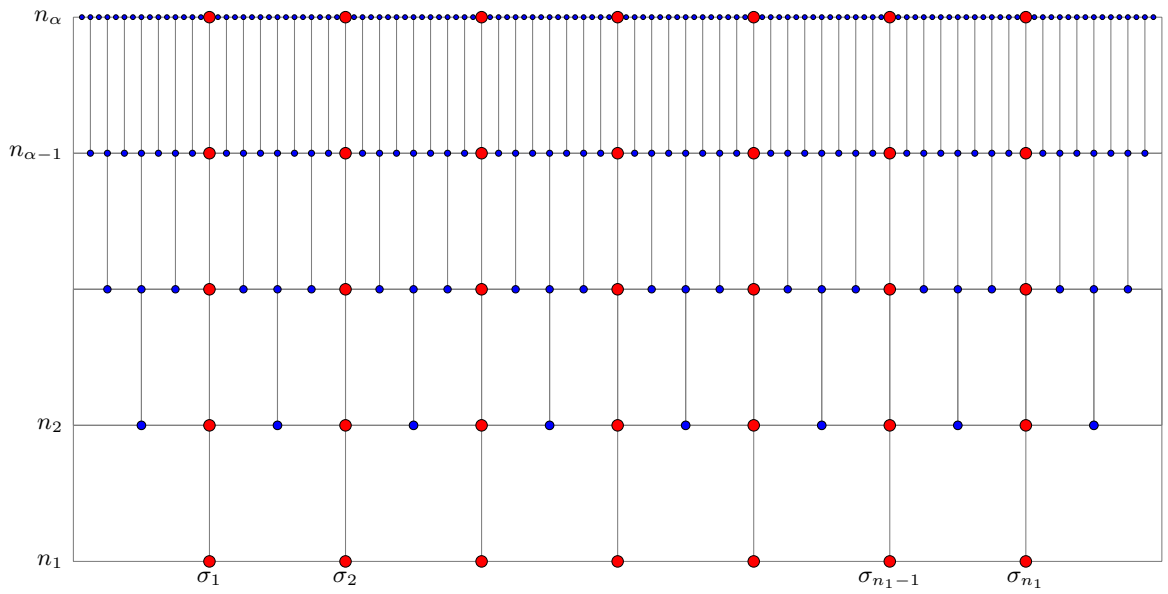
$$\theta_{j,n} = \frac{\pi j}{n+1} = \frac{c\pi j}{c(n+1)} = \frac{\pi(cj)}{(c(n+1)-1)+1} = \theta_{cj,c(n+1)-1}.$$

Showing that for any $c \in \mathbb{Z}^+$ we have the preceding equality. Thus, if we let $c_k = 2^{k-1}$ for $k \in \mathbb{Z}^+$ we obtain an increasing sequence of matrix sizes, whose eigenvalues will be computed and used in the precomputing phase. In the following we mirror notations for n and h, i.e, h_k corresponds to $1/(n_k + 1)$. For the precomputing phase we start with an initial and fixed n_1 and a positive integer α which gives us information about how many matrices we will be using in this phase, since for every $1 \leq k \leq \alpha$ we compute the eigenvalues of X_{n_k}, where $n_k = c_k(n_1 + 1) - 1 = 2^{k-1}(n_1 + 1) - 1$. For $1 \leq j_1 \leq n_1$ let $j_k = c_k j_1 = 2^{k-1} j_1$. Therefore, with the current notations we have,

$$\theta_{j_1,n_1} = \theta_{j_2,n_2} = \cdots = \theta_{j_{n_\alpha},n_\alpha}.$$

Let σ_{j_1} denote this number since it changes according to j_1 because n_1 is fixed.

In the following grid we can graphically get an idea of the method described above. We call each row a level and levels corresponds to matrix sizes, in each one of these we will find then, the eigenvalues of X_{n_k} in blue and red dots, which we have from the precomputing phase. In the extrapolation phase we will find an approximation for each function ρ_k in the red points σ_{j_1}. After this, we use interpolation for knowing the value of ρ_k in the interval $(0, \pi)$, since having this would mean we can then calculate $s_{j,n}$ and according with the conjecture this implies we can also find the j-th value of X_n with the equation $\lambda_j(X_n) = f(s_{j,n})$. This explains roughly the method and how the figure helps in understanding it.



Continuing with the method, we use the equation in conjecture 4.0.1 with $n_1, \ldots, n_\alpha$ to form a system of linear equations in the variables $\rho_k(\sigma_{j_1})$.

$$\begin{aligned}
s_{j_1,n_1} - \sigma_{j_1} &= \rho_1(\sigma_{j_1})h_1^\beta + \rho_2(\sigma_{j_1})h_1^{2\beta} + \cdots + \rho_\alpha(\sigma_{j_1})h_1^{\alpha\beta} + E_{j_1,n_1,\alpha},\\
s_{j_2,n_2} - \sigma_{j_1} &= \rho_1(\sigma_{j_1})h_2^\beta + \rho_2(\sigma_{j_1})h_2^{2\beta} + \cdots + \rho_\alpha(\sigma_{j_1})h_2^{\alpha\beta} + E_{j_1,n_1,\alpha},\\
&\vdots\\
s_{j_\alpha,n_\alpha} - \sigma_{j_1} &= \rho_1(\sigma_{j_1})h_\alpha^\beta + \rho_2(\sigma_{j_1})h_\alpha^{2\beta} + \cdots + \rho_\alpha(\sigma_{j_1})h_\alpha^{\alpha\beta} + E_{j_1,n_1,\alpha}.
\end{aligned}$$

If we let $\hat{\rho}_k$ be the approximation of ρ_k then we can see the system as follows:

$$\begin{pmatrix} h_1^\beta & h_1^{2\beta} & \cdots & h_1^{\alpha\beta} \\ h_2^\beta & h_2^{2\beta} & \cdots & h_2^{\alpha\beta} \\ \cdots & \cdots & \cdots & \cdots \\ h_\alpha^\beta & h_\alpha^{2\beta} & \cdots & h_\alpha^{\alpha\beta} \end{pmatrix} \begin{pmatrix} \hat{\rho}_1(\sigma_{j_1}) \\ \hat{\rho}_2(\sigma_{j_1}) \\ \vdots \\ \hat{\rho}_\alpha(\sigma_{j_1}) \end{pmatrix} = \begin{pmatrix} s_{j_1,n_1} \\ s_{j_2,n_2} \\ \vdots \\ s_{j_\alpha,n_\alpha} \end{pmatrix} - \sigma_{j_1} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

and since $1 \leq j_1 \leq n_1$ we will find the value of $\hat{\rho}_k$ in n_1 points. The system above can be

written as follows in a simplified notation. First, let

$$S = \begin{pmatrix} s_{1,n_1} & s_{2,n_1} & \cdots & s_{n_1,n_1} \\ s_{2,n_2} & s_{4,n_2} & \cdots & s_{2\cdot n_1,n_2} \\ \vdots & \vdots & \ddots & \vdots \\ s_{2^{\alpha-1},n_\alpha} & s_{2^{\alpha-1}\cdot 2,n_\alpha} & \cdots & s_{2^{\alpha-1}\cdot n_1,n_\alpha} \end{pmatrix}.$$

The matrix S has as rows the vectors s_{j_k,n_k} for $k = 1, \ldots, \alpha$, and thus, it has α rows. Now, if we write S_{j_1} for the j_1-th column with $j_1 = 1, \ldots, n_1$, then we can write the system as

$$H^\beta \hat{\rho}(\sigma_{j_1}) = S_{j_1} - \sigma_{j_1} 1,$$

where 1 is the vector having one in all components. Note that we would solve this system for each column, i.e, n_1 times.

After having this, we move to the next stage which corresponds to the interpolation part in each level n_k of the grid. If we want to find the value of $\rho_k(\theta)$ for $\theta \in (0, \pi)$, if $\theta = \sigma_{j_1}$ then we already know the value of $\rho_k(\theta)$ from the extrapolation phase. Otherwise, we will pick the m closest points in $\{\sigma_1, \sigma_2, \ldots, \sigma_\alpha\}$ to θ and we interpolate with $\{\rho_k(\sigma_1), \rho_k(\sigma_2), \ldots, \rho_k(\sigma_\alpha)\}$. The value of m is usually calculated with the formula $\alpha - k + 12$. However, we did it by trying/error in the code. Finally, our approximation of eigenvalues with α terms will be given by the formula of conjecture 4.0.1

$$\lambda_j(X_n) = f(s_{j,n}) \approx f\Big(\theta_{j,n} + \sum_{k=1}^{\alpha} \hat{\rho}_k(\theta_{j,n}) h^k\Big).$$

It is also worth noting that the numerical results achieved by using this approximation are extremely precise (as stated in paper [2]).

We also want to highlight the importance of having a formula for approximating eigenvalues since it gives a way to find any eigenvalue of X_n for arbitrary n without having to compute all the eigenvalues of X_n, as would be the case if we did it in any eigensolver. Computing eigenvalues for large matrices presents significant challenges. For example, the computational resources needed for computing the spectrum of a matrix, increase significantly as its size does it too. Moreover, if we are dealing with very large matrices, the memory capacity of the computer may be exceeded, preventing the computation from even starting or stopping halfway. Even if the computation is feasible in terms of memory, the time required to calculate all eigenvalues can become absurdly long, making the process impractical. Furthermore, when working with eigensolvers the accumulation of rounding errors during the computation can lead to inaccurate or even completely incorrect eigenvalue

estimates, especially for very large and potentially ill-conditioned matrices. Therefore, having an approximation formula allows us to target specific eigenvalues of interest without having to compute the entire spectrum, saving both computational resources and avoiding large numerical errors.

5

NUMERICAL IMPLEMENTATION

In the present chapter we will explain as much as possible the implementation of the code which was done in Mathematica for the method described in the previous chapter.

We divided the code in 4 parts, precomputing, extrapolation, interpolation and error. The following corresponds to the first part. We started by defining some functions we will use later on. The h, θ from (4.1), the a_{ij} which are the Fourier coefficients of $\cos(j\theta)$, and j_k, n_k which are the index and matrix size respectively.

```
h[n_] := 1/(n + 1);
\[Theta][j_, n_] := \[Pi]*j*h[n];
a[i_, j_] := Switch[i, j, 1/2, -j, 1/2, _, 0];
j[k_, j1_] := 2^(k - 1)*j1;
nk[k_, n1_] := 2^(k - 1)*(n1 + 1) - 1;
```

For this implementation we will work with the following functions (and their inverses):

$$f_1(x) = \frac{2 - \cos(x) - \cos(2x)}{3 + 2\cos(x)}, \qquad f_2(x) = \frac{40 - 15\cos(x) - 24\cos(2x) - \cos(3x)}{1208 + 1191\cos(x) + 120\cos(2x) + \cos(3x)},$$

which we wrote in the code as:

```
l1[x_] := 2 - Cos[x] - Cos[2 x];
g1[x_] := 3 + 2 Cos[x];
f1[x_] := l1[x]/g1[x];
l2[x_] := 40 - 15 Cos[x] - 24 Cos[2 x] - Cos[3 x];
g2[x_] := 1208 + 1191 Cos[x] + 120 Cos[2 x] + Cos[3 x];
```

```
f2[x_] := l2[x]/g2[x];
f1inv[x_] := ArcCos[1 - x];
f2inv[y_] := SetPrecision[x /.
    FindRoot[f2[x] - y, {x, \[Pi]/2*Sqrt[17 y]}], 40];
```

then, we built the corresponding matrices $X_n = T_n^{-1}(g)T_n(l)$ for these functions:

```
Xf1[n_] :=
 Inverse@ToeplitzMatrix@Table[6 a[0, j] + 2 a[1, j],
            {j, 0, n - 1}] .
  ToeplitzMatrix@Table[4 a[0, j] - a[1, j] - a[2, j],
            {j, 0, n - 1}]
Xf2[n_] :=
 Inverse@ToeplitzMatrix@
    Table[2416 a[0, j] + 1191 a[1, j] + 120 a[2, j] + a[3, j],
            {j, 0, n - 1}] .
  ToeplitzMatrix@
   Table[80 a[0, j] - 15 a[1, j] - 24 a[2, j] - a[3, j],
            {j, 0, n - 1}]
```

and calculated the eigenvalues we needed with the following function:

```
\[Lambda]f1[n_, r_] := Sort@Eigenvalues@SetPrecision[Xf1[n], r];
\[Lambda]f2[n_, r_] := Sort@Eigenvalues@SetPrecision[Xf2[n], r];
```

after saving the eigenvalues for n =100, 201, 403, 807, 1615 with a precision of $r = 40$ we applied the inverse of f_1 and f_2 accordingly obtaining $s_{j,n}$ for building the matrix S. Since we will be working with $n_1 = 100$ we formed also the matrix H and built the function for calculating σ_{j_1}:

```
H = SetPrecision[N@Table[h[nk[j, 100]]^k, {j, 1, 5}, {k, 1, 5}] ,
   40];
IH = N@Inverse[H];
```

```
\[Sigma][n1_] :=
  SetPrecision[Table[\[Theta][j1, n1], {j1, 1, n1}], 40];
```

For the extrapolation we programmed a function for finding the solution of the system of linear equations given the matrix S, the initial matrix size n_1 and the index j_1 indicating which column of S we will use.

```
Linv[matrix_, j1_, n1_] :=
  IH . (matrix[[All, j1]] - \[Sigma][n1][[j1]]);
```

and then use this function to create another one in which we could organize each approximation $\hat{\rho}_k(\sigma_{j_1})$ in a matrix:

```
\[Rho]matrix[matrix_, n1_] := Module[{solutions},
   solutions = Table[Linv[matrix, i, n1], {i, 1, n1}];
  Transpose[solutions]]
```

Finally, a function for graphing each row of this matrix.

```
graphbyrow[matrix_] :=
  Module[{nFilas = Length[matrix]},
   Table[ListPlot[
     Transpose[{Range[Length[matrix[[i]]]], matrix[[i]]}],
     PlotLabel -> "Fila " <> ToString[i]], {i, 1, nFilas}]];
```

Next, we test how the code is working so far with an example using f_1 and f_2. By showing the graphs of each row of the folowing matrices $f1M$, $f2M$.

```
f1M = \[Rho]matrix[Sf1, 100];
f2M = \[Rho]matrix[Sf2, 100];
```

In $f1M$ and $f2M$ we have the values $\hat{\rho}_k(\sigma_{j_1})$ (k-th row and j_1-th column) found in the extrapolation part for the function f_1 and f_2 respectively.

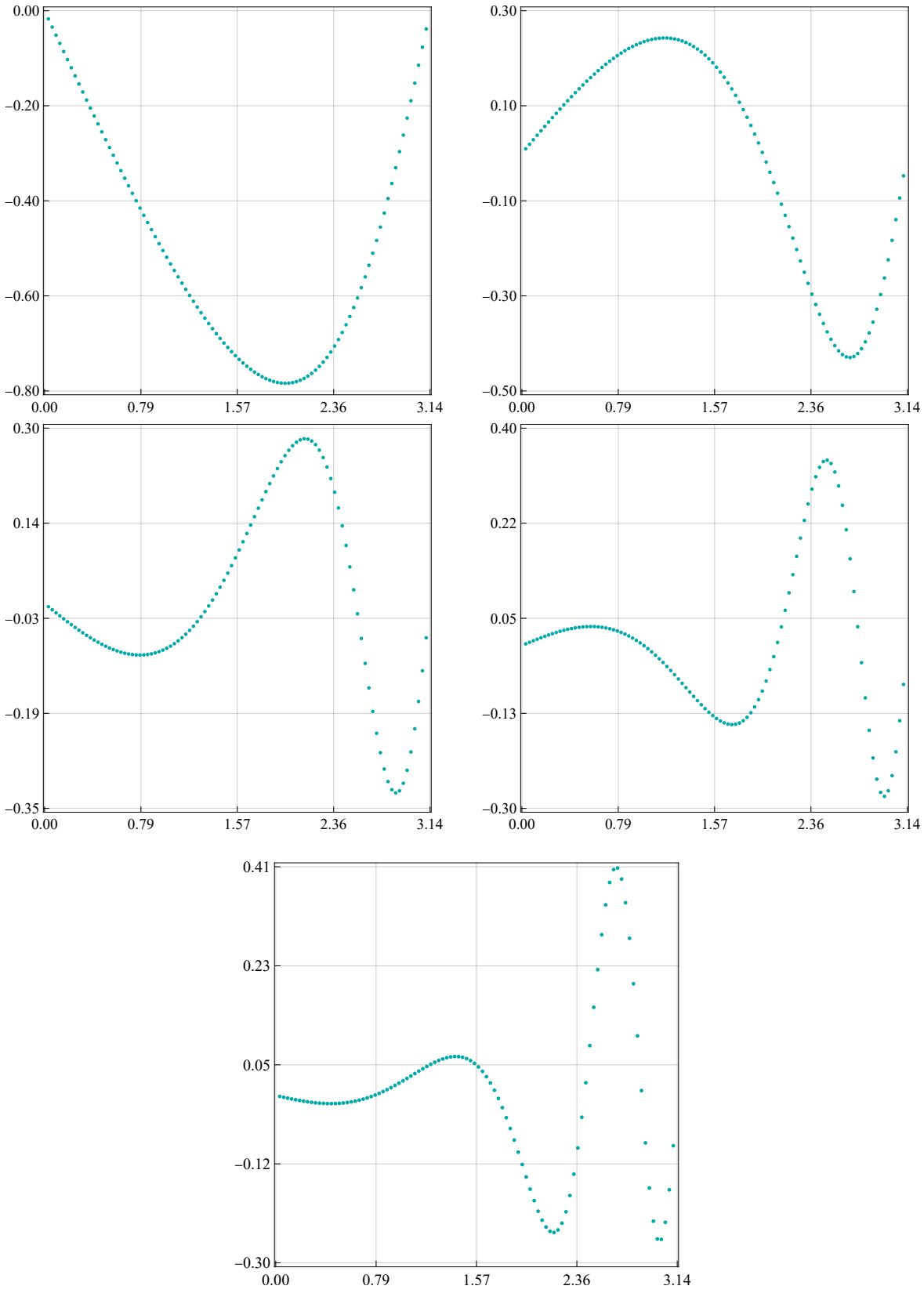


Figure 5.1: Extrapolation graphs of $\hat{\rho}_k$ for $k = 1, 2, 3, 4, 5$ using a grid of $n_1 = 100$ for the matrix $X_n = T_n^{-1}(g)T_n(l)$ with $l(x) = 2 - \cos(x) - \cos(2x)$ and $g(x) = 3 + 2\cos(x)$.

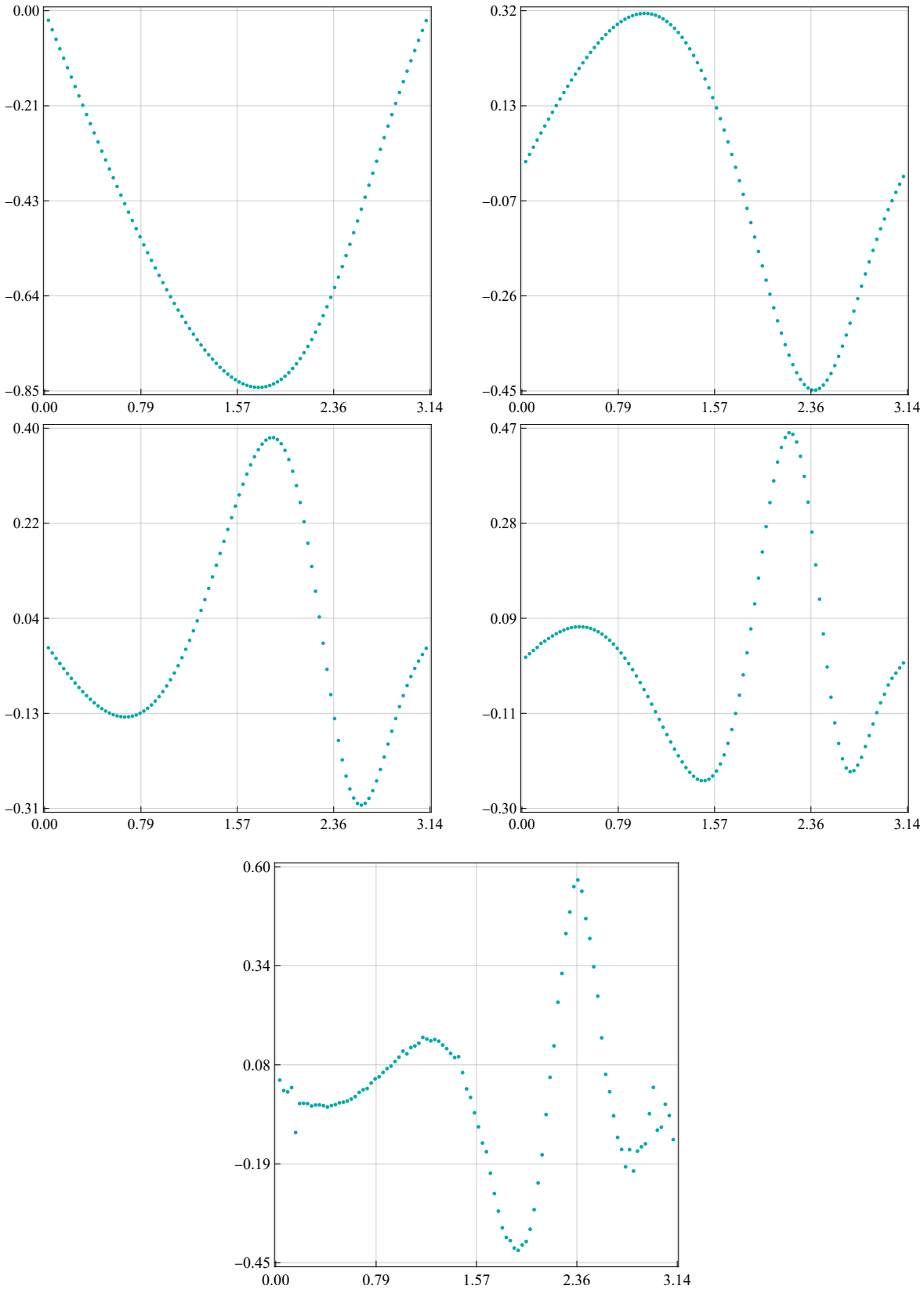


Figure 5.2: Extrapolation graphs of $\hat{\rho}_k$ for $k = 1, 2, 3, 4, 5$ using a grid of $n_1 = 100$ for the matrix $X_n = T_n^{-1}(g)T_n(l)$ with $l(x) = 40 - 15\cos(x) - 24\cos(2x) - \cos(3x)$ and $g(x) = 1208 + 1191\cos(x) + 120\cos(2x) + \cos(3x)$.

We continue with the final part which is the interpolation. We create a function for doing the interpolation described above with a number of points of interpolation given by "ninter".

```
interpol[matrix_, list_, ninter_, x_, k_] :=
 Module[{distances, pol, puntos},
  distances = Abs[list - x];
  o = Ordering[distances];
  puntos = Table[{list[[o[[j]]]], matrix[[k, o[[j]]]]},
                 {j, 1, ninter}];
  pol[z_] = InterpolatingPolynomial[puntos, z];
  pol[x]]
```

Again, we test the code to check if it follows the expected dotted pattern for ρ_1 with f_1 and f_2 as shown in the graphs below.

since the graphs overlap properly then we proceed to define $\hat{\rho}_k$ using the interpolation function:

```
\[Rho][matrix_, x_, k_, ninter_] :=
 interpol[matrix, \[Sigma][100], ninter, x, k]
```

Lastly, we make the function for calculating $s_{j,n}$. Note that we will be able to control how many addends we have by changing α.

```
\[CapitalSigma][matrix_, x_, n_, \[Alpha]_, ninter_] :=
 Module[{s = 0, i = 1},
  While[i <= \[Alpha],
   s += \[Rho][matrix, x, i, ninter]*h[n]^i;
   i++];
  s]

Sjn[matrix_, j_, n_, \[Alpha]_, ninter_] :=
 SetPrecision[\[Theta][j, n], 40] +
  SetPrecision[\[CapitalSigma][matrix, \[Theta][j, n] , n,
                \[Alpha], ninter], 40]
```

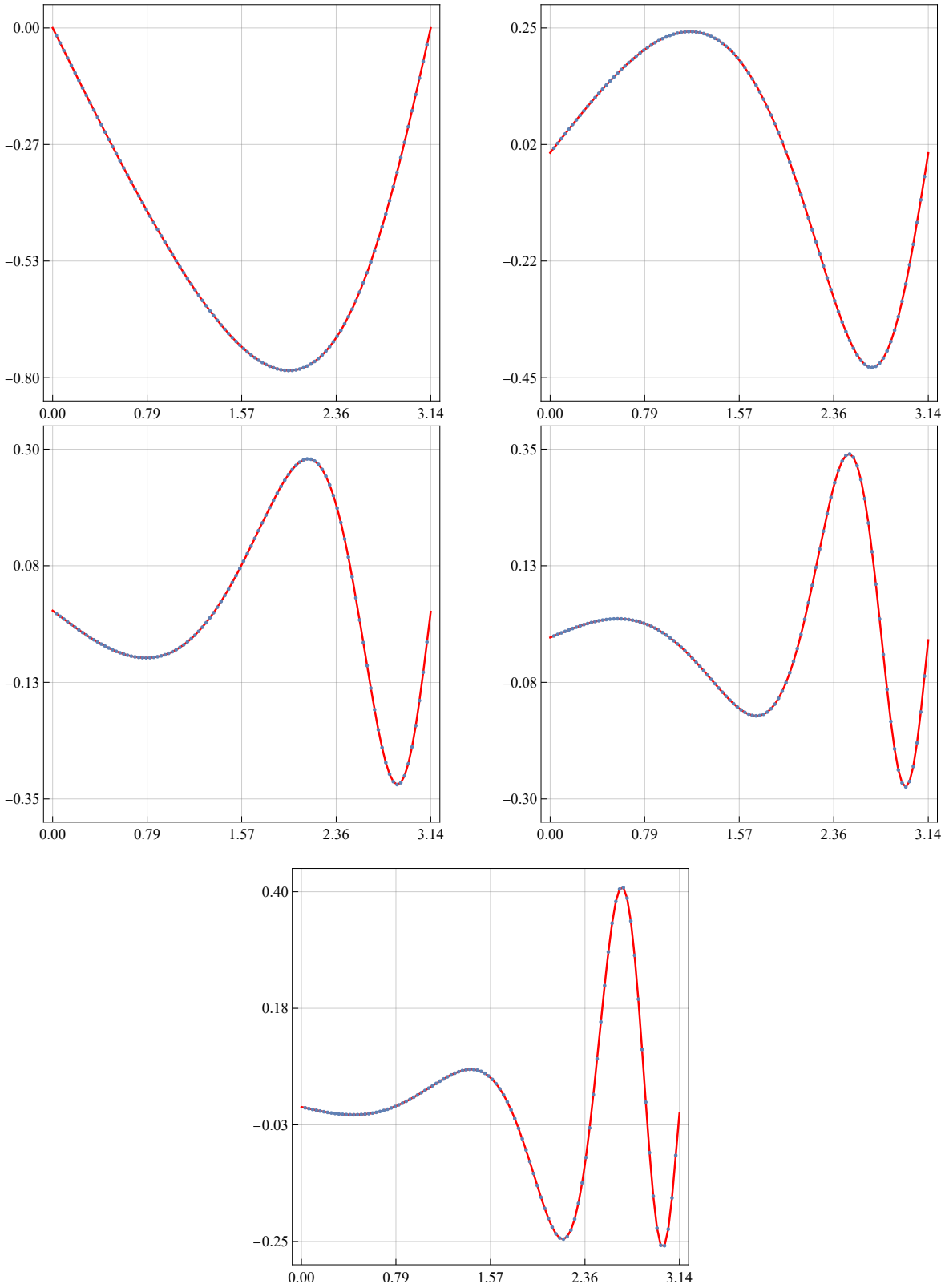


Figure 5.3: Polynomial interpolation (with 2 closest points) graphs of $\hat{\rho}_k$ for $k = 1, 2, 3, 4, 5$ grid of $n_1 = 100$ for the matrix $X_n = T_n^{-1}(g)T_n(l)$ with $l(x) = 2 - \cos(x) - \cos(2x)$ and $g(x) = 3 + 2\cos(x)$.

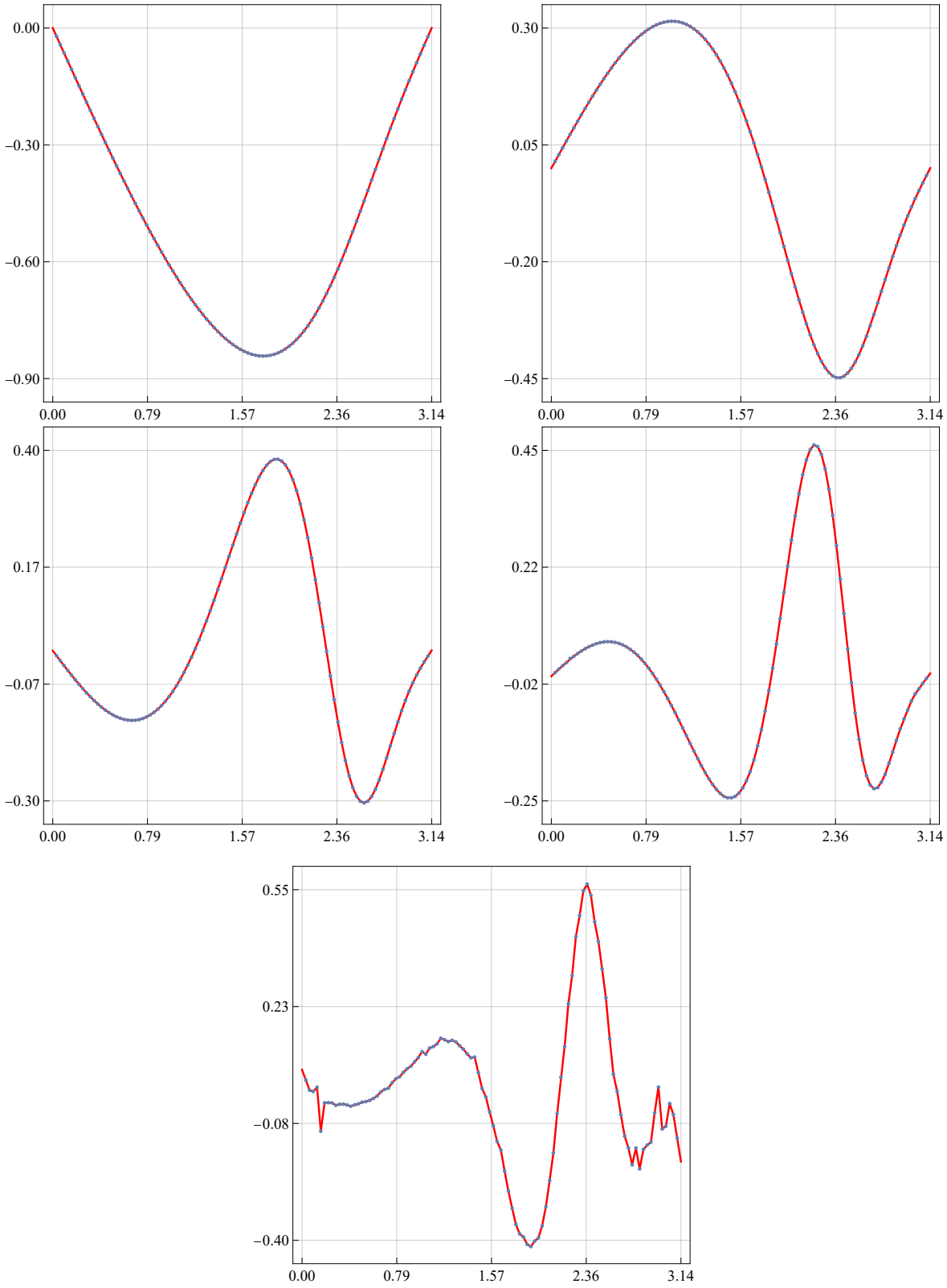


Figure 5.4: Polynomial interpolation (with 2 closest points) graphs of $\hat{\rho}_k$ for $k = 1, 2, 3, 4, 5$ using a grid of $n_1 = 100$ for the matrix $X_n = T_n^{-1}(g)T_n(l)$ with $l(x) = 40 - 15\cos(x) - 24\cos(2x) - \cos(3x)$ and $g(x) = 1208 + 1191\cos(x) + 120\cos(2x) + \cos(3x)$.

The last part of the code corresponds to analyzing the error between the eigenvalues calculated from this algorithm and the actual eigenvalues. For this we computed the eigenvalues of X_n for n =1024, 2048 and 4096. Then we compare them to the numerical approximation given by the code and we got the following graphs using $\alpha = 1, 2, 3, 4, 5$ (with colors red, blue, orange, cyan and black, respectively). Figures 5.6 and 5.8 show the absolute errors $AE_{j,n}^{N(k)}$ for $f1$ and $f2$ respectively and for different k levels where $N(k)$ represents the numerical approximations with k terms in $s_{j,n}$. The choice of colors is as follows:

- $k = 1$.
- $k = 2$.
- $k = 3$.
- $k = 4$.
- $k = 5$.

More can be said about the figures, for example, our code performs well for $k = 1, 2, 3$ but for $k = 4, 5$ the errors are unstable and irregular, we can not see a pattern in there as we do in the former case. For this reason we chose to do the error tables just for $k = 1, 2, 3$ since in these three cases we got the expected results. For tables 5.1 and 5.2 we use the following notation for errors,

$$\mathrm{AE}_{j,n}^{N(k)} = |\lambda_j(X_n) - \lambda_j(X_n)^{N(k)}|, \qquad \mathrm{NE}_{j,n}^{N(k)} = (1+n)^k \mathrm{AE}_{j,n}^{N(k)}, \qquad \mathrm{RE}_{j,n}^{N(k)} = \frac{\mathrm{AE}_{j,n}^{N(k)}}{|\lambda_j(X_n)|}$$

and

$$\mathrm{AE}_{n}^{N(k)} = \max_j \mathrm{AE}_{j,n}^{N(k)}, \qquad \mathrm{NE}_{n}^{N(k)} = \max_j \mathrm{NE}_{j,n}^{N(k)}, \qquad \mathrm{RE}_{n}^{N(k)} = \max_j \mathrm{RE}_{j,n}^{N(k)}.$$

AE is called absolute error and it measures how far is the actual eigenvalue with the numerical approximation. NE is the normalized error and it is a scaled version of the absolute error, in this case the normalization parameter is $(1+n)^k$ and it is for adjusting the scale of the error. However, depending on the method, another parameter will be used as in the paper [1], in which the scale parameter is n^4/j^k. This error is useful for studying the convergence of errors and the expected stable behavior of errors. RE is the relative error and since it is divided by the absolute value of the each eigenvalue it provides a way to analyze the error independently of the magnitude. It does not depend on the scale of the eigenvalue so we will have a "standard" way to evaluate the size of the error, for example, a low relative error is strongly related with a precise approximation.

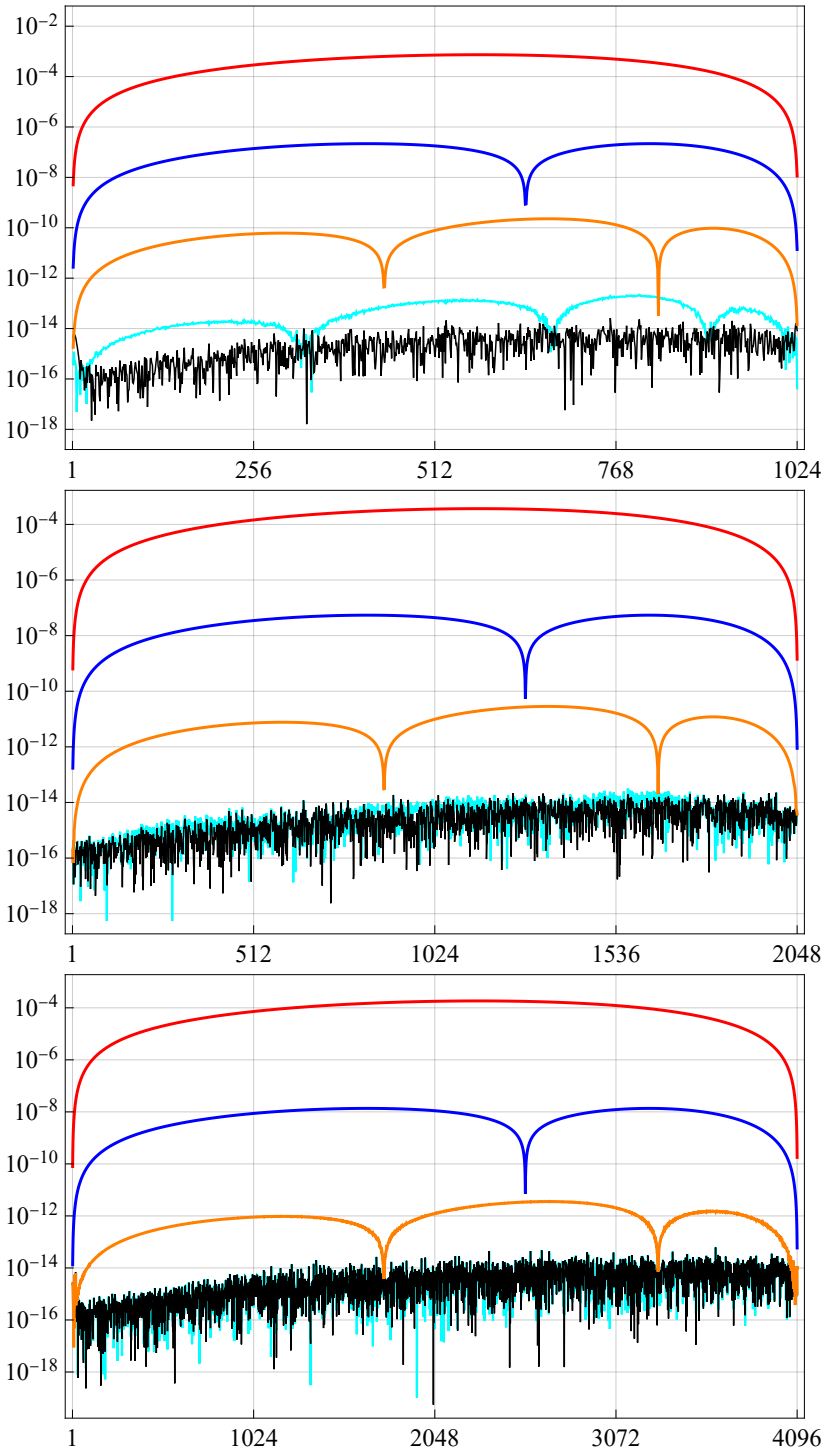


Figure 5.5: Base 10 logarithm scale for the errors $\mathrm{AE}_{j,n}$ for the matrix $X_n = T_n^{-1}(g)T_n(l)$ with $l(x) = 2 - \cos(x) - \cos(2x)$ and $g(x) = 3 + 2\cos(x)$. The horizontal axis corresponds to the j-th eigenvalue. A grid size of $n_1 = 100$, $\beta = 1$, 40 digits of precision for computations. Different $\alpha = 1, 2, 3, 4, 5$ levels in ascending order from red to black. In order, the matrix sizes used were $n = 1024, 2048, 4096$.

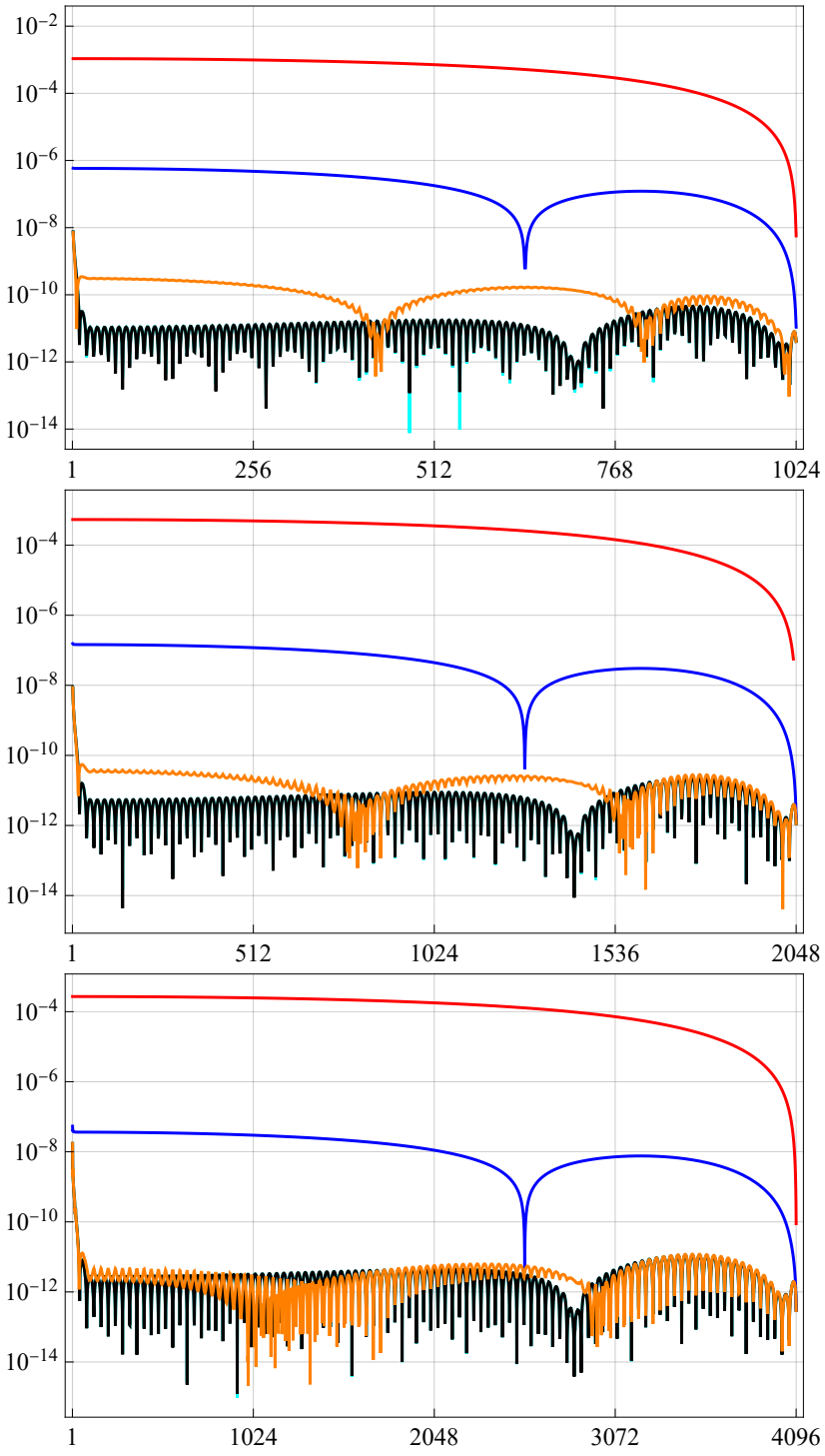


Figure 5.6: Base 10 logarithm scale for the errors $\mathrm{RE}_{j,n}$ for the matrix $X_n = T_n^{-1}(g)T_n(l)$ with $l(x) = 2 - \cos(x) - \cos(2x)$ and $g(x) = 3 + 2\cos(x)$. The horizontal axis corresponds to the j-th eigenvalue. A grid size of $n_1 = 100$, $\beta = 1$, 40 digits of precision for computations. Different $\alpha = 1, 2, 3, 4, 5$ levels in ascending order from red to black. In order, the matrix sizes used were $n = 1024, 2048, 4096$.

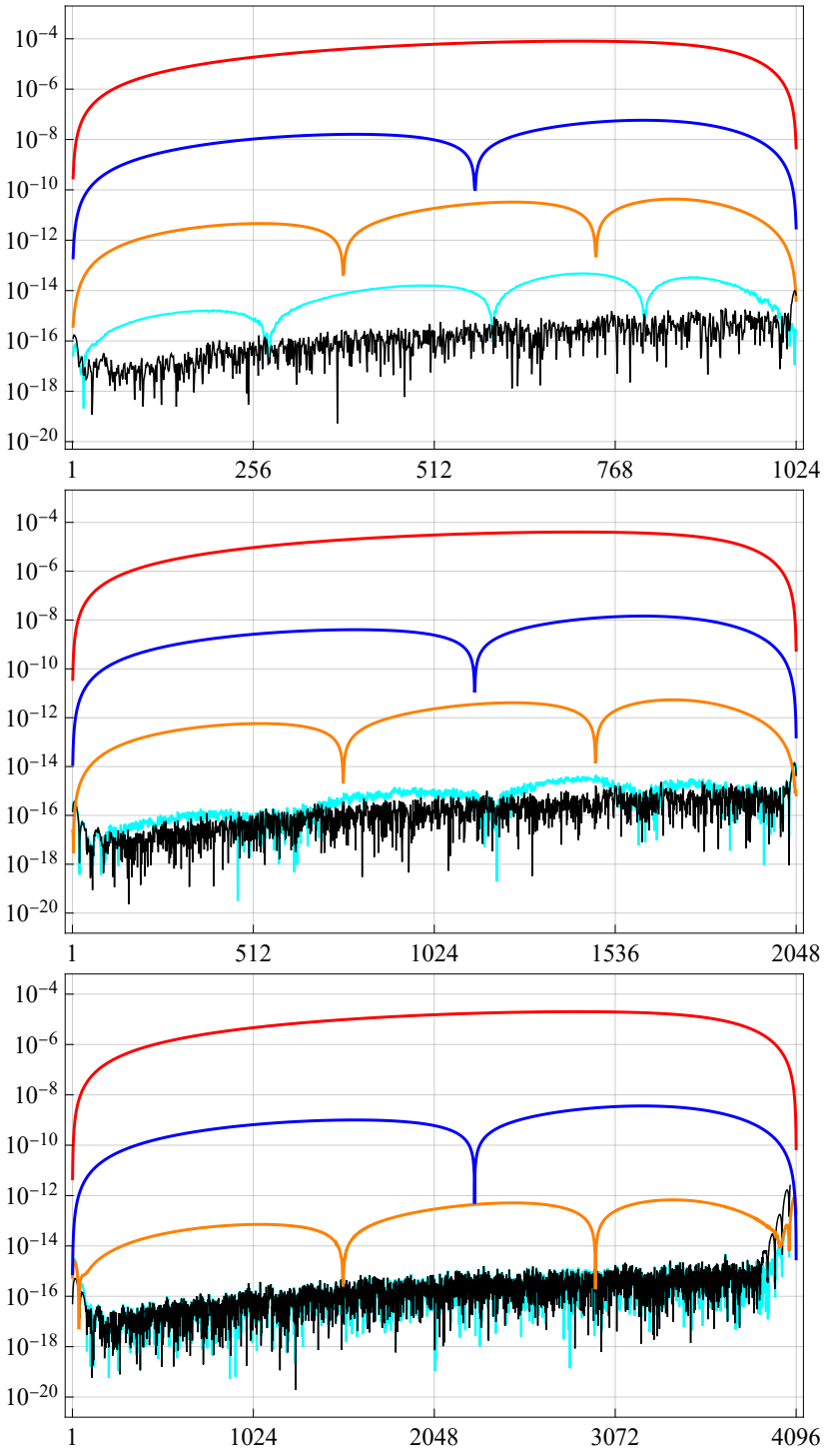


Figure 5.7: Base 10 logarithm scale for the errors $\mathrm{AE}_{j,n}$ for the matrix $X_n = T_n^{-1}(g)T_n(l)$ with $l(x) = 40 - 15\cos(x) - 24\cos(2x) - \cos(3x)$ and $g(x) = 1208 + 1191\cos(x) + 120\cos(2x) + \cos(3x)$. The horizontal axis corresponds to the j-th eigenvalue. A grid size of $n_1 = 100$, $\beta = 1$, 40 digits of precision for computations. Different $\alpha = 1, 2, 3, 4, 5$ levels in ascending order from red to black. In order, the matrix sizes used were $n = 1024, 2048, 4096$.

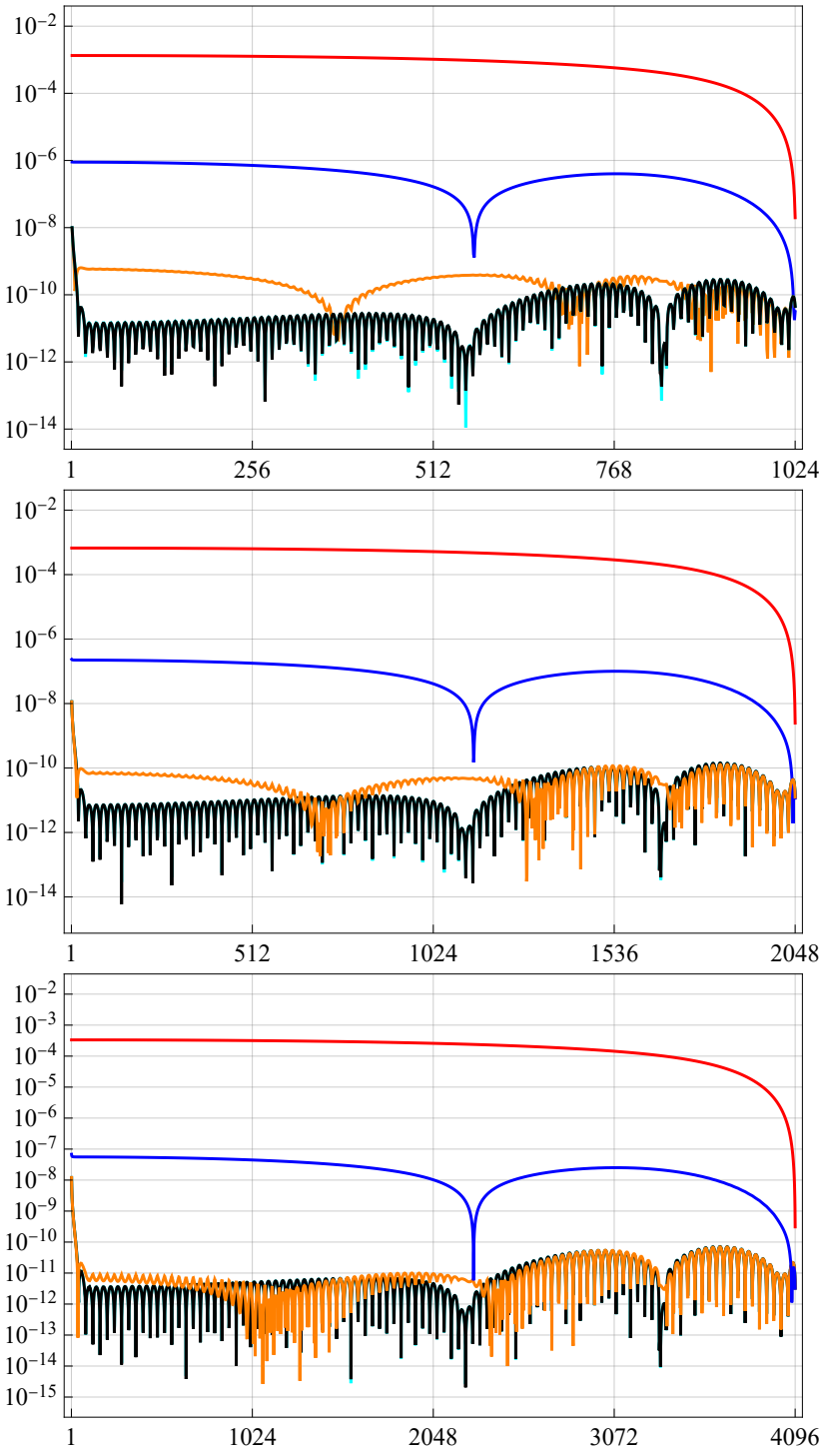


Figure 5.8: Base 10 logarithm scale for the errors $\mathrm{RE}_{j,n}$ for the matrix $X_n = T_n^{-1}(g)T_n(l)$ with $l(x) = 40 - 15\cos(x) - 24\cos(2x) - \cos(3x)$ and $g(x) = 1208 + 1191\cos(x) + 120\cos(2x) + \cos(3x)$. The horizontal axis corresponds to the j-th eigenvalue. A grid size of $n_1 = 100$, $\beta = 1$, 40 digits of precision for computations. Different $\alpha = 1, 2, 3, 4, 5$ levels in ascending order from red to black. In order, the matrix sizes used were $n = 1024, 2048, 4096$.

Table 5.1: Errors for $k = 1, 2, 3$ levels with different matrix sizes n, corresponding to the matrix $X_n = T_n^{-1}(g)T_n(l)$ with $l(x) = 2 - \cos(x) - \cos(2x)$ and $g(x) = 3 + 2\cos(x)$. A grid size of $n_1 = 100$, $\beta = 1$ and 40 digits of precision were used for computations.

$k = 1$

n	$\text{AE}_n^{N(1)}$	$\text{NE}_n^{N(1)}$	$\text{RE}_n^{N(1)}$
16	4.41×10^{-2}	7.49×10^{-1}	6.58×10^{-2}
32	2.28×10^{-2}	7.52×10^{-1}	3.37×10^{-2}
64	1.16×10^{-2}	7.53×10^{-1}	1.71×10^{-2}
128	5.84×10^{-3}	7.54×10^{-1}	8.59×10^{-3}
256	2.93×10^{-3}	7.54×10^{-1}	4.31×10^{-3}
512	1.47×10^{-3}	7.54×10^{-1}	2.16×10^{-3}
1024	7.36×10^{-4}	7.54×10^{-1}	1.08×10^{-3}
2048	3.68×10^{-4}	7.54×10^{-1}	5.40×10^{-4}
4096	1.84×10^{-4}	7.54×10^{-1}	2.70×10^{-4}

$k = 2$

n	$\text{AE}_n^{N(2)}$	$\text{NE}_n^{N(2)}$	$\text{RE}_n^{N(2)}$
16	8.88×10^{-4}	2.33×10^{-1}	2.89×10^{-3}
32	2.13×10^{-4}	2.32×10^{-1}	5.60×10^{-4}
64	5.47×10^{-5}	2.31×10^{-1}	1.45×10^{-4}
128	1.38×10^{-5}	2.30×10^{-1}	3.67×10^{-5}
256	3.47×10^{-6}	2.29×10^{-1}	9.25×10^{-6}
512	8.69×10^{-7}	2.29×10^{-1}	2.32×10^{-6}
1024	2.18×10^{-7}	2.29×10^{-1}	5.82×10^{-7}
2048	5.44×10^{-8}	2.29×10^{-1}	1.46×10^{-7}
4096	1.36×10^{-8}	2.28×10^{-1}	4.48×10^{-8}

$k = 3$

n	$\text{AE}_n^{N(3)}$	$\text{NE}_n^{N(3)}$	$\text{RE}_n^{N(3)}$
16	5.82×10^{-5}	2.47×10^{-1}	6.73×10^{-5}
32	6.87×10^{-6}	2.47×10^{-1}	9.35×10^{-6}
64	8.95×10^{-7}	2.45×10^{-1}	1.23×10^{-6}
128	1.14×10^{-7}	2.45×10^{-1}	1.57×10^{-7}
256	1.44×10^{-8}	2.45×10^{-1}	1.99×10^{-8}
512	1.81×10^{-9}	2.45×10^{-1}	2.50×10^{-9}
1024	2.27×10^{-10}	2.45×10^{-1}	3.82×10^{-10}
2048	2.84×10^{-11}	2.45×10^{-1}	1.63×10^{-10}
4096	3.60×10^{-12}	2.48×10^{-1}	8.43×10^{-9}

Table 5.2: Errors for $k = 1, 2, 3$ levels with different matrix sizes n, corresponding to the matrix $X_n = T_n^{-1}(g)T_n(l)$ with $l(x) = 40 - 15\cos(x) - 24\cos(2x) - \cos(3x)$ and $g(x) = 1208 + 1191\cos(x) + 120\cos(2x) + \cos(3x)$. A grid size of $n_1 = 100$, $\beta = 1$ and 40 digits of precision were used for computations.

$k = 1$

n	AE1	NE1	RE1
16	4.92×10^{-3}	8.36×10^{-2}	8.19×10^{-2}
32	2.51×10^{-3}	8.29×10^{-2}	4.19×10^{-2}
64	1.27×10^{-3}	8.26×10^{-2}	2.12×10^{-2}
128	6.39×10^{-4}	8.25×10^{-2}	1.06×10^{-2}
256	3.21×10^{-4}	8.24×10^{-2}	5.33×10^{-3}
512	1.61×10^{-4}	8.24×10^{-2}	2.67×10^{-3}
1024	8.03×10^{-5}	8.23×10^{-2}	1.34×10^{-3}
2048	4.02×10^{-5}	8.23×10^{-2}	6.68×10^{-4}
4096	2.01×10^{-5}	8.23×10^{-2}	3.34×10^{-4}

$k = 2$

n	AE2	NE2	RE2
16	2.05×10^{-4}	5.93×10^{-2}	3.21×10^{-3}
32	5.62×10^{-5}	6.12×10^{-2}	8.58×10^{-4}
64	1.44×10^{-5}	6.10×10^{-2}	2.22×10^{-4}
128	3.66×10^{-6}	6.10×10^{-2}	5.63×10^{-5}
256	9.23×10^{-7}	6.09×10^{-2}	1.42×10^{-5}
512	2.31×10^{-7}	6.09×10^{-2}	3.56×10^{-6}
1024	5.80×10^{-8}	6.09×10^{-2}	8.92×10^{-7}
2048	1.45×10^{-8}	6.09×10^{-2}	2.24×10^{-7}
4096	3.63×10^{-9}	6.09×10^{-2}	5.63×10^{-8}

$k = 3$

n	AE3	NE3	RE3
16	9.56×10^{-6}	4.69×10^{-2}	1.27×10^{-4}
32	1.28×10^{-6}	4.60×10^{-2}	1.77×10^{-5}
64	1.70×10^{-7}	4.67×10^{-2}	2.33×10^{-6}
128	2.17×10^{-8}	4.66×10^{-2}	3.00×10^{-7}
256	2.74×10^{-9}	4.65×10^{-2}	4.12×10^{-8}
512	3.44×10^{-10}	4.65×10^{-2}	7.48×10^{-9}
1024	4.31×10^{-11}	4.64×10^{-2}	1.71×10^{-9}
2048	5.40×10^{-12}	4.64×10^{-2}	4.14×10^{-10}
4096	6.76×10^{-13}	4.65×10^{-2}	1.42×10^{-9}

BIBLIOGRAPHY

[1] M. Bogoya, A. Böttcher, and S. Grudsky. Eigenvalues of toeplitz matrices emerging from finite differences for certain ordinary differential operators. *Linear Algebra and its Applications*, 706:24–54, 2025.

[2] M. Bogoya, S. Serra-Capizzano, and P. Vassalos. Fast toeplitz eigenvalue computations, joining interpolation-extrapolation matrix-less algorithms and simple-loop theory: The preconditioned setting. *Applied Mathematics and Computation*, 466:128483, 2024.

[3] A. Böttcher and S. M. Grudsky. *Toeplitz matrices, asymptotic linear algebra and functional analysis*, volume 67. Springer, 2000.

[4] I. Gohberg, S. Goldberg, and M. Kaashoek. *Basic classes of linear operators*. Springer Science & Business Media, 2003.

[5] A. N. Kolmogorov and S. V. Fomin. *Introductory real analysis*. Courier Corporation, 1975.

[6] E. Kreyszig. *Introductory functional analysis with applications*, volume 17. John Wiley & Sons, 1991.

[7] G. J. Murphy. *C*-algebras and operator theory*. Academic press, 2014.

[8] F. Olver. *Asymptotics and special functions*. AKP Classics. A K Peters, Ltd., Wellesley, MA, 1997.

[9] G. Szegő. Ein Grenzwertsatz über die Toeplitzschen Determinanten einer reellen positiven Funktion (German). *Math. Annal.*, 76:490–503, 1915.